Bounded Rationality as a Strategy for Cognitive Science

by

Tyler Brooke-Wilson

B.A. Philosophy and Psychology
City University of New York, 2017

Submitted to the Department of Linguistics and Philosophy
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2023

Author _____

Department of Linguistics and Philosophy
September 8, 2023

Certified By _____

EJ Green
Associate Professor
Thesis Co-Supervisor

Certified By _____

Josh Tenenbaum
Professor
Thesis Co-Supervisor

Accepted By _____

Brad Skow
Laurance S. Rockefeller Professor
Chair of the Committee on Graduate Students

Abstract

This thesis has two themes. The first is the structure of the mind – how does the mind break down into parts, and how do those parts relate to one another? The second is the computational complexity inherent in seeing and thinking – how are we able to see and think when these problems appear to be computationally intractable? The motivating idea is that managing this computational complexity is one of the central forces shaping the structure of the mind. We see and think in the ways we do because these are rare solutions to hard problems. This means we can use the theory of computational complexity to shed light on the structure of the mind – ruling out theories that fail to account for the computational tractability of the mind and focusing attention on those that might. In that spirit, Chapter 1 lays out a framework for thinking about the complexity of an important class of mental processes and applies the framework to perception, using it to make progress on the question of how what we think influences what we see. Chapter 2 argues that perception (e.g. seeing and hearing) and cognition (e.g. reasoning and planning) take different strategies to tame the complexity they face, and uses this fact to explain otherwise puzzling differences between seeing and thinking. Chapter 3 outlines the unique computational challenge facing thinking and proposes a novel account of how we think designed to meet that challenge, highlighting key points of similarity and difference between our minds, the computational models of cognitive science, and the neural networks of contemporary AI. In each chapter, inquiring into the structure of the mind via computational complexity helps us see both how the mind works and why it works that way. In this way we begin to reverse engineer the blueprint of the mind.

Acknowledgments

I have many people to thank.

First, to my friends in graduate school, who made the last few years some of the best imaginable, contributed to my thinking on so many topics, and made Cambridge feel like home, I'd like to thank João Loula, Jon Gauthier, Luke Hewitt, David Builes, Quinn White, Verónica Gómez Sánchez, Ishita Dasgupta, Matthias Hofer, Maddie Cusimano, Andrés Campero, Lionel Wong, Alex Lew, Tan Zhi-Xuan, Dae Houlihan, Johannes Mahr, Yunu Kwon, Reuben Cohn-Gordon, Sandra Romero Pinto, Michele Odisseas Impagnatiello, Andrew Bahle, Junyi Chu, Ethan Wilcox, Carley Ruemmele, Ryan Ravanpak, Sophie Gibert, and Abby 'Socks' Finer. I wish I could list all the things I've learned from you all.

To members of the MIT philosophy department, with special thanks to Kevin Dorst, Agustin Rayo, Bob Stalnaker for helpful comments and discussions and to Steve Yablo for inspiration and critical encouragement in my first years at MIT.

To those in New York who set me on this path, I'd like to thank Jake Quilty-Dunn and Eric Mandelbaum, who taught me to ask the big questions of cognitive architecture, and Javier Gomez-Lavin, who taught me to navigate academia even while he urged me to flee.

I was lucky to have a big committee and every one of them has shaped this work.

To Ned Block for raising the big questions of seeing and thinking and helping me see what answers could look like.

To Laurie Paul for tying mind to experience, epistemology, and metaphysics and pushing me to do the same.

To Alex Byrne for teaching me to start with a gripping example and to not lose sight of the fact that seeing and thinking are familiar things.

To Jack Spencer for tightening every argument and thesis and getting me unstuck just about every other week. This dissertation could not have been written without you.

And to my co-advisors...

To Josh Tenebaum for the breadth of your vision of the mind and the breadth of your support as an advisor and a friend.

To EJ Green for your guidance at every step of the way, which kept me grounded in the philosophical tradition, perception, and a reasonable timetable and which allowed me to explore an exciting idea without getting lost in it.

Finally, to my family Tom, Patti, Jules, Jenna, Trev, and my partner, Madeline, this would not have been possible without your loving support.

Table of Contents

Introduction

When people learn of an event in the news, it can influence their voting behavior, their investment decisions, or their vacation plans, often in sensible ways. But how are people able to recognize such wide ranging consequences of new information, when the potential connections are nearly endless? Similarly, when people open their eyes, they effortlessly see the 3D world before them, despite the fact that the light hitting the retina is compatible with an infinite number of different 3D scenes. How do we do it? In both cases, the space of ways the world could be is vast, but must be navigated quickly in order to see and think. The challenge here is so great that, when viewed through the lens of theoretical computer science, the problems the brain solves seem like they should be impossible. This threat of *computational intractability* has at various times been taken to entail very controversial theses about the mind – that the mind could not be unified, that thinking could not be realized by symbolic operations, that true artificial intelligence is impossible, or that the computational theory of mind must be false. Whether any of these arguments are sound is unclear. What is clear is that explaining computational tractability of the mind is a central challenge for cognitive science. This dissertation addresses several aspects of this challenge and uses it to shed light on how the mind works.

The threat of computational intractability touches on many different themes – from questions in the philosophy of mind and epistemology (questions like, how rational are we? And, how rational can we be?) to questions in the philosophy of AI (What kinds of minds are possible? How do different AI methods deal with intractability?). The central focus of this dissertation are the consequences of tractability for *cognitive architecture*. Cognitive architecture is the study of how the mind breaks down into parts and how those parts relate to one another. These are questions that have been with philosophy for a long time – from Plato's tripartite theory of the soul to the faculties of the early moderns[1] – and which are being investigated using empirical, computational, and theoretical methods today. The central thought behind this dissertation is that computational tractability is one of the chief challenges facing the mind, and cognitive architecture is one the chief tools that allows the mind to

[1] Architectural distinctions play a central role in many key projects of the early moderns, such as the distinction between the Will and the Intellect in Descartes' Meditations or between Reason, the Senses, and the Imagination in Hume's argument for skepticism about causation.

tame intractability. By investigating how that is done, we uncover important truths about the structure of the mind and explanations as to why it is that way.

The history of tractability in computational approaches to intelligence goes back to the beginning of both cognitive science and AI. The intractability of important classes of problems solved by the mind has been the chief force holding AI back since its inception. Early AI pioneers were buoyed by initial successes in automating reasoning and planning that many in the 60's thought that human-level intelligence was just a decade or two away.[2] What ultimately held them back was that the methods that worked for very small scale problems became infeasible when applied to larger problems because of the ways computational costs scale. This left real world problems far out of reach. The infamous Lighthill report, penned by British mathematician James Lighthill in 1973, which led to a dramatic drop in funding for the field, an 'AI Winter', flagged tractability as the core challenge facing AI. It pointed out that the field had not made discernable progress on that problem in its first 15 years of life (Lighthill 1973). Writing a year earlier in 'What Computers Can't Do', Hubert Dreyfus also highlighted this same pattern of early successes on small problems followed by failure to scale methods to larger instances. While Dreyfus lacked the computer science concepts to make the argument precise, he highlighted several ways the human mind comes equipped to deal with tractability, all of which had so far eluded automation (Dreyfus 1972).

These problems that haunted classical AI methods reappear in a new guise for the models in vogue today. Computational costs can be traded off against one another – for example, run time costs against training costs and data requirements. Contemporary AI methods do just this, trading speed at solving hard problems for massive data and training. For example, today's language models – AI systems for processing language – are trained on upwards of 10,000 times more linguistic data than are people (Wardstadt & Bowman 2022). Training costs for the most competent models run in the hundreds of millions of dollars (Cottier 2023). Discussion has turned to the possibility of billion-dollar training runs (Amodei 2023) and some have made the case that the availability of data may soon

[2] Claude Shannon, for example, predicted in 1961 that the 'robots of science fiction fame' are 10-15 years off, while Herbert Simon predicted that "machines will be capable, within twenty years, of doing any work a man can do" (1965, p. 96). Alan Turing more conservatively predicted that we would have a machine that passed the Turing test by the year 2000 (1950).

become a bottleneck (Villalobos et al. 2022). Undaunted by these challenges, leaders in the field have again started to predict that human-level intelligence is imminent, with predictions ranging from 2 years off (Amodei 2023) to 5-20 (Bengio 2023). Will this be like last time? Or are these problems now surmountable? Both are live options. Understanding where we are requires careful thought about computational tractability.

Much like for AI, tractability is a central challenge for cognitive science. This isn't surprising – AI systems and the computational models of mental processes that cognitive science seeks to develop are two sides of the same coin. An AI system that can do some task is thereby a potential hypothesis about how the human mind does it, while a computational model of human performance on that task represents a potential AI solution to it. Tractability problems that haunt AI show up in similar ways in cognitive science. Many computational models that beautifully capture human behavior on some toy problems cannot be scaled up to model human behavior on similar but larger problems out in the world. This fact has been appreciated by many working in the field. Limitations on computational resources have been offered programmatically as a general constraint on theories of mental processes (Icard 2018; Griffiths et al. 2015; Lewis et al. 2014). Intractability has been used to argue against broad classes of cognitive models (Kwisthout 2011) and to motivate 'resource rational' explanations of various effects in the field (Lieder et al. 2014). Most ambitiously, tractability has been used to argue that mental processes can't be symbolic in nature (Churchland 1989), that they can't be probabilistic (LeCun 2022), and even that they can't be *computational* (Fodor 2000). Here again, we're left with many questions – Which of these arguments hold water? And *how* can we use tractability as a constraint on our theories of the mind? Answering these requires building a framework for thinking about computational tractability as it applies to the mind.

A final theme in this dissertation are questions of cognitive architecture. Inquiry here has historically foregrounded two broad categories of questions. The first relates to the difference between perception (e.g. seeing and hearing) and cognition (e.g. reasoning and planning). Perception and cognition have very different psychological properties – perception tends to be fast (we see things without any noticeable delay), automatic (we see whatever is in front of us when our eyes are open),

and autonomous (we see regardless of whatever else is going on in the mind). In contrast, cognition tends to be slow (happening with noticeable delays), deliberate (we choose what to think about), and dependent on what else is going on in the mind (we can get distracted or have to stop thinking to think about something else). Much work in cognitive architecture has been dedicated to understanding these differences, asking questions like: what marks the difference between perception and cognition? What explains why more superficial differences exist? And, how do these two systems interact? Views in this category highlight potential differences in the formats the two use to represent (Block 2022), in their core functions (e.g. Beck 2018), in the way in which information is accessed (Fodor 1983), or in the contents they can represent at a given time (Green 2020). The second broad category of questions in cognitive architecture deals with the structure of cognition – Is the part of the mind responsible for reasoning and planning unified or disunified? In what ways? And, are there different kinds of cognition? A large space of views have been explored here, with arguments that the mind is unified (e.g. Fodor 2000), composed of isolated parts hardwired by evolution (e.g. Pinker 1997, Carruthers 2007), or broken down into separate belief stores acquired through experience (Bendaña & Mandelbaum 2021), as well as views about the various kinds of thought, including different kinds of belief (e.g. Van Leeuwen & Lombrozo 2023) and different kinds of reasoning and planning, e.g. the famous System 1 – System 2 distinction due to Kahneman and Tversky (Kahneman 2011). Questions of cognitive architecture have consequences throughout cognitive science (what mental processes can we avail ourselves of when explaining various phenomena?), AI (might artificial minds have parts for much the same reason that human minds do?), and decision theory and epistemology (how ought we to reason with minds like ours?).[3]

Where tractability has met cognitive architecture in the past, its treatment has often been driven by intuitions about what makes a problem hard or easy, rather than systematic treatment. This has led to hasty, and, at times, very strange conclusions. For example, previous work has operated under the assumption that information access as a key driver of computational costs. This has led different

[3] Architectural questions relevant for epistemology and decision theory include the extent the mental processes that are rational evaluable (Siegel 2012, Jenkin 2020), the nature of belief and introspection (e.g. Schwitzgebel 2008, Icard 2013), and how we ought to reason with minds like ours (e.g. Rayo & Elga 2021, Friedman 2020).

theorists to conclude that perception must be informationally encapsulated (cut-off) from cognition, that cognition must be broken up into many parts operating over isolated bodies of information, or, as noted above, that the computational theory of mind must be false. These are clear examples of what goes wrong when we lack a systematic understanding of computational tractability as it is relevant for the philosophy of mind. A more systematic framework, like that offered in Chapter 1, changes the landscape significantly. Many of these previous conclusions can be outright rejected (see Chapters 1 & 3). Some are turned on their head – e.g. finding that cognitive influences may help, rather than hinder, perceptual tractability (Chapter 1). And new insights into the mind can be gleaned – e.g. insights into how differences in the information available to perception and cognition can explain differences in their speed and accuracy (Chapter 2), or into how cognition can approximate reasoning over very large bodies of beliefs by reasoning in principled ways over just the most relevant beliefs (Chapter 3).

Now is the time to be thinking about the issues. Progress in AI and computational modeling has reshaped the field in several relevant ways. First, better AI methods are leading to more successful and more diverse models of human mental processes, offering both the opportunity to understand the differences between them and the tools to explore new directions. Second, AI systems are permeating our lives; we increasingly need to be able to reason about what they can and can't do. Our intuitions are all too often informed by glib comparisons between machines and people that don't hold up in practice – an AI system, for example, that can ace a medical exam may nevertheless fail in ways human doctors never would. Systematic comparisons of AI systems and human mental processes can help us build understanding here. Finally, as gestured above, human minds and AI systems face some of the same problems of tractability. These include challenges that the field of AI is starting to grapple with in the pursuit of human-level intelligence. Lessons about how human minds are tractable can help in building such systems as the challenges of scale become apparent.

Chapter 1: How Is Perception Tractable?

Can what we think influence what we see? When we believe or desire something, can that change the way the world looks to us? This debate matters for both foundational questions in

cognitive architecture and for epistemology. Various authors have proposed that immunity to influence from cognition is what differentiates perception from cognition. That is, the two are distinguished by a firewall that permits information to flow only in one direction. If true, this gives us our first handhold on a big division in the mind. On the epistemology side, this question of influence matters for how we can trust perception to supply new evidence for our beliefs. In general, perception seems to be a good way to learn about the world. But if a bottle of mustard looks to me like a lemon only because I believed there was a lemon there, then treating my perception as support for my belief amounts to a kind of circular reasoning.

The empirical literature on whether perception can influence cognition is decidedly mixed. There is a large body of studies, spanning decades, purporting to show effects of cognition on perception. But there have also been large scale rebuttals, demonstrating important flaws that may show up in most or even all of these studies. When the empirical findings are so hotly contested, theoretical arguments take on an important role. One of the core theoretical motivations for the encapsulation of perception is a tractability argument. The thought behind these arguments is that if vision had to wade through the contents of cognition before deciding e.g. if there was a panther in a tree, then vision couldn't happen on the timescales needed for survival (Fodor 1983, Pylyshyn 1999, Mandelbaum 2017, Quilty-Dunn 2019). The number of computational steps needed for a cognition-informed perception then would make perception too slow to be usable.

In the first chapter, I argue that this isn't right. I offer a framework for thinking about computational tractability in cognitive science, drawing on concepts from computational complexity theory, psychophysics, and observations of engineered AI systems. I then use it to show that the computational costs *intrinsic* to visual processing are great enough on their own that accessing information from cognition does not make a meaningful difference to the costs of perception. More specifically, perceptual inference problems (e.g. recovering 3D scene from 2D retinal image) have costs that grow rapidly with the number of independent variables (shape, color, edges) that must be recovered jointly, while psychophysical results suggest that many dimensions are in fact recovered jointly. (One example of this kind of joint recovery is an effect known as Shape from Shading, where

the luminance gradient on a 2D image suggests a 3D shape. There are many such examples.) We can think about the size of perceptual inference problems as lower bounded by the number of jointly recovered variables. This perspective suggests that most of the costs of perception are the costs of perceptual inference. I argue that this theoretical perspective fits with our experience of engineered systems. While inference is often prohibitively expensive and is frequently the bottleneck for intelligence systems, search is cheap and fast, even over large databases. A typical Google search, for example, searches Google's copy of the internet and takes a mere 500ms to run. If this is right, there is no good tractability argument against cognitive effects on perception.

The framework gives us insight into the origin of the computational costs of intelligent behavior, but also into how those costs can be tamed. One of the key resources for managing the computational complexity of inference is prior information about where in the space of possible solutions an algorithm ought to search. I show how this general principle turns the encapsulation view on its head. There may be tractability arguments *in favor* of cognitive effects on perception, where such information can help perception operate quickly.

Chapter 2: Why Is Seeing Fast And Thinking Slow?

Seeing is fast, while thinking is slow. When we open our eyes, we see without any noticeable delay. Canonical cases of cognition take more time, often with noticeable delays (we catch ourselves midthought, and we watch the gears turn in somebody's head). The existence of such a speed difference is widely believed in psychology and philosophy. It is often invoked as the basis for diagnosing new contents in perception – such as arguments that we *see* demographic features, such as age, race, and gender, rather than inferring these from what we see (Colombatto et al. 2021). In light of the failure of encapsulation however (see Chapter 1), we lack an account of *why* seeing is fast and thinking is slow. This puts arguments about perceptual contents on uncertain theoretical footing and deepens the mystery about why perception and cognition should have such different high-level properties. This chapter proposes and defends a computational explanation of the speed difference, using it to shed light on otherwise puzzling findings in psychology and neuroscience.

To start with, I look at the evidence that there is a speed difference between perception and cognition. The idea that there is a difference has broad support in the field, but has not been carefully investigated. The empirical evidence is weaker than one might think. It is extremely clear that perceptual processing happens very fast, but comparable processes in cognition are much less well-studied. I offer what I believe to be the most thorough defense of the speed difference in the literature, while flagging several places where more research is needed to tighten the case.

After exploring the evidence for the speed difference, I turn to possible explanations. Using several case studies as a way to formally ground the argument, I show that several straightforward analyses based on computational complexity fail to explain the speed difference. These analyses only deepen the mystery. It seems perception solves problems that are significantly *harder* than those solved by cognition in important respects, and yet solves them more quickly.

I propose a new computational explanation for the speed difference. The key idea is that there is a trade-off between memory and online computation. When problems are similar enough and have been encountered often enough, relatively more of the work of online processing can be offloaded to learned data structures. I make this idea formally precise drawing on information theory and probabilistic inference methods. Perception and cognition strike different points in this trade off – with perception relying relatively more on memory to solve its problems. Extensive prior exposure to similar problem instances across ontogeny and phylogeny allows perceptual processing to be initialized closer to solutions, requiring much less online computation to deliver highly accurate answers to demanding problems. This strategy is unavailable to cognition, which must maintain a higher degree of flexibility to solve a wider diversity of problems. This gives us a deeper understanding of why perception and cognition differ and adds to our repertoire of hallmarks for telling between perception and cognition. I end by showing how this computational difference helps to shed light on findings in perceptual neuroscience.

Chapter 3: An Architecture for Central Cognition

When we learn new information we can recognize the implications of that information for disparate areas of our lives. When we see an event on the news, for example, we can recognize its significance for our voting behavior, travel plans, or the well-being of a friend. The commonsense view of cognition is that the mind is unified in the sense that we can consider connections between any of our beliefs and are at least reasonably good at noticing relevant connections. On the face of it, computational tractability considerations suggest this commonsense picture ought to be impossible. There are so many possible connections to make between beliefs and only a negligible proportion of these are relevant on any given occasion. Moreover, which connections are relevant is highly context-sensitive – governed by ever changing background beliefs and auxiliary hypotheses. Determining which connections are relevant to draw on a given occasion would seem to require nothing short of evaluating each one – an impossible task. The resulting tractability problem for cognition is sufficiently dire that it has been taken at various times to establish several surprising theses, including that the computational theory of mind must be false, that human-level AI is impossible, or that cognition must be fundamentally disunified, composed of parts dedicated to processing different kinds of contents rather than a singular mind.

This chapter aims to reconcile the commonsense view of cognition with the computational theory of mind. I start by making the above impossibility arguments more precise, showing how they rely on a critical premise, that what is relevant to think about cannot be tractably computed. I then show how certain methods in contemporary AI, large language models (models trained on large bodies of text to produce language semi-cogent prose) provide a counterexample by tractably computing approximate relevance. They do this by trading off extensive prior exposure to the domain with online computation (see Chapter 2).

The ability to tractably compute relevance represents an important step in the direction of an account of tractable cognition, but it is not enough on its own. An important further requirement is the ability to reason in principled ways over relevant considerations. I argue using a range of case studies that language models are lacking here – they fail to reason normatively in many places where

people succeed. This suggests a deep difference in the way that people and language models think and places a further requirement on an account of central cognition.

I propose a new architecture for cognition, which builds on these successes of connectionist models while acknowledging their shortcomings. The key is to exploit the ability to tractably compute relevance to build small, bespoke models tailored to individual tasks. When these models are small enough, principled operations for reasoning (such as bayesian inference or planning algorithms) can be tractably computed over them. Since reasoning or planning over just a small set of highly relevant considerations can approximate a solution to reasoning or planning over a much larger body of beliefs in many cases, an architecture of this kind can approximate the reasoning of a system with much less stringent computational limitations.

A theory of how the mind could be both unified and tractable allows us to reconcile the commonsense view of the mind with a computational theory of mind. It sheds light on the ways in which our minds are both unified and disunified. To a first approximation, we are unified in the sense that we can reason about anything. And disunified in the sense that we can reason only about a small fraction of things at once. I end by drawing out some consequences of this view for epistemology, decision theory, and the philosophy of AI.

Chapter 1: How Is Perception Tractable?

Abstract: Perception solves computationally demanding problems at lightning fast speed. It recovers sophisticated representations of the world from degraded inputs, often in a matter of milliseconds. Any theory of perception must be able to explain how this is possible; in other words, it must be able to explain perception's *computational tractability*. One of the few attempts to move toward such an explanation has been the information encapsulation hypothesis, which posits that perception can be fast because it keeps computational costs low by forgoing access to information stored in cognition. I argue that we have no compelling reason to believe that encapsulation explains (or even contributes to an explanation of) perceptual tractability, and much reason to doubt it. This is because there exist much deeper computational challenges for perception than information access, and these threaten to make the costs of access irrelevant. If this is right, it undermines a core computational motivation for encapsulation and sends us back to the drawing board for explanations of perceptual tractability.[4]

I. Introduction

Perception is hard. It is so hard that one of the main challenges of philosophy of cognitive science is to account for how perception, of the kind seen in people, is possible at all. But why is it so difficult? One reason is that perception solves problems with staggering computational requirements. It delivers reasonable solutions to these problems most of the time. And it does this all with very little of the resources central to computation: very little time, very little energy, very little data. No method in contemporary AI approaches these capabilities.[5] We can call this dramatic efficiency the *computational tractability* of human perception.

---

[4] I am indebted to many people for help developing the ideas in this paper. For comments on (sometimes multiple) earlier drafts of this paper, I'd like to thank EJ Green, Jack Spencer, Alex Byrne, Laurie Paul, Ned Block, Agustín Rayo, Josh Tenenbaum, Bob Stalnaker, Kevin Dorst, and three anonymous referees for this journal. For help editing, I'd like to thank Madeline Medeiros Pereira. For discussions of these ideas and others, I'd like to thank Luke Hewitt, Jon Gauthier, Eric Mandelbaum, Johan Kwisthout, Scott Aaronson, and many others.

[5] For comparisons between human abilities and those of contemporary AI systems, see Lake et al. 2017, Kim, Ricci & Serre 2018, Marcus 2020, Firestone 2020, Jacob et al. 2021.

A theory of perception should explain how perception is computationally tractable. (Previous work in multiple traditions have defended tractability as a general constraint on mental processes and such arguments form the basis for research programs such as bounded rationality (Simon 1997), ecological rationality (Gigerenzer 2011), the tractable cognition thesis (Van Rooij 2008, Kwisthout 2011, 2018, Szymanik & Verbrugge 2018), and Bayesian resource rationality (Griffiths et al. 2015, Gershman et al. 2015, Icard 2018)). To date, however, few potential explanations have been offered. One notable exception is Information Encapsulation Hypothesis, which holds that perception is barred from accessing information stored in cognition. Proponents of the information encapsulation hypothesis often offer a computational motivation, suggesting that encapsulation helps account for the tractability of perception (Fodor 1983, Pylyshyn 1999, Mandelbaum 2017, Quilty-Dunn 2019).[6] Encapsulation, it is thought, explains (or partially explains) tractability by ensuring that perceptual processing does not incur the computational costs of search through large stores of information in cognition, as it would if perception were unencapsulated. Call this the Encapsulation Explanation of Tractability (EET).

In this paper, I argue that we have no positive reason to believe the EET, and many reasons to doubt it. Given what we know about the science of computational costs, information encapsulation seems to be the wrong kind of thing to explain the computational tractability of perception. In particular, encapsulation is ill-equipped to account for computational tractability because there exists a vastly larger problem for perceptual tractability than the cost of information access. I argue that, in light of the true landscape of computational costs inherent in perception, encapsulation can be neither necessary nor sufficient for tractability and is unlikely to even be a difference maker.[7]

If this is right, the implications are threefold. First, while it is still an open empirical question whether perception is informationally encapsulated from cognition, a core motivation for the thesis is cut adrift. This leaves the thesis more dependent on the weight of the psychophysical and

[6] Tractability is not the only motivation for information encapsulation – encapsulation has also been offered as an explanation for striking psychophysical data, such as persistent illusions (see Muller-Lyer illusion).

[7] I'll consider that if encapsulation is any of these (necessary, sufficient, or a difference maker) then it is an explanation of perceptual tractability.

neuroscientific evidence, bereft of a computational *raison d'être*. Since the empirical question is hotly debated (Macpherson 2011, Firestone & Scholl 2016, Lupyan 2017, Quilty-Dunn 2019, Green 2020), the loss of computational motivation matters a great deal to how we view the thesis. At stake in the encapsulation debate more broadly are issues of central importance to epistemology, such as whether perception can be treated as justificatory bedrock (Siegel 2012, 2017, Silins 2016, Jenkin 2020), and to philosophy more broadly, such as whether any distinction can be drawn between perception and cognition at all (Clark 2013) and how that distinction is to be spelled out if so (Phillips 2019, Green 2020).[8]

Second, our discussion places a strong constraint on future theories of perception. At the end of the day, we do not know how perception is computationally tractable, but a deeper understanding of the problem provides a better understanding of what a future solution must look like. I discuss constraints on a future theory of perceptual tractability in Section (VI).

Finally, revisiting tractability arguments for information encapsulation has ramifications for theories of cognition more generally. If systems that are unencapsulated are *thereby* computationally intractable, then the traditional view of a unified mind post-perception is incompatible with the computational theory of mind; the view that mental processes are computational processes.[9] It would follow that either central cognition too must break down into parts that are encapsulated from one another (Tooby & Cosmides 1992, Pinker 1997, Carruthers 2004) or that the computational theory of mind must be abandoned (Fodor 2000). A re-evaluation of the connection between encapsulation and tractability will shed light on what is right, and what is wrong, with such arguments.

The paper proceeds as follows. Section (II) motivates the problem of computational tractability as it pertains to perception. Section (III) presents the solution offered by information encapsulation and the classical arguments for it. A formal definition of computational tractability as it is relevant to debates in the science of mind is developed in Section (IV). Sections (V) and (VI) argue

[8] Information encapsulation is one way in which perception might be modular, but there are others.

[9] I.e. processes characterized by an abstract causal organization that mirrors the stages of a formal computational process (Chalmers 2011), in tandem with whatever relations to the environment are necessary to make some of those states representations (Fodor 1975).

that there exists a much deeper problem of computational tractability than the one encapsulation was designed to solve, while Section (VII) argues that, in light of this, encapsulation can be neither necessary nor sufficient for tractability, and is unlikely to even be a difference maker. Some implications of this for the future of tractability arguments are presented in Section (VIII). Section (IX) concludes.

II. Why There is a Problem of Tractability

A venerable tradition in philosophy and psychology holds that perception is computationally tractable because it is informationally encapsulated from cognition (Fodor 1983, Tooby and Cosmides 1992, Pinker 1997, Pylyshyn 1999, Fodor 2000, Mandelbaum 2017, Quilty-Dunn 2019). In a moment we'll look at what information encapsulation is and how it is meant to address issues of tractability, but before evaluating potential answers, we should get clear on the question. Why should we think that perception has a computational tractability problem in the first place?

For the purposes of this paper, perception is the set of mental processes dedicated to gathering information by way of the sensory surfaces (such as the retina for vision or the cochlea for audition). This includes the final stages of these processes, the perceptual outputs.[10] A good part of what perception does is solve *inverse inference problems*, in which latent causes are recovered or 'inferred' from their proximal effects. In the case of human perception, the latent causes are distal objects and their properties, and their proximal effects are their effects on the sensory surfaces, such as the retina, skin, or cochlea. In the particular case of vision, a set of properties including the shape, orientation, color, and distance of an object must be inferred from their joint effect – an image of colored light projected onto the retina. In nearly all real world cases of inverse inference, the proximal effects underdetermine the distal causes.

[10] For many authors, these outputs are synonymous with perceptual experience, see e.g. Firestone & Scholl 2016, p.1: 'There is a deep sense in which we all know what perception is because of our direct phenomenological acquaintance with percepts – the colors, shapes, and sizes (etc.) of the objects and surfaces that populate our visual experiences. Just imagine looking at an apple in a supermarket and appreciating its redness (as opposed, say, to its price). That is perception... Throughout this paper, we refer to visual processing simply as the mental activity that creates such sensations; we refer to percepts as the experiences themselves, and we use perception (and, less formally, seeing) to encompass both (typically unconscious) visual processing and the (conscious) percepts that result.'

Inverse inference shows up everywhere in the mind, not just in vision. Audition performs inverse inference when it separates out particular voices or other auditory objects from an undifferentiated stream of vibrations, as when listening to someone talk in a crowded room. It is not limited to individual senses either. Perceptual inferences that recover the events associated with sounds take inputs from both audition and vision (a fact responsible for the ventriloquism effect, see Alais and Burr 2004), while the inferences that recover the shapes of objects take inputs from both vision and touch (Ernst and Banks 2002). Nor is it unique to perception. Cognition, by which I mean the set of mental processes of which reasoning and planning are paradigmatic examples, solves similar problems. When we infer that it rained from the fact that the ground is wet (when it could have been the sprinklers), that the neighbor is home from the fact that their car is outside (when they could have left on foot or bike), or the identity of a criminal from the evidence at a crime scene (which is consistent with any number of identities and scenarios), we are solving inverse inference problems. Other examples are less obviously causal, but are formally homologous, such as learning concepts from a finite set of examples, consistent with multiple hypotheses about their content (Feldman 2000, Xu and Tenenbaum 2007, Goodman et al. 2008) or learning the theoretical relations that govern a novel domain (Gopnik et al. 2004, Tenenbaum et al. 2011, Ullman et al. 2012).

The fact that perception and cognition solve inverse inference problems is interesting because these problems are *hard*. They're hard enough that current methods for solving real-world inference problems either take a very, very long time to run (Sokal 1997, Park & Haran 2018) or huge amounts of time, energy, and data to train (Marcus 2020). The most recent work in AI illustrates these challenges. Training state-of-the-art language models (which infer likely completions from portions of sentences, for example, requires data sets on the orders of trillions of words (Brown et al. 2020) and days or weeks of computing time on hundreds or thousands of machines (Narayanan et al. 2021, Chowdhery et al. 2022). Contemporary vision models, which infer 3D-scene properties from images, are similarly compute intensive (e.g. Karpathy 2021).

In contrast, people solve inference problems quickly, cheaply, and with little training data (Lake et al. 2017, Marcus 2020). Why is the human mind so startlingly efficient? How is inference in

the mind possible on the timescales that human-beings solve them? This is the first question of the tractability of the human mind. Call it the question of *absolute* tractability. Answering this question should tell us how computational systems could operate to solve inference problems in roughly the neighborhood of how long people take on those problems. Things that people solve in milliseconds should not take days of compute time. Things that take humans hours to learn should not take months. The question of absolute tractability applies equally well to both perception and cognition.

There is also a second question of tractability which is unique to perception. Even against the backdrop of the computational efficiency of cognition, perception stands out. While cognitive inference problems, such as concept learning, take many trials, encompassing seconds or minutes in the lab (Kemp et al. 2012) or hours or days in classroom and developmental settings (Carey 2009, Ullman 2012), perception solves its inference problems in record speed. For example, perceptual categorization of natural scenes on the basis of category (in this case, 'animal present' or 'animal absent') can be made within 150 milliseconds, as detected by EEG (a measure of the brain's electrical activity; Thorpe et al. 1996), while rapid eye movements or 'saccades,' which require motor planning as well as perceptual processing, can be made on the basis of similar categories in a few hundred milliseconds (Kirchener & Thorpe 2006). Changes in the neural decodability of stimulus information shows that by 350 milliseconds processing in visual areas has largely run its course, with perceptual outputs passed on to frontal, cognitive areas (Marti & Dehaene 2017).

Of course, speed alone is not impressive. It's easy to answer a problem quickly if one is willing to sacrifice performance. In the limit, problems can be answered randomly as quickly as one can roll an internal die. What is remarkable is perception's combination of speed *and performance*. Findings that support the optimality of perception (performance that reaches theoretical limits) are common in the field (e.g. Kording & Wolpert 2004, Ernst & Banks 2002, Weiss et al 2002, see Ma 2010 for a review). Other authors push back (see e.g. Rahnev and Denison 2016). Far less controversial is that perception's accomplishments are both impressive and unparalleled. It represents the world accurately enough that we get by in the myriad tasks we undertake and the diverse and open-ended environments in which we do them. Human beings rarely look at familiar objects and wonder what they are. We can pick out

objects from a crowded visual field, recognize their distances and navigate to them, avoiding obstacles in the process. And we do this in all manner of circumstances: in various weather and lighting conditions, when viewed from different angles, and in novel surroundings. While contemporary machine learning systems can often beat human beings by a few percentage points in speeded classification tasks (Dodge and Karam 2017, Geirhos et al. 2018), human beings are unparalleled in their ability to recover 3D-scene geometry (Spelke and Kinzler 2007), object parthood (Green 2017), physical and relational properties (Wu et al. 2015, Hafri et al. 2013, Little & Firestone 2021), and the consequences of these for high-level features such as stability (Battaglia et al. 2013, Ullman et al. 2017, Hafri & Firestone 2021). Reproducing such accomplishments is the holy grail of computer vision.

There are then two distinct problems of the tractability of perception. The first, the problem of *absolute* tractability, is common to both perception and cognition. This is the problem of how either system is able to accomplish inverse inference on human-like timescales despite theoretical costs and engineered systems that suggest compute times well beyond this (much more on this to come). The other is the question of how perception manages to be so much more efficient (more tractable) than cognition, clocking in at speeds orders of magnitude faster than comparable processes in cognition. We can call this the *relative* tractability of perception (relative to cognition). Theories of perception must explain both the absolute and relative tractability of perception. This is a demanding requirement, but pursuing it vigorously is likely to be productive. Insofar as most theoretical frameworks for perception fail to account for tractability, insisting that a theory does so will help us cull the space of hypotheses as to how perception works.

III. The Encapsulation Explanation of Tractability

An architecture of perception must explain perception's impressive combination of speed and performance, both absolute and relative to cognition. Proponents of information encapsulation endorse the following explanation:

> **Encapsulation Explanation of Tractability (EET):** The computational tractability of perception is explained by the information encapsulation of perception from cognition.

The EET invokes the key concept of information encapsulation, but what exactly is this? Proponents of the thesis write:

> Looked at this way, the claim that input systems are informationally encapsulated is equivalent to the claim that the data that can bear on the confirmation of perceptual hypotheses includes, in the general case, considerably less than the organism may know. (Fodor 1983, p. 69)

> This target article ... defends the position that an important part of visual perception ... is prohibited from accessing relevant expectations, knowledge, and utilities in determining the function it computes – in other words, it is cognitively impenetrable. (Pylyshyn 1999, p.1)

> [This article focuses on]... traditional questions of whether visual perception is modular, encapsulated from the rest of cognition, and "cognitively (im)penetrable." At issue is the extent to which what and how we see is functionally independent from what and how we think, know, desire, act, and so forth. (Firestone & Scholl 2016, p. 2)

> With all this in the background, we can give a more precise characterization of encapsulation: System A is encapsulated from System B when A has a proprietary store of information that excludes information stored in B. (Quilty-Dunn 2019, p.3)

Each of these quotes seem to turn on some common idea, but there is also significant ambiguity. Fodor writes only that the information available to perception is 'considerably less' than is available to the entire organism. Pylyshyn prohibits access of the 'relevant' expectations, knowledge, and utilities, although it seems unlikely that he thought the irrelevant varieties of these could be accessed. Firestone

and Scholl are interested in the 'extent' to which what and how we see is independent from what and how we think, know, and desire, while Quilty-Dunn endorses the generic, that 'system A's information store excludes information (Some of it? All of it?) stored in B.'[11]

While the specifics might be hazy, the gist is clear – the encapsulation of perception means that information in cognition is *verboten* for perception. Fodor thinks that the information available to perception is considerably less than the organism may know because none of the information in cognition is available to it. Pylyshyn points out that the relevant expectations, knowledge, and utilities are prohibited because all of the expectations, knowledge, and utilities are prohibited.[12] Firestone and Scholl are interested in the extent to which what and how we see is independent from what and how we think, know, and desire because they want to defend the view that perceptual processing is not influenced by any of these. Information encapsulation then is a relational property. One system is informationally encapsulated from another when the first is barred from accessing the information in the second. In this case, the relevant kind of information encapsulation is the encapsulation of perception relative to cognition.

The strongest version of the thesis is that none of the information in cognition is accessible to perception. This universal reading is reasonably natural and satisfies the conditions given by each of the quotes. It is further motivated by the content of the papers and chapters from which these quotes are drawn.[13] It is also possible that the universal reading is too strong. Perhaps it is enough if *most* of the information in cognition is barred from access by perception. If this were an empirical paper, with the aim of providing a counterexample to encapsulation, a lot would turn on whether encapsulation theorists are committed to the universal thesis. As it stands, the project of this paper is to show that information encapsulation makes at best a negligible contribution to an explanation of the

[11] In context, it's clear that the Quilty-Dunn quote should receive the universal reading.

[12] At least for early vision.

[13] In each case, what follows are arguments challenging a swath of psychological results that have been taken to evince the the accessibility of information in cognition on perceptual processing, either because such effects are consistent with an explanation citing only information available in perception (Fodor 1983, Pylyshyn 1999), because the effects can be reproduced where the theory of cognitive penetration predicts they should not be (Firestone and Scholl 2016), or because effects that might look like cognitive penetration are in fact mediated by attention, rather than access (Quilty-Dunn 2019).

computational tractability of perception. For this, the strongest version of the thesis will do just fine. If a prohibition on all of the information in cognition is not enough to meaningfully impact tractability, a weaker version of the constraint is unlikely to fare better.

The next question is, *how* is information encapsulation meant to explain tractability? Proponents of the EET write:

> ... speed is purchased for input systems by permitting them to ignore lots of the facts. (Fodor 1983, p. 70)

> One of the reasons theorists have been drawn to modularity theory is its evolutionary rationale.... Roughly, the intuition is that during panther identification what really matters is accomplishing such identification quickly... Searching through everything we know about panthers in order to make an identification would be extremely time consuming. (Mandelbaum 2017, p. 10)

> How are perceptual processes computationally tractable? ... If the processes that solved these problems had to sift through all information stored in central cognition, they would face an unwieldy computational burden... If instead perceptual processes are encapsulated, then they need only check input against their proprietary stores of information ... Encapsulation can therefore provide a unified account of perceptual processes as computationally tractable operations that occur outside of central cognition. (Quilty-Dunn 2019, p. 5)

The thought seems to be that retrieving information takes time, retrieving information from larger stores takes longer, and retrieving the relevant information from a store of information as large as cognition would take *too* long. By foregoing this expense, however, perception can be accomplished tractably.

We can call this basic idea the *Haystack Idea*. Finding a needle in a haystack is a hard problem (hard enough that it has become an idiom for difficulty) and finding relevant information in a huge store of information poses a similar problem. Moreover, needle-in-a-haystack problems get harder as the haystack gets larger.

If we run with this idea for just a moment, we can also get a sense for how the problem 'scales,' or gets harder, as the number of inputs changes. Intuitively, every additional entry makes the problem a little bit harder in expectation. For concreteness, we can think of the set of possibly relevant entries as a list. Entries on this list are information in whatever format one thinks that the mind represents it – this could be a list of beliefs written in the language of thought (Fodor 1975, Goodman et al. 2015), natural constraints (Marr 1982), or parameter values of graphical models (Danks 2014), to name just a few possibilities. Under the pessimistic assumption that this list is unordered – that is, that we do not know in advance where relevant information is to be found (Fodor 1983, 2000) – the expected number of steps needed to find relevant information grows linearly with the number of entries.[14] If the haystack is large enough, search will take too long. Under these conditions, an artificial limit in the size of the haystack, the encapsulation of perception relative to cognition, could explain the tractability of perception. This then is the key idea motivating the EET. The EET holds that the tractability of perception can be explained by avoiding the linear costs of search through the information stored in cognition.[15]

A few clarifications about the EET are in order before we continue. These concern the kind of tractability (relative or absolute) that the EET is meant to explain, the kind of explanation the EET is meant to offer (whether a sufficient condition for tractability, a necessary condition, or a difference

[14] If sampling randomly over the unordered list, the growth in expectation of a geometric distribution with probability of success i/n, where n is the number of entries where the information can be found and i is a constant. Deterministic search over an unordered list (i.e. a list where order is independent of relevance) is equivalent to sampling randomly without replacement, an unnamed distribution which also scales linearly in expectation as n grows. The costs are also linear in the worst case, when all the information must be accessed, in which case the costs increase at a rate of exactly N steps per entry, where N is the number of steps required for access.

[15] Other operations, other than search, scale non-linearly, either in the number of entries or in other parameters. We'll do a deep dive into such costs in the sections to come. For our purposes now, the essential takeaway is that the costs of search scale at worst linearly, even under strong pessimistic assumptions about the efficiency of that search.

maker), and the empirical assumptions the EET requires to get off the ground. I'll look at each of these in turn.

First there is the question of the target of explanation. We noted above that there are two questions of tractability. One asks how perception could be tractable relative to cognition – that is, why perception is orders of magnitude faster than cognition, despite solving mathematically similar problems. The other wonders how perception could be tractable in absolute terms, which is to say, how perceptual processing can happen on roughly human timescales. From what we've seen so far, it might seem that the EET is best suited as an explanation of the *relative* tractability of perception. This is a modest version of the thesis. On this interpretation, the EET is silent on the question of how cognition and perception are accomplished with merely human levels of compute.[16] Instead it merely aims to explain the speed of perception relative to cognition (the difference between, say, minutes for thought and milliseconds for seeing).

Some proponents of EET likely understand the thesis in its modest version. This is just as well, as the immodest version of the thesis leads to some strange consequences. For example, if the price of *absolute* tractability is forgoing access to information stores on the scale of those that exist in cognition, then it follows that cognition itself must be divided into parts, none of which exceeds that critical threshold, or that cognition must be computationally intractable. Interestingly, *both* of these consequences have been endorsed by theorists working in this tradition. Massive modularists, such as Tooby and Cosmides (1992), Pinker (1997), and Carruthers (2007), give up on the idea of a unified central cognition, citing tractability arguments, among others.[17] These theorists prefer a view of cognition on which the mind is really a bundle of independent cognitive entities, each working on certain ecologically salient problems. Opting for the other horn, Fodor himself held that the integrated

[16] Rather than the industrial levels of compute required by today's AI (see Section II) or the astronomical levels of compute suggested by theoretical analyses (more on this in a moment).

[17] Carruthers (2007, p. 44-52) offers the clearest defense of massive modularity on tractability grounds. See also Tooby & Cosmides (1992, p. 106).

nature of central cognitive processing is undeniable and argued on these grounds that a computational theory of mind could never include central cognition![18][19]

How should we understand the EET then? As an explanation of relative or of absolute tractability? For the purposes of this paper, we won't ask the proponent of the EET to commit one way or the other. The argument developed below will show that information encapsulation is not the right place to look for an explanation of either kind of tractability.

Next, there is the question of what kind of explanation the EET is meant to offer. There are a few options here. One could hold that encapsulation is sufficient for tractability: that is, that perception must be tractable if it is encapsulated. Or that encapsulation is necessary for tractability: that perception could not be tractable unless encapsulated. Finally, a weaker version of the EET might grant that encapsulation is neither sufficient nor necessary for tractability, but maintain that encapsulation is nevertheless a difference-maker: that is, that perception would not be tractable were it not encapsulated. (This is different from either necessity or sufficiency. For example, striking a match is not sufficient for lighting a match, since there must be oxygen and the room. Nor is it necessary, since a match can be lit by other means. It is, nevertheless, a difference maker – holding fixed all else about the system, the match would not have been lit but for the striking.) If encapsulation explains tractability, then absent systematic overdetermination, it must at least be a difference maker. Here again, I won't try to pin down exactly which of these versions of the EET proponents have in mind. Instead, I'll argue against all three versions of the thesis. That is, I will argue that encapsulation is neither sufficient nor necessary for tractability, and that there is no positive reason to believe it is even a difference-maker.

Finally, a few words about the empirical assumptions that the EET requires in order to get off the ground, which I'll be granting for the sake of argument. These assumptions fall into two categories.

---

[18] Fodor writes, "Indeed, I am inclined to think that, sooner or later, we will *all* have to give up on the Turing story [of computation] as a general account of how the mind works…" (p. 47). Why? Because "…the computational theory of mental processes doesn't work for abductive inferences" (p. 41). This means that "… a cognitive science that provides some insight into the part of the mind that isn't modular may well have to be different, root and branch, from the kind of syntactical account that Turing's insights inspired." (2000, p. 99)

[19] Examples like these illustrate that some theorists clearly have an immodest version of the EET in mind, but not all versions of the EET lead to this dilemma. If the EET is meant to explain only the relative tractability of perception, then no such conclusions about cognition follow.

One set of empirical assumptions has to do with the distribution of information between perception and cognition. If the encapsulation of perception from cognition is meant to explain the tractability of perception in any of the senses discussed above, then such an explanation turns on contingent facts about just how much information is stored in each. If there is too much information stored in perception, for example, then perception will be intractable regardless of whether it is encapsulated (and so encapsulation cannot be sufficient for tractability). Conversely, if there is too little information in cognition, then eschewing access to such information will be neither necessary for tractability, nor a difference maker. The interest of the thesis therefore depends on facts about the relative amount of information in perception and cognition.

In what follows, I'll grant the encapsulation theorist the empirical facts that they seem to believe: that perception's proprietary store of information is small enough to be tractably searched in the time it takes for perceptual processing to unfold, and that cognition's store is meaningfully larger, such that searching cognition would represent a significant multiplier on the work involved in searching perception alone. (My main interest in this paper will not be in challenging any of these facts, but rather in taking issue with the underlying view of tractability that makes such facts relevant.)

The other set of empirical assumptions required to warrant interest in the EET has to do with the connection between information access and information search. Encapsulation bars information access, and the EET holds that foregoing such access explains tractability by keeping computational costs low. Strictly speaking however, information access costs hardly anything at all – all the relevant costs are the costs of search. To put the idea bluntly: finding a needle in a haystack can be challenging, but if someone gives you a needle from a haystack, receiving it is not difficult. Why couldn't cognition simply send the relevant information for perceptual processing to prime perception in the next moment, obviating the need for an expensive search on perception's part? Cognition could, for example, send a relevant color memory (Hansen et al. 2006, Machperson 2012) or expectation about some other low-level feature (Kok et al. 2012). While this would be a violation of encapsulation, it wouldn't require anything like a perception-initiated, real-time, or full-scale search through cognition, and wouldn't require anything that any party to the debate currently believes to be intractable (no one

denies that people can recall the approximate colors of objects from long-term memory or notice a pattern in the features of serially presented stimuli!). In such a case, the computational costs of a violation of encapsulation would be near-zero.

To connect the negligible costs of access to the more considerable costs of search requires some argument. Maybe evolution opted to prevent all cognitive influences, including ones that are obviously cheap, in order to avoid the costly ones? Such a scenario would be plausible if we assumed that evolution faced the choice between either barring all cognitive influences or barring none, but this idea rests on a strangely dichotomous view of the computational options available. After all, there are many ways in which access could be consistent with non-exhaustive search (some of which are discussed in Section VII) and no a priori reason to think such intermediate solutions are inaccessible to evolution. Be that as it may, we will assume that there is some argument of this type available to the proponent of the EET, as we can make sense of the view only if information access can be wedded to the computational costs of search.

We now have a sense of the breadth of versions of the EET and the empirical assumptions on which the plausibility of the EET depends. In the next section, we'll analyze the concept of computational tractability at work in the EET.

IV. Tractability as an Empirical Bound

We can begin with some points of agreement between all parties. If the mind is computational, then it has some basic operations for manipulating information.[20] These operations could be manipulations of explicit symbols according to rules, as in traditional computers, or the transformation of large vectors of inputs by matrix multiplication, as in contemporary neural networks, or something else besides. Because these operations are implemented in a physical substrate, each instance of an operation takes up some fixed, finite amount of time. It follows that doing too many such operations will take too long; i.e. will render a computation *intractable*. This line of reasoning gives us a simple account of computational tractability as it applies to the mind, which is

[20] See Section I, p. 5 and footnote 6, for discussion of the computational theory of mind.

common to proponents of EET and their detractors.

> **Tractability:** A computational procedure is tractable when it can be completed in fewer than K steps.

A few clarifications are in order. A computational procedure is a finite set of instructions and basic computational operations which define a series of applications of those operations for each instance of a given class of inputs, delivering an output. Crucially, computational procedures must function without recourse to anything but their inputs, instructions, and basic operations (see Turing 1936). A computational procedure may be *branching* in the sense that it doesn't have to execute the same series of operations every time. It can treat different inputs differently (say, running a distinct series of operations for odd numbered inputs as for even), and could even be stochastic, making random choices at predefined points in its execution.

Computational procedures are the only way we know of to solve computational problems. A computational *problem* is a set of inputs, a set of outputs, a set of ordinal or metric structures over those outputs, and a mapping from inputs to a given structure over outputs. The metric or ordinal structure over the set of outputs reflects the fact that answers to computational problems are not always right or wrong, but are often better or worse than one another. Better procedures are those that deliver better performance on a problem.

A computational problem can also be tractable or intractable. A computational problem is intractable, relative to a performance criterion, when no procedure can tractably solve that problem to that performance specification. This bit of relativism is necessary for a meaningful notion of tractability, as there are few limits on how quickly an answer can be computed in the absence of non-trivial criteria for how good an answer it has to be (see Section II). Criteria may include how close one is to the right answer, how often one gives the right answer, the class or proportion of problem instances for which one gives the right answer, or any combination of these.[21] The performance

[21] An acceptable criterion for performance for visual estimation of distance for example, could be that the visual system delivers an answer within 20% of the true value, 90% of the time, when presented with an object in good light.

criterion relevant for our purposes is that of *human-level* performance. To explain how, say, visual inference is tractable, is to explain how it could be computed to human-level performance in fewer than K steps.

Finally, we might wonder, what is K? For our purposes we can imagine that, for a particular problem and performance criterion, K is some fixed number, determined by how long it takes a human being to solve that problem to that performance criterion. For a given problem and performance criterion, there are many things that might affect how big K is. One is the speed of the relevant basic operations in the human brain. Faster operations permit a higher K. Another is parallelization. Some problems have parts that can be solved by parallel trains of operations, increasing the number of operations that can be packed into a unit time. The speed of basic operations in the brain and the extent to which parallelization is employed are both questions outside the scope of this paper. In light of substantial uncertainty about these values we should err on the side of liberality when setting K, so as not to prematurely eliminate hypotheses about the mind that we can't be sure are intractable. In other words, we should allow that K for many human perceptual processes may be quite large. It will not, however, be *astronomically* large.

What counts as astronomically large? We can gain something of a foothold on this concept by starting with the capacities of today's supercomputers. Today's fastest supercomputers perform on the order of 10^17 operations per second. To do this, they run thousands of processors, occupy whole complexes, and consume vast amounts of power. For the purposes of this paper, we'll say that anything that would take one billion supercomputers one billion seconds (~115 days) to compute (that is, >10^30 operations) is 'astronomical.' Trivial as it sounds, we will see that the requirement that perceptual processing not require astronomically many steps will turn out to be a constraint with some teeth.

Taken together, an explanation of the tractability of perception is an explanation of how perception is accomplished without astronomical costs. Over the next two sections, however, we'll see that plausible assumptions about the costs of perceptual inference actually do entail astronomical costs. To explain the tractability of perception then, a theory must explain how these assumptions can

be denied and these costs avoided. This will give us a positive framework for thinking about tractability. Finally, we'll see in Section (VII) that information encapsulation is ill-equipped to contribute meaningfully to an explanation of tractability in light of all this.

V. Inference & Scaling Behavior

A theory of perceptual tractability must explain how perception is accomplished, relative to human-level performance, by some computational procedure that takes less than astronomically many steps. If we call the number of steps needed to solve a problem to the relevant performance criterion M, then a perceptual inference is tractable when $M < K$. But what properties of a problem contribute the most to M for the computational procedures that solve it? And, in particular, what properties put us at risk of astronomically large M?

Theoretical computer science can be a source of insight here. One branch of theoretical computer science, Computational Complexity Theory (CCT), reasons about computational costs through the lens of how M grows, or 'scales,' with different properties of a computational problem, including most famously the number of inputs to a problem, but with theoretical extensions to include considerations of performance criteria and distributions over inputs. Co-opting some of the core concepts of this field will help us better understand our own notion of tractability. (For more detailed introductions to CCT then I can provide here, see Sipser 2012, Arora and Barak 2007, or Goldreich 2008.)

CCT and Scaling Behavior

CCT taxonomizes computational problems according to the functional form of the 'scaling behavior' of a problem on the number of inputs. The scaling behavior of a computational *procedure* is the relationship between the number of inputs and the number of steps the procedure goes through for those inputs, while the scaling behavior of a *problem* is the behavior of the most efficient procedure for solving it. To get a feel for how scaling behaviors differ between problems, imagine attempting to plan a wedding given a guest list. If you want to know whether you'll need chairs or not, all you have to

do is check whether the list is non-empty. This is a constant time operation; and this takes equally many steps no matter how long the list is. If you want to know how many chairs you'll need, by contrast, then you need to count the number of names on the list. This is a linear time operation; it requires a number of steps that is a linear function of the length of the list. Finally, if you want to know what seating arrangement will maximize the well-being of your guests, allowing old friends to catch up, kindling new romances, and avoiding explosive tiffs, you'll need to consider every way your seating chart could be arranged. This is an exponential time operation. Such operations tend to be sticking points in our lives, as in the lives of computers. No one complains about having to count the number of guests on a list, but planning the seating can be a nightmare.

CCT treats this difference between exponential and sub-exponential scaling as a difference in kind rather than degree. It does this because, for moderately sized inputs and beyond, the contrast between exponential and sub-exponential scaling often separates operations that can be feasibly computed, even at significant cost, from those that cannot. For example, in the wedding case above, if your guest list contains 90 people, then checking whether you'll need seating takes 1 step, while counting how many chairs takes 90. If you can set up at most 10 tables of variable size for your guests, then finding the optimal seating arrangement requires on the order of 10^90 steps, at least one for each of the unique possibilities that must be considered. 10^90 is a big number; it's more than the number of atoms in the known universe. For the purposes of practical computation, it might as well be infinite. That's why CCT treats this difference in degree as an effective difference in kind.

A few clarifications. First off, not all inputs to a computational problem contribute equally to the cost of solving it.[22] We'll discuss this at length in the case of perceptual inference in the next section. Second, CCT makes strong assumptions about the performance criteria relative to which costs are assessed – most famously requiring guaranteed performance on all (and therefore the most difficult)

[22] For example, the computational costs of determining whether a formula in propositional logic is satisfiable is exponential in the number of literals that appear in the formula, but not exponential in the length of the formula. More detail on this area of research, known as Parameterized Computational Complexity Theory, can be found in Downey & Fellows (2013) and Flum & Grohe (2006). See Kwisthout (2011; 2018) for an overview of key results related to cognitive science.

problem instances — and this limits its relevance to our project.[23] Relatedly, CCT doesn't model aspects of problem structure that might make certain problem instances easier or harder. Where particular classes of problem instances have additional structure, that structure can sometimes be exploited to make a problem in that class easier than the complexity of its super-class would suggest.[24] When it comes to the project of understanding computational complexity as it applies to theories of the mind, we'll take what we can use from CCT and leave what we can't.

The key thing we will keep is CCT's focus on scaling behavior. This simple idea is both deep in what it reveals about the nature of computational problems and crucial for the task at hand. To be fit for our purposes, however, the concept of scaling behavior drawn from CCT will have to be both broadened (to include more diverse performance criteria) and refined (so as to be applied to classes of instances that have exploitable structure). Our CCT-inspired examination of scaling behavior will give us a place to start in examining what properties of perceptual inference problems entail which computational costs.

The argument will proceed as follows. We'll begin by establishing some general facts about the way that the *hypothesis space* of an inference problem grows as a function of the *dimensionality* of that problem. What we mean by these words will be made clear in due course. We'll find that the hypothesis space grows *exponentially* as a function of dimensionality. Under some simple starting assumptions, this exponential growth translates into exponential growth in the costs of computing inference. We'll then see that these assumptions can be substantially weakened, leaving the main result intact. In the following section, I'll argue that this exponential growth in the costs of perceptual inference, combined with the actual dimensionality of real life perceptual inference problems, suggest astronomical costs for perceptual inference. These costs dominate anything else in perceptual processing. One consequence of

[23] Other branches of CCT look at average performance assuming simple distributions over inputs. This too is unlikely to be the kind of performance criterion most relevant to a computational theory of the mind, since ecologically realistic distributions over inputs are often not simple and generally lack closed form expression.

[24] Backtracking Satisfiability (or 'SAT') solvers (e.g. Davis & Putnam 1960, Davis et al 1962) are a classic example of a strategy that exploits local problem structure to find a solution more quickly. Since a minority of SAT cases exhibit global structure that frustrates such strategies, SAT exhibits a disconnect between theoretical intractability and computational procedures that are tractable for most purposes.

this is that explaining the tractability of perception requires explaining how some of these assumptions can be credibly rejected so as to avoid astronomical costs. Another is that avoiding the costs of information access (the focus of the Haystack Idea) is unlikely to make an important contribution to the tractability of perception.

Scaling Behavior of Inference

The first concepts we'll need are those of a *hypothesis space* and the *dimensions* that define it. Solving an inference problem requires finding one or more good hypotheses about how the world might be from the set of all the ways the world could be, at least by the lights of that inference problem. In the wedding planning example, 'all the ways the world could be' includes all the ways that people at your wedding could be seated. (The problem is 'blind' to many other ways the world could be, such as how the astronauts in the International Space Station might be seated.) Hypotheses, or candidate solutions to the inference problem, differ from one another in their assignments of values to variables, such as people to tables. These variables can be thought of as the 'dimensions' of a space (the 'hypothesis space'), the values as coordinates along those dimensions, and hypotheses as unique points in the space.[25] In the wedding planning example, each attendee is a 'dimension' which must be assigned a value, in other words, a table. In the case of visual, perceptual inference, which we'll get to shortly, the hypothesis space is given by all the ways the objects in a scene could be – their colors, shapes, locations, etc.

One thing to notice is that the dimensionality of an inference problem (which dictates what hypotheses can be represented) comes apart from the information we bring to bear in solving that

[25] Here I am eliding the question of whether to think of the hypothesis space as a semantic feature of the problem (ways the world could be) or as a syntactic feature of the representation of the problem (ways the world could be from the perspective of a procedure). In the absence of meaningless or synonymous expressions and assuming that all relevant hypotheses are expressible, there will be a 1-to-1 correspondence between hypotheses in the syntactic sense and in the semantic sense. Assuming that a problem can be fully represented then, deviation from this 1-to-1 correspondence comes when there are more syntactic hypotheses than there are genuine possibilities. Since a computational procedure can only operate over syntactic hypotheses, such deviations create additional costs. On the assumption that all semantic differences can be represented then, we can treat the semantic dimensionality of a problem as a lower bound on the cost-driving, syntactic dimensionality of a computational procedure for solving it.

problem (what is known about those hypotheses). Adding a new dimension – say, kind of chair – allows the system to formulate new hypotheses (should I seat Veronica and Ezra in bean bags?), while information changes the assessment of quality of various hypotheses (I might know that Matthias would not like being sat at the kids' table). The distinction here is not idle. While being able to represent a dimension offers a natural way to represent information about that dimension, a system can also represent a dimension without having any information about it.[26] Similarly, a system can make use of information that is encoded in dimensions it does not represent. An example of the first would be if the visual system could represent colors and object categories, but was encapsulated from relevant information in cognition about the colors of known objects. In this case, vision possesses the dimension of color, but lacks information about it. An example of the second would be if the visual system could represent color and face identities, but not party affiliation. Cognition, for its part, might know that a particular person is a republican and that republicans are likely to wear red ties. We can imagine that cognition sends a visual expectation down to vision about the color of a tie in response to a perceptual output recognizing the face. In this case, vision would possess information about the dependency between identity and tie color, while lacking the dimension of party affiliation that introduces the connection. This distinction, between information and dimensions, will be critical in what's to come since the costs of inference are sensitive to the two in very different ways.

How exactly are the costs of inference related to dimensionality? To ease into this, think first about how the set of hypotheses grows as new dimensions are added to the space. When we add a new dimension, each possible value of the new dimension combines with every previously complete hypothesis to deliver a new set of unique hypotheses. So, for example, if we add 'kind of chair' to our wedding planning problem, then where we previously had a single complete hypothesis (a total assignment of people to tables), we now have a set of hypotheses; every possible combination of assignments of people to types of chairs, consistent with a table assignment. Just like adding a new dimension to a real coordinate space, this produces exponential growth in the set of hypotheses.

[26] If some information is necessary for concept possession, then read this as 'without any further information than is necessary for representing the dimension'.

So our set of hypotheses grows exponentially as the dimensions of the problem are increased. But how is this tied to the actual costs of performing inference? Some simple assumptions will deliver the result that exponential growth in the hypothesis space produces exponential growth in the costs of inference.[27] I'll first present these assumptions and then look at ways they might be weakened.

First, consider performance criteria. We saw that computing inference requires finding 'good' hypotheses from within an exponentially growing hypothesis space, where the goodness of a hypothesis consists in its probability, plausibility, or explanatory import. For the moment, take finding a 'good' hypothesis to mean finding the 'best' hypothesis. Next, assume that it costs at least one computational step to evaluate a hypothesis and only one hypothesis can be evaluated at a time.[28] Finally, assume that nothing is known beforehand about the relative or absolute distribution of good hypotheses throughout the space. That is, the only way to find out whether a hypothesis is any good is to evaluate its plausibility relative to a prior and the data.

When these assumptions are met, the computational costs of doing inference grow linearly with the number of hypotheses and therefore exponentially with the number of dimensions defining the hypothesis space. This is because hypotheses must be evaluated in order to determine their performance, and must be evaluated in some order that is independent of the performance of the hypotheses (since having access to an order that privileges better hypotheses would violate the assumption that nothing about hypothesis performance is known beforehand). This entails that the number of hypotheses that must be evaluated grows linearly in expectation with the number of hypotheses in the space, and therefore exponentially in the dimensionality. Note that this holds

[27] The thought here is that the costs of inference are an exponential function of the dimensionality of the problem (I'll show later that this, combined with the (possibly fixed) dimensionality of perceptual inference problems is sufficient to put the costs of inference in astronomical territory). The talk of exponential 'growth' in costs is merely meant to bring out this functional relationship. It will not be important to my argument whether the dimensions of perceptual inference can continue to be increased (in fact, they may be architecturally barred from doing so for just this reason; See Section VIII).

[28] We can actually get by with a much weaker assumption, i.e. the assumption that there are no exponential speed-ups in the number of hypotheses that can be evaluated at a time (either as a function of the amount of time spent reasoning or the number of dimensions in the hypothesis space). That is, we can get by with the assumption that there can be no finite equivalent of super-tasking: evaluating one hypothesis in c steps, the next in ½c steps, the next in ¼c steps, and so on. I use the stronger assumption that evaluating a hypothesis costs one computational step because it will considerably simplify the presentation of the argument in the following section.

whether we are sampling randomly (with or without replacement[29]) or evaluating hypotheses in a predetermined order (which, since it cannot be relied on to privilege the best hypotheses in the general case, might as well be a random order). Finally, since the number of hypotheses that must be evaluated in expectation grows exponentially as a function of the dimensionality, and since the costs of evaluating a hypothesis are constant, the costs of evaluation grow exponentially as well.

This line of reasoning establishes exponential growth under these assumptions, but some may find the assumptions troubling. The performance criterion is a particular sticking point. While many have argued that human perception is optimal, in the sense of finding the best solutions to its inference problems (Ernst & Banks 2002, Weiss et al. 2002, see Ma et al. 2010 for a review), others have argued against this perspective (e.g. Rahnev & Dennison 2018). We can, however, weaken the performance criterion in reasonable ways while maintaining the main result. Imagine, for example, that instead of finding the best hypothesis for a given problem, human perception finds hypotheses that are merely 'good enough', in the sense that they are close enough in value to the best hypothesis along each of the dimensions of the problem. In this case, we might count as a satisfactory answer any hypothesis within 3% of the value of the best hypothesis along each of the relevant dimensions. This has the effect of turning a solution set from a point to a contiguous region in the hypothesis space. Such relaxations would certainly make these problems easier to solve, but they do not resolve the more fundamental issue of exponential scaling. To see this, imagine solving an inference problem to this 'good enough' standard of performance. Even in this case, the proportion of hypotheses meeting this criterion shrinks exponentially as the dimensionality of the space increases. For one dimension, 3% of samples will meet this criterion. But for 3 dimensions, that proportion is 0.0027%, for 6 it's 7.3 x 10^-8 %, and so on.[30] Here, as above, the proportion of good hypotheses becomes vanishingly small, and reasonable assumptions about the costs of evaluation will entail astronomical costs for finding those hypotheses. (This example also illustrates how exponential scaling generalizes to continuous hypothesis spaces,

[29] Assuming, for simplicity, a uniform distribution over hypotheses, the expected number of hypotheses sampled before finding the best corresponds to the Geometric distribution with exponentially decreasing probability of success as dimensions are added. Sampling without replacement yields another distribution that also grows linearly in expectation as the number of non-best hypotheses grows, and therefore exponentially in dimensionality.

[30] That proportion is given by the equation (3/100)^n.

where in the continuous case, as in the discrete case, the proportion of the measures of the solution set and the problem set shrinks exponentially.[31])[32]

There are, of course, many ways to weaken the performance criteria, and we've only looked at one. It may be that some of these ways avoid exponential growth in the costs of inference while still delivering human-level performance. This is, however, not where I'd put my money. Human performance on perceptual inference tasks is excellent (see Section II). It seems for this reason that weakening the performance criteria to such an extent that hypotheses that meet those criteria will be easy to come by, even in astronomically large hypothesis spaces, is a non-starter. Instead, we'll have to ask which of our other assumptions can be given up, in particular the assumption that nothing is known in advance about the distribution of promising hypotheses. This will be a focus of a later section (Section VII). For the time being, we need to show that this theoretical result of exponential scaling actually is sufficient to push the costs of perceptual inference into astronomical territory when certain empirical facts about the dimensionality of perceptual inference are considered.

VI. Dimensionality of Perceptual Inference

I've argued that, under reasonable assumptions, the costs of inference scale exponentially in the dimensionality of the problem. But what does all this mean for the tractability of real world perceptual inference? To know what conclusions we should draw requires developing a rough idea of the dimensionality of perceptual inference and the proportion of perceptual hypotheses that satisfy human-like performance. I'll argue that conservative assumptions about dimensionality, and liberal assumptions about the proportion of hypotheses that satisfy human-like performance, combined with

[31] The challenges of exponential scaling are also robust to reparameterization. While in either discrete or continuous cases one could always map the hypotheses from N dimensions onto a single dimension, such a trick would change neither the relative cardinalities of the solution set to the problem set in the discrete case nor the proportion of the measures in the continuous case. (In general, a syntactic representation of the problem that differs from the problem's intrinsic dimensionality can make the solving the problem more difficult, but it cannot reliably make it easier. See footnote 22.)

[32] Note that the astronomical costs of inference hold even if, as is believed, the brain is massively parallel (see Section III, p. 30). Since parallelization can deliver at best a factor of N speed-up, where N is the number of parallel processes, the exponential increase of costs is maintained regardless. Unless the number of parallel processes is itself astronomical (much greater than the estimated 10^7 neurons in the brain), parallelization won't make a difference in the analyzes to come.

our assumptions so far, entail astronomical costs for perceptual inference. I'll make this case by presenting a toy visual inference problem, involving just a few of the many dimensions that vision represents.

A Toy Inference Problem

Consider a scenario in which I open my eyes to see a simple scene of static objects. Each object has a color, a lighting condition, a location in three dimensions, and a shape. We can set some numbers to these possibilities. Perhaps there are a million (10^6) colors we can see,[33] another million (10^6) ways the lighting could be (Tokunaga and Logvinenko 2010), perhaps a billion possible locations (10^9), and another billion (10^9) possible shapes.

These are conservative figures. Stipulating a billion possible locations amounts to assuming that there are a mere 1000 just noticeable differences in location across each of 3 dimensions – a modest estimate of human spatial acuity.[34] In the case of shapes, a mere billion discriminable possibilities across all the configurations of shapes and sizes perceptible to human beings is a gross underestimate. Even with such conservative numbers, however, the different combinations deliver 10^30 ways an object could be. If there are 3 objects in a scene, the number of possible scenes is 10^90. Here again, this number is greater than the number of atoms in the known universe. For practical purposes, it may as well be infinite.

We saw earlier that loosening up the performance criteria within reason does not change the exponential decrease in the proportion of viable solutions. But where do these considerations get us in the case of perceptual inference? We already assumed that the divisions were coarse-grained (with just

[33] Estimates range from 1-10 million.

[34] Just Noticeable Differences (JNDs) are the smallest differences that provoke above chance discrimination. Here I'm assuming for the sake of simplicity that there are an equal number of JNDs across each dimension. Distortions in visual space may mean that this is not quite right (Green & Rabin 2019). Note that the connection between perceptual hypotheses (distinct internal representations) and discrimination is not direct – distinct representational states are the competence to discrimination's performance. Discrimination is accomplished by mapping equivalence classes of stimuli to distinct representational states. Distinct representational states can, however, exist without showing up in discrimination, say if insufficient light, damage to the retina, or other peripheral constraints impair performance. Discrimination then places a lower bound on the number of distinct representational states; the value relevant for inference.

1,000 JNDs in location along each spatial dimension). But we can go further. Let's say that any hypothesis is acceptable so long as it falls within a range of 3% of the best hypothesis, along each dimension in the hypothesis space. Assuming that color, lighting color, location, and shape each involve three dimensions, the proportion of hypotheses satisfying this condition for a scene with three objects would be on the order of 1 in 10^54 – still well within astronomical territory.[35]

In setting this up I have said nothing of numerous other dimensions represented in vision, including low-level dimensions such as edges, as well as many high-level contents, such as motion (Weiss et al. 2002),[36] object identity (Quilty-Dunn 2019), causality and animacy (Scholl & Tremoulet 2000), or hierarchical part structure (Green 2017). I have also neglected dimensions from other modalities which participate in inference in cross-modal perception (Green 2021) and cue integration (e.g. Ernst and Banks 2002). Each additional dimension should be expected to make an exponential contribution to the problem size.

Individuating Inferences

One thing we haven't discussed yet is how to individuate inference problems. As it turns out, this question matters a great deal. This is because inference problems are *much* more than the sum of their parts. So far we've been assuming that if perception represents the dimensions of color, lighting condition, shape, and location, then it must recover these in a single inference problem. But recovering them in a set of smaller inference problems is exponentially less costly.[37] Imagine, for example, that perception were to solve two inference problems, one to recover the color and lighting condition of an object and another to recover its shape and location. Using the same figures we used above, but recovering the surface color and lighting color for three objects and, separately, three object's shapes and locations, would deliver a hypothesis space of approximately 10^51 hypotheses; about a million

[35] Here I am calculating (3/100)^36 – or 12 dimensions over 3 objects. This assumes that shape representations are parameterized along 3 natural, continuous dimensions. This is almost certainly not the case. The actual dimensionality will be higher, and the resulting proportion of acceptable hypotheses will be smaller. We are considering 'astronomical' anything > 10^30, see Section IV.

[36] Which is not just successive location (see the waterfall illusion).

[37] Based on our assumptions so far.

trillion times fewer than the 10^90 that results if we jointly solve for all of these dimensions.[38] These savings only get more dramatic as the overall dimensionality grows.

One might wonder whether perception could limit the costs of inference by adopting a divide and conquer strategy of this kind, in effect holding that perception is composed of many distinct modules responsible for each of the different sets of properties discussed above. This broad outlook on vision was made famous by Marr's foundational work on vision (Marr 1982) and has many contemporary adherents.

The problem with such an architecture is what is lost when larger inference problems are broken up into smaller problems in this way. In such cases, the sum of inference problems is no longer sensitive to the *dependencies* between the dimensions housed in separate problems (more on this in a moment). The loss of sensitivity to these dependencies matters because human-level performance requires this sensitivity (otherwise, color, shape, and location cannot be accurately recovered), and, unsurprisingly, human vision empirically exhibits it (as will become clear shortly). The rest of this subsection will spell out this reasoning more carefully.

Dependencies, as may be clear from the above, are the relationships between dimensions such that information about one dimension bears on the probable values of another. Sensitivity to dependencies is necessary if inference is to arrive at an internally consistent percept. For example, if one large object stands between another object and a scene's source of illumination, then the second object is likely to be cast in the first's shadow. This in turn influences how the intensity (and spectral profile) of the light reflected off the second object is interpreted, as object color or lighting condition. Conversely, if the light bouncing off an object of unknown location is reflecting light that is darker than expected, this could be evidence that the object is in shadow, providing information about its location. Such dependencies between dimensions (in this case location and lighting condition) are invisible when inference problems are broken up into their component parts. In such cases, assignments of probable values of color must be made independently of assignments about location, leading to inconsistency.

[38] For color and lighting color, that's ((100)^6)^3 = (10^11)^3 = 10^33. For shape and location, ((1000)^6)^3 = (10^17)^3 = 10^51. 10^33 + 10^51 ≈ 10^51.

If such inconsistency is kept to modest levels, it might be a reasonable price to pay for tractable inference, but it does not seem to be the strategy that human perception takes. This is because perception is, in fact, sensitive to a great many dependencies between perceptible dimensions, including dependencies between all of the dimensions used in the toy example above. Sensitivity to these particular dependencies can be seen through a series of established psychophysical results. (Such results will naturally not show that perception is sensitive to the dependencies between all of the dimensions it represents, but will show that a divide and conquer strategy is insufficient for tractability, as the dependencies which are represented are sufficient to establish astronomical costs given our other assumptions.)

Start with color constancy – Objects in the world are seen as having a stable color, despite changes in lighting condition between indoors and out, across changes in weather and time of day. This fact is quite surprising when one considers just how much the light hitting your eye differs under these conditions. A lump of coal in bright sunlight reflects about as much light as white chalk indoors, but the chalk appears bright white and the coal jet black. This phenomenon, known as color constancy, is accomplished by jointly inferring color and lighting condition so as to find a consistent assignment of values to those dimensions (Tokunaga & Logvinenko 2010). If a lot of light is hitting the retina, for example, this could be because the object reflects most light (as in the case of chalk) or because it is intensely lit (as in the case of coal in bright outdoor light). By doing joint inference over these dimensions, perception can ensure that it is not double counting the properties of the proximal stimulus – which might result in seeing the coal outside as bright white. Color constancy then, is perceptual sensitivity to the dependency between color and lighting condition.[39]

Just as with color and lighting condition, all four of the dimensions we've discussed so far are jointly confounded in the retinal stimulus and so conditionally dependent on one another. For example, different shapes in different lighting conditions give rise to different patterns of coloration across an object. If information about probable lighting sources is present, either from a prior or from further cues in a scene, then the pattern of coloration can be used to infer the object's shape. In a

[39] Really, the conditional dependency between color and lighting condition, conditional on a given retinal input.

phenomenon known as 'Shape from Shading,' the visual system does just this. A classic study showed participants 2D shaded circles, either darker on the bottom and lighter on top or vice versa. Participants saw the light-on-top circles as convex 3D reliefs while seeing the dark-on-top circles as concave recesses, demonstrating both a visual prior that light comes from above and a sensitivity to the dependency between lighting condition and shape (Ramachandran 1988, see Figure 1 for illustration). Sensitivity to the dependencies between color, lighting condition, and shape extends to the location dimension and to the properties of other objects as well. When multiple shaded objects provide further cues to lighting direction, participants can be induced to assign different locations to an unobserved lighting source (Morgenstern et al. 2011). Similarly, scenes with cues suggestive of multiple lighting sources induce global percepts of objects with shape properties consistent with those lighting sources (Wilder et al. 2019).

Collectively, these effects illustrate perceptual sensitivity to the dependencies that exist between color, lighting condition, shape, and location. Insofar as color and lighting condition are jointly dependent on one another (by color constancy), lighting condition is dependent on shape and location (by light source, shadow, and mutual illumination), and the locations of objects and light sources are dependent on the shape and color of objects (by the flexibility of the illuminant prior), there cannot be any consistent independent recovery of these attributes. Rather, they must be recovered jointly.

A vivid illustration of this joint inference can be found in the bistable chromatic Mach card (Bloj et al. 1999, Harding et al. 2012). In this effect, a folded card with two colored sides is shown to participants. One side is painted white and the other magenta. The card is folded in a concave fashion, with the edges of the paper protruding, and presented to the viewer head-on. Viewed at this angle, the card can be seen as either concave or convex. Because the card is actually concave, the two sides mutually illuminate, with light from the magenta side casting a pink glow on the white side. When participants see the card as concave, all of this is perceived veridically – the card looks concave, the sides white and magenta, and the white side cast in pinkish light. When participants see the card as convex however, one side is perceived as magenta and the other side as light pink (i.e. having a light pink surface color). In this case, the pinkish coloring that vision had originally attributed to mutual

illumination between two facing sides is now seen as the much darker surface color of a second painted side. The bistability of the chromatic Mach card vividly illustrates human visual sensitivity to the dependencies between color, lighting condition, shape, and location (mediated by mutual illumination).

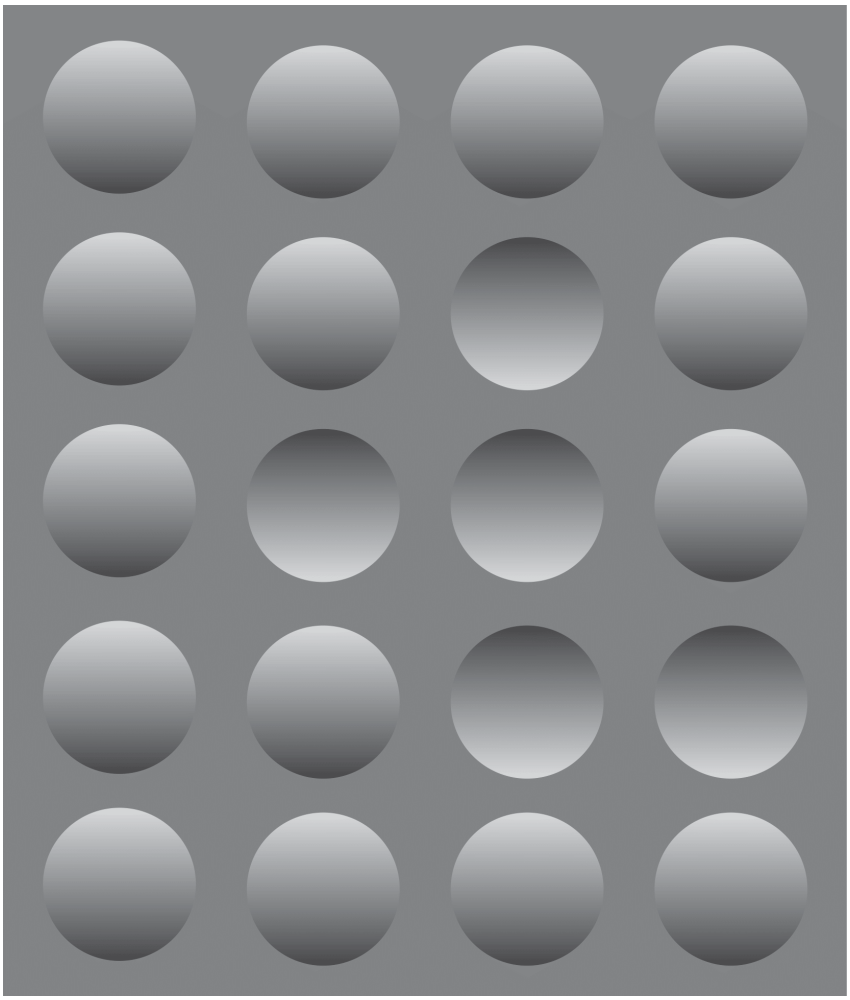


Figure 1: Typical shape from shading stimuli – Shape (either convex or concave) is assigned to multiple objects under the assumption of a single illuminant governing them all. This assumption is defeasible, as discussed in (Morgenstern et al. 2011, Wilder et al. 2019).

Sensitivity to these dependencies shows that human perception cannot be using a simple divide and conquer strategy to head off exponential costs. But what about a modular strategy followed by a

recombination stage? There are lots of ways that such a strategy could work, but they all fall into two broad categories – independent computation of dimensions followed by *principled* combination of those values into a coherent hypothesis, and independent computation followed by *heuristic* combination. We'll look at each of these in turn.

Take the first case, of independent recovery followed by principled recombination. When this strategy is deployed, the problem is first broken up into small subsets of dimensions which are jointly inferred, with exponential savings for breaking up the larger inference problem. The outputs of these sub-inferences are then recombined into a full hypothesis in some principled fashion, such that the end result is the same as if inference had originally been computed over the full set of dimensions. Illustrative examples of this approach come from the literature on 'Bayesian cue combination.' In a typical Bayesian cue combination study a model is proposed on which independent measurements of some perceptual dimension are combined in a way that is sensitive to the uncertainties in each measurement. These independent measurements are then combined analytically, often by multiplying gaussians. In one famous study, due to Ernst and Banks (2002), subjects were asked to assess the height of an object presented to them both visually and haptically. This was accomplished by allowing subjects to simultaneously touch an object with their hands while viewing it through a window of varying opacity, blurring the image of the object beyond. The authors showed that subjects' judgements of size reflected information from both vision and touch. Intriguingly, subjects' final judgements were also sensitive to the uncertainty in each of the input modalities, with the more certain (lower variance) channel having a greater 'weight' in the final judgment. Vision was relied on more by default, but subjects' judgements reflected greater weight placed on haptic information as visual inputs were made noisier (by increasing opacity of the viewing window). Finally, the uncertainty of subjects' final judgements was always less than the uncertainty of the measurement from the more reliable modality, suggesting that information from both channels was in fact being integrated, rather than information from the less reliable modality being thrown away.

What's interesting about this work for our purposes is the way in which information is integrated. In these models, inference (the process of considering and evaluating hypotheses discussed

above) is entirely eschewed. Rather, information is combined analytically – in this case by multiplying two normal distributions representing independent visual and haptic measurements of the relevant value.[40] When measurements are combined analytically in this way, the full costs of inference are avoided, leaving only the costs of inference over the subsets of dimensions combined together and the trivial cost of multiplying gaussians.

Despite the promising start, approaches of this kind face several problems that severely limit their generality, and hence their viability as models of perceptual inference.[41] Here I'll focus on just one such problem. In cases of Bayesian cue integration, an analytic solution to integrating the outputs of partial inferences is available only when integration is mandatory. So, in the case of Ernst and Banks above, subjects' perceptual systems were able to recognize that the haptic and visual input came from the same object, and so it made sense to integrate information from both senses. But we often find ourselves touching and viewing distinct objects, and in these cases we do not integrate information from haptic and visual channels (Kording et al. 2007). The question then is, how does perception know which case it is in (whether the objects are distinct or identical) and therefore whether it should integrate? Models of this integration-decision require nothing less than full inference over the relevant dimensions to determine whether a single cause of haptic and visual inputs, or distinct causes, is more likely (see Kording et al. 2007, Beierhold et al. 2007). In this case, the exponential costs of inference cannot be avoided by analytic integration.

The challenge for the approach above is that delivering the outputs of inference in the general case seems to require inference. A natural thought at this point is that there might be some heuristic method for integrating disparate sub-inferences – here a heuristic method is defined as one that

[40] Indeed, Bayesian cue combination is typically framed as independent measurements of a single dimension, rather than inference over multiple dimensions. What it mimics is true inference over low level haptic and low level visual dimensions in order to recover the height of an object. Cue combination and inference output the same value if the relevant uncertainties over size actually are independent and gaussian in the full model.

[41] It's unclear if the authors of studies of this kind ever intend their models to be understood in this realist way, as models of the actual computational processes by which the brian solves these inference problems, rather than as demonstrations of the optimal use of information by whatever process the brain actually implements (that is, whether the models are ever intended as Marr algorithmic level models). The concerns I'll offer here give us reason to doubt that the brain actually computes inference in the way described by these models, but not to doubt that the brain is sensitive to dependencies in the ways the model describes.

integrates sub-inferences well enough to meet the needs of human vision, but is not guaranteed to work in all cases. Delivering such a heuristic is easier said than done. To get a sense for the difficulty, consider what heuristic means of integration would give rise to behavior exemplified by the Mach Card described above. What general heuristic tells us when colored light should be seen as part of the lighting condition, rather than object color? Or could tell the visual system how to update its assessment of an object's shape as a function of those assignments? Or recover the number and location of lighting sources based on the shadows cast on objects of disparate shape? The sheer number of ways that the dimensions of shape, color, location, and lighting condition might depend on one another makes the prospect of a heuristic method of integration adequate to human vision itself an exponentially vanishing prospect. At a minimum here, we can note that no such general heuristic method of integration has been proposed in the literature.

Our assessment of the viability of these proposals is, of course, subject to change. Perhaps a heuristic approach to the problem of inferential integration will come along, and one should be taken seriously if and when it does appear. For the moment, however, there does not seem to be an alternative to doing inference, which minimally must respect the dependencies described in our toy inference problem and illustrated by the bistability of the Mach Card. If this is right, then the assumptions we've explored so far are sufficient to land the costs of perceptual inference in astronomical territory. Any would-be explanation of the tractability of perception must then account for how those assumptions can be challenged, allowing such costs to be avoided. In the following section, we look at what it would take to provide an explanation of the tractability of perception along these lines.

VII. How (And How Not) To Explain Tractability

How to Explain Tractability

It would seem then that we've reached an impasse. By our lights the intrinsic costs of perceptual inference scale exponentially in dimensionality and a mere subset of the dimensions

involved in perceptual inference run those costs into astronomical territory. For perception to be tractable, however, the costs of performing inference must not be astronomical. At this point we need to stop and take stock of the assumptions that got us here and ask ourselves if any of them can reasonably be denied.

As a reminder, these assumptions were threefold: (1) that good hypotheses are found, relative to a reasonable performance criterion. (2) that the costs of evaluating hypotheses are relatively fixed. And (3), that nothing is known about the distribution of good hypotheses in the hypothesis space. We discussed (1) and (2) at length in Section V.[42] That leaves (3). For (3) to be false would mean that perception has information, in advance of inference, about the distribution of plausible hypotheses in the hypothesis space. If perception *were* to have such prior information, this information could be used as a guide when exploring the hypothesis space. While drawing hypotheses randomly entails exponential growth in the expected number of hypotheses sampled before finding a good one, guided exploration of the space does not – in the guided case, the costs would depend straightforwardly on the quality of the information used as a guide.[43] To deliver tractability, this information must be good enough to find criterion-meeting hypotheses from among astronomical numbers of options in fewer than K steps.

Take 'sampling' to describe the choice that any inference algorithm must make as to where to look for good hypotheses in the hypothesis space.[44] We can call the outcomes of these decisions an algorithm's 'sampling dispositions.' When these dispositions are informed by information about the distribution of good hypotheses in the space, we'll call them *intelligent sampling dispositions* (ISDs). With this concept in hand, we can now offer a precise statement of the problem of the tractability of inference:

---

[42] P. 30ff

[43] See Chatterjee & Diaconis (2018).

[44] In this case, we're using the term to describe something broader than sampling in the technical sense that is relevant to the Monte Carlo inference methods that might be familiar to some readers. Sampling in our sense includes any way that an inference algorithm might go about selecting promising portions of the hypothesis space, including those in non-Monte-Carlo inference algorithms, such as variational methods.

> **The Challenge of Tractable Inference:** The challenge of explaining how perceptual inference is tractable by accounting for the intelligent sampling dispositions at work in perceptual processing.

For the rest of this paper, I will defend the claim that the challenge of explaining the tractability of perception is the challenge of explaining how perception comes to have intelligent sampling dispositions (ISDs) sufficient to avoid astronomical costs when performing inference in an astronomical space of options.[45] While candidate ISDs abound, delivering on such an explanation that is up to the task of perceptual inference is far easier said than done.[46] What makes it difficult is that the location of plausible hypotheses is not fixed, but is rather sensitive to the specifics of the problem instance at hand. We see very different scenes in the course of our lives, and which scene we're looking at on any particular occasion dictates where the plausible hypotheses are to be found.

To see why delivering such intelligent sampling dispositions is difficult, it is helpful to see why one popular idea, that the perceptual system embodies 'natural constraints' on perceptual scenes, is not a solution.[47] The idea of natural constraints is the idea that the perceptual system has access to (implicitly or explicitly represented) information about how the world typically is. The canonical example here is the visual system's sensitivity to the fact that light typically comes from above (see discussion of shape from shading in Section VI above). That the visual system possesses such a prior may be true as far as it goes. But such a prior, even if it's used to inform sampling, is unlikely to address the issues of computational tractability discussed here. This is for the simple reason that human vision in fact recovers any number of different lighting sources and lighting directions, and recognizes

[45] See Schulz (2012) for a similar thought in the case of cognitive inference in children.

[46] Any inference algorithm that delivers a speed advantage over exhaustive search or uniform sampling will have some ISDs that are responsible for its speedup. This includes algorithms making use of the idea that good hypotheses tend to be near one another (e.g. local MCMC), that good assignments of values to variables will be high probability conditional on good assignments to other variables (e.g. Gibbs sampling), that the posterior landscape is smooth (e.g. Hamiltonian MCMC, Variational Methods), etc.

[47] Or at least not a solution on its own. Note that many of the proponents of information encapsulation are also proponents of natural constraints (the information in perception has to come from somewhere, after all) and so already accept that perception has prior information about its domain. I expect for this reason that many will be broadly sympathetic to the idea that more information is present in the form of ISDs.

fine-grained local differences in lighting condition, such as shadow and mutual illumination (all while respecting the dependencies between these dimensions and many others, see Section VI). The mere starting assumption that lighting is singular and comes from above does not save vision from the requirement to be sensitive to a vast number of other ways that lighting could be, including more fine-grained ways consistent with light coming from above, and it is this requirement that entails astronomical computational costs.

While natural constraints are not themselves enough to deliver computational tractability, they are the right kind of thing. That is, they are sources of information, prior to inference, about which perceptual hypotheses are likely to be good. What's needed to account for tractability is much stronger sources of this kind of information. In contrast to natural constraints, which embody information about which hypotheses are plausible *in general*, what is needed for tractability is more fine-grained information about the distribution of plausible hypotheses for the problem instance at hand.[48]

Why Information Encapsulation Does Not Explain Tractability

Now that we better understand the sources of intractability in perceptual processing and what is needed to avoid astronomical computational costs, we're also better able to see why information encapsulation is not an explanation of tractability. The main idea here is that the costs intrinsic to perceptual processing are vastly larger than those associated with information access, and this difference in size undermines any intimate explanatory connection between encapsulation and perceptual tractability. This is the main idea, but I don't expect the reader to be convinced just yet. As always, the devil is in the details. In what follows, we'll go through a series of things it might mean for information encapsulation to explain tractability – including the possibility that information encapsulation is sufficient for tractability, that it is necessary, or that it is a difference maker. We'll see how our new appreciation of the challenge of accounting for perceptual tractability allows us to definitively rule out versions of the EET on which encapsulation is necessary or sufficient for tractability, while leaving us with strong reasons to be skeptical that it might be difference maker.

[48] That is, not merely a good prior, but a good estimate of the posterior. For the recurring distinction between dimensions and information, see p. 38.

Start with sufficiency. Could avoiding the costs of information access by way of encapsulation be *sufficient* for the computational tractability of perception? Based on what we've said so far, the answer to this is clearly no. This is because ISDs are necessary for computational tractability, and a perceptual system could be encapsulated from a cognitive system without also possessing ISDs. For example, a simple model aimed at doing the inference described in Section VI might receive no inputs from any external computational system (and so be encapsulated) and yet lack any ISDs. In the simplest case, it could perform inference by sampling randomly from the space of possibilities. Such a model would be encapsulated, but inference in it would be straightforwardly intractable, running up against the astronomical costs of inference. So encapsulation is clearly not sufficient for computational tractability.

How about necessity? Could information encapsulation be necessary for computational tractability? Here too I think the answer is no, but before arguing for this, it's worth first seeing why this idea commands so much appeal. There is a ton of information in cognition, from random facts about people, such as names and political persuasions, to the habitats of animals, to memories of your grandmother's garden. Perception, for its part, has to operate very fast, on the order of tens or hundreds of milliseconds, as we saw before. What's more, some kinds of very demanding search are certainly intractable. Take, for example, what we might call 'full relevance search.' By full relevance search, I mean sorting a list of information into those entries that are relevant to an inference problem and those that are not. In the limit, this requires performing the full inference problem once with each subset of the entries on the list and comparing the results to see which entries make a difference (in different combinations) to the outcome of the inference, in order to determine which entries are relevant to the task at hand. Such an operation is likely to scale super-exponentially, since it involves inference (which scales exponentially with dimensionality) being performed as a subroutine exponentially many times (as a function of the size of the list). Search of this kind would of course be intractable. If perception were required to do an exhaustive relevance search through cognition in the course of each perceptual inference, then there can be little question that it would be intractable.[49]

[49] Full relevance search of this kind seems to be what Fodor (1983) has in mind when he writes, "the point of the informational encapsulation of input processes is not—or not solely—to reduce the memory space that must be searched to

This is all true as far as it goes. But it is also very far from establishing that encapsulation is necessary for tractability. This is for three reasons. First, search does not have to be exhaustive, going through the entire mind to guarantee that it has returned all of the relevant information, in order to violate encapsulation. Search methods might search some portion of the database that merely sometimes has relevant information (say, 'search only memories from the past 24 hours'), or might search in a way that could access the entire database, but with a limited amount of time in which to do so ('search everything but stop after 100 milliseconds'). Other ways of limiting search exist as well. Instead of circumscribing search on the basis of the store or the duration of the search process, search could be limited by properties of the information being accessed, say returning values based on their place in the full list of entries (even very simple organizations of lists keep search costs sub-linear) or on the basis of their syntactic features (say, 'return only those memories that explicitly encode the color of this object'). If perception does in fact search through cognition, it could limit its search in any of these ways, making the costs of exhaustive relevance search irrelevant.[50]

Second, not all kinds of search that return some relevant information require sorting that information into relevant and irrelevant entries, and it is often better not to do so. Consider an over eager search strategy that returns some relevant information and much that is irrelevant. If we do inference with this information, the outcome is the same as if we'd done inference without the irrelevant information (that's what it is for the information to be irrelevant!). As for computational costs, the costs are no more than if we'd first sorted the list into relevant and irrelevant entries and accessed only the relevant ones (since the information has to be accessed in both cases – either to be fed directly into inference or to be sorted) and are often much less (since the super-exponential costs of sorting are neatly avoided in the over eager case). Inference itself is not more expensive with the

find information that is perceptually relevant. The primary point is to restrict the number of confirmation relations that need to be estimated as to make perceptual identifications fast" (p.71). See Fodor (2000) for similar arguments about the intractability of relevance search.

[50] Note that in Section III we assumed for the sake of argument that information access could be tied to the costs of search, despite the possibility of search without access (say, if cognition does search and sends information to perception as an expectation prior to inference, see Kok et al. 2012). I am not reneging on this deal – the costs of search are still at issue – but rather pointing out that search does not entail exhaustive search.

irrelevant information, since the costs there are dictated by dimensionality, not information.[51] So search for relevant information need not be the super-exponentially scaling relevance search of the kind envisaged above.

Finally, search strategies that reliably return relevant information without exhaustive relevance search are not an idle theoretical possibility. Rather, search strategies of this kind are a fixture of the modern era, making searching through even extraordinarily large databases fast and efficient. A typical Google search, for example, searches Google's copy of the internet, an enormous body of information, and returns general relevant results at an average latency of 500 milliseconds.

With all of this in mind, we can now see why encapsulation cannot be necessary for perceptual tractability. Consider a perceptual system with the following property: after coming up with an initial guess as to the identity of an object, it runs a Google search to find the typical color of that object, and uses this as an additional input into color and identity processing. This system would be unencapsulated, in the vein of anti-encapsulation interpretations of color processing effects in people (MacPherson 2012). More importantly for our purposes, if inference in this system was tractable before adding the search, then it will be tractable afterward. The possibility of such a case shows that, at least based on our current evidence, encapsulation can't be necessary for tractability.

Here I want to be clear about what I am saying and what I am not. The point is not that search in the mind might work just like Google search – very likely this is an unrealistic model. The point is rather that Google search gives us a proof of concept that some searches over very large databases are nevertheless very cheap. In a few short decades of computer science, human ingenuity has already hit upon cheap ways of doing large scale search. In light of that, we would need a very strong argument to convince us that cheap ways of doing search were out of reach for evolution. And without such an argument, we should not believe that avoiding the costs of search is necessary for tractability (cp. Clark 2002 for a similar point).

If encapsulation is neither necessary nor sufficient for tractability, then in order for the EET to be true encapsulation must at least be a difference maker. To be a difference maker it must be the case

[51] See Section V, p. 37 for the distinction between information and dimensions.

that, given all the facts on the ground, if perception were unencapsulated, then it would be intractable.[52] Here the thought would be that engaging in information access is a discretionary line item in the brain's computational budget for perception, and one that pushes perception over budget after all the essential line items are paid for. The question then is, what reasons could we have for believing that information access is such a decisive line item? These reasons break down into two categories. First, we could have reason to believe that those costs are a large part of the final budget for tractability, once all the strategies that evolution has employed to keep costs low in search, inference, etc. have been taken into account. (This would mean that the costs of information access would also be a big part of the final budget of K once a much tighter bound had been set on K). Second, even if the costs of information access are not a big part of the budget, we might nevertheless have reason to believe that they are a small, but critical part of the budget – the final line item that just tips the balance and pushes us over budget. Here, as above, I'll argue that the vast difference in scale between the problems of inference and access undermines either case for believing that avoiding the costs of access will be a difference maker.

Take the first possibility. Do we have any reason to believe that the costs of access will be a large part of the final budget, once we've figured out all of the optimizations evolution has employed to keep the costs of both access and inference down? In evaluating this admittedly very challenging question, start with a sociological fact: When computer scientists evaluate the complexity of a program which is the sum of a non-exponential term and an exponential term, they tend to ignore the non-exponential term. They don't do this out of a sense of wanton violence toward an accurate representation of the complexity of the program, but rather because, empirically, when computational costs are the sum of a non-exponential and an exponential or greater term, the contribution from the non-exponential term

[52] That is, for a natural analysis of what it is to be a difference maker, which is to be a necessary condition holding everything else about the system fixed. See p. 21 for a brief discussion of what it is to be a difference maker and how this differs from both necessity and sufficiency.

tends to be negligible. That is, if the costs of running a program are n^x + 5x, this is typically very well approximated by n^x.[53]

New insights into how full inference and exhaustive search might be approximated could, of course, change this. If evolution has been tremendously successful at keeping the would-be exponential costs of inference low, while finding few or no strategies to lower the would-be linear costs of search, then this could change the relative proportions of the budget that go to each term, leaving search and access the larger chunk of the final budget. There is no doubt that this is possible. But it does not seem particularly likely. For one, the starting costs are so different that the successes in lowering costs would have to be remarkably one-sided. For another, as we saw just above, the current state of affairs paints just the opposite picture – we currently know of many methods for making theoretically cheap search even cheaper, while we have very few insights into how theoretically expensive inference could be made much less expensive. At the very least then, we have no positive reason to believe that encapsulation will turn out to be a difference maker for perceptual tractability by way of being a large portion of the final budget for perceptual processing.

If the costs of information access are not a large part of the budget, we could still have reason to believe they are a difference maker if we have reason to believe that they are a small but critical part of the budget – the final expense that pushes us over budget once all essential operating costs have been paid. In evaluating this possibility, consider one final time the difference in size between our exponential and our linear terms in the theoretical costs of unencapsulated perceptual processing. If our thinking so far in this paper is on the right track at all, then the final costs of information access are likely to be a drop in the bucket compared to the final costs of inference. The theoretical result (based on the difference in theoretical costs) is clear cut, and the convergent empirical evidence (based on the current ease of search, discussed above, and current difficulty of inference, briefly surveyed in Section II) is at least suggestive of a significant difference in the computational costs associated with these two

[53] The exception to this is when x is small, in which case the linear term will dominate. That does not arise in this case, since the variables are distinct (the exponential term is exponential in dimensionality, while the linear term is linear in the number of entries that must be searched) and the exponential term is fixed empirically by the dimensionality of the inference problem (see Section VI).

problems. Given our current evidence then, believing that the costs of information access are a small but still critical portion of perception's computational budget would require believing that these costs are going to be the proverbial drop that makes the bucket overflow. This is certainly possible, but we have no positive reason to believe it!

We've seen that the difference in size between the problem of inference and the problem of access rule out certain versions of the EET (that avoiding the costs of access is necessary or that it is sufficient) and give us considerable reason for skepticism about others (that avoiding the costs of access is a difference maker). Thus we've seen that the original motivation for the EET, what we earlier called the Haystack Idea, can be safely laid to rest. There is, however, one final difference making role for encapsulation that should be examined. This is the possibility that encapsulation might be a difference maker, not by allowing perception to avoid costly search, but rather by being a crucial part of the ISDs which are themselves critical to the tractability of perceptual inference. In this case, proponents of encapsulation would acknowledge that the costs of search are likely insignificant in explaining tractability, but would turn to seeing encapsulation as a plausible contributor to limiting the actual costs of inference. I'll briefly argue that even this revitalized version of the EET lacks sufficient motivation.

As a way of easing into it, start with a simpler thesis, that information encapsulation is necessary for ISDs. Based on what we've established so far in this paper, we know that this can't be the case. This is because having an amazing perceptual prior, one with strong expectations about what you're likely to see when,[54] is sufficient for ISDs. And it's possible to have an amazing prior while not being encapsulated from cognition. If a perceptual system came equipped with such a prior, say by way of evolution or perceptual learning, but were unencapsulated, accessing select information from cognition (say, just color memories), then this system would exhibit strong ISDs despite being unencapsulated. At least based on what we know right now then, encapsulation can't be necessary for ISDs.[55]

[54] And therefore a good approximate posterior.

[55] There is also no question that it's not sufficient for ISDs, since it's categorically the wrong kind of thing to deliver ISDs – not a source of information, but merely a prohibition on one.

Now consider the possibility that encapsulation might be a difference maker – making some critical contribution to perceptual ISDs, holding all other facts about the system fixed. To evaluate whether this is plausible, consider how it is that ISDs provide for tractability. They do so by helping to locate plausible hypotheses from among an astronomically large expanse of random possibilities. At first pass then, any information that might help the system locate plausible hypotheses in this space is likely to result in computational savings. In the next section, I'll argue on these grounds that there are likely to be many cases where information from cognition could significantly reduce the costs of perceptual processing. The gist of those examples is that cognition can sometimes propose reasonable solutions to perceptual inference problems, in virtue of sometimes possessing veridical information about what we are likely to be seeing. For the moment however, let's just consider what it would take for information from cognition to be *harmful* to tractability. In order to significantly increase the costs of perceptual processing by way of influencing ISDs cognition would have to contribute information that is not just sometimes wrong about what you are seeing, but rather systematically and relentlessly misleading about what you might be seeing.

To wrap our minds around this point, consider a related downside to cognitive influence on perception that has been discussed in this literature. Some authors have argued that human perception is susceptible to cases of 'wishful seeing' in which someone has some idea of what they would like to see, and this very idea influences the perceptual interpretation of ambiguous stimuli. So, for example, if I am looking for the mustard in my fridge, I might briefly misperceive a lemon in the fridge as mustard (Siegel 2017). Wishful seeing, if it happens, influences the outcome of perceptual inference – the lemon looks to me (perhaps briefly) as if it were mustard. When the outcome of perceptual inference is less veridical then it would have otherwise been, wishful seeing has an epistemic cost. What we're imagining here is slightly different. We're imagining a case where cognitive influences have a

computational cost.[56] One way to get onto this is to take a case that is like wishful seeing, but which holds fixed the final outcome of perceptual processing. So, in our case, cognition would first offer the mustard hypothesis, and perception would check it against the data, perhaps rejecting it because the mustard hypothesis sits poorly with the absence of any noticeable label or cap on the yellowish figure. Finally, perception settles on the lemon hypothesis, as it would have if there had been no effect of cognition.

In this case, even though perception avoids any epistemic cost, ultimately settling on the same output hypothesis, the proposal of the mustard hypothesis creates unnecessary computational costs – the costs of evaluating and rejecting the falsidical hypothesis. It stands to reason then that, if there were a sufficient number of such unnecessary hypotheses proposed by cognition, this could run up the costs of perceptual inference, eventually pushing it over-budget. In such a case encapsulating perception from cognition's misleading proposals could be an important part of how the ISDs deliver tractability.

The problem with this line of reasoning should already be apparent. If, on the one hand, the implausible hypotheses proposed by cognition are just one or a few, then the costs they impose are negligible relative to the size of the perceptual inference problem. Since we should believe that even with great default ISDs perception is likely to have to evaluate many hypotheses, the addition of a handful from cognition seems unlikely to be difference making. If, on the other hand, the hypotheses are not implausible, then evaluating them may not be unnecessary for veridical perceptual inference (and in some instances may even help). In order for cognitive influences to pose a threat to tractability by this route then, the proposals from cognition must be both highly numerous *and* systematically misleading. While it may be easy to imagine that cognition, were it allowed to, might occasionally send perception an implausible proposal, the idea that it might be a source of of implausible proposals on the scale needed to threaten perceptual tractability, where the default requirement is to navigate an

[56] Of course, both epistemic and computational costs are relevant to tractability, since tractability is relative to both a budget, K, and a performance criterion (see Section III). If cognitive penetration would bring with it a sufficient decrement to performance by way of effects like wishful seeing, then technically such penetration could make perception intractable by way of decreasing performance beneath human levels, rather than increasing costs above human levels. I take it that the core idea behind the EET as it has been defended is that cognitive influences would make perception more expensive, not less accurate, so I won't consider this back route to the thesis here.

exponentially large hypothesis space, is a heavy lift. Here again, at least absent some positive argument in its favor, we should not believe that this is the case.

I'll say more about the relationship between encapsulation and ISDs in the following section. At this point however, we should take stock of where we've gotten. We've seen that encapsulation is neither necessary nor sufficient for computational tractability – not sufficient because ISDs are necessary and a system can be encapsulated without exhibiting any ISDs, and not necessary because the kinds of information that allow for tractability can exist in perception even if it is unencapsulated. We've also seen that the costs of information encapsulation are unlikely to even be a difference maker to perceptual tractability, in light of the massive difference in size between the costs of information access and the costs of inference itself. After establishing that the costs of information access are unlikely to motivate encapsulation, we finally asked if encapsulation might be critical to tractability by way of being critical to ISDs, and found that defending such a view requires believing the proposals influenced by cognition are not merely often non-actual, but both numerous and systematically implausible. The idea then that tractability considerations motivate encapsulation should be laid to rest. Many questions remain, however. In the next section, I use some of the tools we've laid out in this paper to explore the future of tractability arguments in this area.

VIII. The Future of Tractability Arguments

The Dimensionality Restriction Hypothesis

At the end of the day, we don't know how perception is tractable, and this limits what we can say with confidence about the bearing that various cognitive effects might have on tractability. But we are not totally in the dark either. For example, we have an understanding of the sources of computational costs and the factors that influence them. Any cognitive effect that threatens to make those prima facie costs exponentially worse should be regarded with suspicion. Similarly, we know the form that a solution to intractability must take – it must offer a theory of the intelligent sampling dispositions that allow perception to navigate its vast hypothesis space. Such dispositions are built on

information. Any potential source of this kind of information, up to and including cognition, should be regarded as potentially part of the solution.

Start with an example of the first kind. The unfortunate truth of computational costs is that, while it can be difficult or even impossible to make a problem easier, it is always possible to make it harder. Certain kinds of cognitive effects could make perception's problem much harder. Take for example the 'enrichment' of perception by cognition. Some authors have suggested that cognition might enrich perception, in the sense of expanding perception's representational capacity to include dimensions previously represented only in cognition, for example in the process of developing expert perception (Siegel 2010). Others argue on empirical grounds that perception is dimensionality restricted, operating over an (at least synchronously) limited set of dimensions, in contrast with cognition, which is dimensionality non-restricted (Green 2020). The ideas laid out in this paper suggest that there may be more a priori considerations relevant to this debate as well. Since the prima facie costs of inference scale exponentially with dimensionality, adding dimensions, whether from cognition, perceptual learning, or by any other mechanism, could dramatically increase the costs of perception. Such an effect could form the basis for a tractability argument against cognitive enrichment effects and in favor of the dimension restriction hypothesis. Making this case rigorously would require careful treatment, but the possibility of such an argument follows naturally from the framework developed here.

Veridical Information From Cognition

Another species of future tractability arguments could put information encapsulation on the defensive. As we saw above, what's needed to account for the tractability of perception are sources of information which help steer perceptual inference toward promising hypotheses. A natural question to ask, then, is whether information *from cognition* could support tractability.

Consider the following case. You are wandering around in a jungle and see an ambiguous form in the branches. Naturally, there are virtually countless possible visual interpretations of the visual scene. Now imagine that you know that you're in panther territory. This key bit of information from

cognition could be used to guide sampling, allowing vision to arrive at an interpretation of the visual scene much more quickly. A tip of this kind could easily be the difference between visually detecting and missing something that was really there.

How, exactly, could the abstract belief that one is in panther territory be used to guide visual inference? The technical proposals are too much to get into here. But the underlying process is not that different from what you would naturally do if I were to ask you to close your eyes and imagine a panther in that tree. Then to reset and imagine another, distinct scene meeting the same constraint. And then another. In each of these cases, you are sampling from a space of possible scenes, under the constraint that they feature a panther in the tree. This cognitively constrained distribution is far more peaked than an unconstrained prior distribution over all possible scenes, with or without panthers, thereby guiding visual inference toward the hypotheses that meet the constraint.[57]

Accounts of roughly this kind have been offered as explanations for the phenomenon of stably resolving ambiguous images (Lupyan 2017, Block 2022). These are images which appear one way at first, say, as an unremarkable brick wall or set of black splotches, but resolve another way when people are given a clue semantically related to the alternative interpretation. Once their more surprising interpretation has been seen, it is often difficult to unsee; a fact which may reflect the visual systems assessment that the new hypothesis offers a better solution to that particular visual inference problem (see Figure 2).

[57] Note that a procedure like this can work even if the visual system doesn't explicitly represent high-level contents such as 'panther', since the relevant distribution could be a distribution entirely spelled out in terms of low-level properties; those that would trigger recognition of panthers.

Figure 2: Visual inference can be affected by information about what one is looking at. Look at the image above and search for anything out of the ordinary before reading this footnote for a hint.[58] Image reprinted from Lupyan 2017 (original photographer unknown). For extended discussion of stable ambiguous images, see Block 2022)

These are not knock down demonstrations of cognition-fed sampling dispositions. Many who have discussed these effects have argued that they are due to attention (Firestone & Scholl 2016, Lupyan 2017, Block 2022 signals openness to this interpretation). Whether attention offers a competing explanation or is merely the mechanism of cognitive penetration is itself an open question (Quilty-Dunn 2019, see Green 2020). Cognitively-driven samples are a computational process while attention is a folk psychological and neuroscientific concept and the relationship between the two is unclear. This is murky territory. My aim in bringing these issues up is not to try to lay them to rest, but merely to illustrate that the tractability considerations that were once taken to require the encapsulation of perception from cognition, may in fact support just the opposite conclusion once the

[58] On first encounter with this image, most people see an small, bluish rock wedged in a stone wall. Given a hint, such as the quip that 'Sometimes a cigar is *just* a cigar', people see the image differently. (If that is not enough of a hint, try seeing the bluish rock and brown space next to it as a single object, protruding outwards from the wall, with the blue tip farthest from the surrounding rock.)

true challenge of perceptual tractability is appreciated. This reversal holds even if, as is likely to be the case, most of the information in the intelligent sampling dispositions is internal to perception.

IX. Conclusion

A theory of the architecture of perception must explain how perception is computationally tractable. This paper has argued that information encapsulation, even if true of perception, does not provide such an explanation. This is because of the significantly greater costs of perceptual inference, as compared to information access, which threaten to make the costs of access a negligible proportion of perception's computational budget. After all this, it remains an open empirical question whether perception is encapsulated from cognition, but the encapsulation thesis has lost its computational *raison d'etre*. As a consequence, we should be more willing to accept some of the psychophysical effects reported in the literature as genuine violations of encapsulation. We are at the very least not bound on computational grounds to find ways in which these effects are not genuine effects of cognition on perception. This is, of course, not to advocate for laxity in our psychophysics or analysis, and alternative interpretations of putative cognitive effects should be carefully proposed and ruled out, but it is an argument for a more even prior between encapsulation and cognitive influence as we approach these debates.

The framework for thinking about computational tractability laid out in this paper also has implications beyond the question of encapsulation. For one, we now have an understanding of the sources of computational costs and the factors that influence them. The things that matter to tractability are things like the dimensions, dependencies, and sampling dispositions involved in inference. Information is a resource for limiting computational costs, rather than a liability. With these factors in mind, novel proposals about the architecture of perception can be evaluated for how they are likely to affect tractability. Proposals to the effect that cognition might expand the range of dimensions perception computes over, thereby increasing the dimensionality of perceptual inference problems and threatening to increase the costs of inference exponentially, have an a priori strike against them, while the alternative, dimensionality restriction, has an a priori consideration in its favor. Going forward, we

should be more skeptical of, and more careful to explore alternative explanations for, psychological effects which purport to evince such dimensionality non-restriction (in effect, saving for dimensionality non-restriction and other exponentially costly architectural theses the jaundiced eye we have hitherto reserved for purported failures of encapsulation.)

We should also look to develop positive accounts of perceptual tractability. Proponents of information encapsulation were right to think that perception faces a threat of intractability and that reflecting on how such a threat is avoided can be a tool in uncovering the architecture of perception. If anything, this is even more true now; with a vastly larger problem of intractability that is integral to perception's essential function, the demand that an architecture allow for tractable inference becomes a powerful constraint, shaping perceptual architecture throughout.

Deciphering what architectures allow for perceptual tractability is a difficult problem, but we've already made a start – spelling out the general form that such a solution must take. Any account must offer a theory of the intelligent sampling dispositions that allow perception to efficiently navigate the vast hypothesis spaces involved in perceptual inference. Such dispositions are built on veridical information about the distribution of plausible hypotheses throughout the space. In order to deliver human-like perceptual competence, including critically the ability to recover a large number of perceptible properties across a vast diversity of scenes, this information must be opinionated (strongly focusing computational work in narrow regions of the hypothesis space), particular (sensitive to the directly measurable properties of the scene, rather than rigid constraints expected to apply across the board), and veridical (concentrating probability mass around genuinely plausible hypotheses). A theory must tell us where this information comes from and what kind of architecture can gather and deploy it. Any source of information of this kind (up to and including cognition) should be regarded as potentially part of this solution, but the key answers are likely to come from a theory of perceptual learning. Mechanisms of such learning, hitherto treated as something of a black box, are likely to be a critical part of the theory.

Chapter 2: Why is Seeing Fast and Thinking Slow?

Abstract: Seeing is fast and thinking is slow. This claim often appears in arguments aiming to identify mental processes as either cognitive or perceptual or to defend the existence of a divide between the two. But is it true, and if so why? In this paper, I look at the evidence for a speed difference and develop a potential computational explanation for it. I then show how the thesis sheds light on some otherwise puzzling phenomena in cognitive science and neuroscience. If this picture is right, it helps us understand one important way in which perception and cognition differ.[59]

I. Introduction

Suppose you are wondering if your partner is home. You notice the keys on the side table, consider the time of day and the light at the end of the hallway, and think about the last time you heard a sound. You conclude that they're likely home. Contrast this with the experience of turning a corner and seeing your partner in the hallway in good light. Reasoning about your partner's presence or absence was slow, while seeing them there was all but instantaneous. This contrast is surprising in light of the fact that both seeing and this kind of reasoning share a fundamental similarity – they are both mental processes that recover the state of the world (in this case your partners presence) on the basis of some data (the keys on the table, or the pattern of light projected onto the retina) by way of the evidential relationship between the two (the keys placed on the side table upon one's return, or the way a 3D form projects onto the retina).

This fact, that seeing is fast and thinking is slow, is often invoked as a premise in psychological methods and philosophical arguments. It has been the basis for classifying ambiguous mental processes as perception, for example, to argue that concepts, confidences, or high-level contents like gender and race are part of perception, given the speed with which such contents are recovered (Mandelbaum

[59] Many people helped me with the ideas in this paper. Many thanks in particular to EJ Green, Jack Spencer, Alex Byrne, Laurie Paul, Ned Block, John Morrison, Kevin Dorst, and Josh Tenebaum for comments on earlier drafts, to members and participants of Chaz Firestone's and Brain Scholl's labs and the philosophy departments at York and U Penn for feedback on talks, and to Maddie Cusimano, Ishita Dasgupta, Martin Schrimpf, David Danks, Bob Rehder, Todd Gureckis, and Ian Phillips for helpful discussion of the ideas in this paper.

2017, Morrison 2016, Colombatto et al. 2021). (Whether these contents are actually perceptual has knock on effects for lots of other debates, about cognitive phenomenology, whether perception is probabilistic, or questions about what role the concepts of race and gender play in our lives.) The speed difference has similarly been recruited in arguments about whether what we think can influence what we see, in particular, as the basis for arguments that cognitive influences on perception are incompatible with the speed with which perception operates (Fodor 1983, 2000, Pylyshyn 1999, Quilty-Dunn 2019, although see Brooke-Wilson *Forthcoming*). The results of these debates matter to epistemology (Siegel 2010) and the philosophy of science (via the theory ladenness of perception, see Churchland 1988, Fodor 1988). Getting clear on whether and why this difference in speed exists is the goal of this paper.

The central thesis is that the difference in speed is due to a difference in inferential strategies available to perception and cognition. In particular, perception exploits mutual information across instances of perceptual problems – in a sense to be made precise – to anticipate the results of perceptual processing. This allows perceptual processing to start closer to its solution. Exploiting this strategy in perception makes perception faster, and also more accurate, on computationally more demanding problems than cognition.

The paper is structured as follows. Section II gets clear on the phenomenon, exploring how the thesis that there is a speed difference between perception and cognition should be understood and making a tentative empirical case for such a difference. Section III offers a hypothesis – that perception exploits prior exposure to its domain in a way that canonical cases of cognition cannot – and shows how this could account for the speed difference. Section IV briefly explores some alternative explanations and argues that they do not account for the speed difference. Section V provides convergent behavioral evidence for a difference in inferential strategies based on the kinds of errors found in perception vs. cognition, while Section VI shows how this thesis dovetails with findings in contemporary neuroscience. Section VII draws out some implications of all of this for understanding of perception, cognition, and attempts to replicate these in AI.

II. Seeing Fast and Thinking Slow

As I'll use the term here, perception is the set of mental processes dedicated to gathering information by way of the sensory surfaces (e.g. the retina for vision or the cochlea for audition). This includes the final stages of these processes, the perceptual outputs, which many believe to be identical to perceptual experience, but also include person-level unconscious perceptual states should such states exist.[60] Cognition is the set of mental processes that sit between perception and motor control, of which reasoning and high-level planning (planning abstracted away from motor details) are paradigmatic examples. Many areas of philosophy and psychology assume that a distinction between perception and cognition exists. It has also been challenged.[61] One of the reasons to accept a distinction is that certain properties seem to cluster together. Mental processes closely tied to sensory stimuli are often automatic, effortless, and relatively limited in the sets of contents they represent, while mental processes more divorced from sensory stimuli tend to be more deliberate, effortful, and relatively unlimited in the contents they represent. Speed is also on this list. Perceptual processes are often seemingly immediate, occurring without any noticeable delay, while cognitive processes are often slow, occurring with noticeable delays. Call the view that there is such a difference in speed between perception and cognition, '*Speed Difference*'.

Speed Difference is rarely explicitly defended, but is invoked in debates about the classification of ambiguous mental processes as either perception or cognition. Morrison (2016) uses speed as a hallmark of perception in making the case for perceptual confidences,[62] while Mandelbaum (2017) uses the speed of processing to make the case that categorization with a limited set of concepts happens internal to perception. Little & Firestone (2018) use speed to tell apart perceptual and cognitive

[60] See e.g. Firestone & Scholl 2016 for a similar usage.

[61] A distinction between perception and cognition is assumed in many debates in epistemology (whether perceptual experience is the basis for all justification; whether perception can be 'taken at face value') and philosophy of mind (e.g. the rich/thin debate about perceptual content or debates about cognitive phenomenology). For one famous challenge to the distinction, see Clark (2013).

[62] For example, on p.19 'First, perceptual confidence is more fully described as the view that confidence is assigned by a state that's conscious, automatic, accessible, dissociable from doxastic states, directed toward perceived objects and properties, and fast enough that we can't detect any delay.'

knowledge of physics, while Colombatto et al. (2021) use the speed of facial categorization to argue that demographic features, such as race and gender, are literally seen.

Before looking at the evidence for a *Speed Difference*, we need to get straight what's been claimed. As any of the proponents of the view would acknowledge, many cognitive processes are fast (e.g. recalling a single fact) and many perceptual processes are slow (e.g. the phase transitions in binocular rivalry). Block (2022) makes this point in justifying skepticism about *Speed Difference* (p. 43). Appealing to a difference in the distributions of processing times won't help without some way to individuate the relevant mental processes (is the relevant cognitive process recalling that a bridge is closed, realizing that a particular route won't work, or planning your errands?). The best way, I think, to understand *Speed Difference* is by reference to a particular family of tasks of similar computational structure, namely, *inverse inference.* Inverse inference is the process of recovering some inaccessible state of the world on the basis of some set of accessible states, or 'data', which suggest, but do not entail, the inaccessible state. Paradigmatic cases of inverse inference are things like inferring the circumstances of a crime from the evidence left at the scene, inferring a speaker's intentions from what is said, or recovering a 3D scene from a 2D retinal projection.

Inverse inference is a central task for both cognition and perception. Visual inverse inference allows the mind to recover the 3D scene before us on the basis of a blurry 2D image. The 2D image serves as evidence for the 3D scene, but does not entail it. Similarly, auditory inverse inference involves recovering acoustic events from the vibrations of the cochlea. Characterized at this level of generality, inverse inference is most, and maybe all, of what perception does. Cognition also does inverse inference. When you infer your partner's presence from the keys on the table, the presence of the keys is a directly accessible fact while your partner's presence must be inferred. Inverse inference is not the only thing that cognition does, things like arithmetic are not inverse inference, but it does capture many things. Understanding pragmatics involves inferring a speaker's intended message from the literal meanings of what was said, reasoning about other minds often involves inferring someone's beliefs and desires from their observable actions, while concept learning involves inferring concept meanings from

sparse evidence, such as examples of a concept's extension or its use in context. (In what follows, I'll often speak just of 'inference' rather than 'inverse inference' for brevity.)

On the face of it, this category of mental processes may seem like they have little in common beyond the highly general definition of inference. But that similarity on its own is important. That's because what makes an inference problem hard or easy, computationally speaking, is similar across disparate domains. What's more, many methods for solving inference are interoperable, in the sense that a method for solving inference in one domain is *ipso facto* a method for solving it in another. Where this is not the case, there are often close analogues of methods that work in one domain in another. In light of computational similarities it would be interesting if, in general, inference in perception is fast while inference in cognition is slow.

Take the claim that 'seeing is fast and thinking is slow' to mean that perceptual inference is fast, while cognitive inference is slow, for inferences of comparable difficulty. Note that this is a claim about typical perceptual and cognitive inference, not a universal claim. There are instances of slow perceptual inference and of fast cognitive inference, both of which I'll discuss below. This is also *not* a claim about what it is to be perception or cognition. There is a lively debate about the necessary and sufficient conditions for perception and cognition and many plausible candidates (cf. Green 2020, Beck 2018, Block 2023). The claim I'm interested in here is that typical cases of perceptual inference are fast, while typical cases of cognitive inference are slow.

There is some evidence to support *Speed Difference* understood in this way. Typical cases of visual inference happen in fractions of a second. Starting with light splashing on the retina, the shapes of objects can be recovered, as measured by masking studies and behavioral responses, in around 110 *milliseconds* (Baker & Kellman 2018). As measured by neural decoding, objects can be classified in between 70-170 milliseconds (Hong et al. 2016). In paradigms that measure the time for both perceptual processing and a motor response, faces can be detected among distractors in realistic images in around ~380ms (Rousselet et al. 2003), while the gender of a face can be recovered in about ~400ms (Barragan-Jason et al. 2012), and a familiar face recognized around ~580ms (ibid.).

Similarly fast estimates hold for non-visual modalities. Take auditory perception. Recognition of sound categories (face, percussion, strings) based on just timbre cues (with pitch duration and power normalized) is estimated to take around 500ms, with accuracy at ceiling (Agus et al. 2012). That 500ms includes the 250ms needed to convey the stimulus and the time to mount a motor response. A paradigm based on RSVP for audition, in which naturalistic sounds were presented in rapid succession, showed recognition for sounds presented at a rate of 30 per second (Isnard et al. 2019).[63] Famously, speech shadowing, which requires recognition of phonemes and the motor planning and execution to reproduce them, can occur with a latency of 150-250ms (Marslen-Wilson 1973, 1985).

We can use these figures to put some round numbers on the speed of typical perceptual inference to facilitate our thinking downstream. In the face studies above, subjects were required to make a motor response to indicate whether they'd seen a target (a face, familiar face, or face of a certain gender). They were able to do this in about half a second (500ms). In the limit, if we imagined that planning and executing a motor response took 0ms, that would give us a 500ms upper bound for an estimate of typical visual processing. On the low end, we saw visual recovery of properties such as shape and category membership approaching 100ms. Going forward, we can use 100-500 ms as an estimated range for the time course of typical perceptual inference.

So perceptual inference is fast. In contrast, typical cases of cognitive inference appear to be quite slow. This is a widely held belief in the field and the circumstantial evidence points in its direction. Take prosopagnosics. These are individuals who lack the ability to perceptually identify individual faces while maintaining otherwise normal perceptual and cognitive function. Subjects with prosopagnosia do not see facial identity the way that neurotypicals do, but can learn to identify faces by a cognitive strategy. These subjects are generally about 10-30 times slower, and significantly less accurate, than neurotypicals at recognizing faces, even after years of practice (e.g. Marotta et al. 2002). In this case we have the very same task being performed by perception and cognition, but an order of magnitude more slowly in cognition than in perception.

[63] Although note that there are concerns with this paradigm (Block 2023).

One well-studied domain of cognitive inference is 'category' or 'concept' learning. In a typical category learning study, subjects are shown a number of examples of a 'category' and tasked with learning the underlying intension that defines the category. Experimenters can then assess which intensions were learned based on subsequent categorization behavior. These tasks have been of interest to psychologists because it's thought that they mirror the concept learning that goes on in early child language acquisition (e.g. learning that a mug is made of ceramic with a cup shape) as well as some adult cognitive inferences.

In a classic study of this kind, Kemp et al. (2012) tasked subjects with learning an underlying category intension defined by boolean combinations of a few salient visual features. Participants were presented with a set of 'in' category objects and another set 'out' of category. Subjects were told to look at the objects during a learning phase before moving onto a test phase, where they were tasked with sorting a random array of objects according to the category. Depending on the complexity of the underlying rule to be learned, participants took between 40 and 70 seconds in the study phase before clicking to the test phase.

We can use these numbers to get some purchase on the idea that cognitive inference is slow. Take 50 seconds as a middle-of-the-road value for learning a concept in this study. That number is considerably larger than the numbers we saw for perceptual inference. Depending on whether we use the low end or the high end of our estimate for typical perceptual processing, the reaction times for this cognitive inference are about 100 to 500 times slower than for perceptual inference alone. It is, of course, not surprising that full reaction times for cognitive inference, which includes the time for perceptual inference, cognitive inference, and motor control and output, is longer than the reaction time for just perceptual inference and motor control. What is surprising is that the reaction time for cognitive inference, minus the time for perceptual inference and motor reactions, is *significantly* longer than for perceptual inference and motor control alone. That is, subtracting 500ms (our estimate for the time course of perceptual inference plus a simple motor output) from 50 seconds, we still get a value about 100 to 500 times slower for cognitive inference alone than for perceptual inference alone.[64]

[64] Put more cleanly, it is not surprising that $RT_C > P + M$. But it is very surprising that $RT_C - (P+M) >> P$.

These numbers offer some support to the idea that perception is fast and cognition is slow. There are, of course, limitations to using these studies for this purpose. Many variables were not controlled for. For example, the motor outputs were not the same (moving a finger in one case vs. dragging a cursor in the other), and we don't know how many perceptual inferences may have been required to do the concept learning task. If the difference in time we were looking at were a question of 100s of milliseconds, rather than 10s of seconds, these kinds of confounds would be very worrying. But a difference in speed of 100-500x offers us some cushion. If the difference in processing times were to be explained by the number of perceptual inferences (people looking at category instances) in the course of category learning, we'd need in the ballpark of 100-500 perceptual inferences as rate limiting steps in the cognitive inference. But there were only 8-24 things to look at, depending on the condition. Similarly for motor outputs, we can draw comfort from the size of the difference. It might have taken subjects a half second longer to click out of a screen than to tap a finger, but it's unlikely to have taken them 20, 30, or 40 seconds...

Massive differences in observed processing times then lend support to the common belief in a *Speed Difference*, despite some confounds. Future work should aim to investigate the question directly, controlling for as many confounds as possible. For now, however, I'll take the *Speed Difference* to be established and turn to possible explanations for it.

III. Amortized Inference in Perception

In this section I offer a thesis about why perceptual inference is fast. Before spelling out the positive thesis, I'll provide some background about what makes inference hard from a computational perspective.

What makes inference hard? As we saw above, inference involves recovering the likely state of the world on the basis of some data, which suggests but does not determine that state. The full landscape of things that contribute to the computational difficulty of this process is both complex and not fully understood. An important part of what makes a particular inference hard, however, has to do with the *size* of the problem. The 'size' of the problem is the space of total states of the world from the

point of view of the problem. Put concretely, if vision has to recover a 3D scene on the basis of a 2D retinal projection, then the size of the problem is, minimally, all the possible 3D scenes that could be seen and the 2D projections that would cause one to see them. Problem size, in this sense, has a close relationship to computational difficulty. Across many different computational methods for doing inference, starting with a problem that we *can* solve and increasing the problem size very rapidly takes us to a problem we can no longer solve. Conversely, for many real world problems we wish we could solve in reasonable time, we find we can solve very restricted versions of the problem that operate over a much smaller hypothesis space.

The reason problem size plays such a big role in problem difficulty is because a method for solving an inference problem has to be sensitive to the full breadth of possibilities. Whether a 3D scene is the most likely scene given an image (or in the top K most likely scenes, or above some threshold probability) all depends on how likely the other scenes are. Reliably delivering good hypotheses requires sensitivity to that space of other possibilities. This space gets very big very quickly. The number of possible 3D scenes is an exponential function of the primitives that make them up (properties like color, shape, location). If we imagine the simplest possible case of visual inference, seeing a single object against a blank background, then if that object can have one of two shapes and one of two colors, the number of possible scenes is 2^2. If the object can also be in one of two possible locations, there are 2^3. For realistic cases of inference, this means that the number of possibilities is enormous (I'll give an example in the next section).

Big problem spaces plus the requirement to be sensitive to them makes inference hard because, in many cases, good hypotheses cannot simply be read off of the image. 3D properties of the scene, for example, are confounded in the retinal image, often in elaborate ways. A splash of green light on the retina could be due to a green object in neutral light, a neutral object in green light, or any number of other combinations of object and illuminant color. Finding a likely 3D scene requires credit assignment – When one fact (the object color) is taking credit for the greenness of the light on the retina, another one (the illuminant color) should not be. To do this in a general way means evaluating hypotheses holistically, rather than piecemeal, evaluating how well a joint assignment of object color

and lighting condition fits with the image. When there is a large space of hypotheses that must be evaluated individually, inference will be very hard.

Other features of a problem, like the particular distribution of good hypotheses in the space (whether they cluster together or are dispersed), the degree to which the values of the hypotheses are confounded in the data (the strengths of the dependencies), or the information that evaluating subsets of hypothesis carries about the larger space, can all make a difference to the difficulty of inference. What all of these have in common is that they force an algorithm to rely on less targeted, and therefore more random, search, in a vast possibility space. So the problem size plays a central role in the computational difficulty of inference.

With that background in hand, we can turn to developing a hypothesis about why seeing is fast and thinking is slow. The thesis I want to defend is that perception is fast and cognition is slow because typical instances of perceptual inference are highly similar to one another, in a way that can be made precise, while typical instances of cognitive inference are not. Solving similar inference problems costs less because more work can be done by traces of previous computations, leaving relatively less work to be done by online computation.

The basic idea, that traces of previous computations can be traded off against online computation, is highly general, but applying it to inference takes some care. There are many ways this trade off can be implemented. A system might store query-answer pairs for particular problem instances its solved in the past. This makes solving 'new' problems trivial if they happen to overlap exactly with stored queries, but provides little help otherwise. More powerful solutions can be devised. A system might store the results of sub-computations that commonly recur across problem instances (such as storing earlier fibonacci numbers when calculating the series), or it might store probabilistic information about what good solutions are likely to look like for different queries.[65]

This last strategy is the one I want to focus on here. Speaking roughly, I want to claim that instances of perceptual inference are relatively similar to one another, in the sense that solutions to instances of perceptual inference carry significant information about the plausible solutions to other

[65] See Dasgupta & Gershman 2021 for discussion and examples of each of these.

instances. This similarity across the space of possible perceptual inferences can be exploited by keeping track of the relationship between queries (e.g. images) and plausible hypotheses (e.g. 3D scenes). Exploiting this similarity allows perception to solve novel instances of inference at a fraction of the computational cost.

In order to spell out this intuitive idea properly, we need to think about the process of inference in formal, information theoretic terms. Bayesian inference tells us how to integrate the prior plausibility of individual hypotheses, a 'prior' distribution P(H), with the probabilistic connection between a given hypothesis and the observed data, a 'likelihood' P(E|H), to deliver a post-update assessment of the plausibility of each hypothesis, the 'posterior' distribution P(H|E).

Viewed as a pure mathematical operation, this is all there is to inference: a prior, a likelihood, and the resulting posterior. As a computational process, however, inference involves another part, which tells the system in what order to consider the hypotheses. We can call this the 'proposal function.' As I'm using the concept here, any method for doing inference that considers individual hypotheses must have a proposal function, because something in the algorithm determines what hypotheses it will evaluate.[66] A proposal function can be deterministic or stochastic, and can be unconditional or conditioned on various aspects of the problem. (When it's stochastic, I'll refer to it as a 'proposal distribution'.)

The proposal distribution doesn't appear in discussions of the pure mathematical inference because it doesn't make a difference to the true posterior. For the mathematical operation, it's as if all the hypotheses are considered once. But in most real world cases of computing inference, where the problem spaces are far too large to consider more than a negligible fraction of the possibilities, the proposal function matters a great deal. Which hypotheses are prioritized will often make a big difference to the runtime, accuracy, and many other features of the computation.

We can ask what makes a proposal distribution better or worse. Unsurprisingly this has to do with the relationship between the proposal and the posterior as well as with what exactly we're trying to accomplish. If we want to approximate the posterior with samples, then a proposal that is as close to

[66] As I'm defining it then, 'proposal function' generalizes a more technical concept which is specific to sampling methods for inference.

the posterior as possible is ideal.[67] If instead we want the most likely hypothesis in as few samples as possible, then a distribution that concentrates proposals around the most likely hypotheses, including prioritizing them even more than the posterior, is better.

A natural place that people look for a proposal distribution when building a system to solve inference is the prior. The prior is, by definition, the epistemically closest distribution to the posterior before the system has received new evidence. We can do inference by using the prior as a proposal distribution and accepting or rejecting samples deterministically depending on whether they are consistent with the evidence. This algorithm, known as 'rejection sampling', converges to the posterior distribution in the limit, but is often impractical. When evidence is surprising (receiving low absolute probability on the prior), most samples will be inconsistent with it. In these cases, the algorithm will have to work very hard – proposing and rejecting many samples from the proposal – in order to obtain a single sample from the posterior.

The way to improve the proposal beyond the prior is to try to take into account the evidence as much as possible, short of doing intractable inference. This can be done by using immediately accessible features of the evidence as inputs to the proposal function; reallocating probability mass in the proposal distribution as a function of which features were present in the evidence. These tailored proposal distributions can place more probability mass on good hypotheses,[68] meaning that fewer hypotheses will have to be evaluated in order to deliver the best ones. Call these tailored proposal distributions 'Evidence-Informed Proposal Distributions', or 'EIPDs' for short (see Figure 1).

[67] The costs of inference for a sampling method is a direct function of the divergence between the proposal and the posterior (Chatterjee & Diaconis 2017).

[68] Hypotheses receiving relatively high probability on the posterior

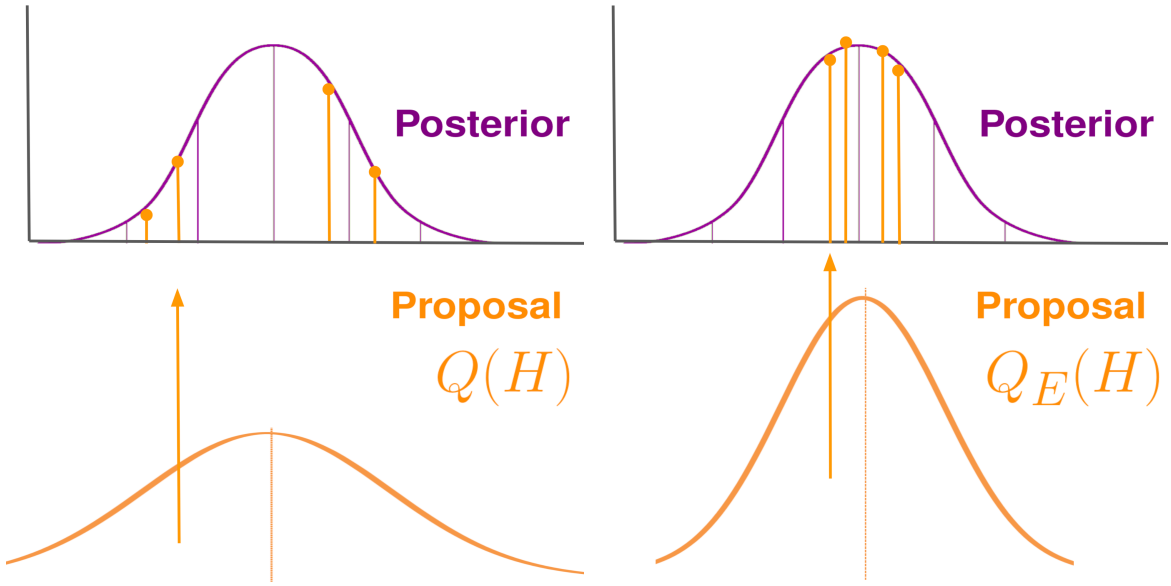


Figure 1. *A proposal distribution Q(H) and an evidence-informed proposal distribution $Q_E(H)$ are both sampled from in order to approximate a posterior distribution or a decision over it. Orange points are samples. The evidence informed proposal distribution places more probability mass around the most likely hypotheses on the posterior, resulting in better samples. (Note that in real world cases of inference posteriors will almost never be Gaussian.)*

[Figure 1]

If a system is solving inference for the first time, there is not a general way (short of doing inference) to know what features in the evidence should be used to inform the proposal distribution or which hypotheses they recommend – that is, no general way to have EIPDs. But EIPDs and the features that inform them can be acquired through prior exposure to the domain. This allows the system to spread out or pre-pay the costs of inference, limiting how many hypotheses must be considered at run time, by drawing on information from prior exposures. Processes of roughly this type have been dubbed 'amortized inference' (Gershman & Goodman 2014) or 'inference compilation' (Le et al. 2017) to invoke these metaphors of spreading out or pre-paying the costs of inference and have been implemented in models for a variety of different tasks, including visual inference (Yildirim et al. 2020).

With these concepts in hand, we can state the thesis precisely:

> **Perceptual Amortization Hypothesis:** Perception is fast and cognition is slow because perception's evidence-informed proposal distributions are more concentrated, placing more probability mass around the most likely hypotheses, than are cognition's evidence-informed proposal distributions.

That this computational difference, between the information theoretic strength of the EIPDs in perception and those in cognition, exists is one of two core claims of this paper. The second is that this difference can explain the *Speed Difference*. As mentioned, when probability mass is concentrated around the most likely hypotheses, effectively prioritizing them for evaluation, many fewer hypotheses have to be checked against the prior and the likelihood in the course of computing inference. We can put a point on this by imagining that we are drawing samples from a proposal distribution until a decision over the resulting approximate posterior will get us within some error, say $\epsilon = 0.03$, of the best hypothesis on the posterior. If our proposal distribution places more probability mass over the more likely portions of the hypothesis space, then fewer draws are needed in order to meet that threshold (see Figure 2).

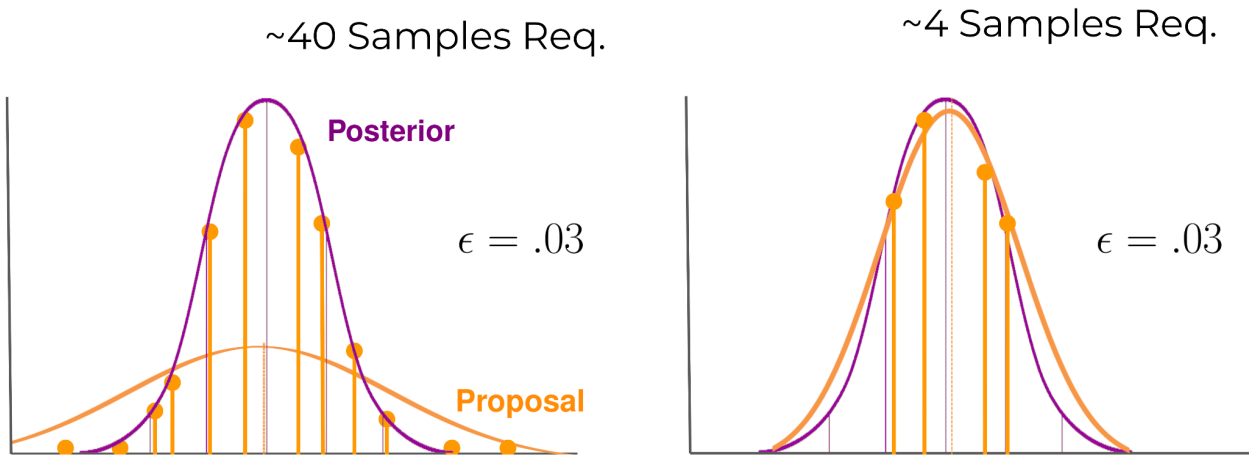


Figure 2. *More concentrated proposal distributions mean that fewer samples are needed in order to get within a given error threshold for successful inference. Orange dots illustrate samples drawn from the proposal distribution but evaluated against the posterior.*

[Figure 2]

The connection between amortization and speed can be demonstrated empirically. Yildirim and colleagues implemented a model featuring amortized inference and compared it to an unaided inference technique based on sampling (Yildirim et al. 2020). The model featuring amortization showed highly accurate responses even when very few samples were taken (see Figure 6). Run on a given input, the amortized model near instantly converges to a highly accurate response, while sampling based inference must be run for hundreds of iterations before it nears the same level of accuracy.

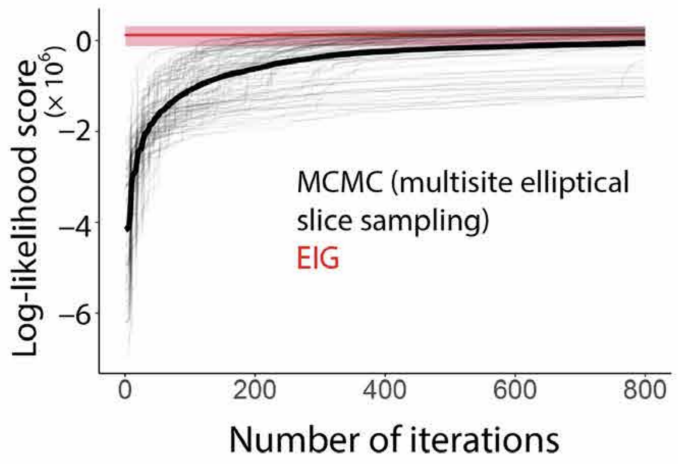


Figure 6. *A model of visual face processing featuring amortized perceptual inference (Efficient Inverse Graphics or 'EIG') in red is compared to a conventional sampling based inference method. (Figure from Yildirim et al. 2020)*

At this point this has all been necessarily abstract. It may be helpful to imagine how these EIPDs are implemented in practice. We can start with an example of the operation of the proposal function in perception. Take a look at Figure 3.

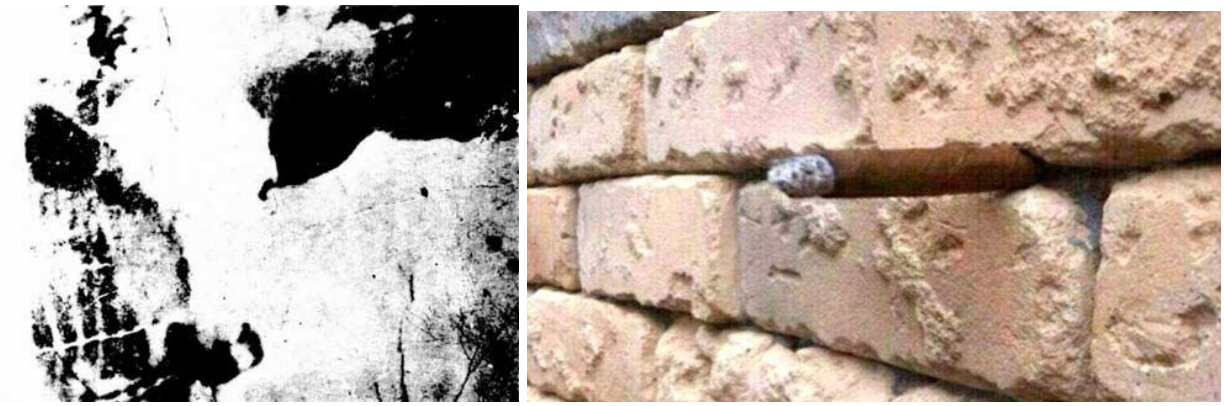


Figure 3. *Upon first exposure to these images most people see some splotches of ink on the left and a brick wall, perhaps with a blue-ish rock wedged in it. See if you can see any more in either of these images before reading the footnote for a hint.*[69]

[Figure 3]

Most viewers will see the images in Figure 3 one way at first and then experience their percept flip after being given a hint as to an alternative interpretation. Most (although not all) will find that they are unable to recover the original percept after undergoing this change. There are different potential explanations for these stably resolving images, but a natural one highlights the role of the proposal function in probabilistic inference. This explanation goes as follows: The percept that the visual system ultimately settles on receives higher probability on the posterior. Because of weak or misleading cues in the image however, this superior hypothesis is not proposed on first encounter with the image, resulting in the first percept. Forcing the hand of the proposal function by giving a hint

[69] With further study or a hint most people will come to see the image on the left as a cow, facing the viewer and leaning over a wire fence. The image on the right can often be resolved to see a cigar, wedged in the wall and jutting outwards, with the blueish 'rock' in the center at the tip of the cigar, with the brown body of the cigar to the right.

results in the superior hypothesis being proposed and evaluated. Because it has higher probability on the posterior, the later interpretation sticks.[70]

A natural follow up question is, what are the cues that serve as inputs into vision's EIPDs? The short answer is I'm not sure. But the place to start is with the features highlighted by vision science. Geometric features such as the presence of parallel lines in an image, filters that can be combined to detect textures or local contours, and complex functions of these, can all be used to prioritize different hypotheses about the 3D form (see Figures 4a and 4b). None of these features *determine* 3D shape – they can all be frustrated by other things going on in the image, from occlusion and overlap, to coloring patterns, atypical textures, or poor lighting – but they all carry information about more or less likely 3D scenes.

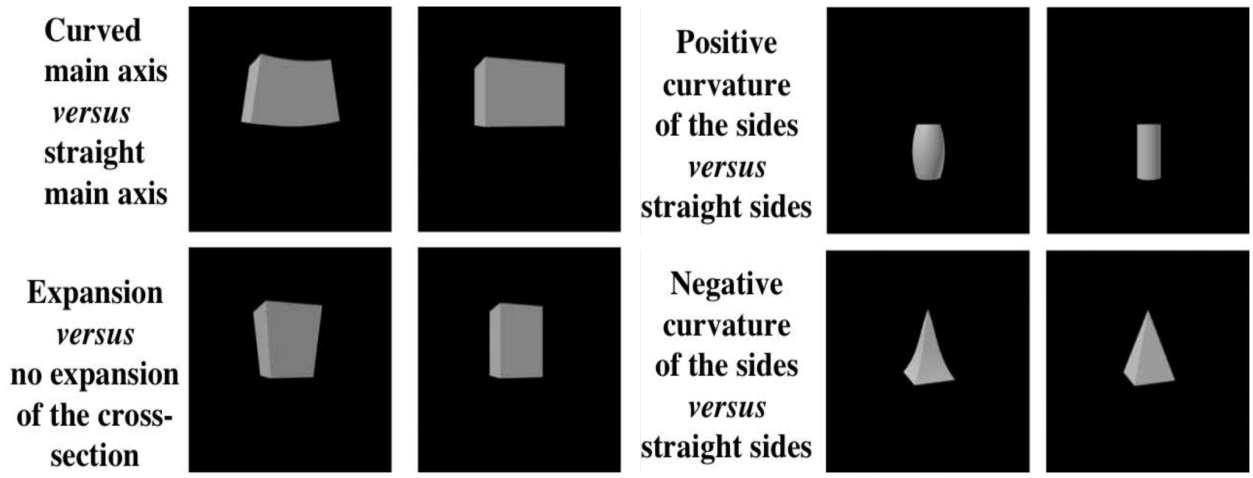


Figure 4a. *Some simple geometric properties of a 3D shape are reliably preserved across many viewing angles. These 'non-accidental' properties do not determine 3D shape – camouflage, overlapping objects of similar color, or merely poor lighting, all mean that even these simple 3D properties cannot be read off the image – but detection of such 2D properties can still help prioritize certain 3D hypotheses (Witkin & Tenenbaum 1983, Biederman 1987; Image adapted from Amir et al. 2012).*

[70] The mode of action of the hint could be cognitive penetration or some low level attentional strategy – this account of stably resolving images is agnostic as to the mechanism by which the proposal function is influenced.

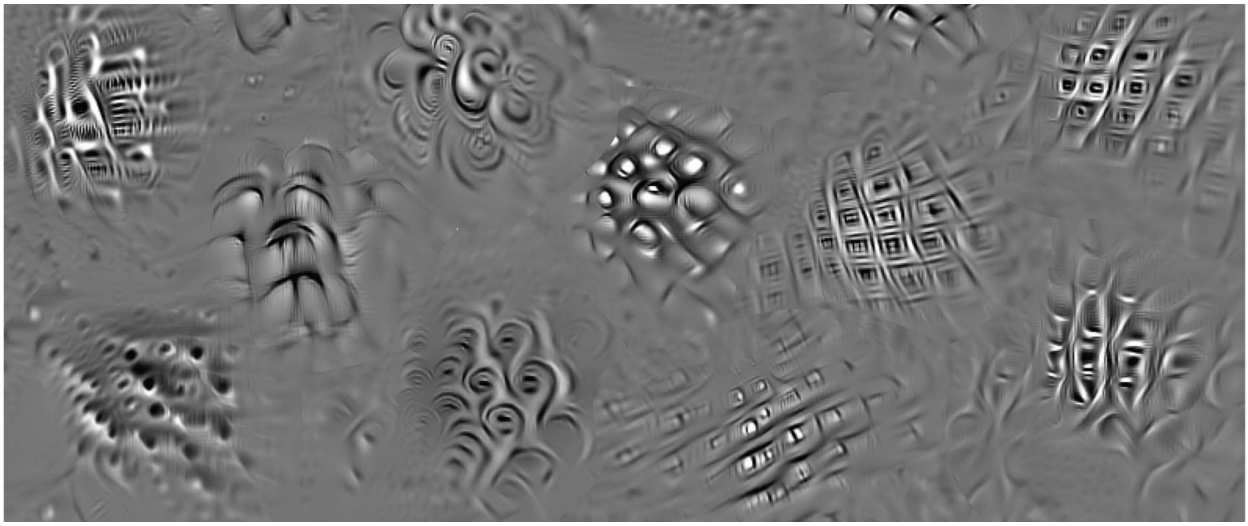


Figure 4b. *Features used to inform EIPDs can be categorical – either present or absent in the image – or graded – more or less present in degrees. Soft features, such as these, may nevertheless carry significant information about plausible 3D scenes, especially when used in tandem with one another. These wild looking features are the result of optimizing for discriminability between images – and certain neural populations appear to be selective for them (Bashivan et al. 2019).*
[Figure 4]

We can use features like those in Figure 4 to tentatively diagnose the failures of vision's proposal distribution in Figure 3. The first image in Figure 3 is highly overexposed, significantly degrading any local geometry or texture information. There are no edges to be detected between large swaths of the cow's head and the background and minimal texture information throughout. In the brick image, the cues are not absent but rather *adversarial* – the rough colinearity of the cigar with the crevice in the brick wall effectively masks one of the image-detectable features that might have otherwise been used to prioritize hypotheses that include a protruding object. (That these cues offer an explanation of why we miss things in these images on first encounter suggests that they are being used to inform visual proposals, while the fact that we ultimately resolve the images even while these features remain degraded or adversarial suggests that vision is evaluating 3D hypotheses against something more than just those features.)

This gives us a sense of how we might think about EIPDs in vision and the features that inform them, but I want to stress that my chief claim is implementation agnostic. One could disagree that these features are represented in vision or that this is the right explanation of stably resolving images and yet still agree that the *Amortized Perception Hypothesis* offers the best explanation of the *Speed Difference*.

I'll end this section by saying a few words about how this idea relates to one other influential idea in vision science, that of Marrian 'natural constraints.' The phrase 'natural constraints' is often used to express the idea that there is implicit information about how the perceptual world works inside of perceptual systems. The canonical example of natural constraints is that vision assumes that 'light comes from above.' This assumption allows otherwise ambiguous shading in the 2D image to be resolved as 3D shape.[71] This idea gets something deeply right about perception. That said, the idea is too vague to do the work we want to do here. For one, it's ambiguous between a claim about the prior and a claim about the proposal. When natural constraints are used to explain why inverse problems have well-formed solutions, that makes 'natural constraints' a claim about the prior (this is sometimes made explicit, e.g. Mamassian & Landy 2001). Since I'm arguing for a claim about the proposal function, this is orthogonal to my thesis. There's room for other versions of the Marrian idea that see natural constraints as contributing to the proposal function instead. In that case, my view is a version of this idea. Even then, the computational and information theoretic vocabulary is needed to make the comparison between perception and cognition that I'm interested in. Cognition's proposal functions are presumably also influenced by the evidence (and so feature natural constraints, in this sense, as well). It's not clear that there are *more* natural constraints in perception than in cognition. The computational and information theoretic vocabulary is needed to state a difference between the two that could explain the *Speed Difference*.

With that clarification out of the way, we can turn to other possible explanations.

[71] A phenomenon known as 'Shape from Shading.'

IV. Alternative Explanations?

So far we've seen some evidence that seeing is fast and thinking is slow and I've offered a hypothesis about why that might be. In this section, I want to look briefly at three alternative explanations and say a bit about why I think each is less promising than the *Perceptual Amortization Hypothesis*. These potential explanations are problem size ('Perception is faster because its problems are smaller'), hardware differences ('Perception is faster because it is implemented by faster hardware'), and accuracy ('Perception is faster because it solves its problems less accurately'). Examining why these alternative explanations do not work will also help us deepen our understanding of the problem and what a potential solution must offer.

Consider problem size. I mentioned that the size of an inference problem is a big part of what makes it difficult. A natural thought then is that if cognition's problems were much bigger than perception's, then this could explain the *Speed Difference*. To put it bluntly, I think this approach is not promising. This is because the empirical facts point in just the opposite direction of what the explanation calls for. Perceptual problems which are solved quickly seem to be much *larger* than cognitive problems that are solved slowly.

Take the category learning experiment discussed in Section II. We saw that subjects take about 50 seconds to learn a category defined by boolean operations and simple quantification over visible properties. The size of this inference problem is given by the number of possible expressions made up of the relevant literals, connectives, and quantifiers. The space of such expressions is in some sense infinite, but we can be confident that subjects are not entertaining arbitrarily long formulas. The model of category learning that the authors used in the paper, which delivered good fits to human learning times, considered logical rules using a maximum of 2 quantifiers and 4 literals, with conjunctions and disjunctions as needed to combine them. Taking this model at face value, we can place a loose upper bound on the number of syntactically distinct hypotheses of about a billion expressions.[72]

[72] Assuming at most four conjuncts or disjuncts, since adding more is guaranteed to be syntactically invalid, the number of syntactically distinct possibilities is bounded by 8^10, or about 1 billion (8 possible syntactic elements and 10 slots). Since that includes meaningless strings like '∧ ∧ ∧ ∧ ∧ ∧ ∧ ∧', semantically redundant strings that include the same literal

A billion is a lot of hypotheses. It's far more than you'd want to write out. But it's also negligible relative to the size of inference problems solved in perception. Take an extremely simple case of visual inference: imagine you are looking at an object against a blank background. The number of distinct hypotheses here is given by the number of different visible properties that the object could have. We'll consider a small subset of these: color, location, shape, and lighting condition. The number of possible 3D scenes in this problem is a function of how many colors, locations, shapes, and lighting conditions we can see. We can put estimates on these. For color, it's estimated that we can see between 1 million to about 10 million. We'll say it's a million (10^6). Lighting conditions can be discriminated along dimensions analogous to object color, the brightness, chroma, and saturation of the light (Tokunaga and Logvinenko 2010) so we might venture that there are another million lighting conditions. For location there are not well documented estimates, but we can conservatively imagine that we can visually distinguish at least a thousand different locations in each direction of visual space.[73] That gives us about a billion possible locations.[74] The last dimension, shape, is the hardest to estimate. How many different objects are discriminable, when shown side-by-side, on the basis of shape alone? Even in just two dimensions the number must be huge – Consider the number of classes of objects that are discriminable based on their silhouettes, from chairs, to keys, splotches of ink, or animal figures, and the number of individual objects that are discriminable within these classes (e.g. the number of chair silhouettes that can be discriminated between). In 3D the number of possible shapes will be even greater. While precise estimates are tricky, I'll offer that there are a billion possible values for visual shape as an extremely conservative estimate.

Multiplying these dimensions together we get a hypothesis space of 10^30 possible ways that an object could be. 10^30 is much bigger than a billion. And this is for a toy visual inference problem. One could try to resist these numbers in various ways, for example, by imagining that vision solves

---

conjoined, and strings differing from the syntactic constraints placed on the authors model (e.g. with more than 2 quantifiers), this number represents a very loose upper bound.

[73] Discriminability places a lower bound on the number of values that can be represented, since discrimination requires distinct mental representation. The number of values could be higher however, if for example, visual noise or lack of cues prevent the optimal use of distinct visual representations for discrimination.

[74] Note that distortions in visual space may mean that there are fewer discriminable locations along certain dimensions, like depth, than others (Green & Rabin 2019).

inference for these different dimensions piecemeal, avoiding the exponential number of interactions. But breaking problems up in this way means giving up the ability to represent dependencies between dimensions of the kind that are widespread in perception. I've chosen these four properties of visual 3D scenes because there is strong evidence that they are jointly recovered, as illustrated by psychophysical effects such as shape from shading (Ramachandran 1988), shading cues to lighting direction (Morgenstern et al. 2011), and the Mach Card (Bloj et al. 1999). While modularity internal to perceptual modalities may play a role, it does not seem like it can tame the vast sizes of perceptual inference problems (See Brooke-Wilson Forthcoming for a more discussion).

So the empirical facts seem to fly in the face of a would-be explanation based on problem size. This argument is quick and there is much, *much* more to be said on the topic, but I think considerations like those above make a prima facie case against the problem size explanation – enough to warrant looking elsewhere.

Another possible explanation of the *Speed Difference* comes from differences in hardware between perception and cognition. There are many possible differences that might be relevant here, but we'll focus on two that researchers have emphasized. One is parallel processing. Many have remarked that much of what goes on in perception happens in parallel, while processing is cognition appears to be more serial. Another possible difference has to do with the time of the basic operations. Areas dedicated to perceptual tasks may be packed more tightly together, in visual or auditory cortex for example, while cognitive processes may involve long range connections that span the brain. Action potentials take time to move through space, and so the added distance might mean that basic operations implementing cognition simply take longer than those implementing perception.

Either of these or other differences in hardware might be part of the explanation of the *Speed Difference*. But they won't be most of the explanation. This is because the kind of speed ups that these differences off are too small relative to the differences in problem size and speed between perception and cognition. Concretely, the most that either of these hardware differences could offer is a factor of N speed up (if perception has N processes running in parallel to cognition's 1, or if perception's basic connections are N times shorter than cognition's). But what's needed to explain the *Speed Difference* is

an exponential speed up. Our toy case of visual inference is 21 orders of magnitude larger than a simple cognitive inference, despite taking two orders of magnitude less time to solve. Very crudely, that means we're looking for something north of a 23 orders of magnitude speed up. Since there are only about 100 billion (10^11) neurons in the brain and many, many fewer in visual cortex, parallel processing can't do most of the work that needs to be done here. Shorter range connections are even more hopeless. The main explanatory speed ups, those doing most of the work in closing the gap between perception and cognition, will have to come from elsewhere.

That last candidate explanation that I'll consider has to do with accuracy. Even very hard inference problems can be solved quickly if they can be solved inaccurately. In the limit, a system can give a *random* answer as fast as it can roll an internal die. Giving more accurate answers often takes more computational work. This is why speed-accuracy trade-offs are ubiquitous in psychology, showing up in perception (Heitz 2014), cognition (Wang & Xu 2015), and in non-human animals (Chittka et al. 2009). They are a common feature of many methods for inference in computational statistics.[75] If perception were much less accurate than cognition, then this could offer a potential explanation of the *Speed Difference*.

I think this too is not promising, largely because, like the problem size explanation above, the empirical facts appear to be just the opposite of what the explanation calls for. It seems perception solves its inference problems very accurately, while typical cases of cognitive inference exhibit striking inaccuracy.

Start with some theory. In inverse inference, the data tend to underdetermine the state of the world, with the former suggesting but not entailing the latter. This means that the best possible performance is often not perfect accuracy, but rather some other limit, defined by the amount of information in the data. Performance that is 'optimal' in this sense – delivering the maximum expected accuracy – offers a natural point of comparison when considering the accuracy of inferential systems. Perception appears to be close to optimal, in this sense, while cognitive inference is comparatively far off.

[75] For example, all sampling-based inference and numerical optimization.

The literature evincing optimality in perceptual inference is vast. To take an example, Brainard et al. (2006) built a bayesian model of color constancy in simple images, using a naturalistic prior over lighting condition to recover a posterior over surface colors. The authors showed that human color judgements were well modeled by optimal responses from this model, including both the successes and failures of color constancy. Similar optimality results have been found for many other perceptual inferences, including inferences to recover contours (Geisler & Perry 2009), orientation in depth (Knill & Saunders 2003), size (e.g. Ernst & Banks 2002), location across a range of modalities (e.g. Van Beers et al. 1999, Kording & Wolpert 2004, Battaglia et al. 2003), and even number of perceptual events (Bresciani et al. 2006) (for reviews, see Ma 2019; Rescorla 2015).[76]

In contrast to the impressive accuracy of perception, cognitive inference exhibits striking degrees of *inaccuracy*. Some of these effects are quite stark – like people's tendency to neglect information about base-rates (Kahneman 2011). Other patterns of inaccuracy are more subtle. Studies of cognitive inference across many domains suggest that subjects give sub-optimal responses with a puzzling kind of regularity. In the category learning studies, for example, subjects appear to 'probability match'. Probability matching describes a pattern of responses in which the frequency of a given response roughly matches the probability of that response on the posterior. Subjects employing this strategy will respond with an answer that has only a 5% chance of being true 5% of the time, will give a value with a 10% chance of being true 10% of the time, and so on. This response strategy leaves considerable accuracy on the table compared to the optimal strategy seen in perception. Apart from category learning, probability matching and similar behaviors have been documented in cases of causal inference (Denison et al. 2013), prediction of chancy events (Vul et al. 2014), and in reasoning about causal interventions (Meng et al. 2018, Nussbaum et al. 2020).

There's much more to be said here, but an initial look at the literature suggests that perceptual inferences are significantly more, not less, accurate than those in cognition. Like for the problem size explanation, the empirical facts are just the opposite of what the accuracy explanation calls for.

[76] Disagreements about perceptual optimality tend to exist on the margins. Rahnev & Denison 2019, for example, argue for several ways in which perception deveats from optimal accuracy, including idiosyncratic ways of setting their cost function, sequential effects, and flag debates about whether confidence ratings are themselves optimal. These are many interesting questions here, but none that change the basic story: That perception is extremely accurate.

Perception seems to perform more quickly, and more accurately, on much larger problems than cognition. There are, of course, other potential explanations for the *Speed Difference*, especially related to other algorithmic strategies for taking advantage of problem structure. For the time being, however, I'll leave things here and turn to some positive evidence in favor of the *Perceptual Amortization Hypothesis*.

V. Behavioral Signatures of Amortization

In this section I argue that the *Perceptual Amortization Hypothesis* is supported by convergent behavioral evidence, as the hypothesis offers an account of both the relative accuracy of perceptual inference as well as the particular pattern of errors seen in cognitive inference. I'll then say something about how the hypothesis fits with the exceptions to the general rule – instances of perceptual inference slower and of cognitive inference that are faster.

First consider the relative accuracy of perception over cognition. The *Perceptual Amortization Hypothesis* claims that perception's Evidence Informed Proposal Distributions (EIPDs) concentrate probability mass more effectively around the most likely hypotheses than do those in cognition. When computational resources are limited relative to the size of an inference problem, sampling from a distribution that places more probability mass over the best hypotheses will tend to increase accuracy. The details of this explanation are just the flip side of the account of speed offered in Section III. If we imagine that we have a fixed set of samples to draw, then drawing from a more concentrated proposal distribution leads to better accuracy in expectation (see Figure 5).[77]

[77] Assuming, for simplicity, that the EIPDs are unbiased.

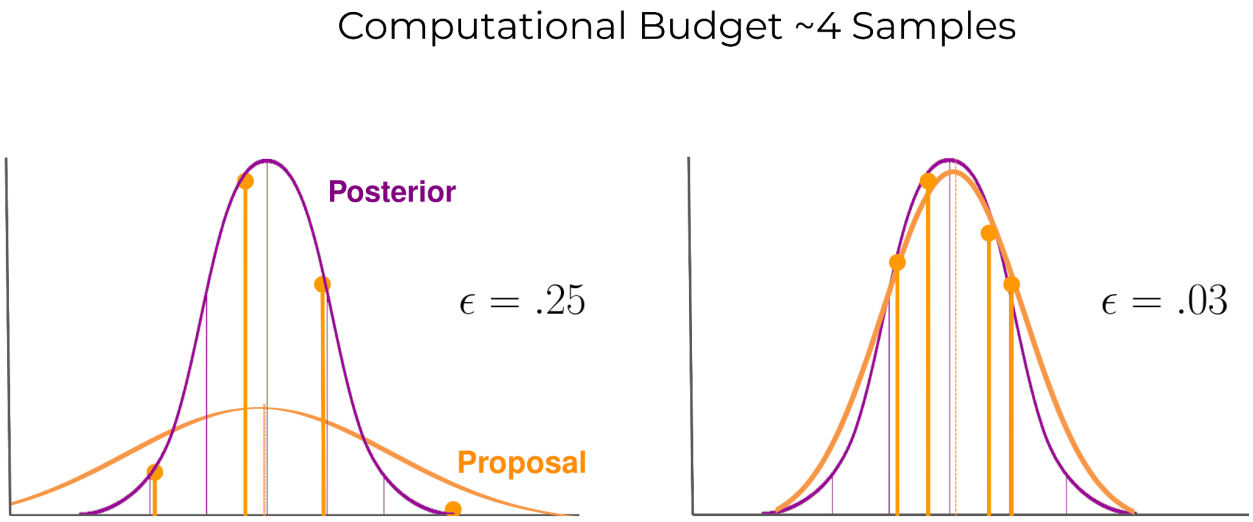


Figure 5. *If, for example, both systems have a computational budget that allows them to evaluate four hypotheses, then the system drawing from a more concentrated proposal distribution will deliver higher expected accuracy (maybe within an error of .03 rather than .25).*

[Figure 5]

The current view can also explain the particular pattern of errors we see in cognitive inference. In the last section I used probability matching as an example of the response patterns seen in canonical cases of cognitive inference. I'll use it as an example here again, but it's important to recognize that the behavior is actually more general. The true explanans is a pattern of responses that is significantly more stochastic than is optimal, but with higher response frequencies associated with higher probability hypotheses.

Probability matching and nearby behaviors are just what you'd expect to see if cognitive inferences involved drawing hypotheses from relatively weak proposal distributions. Drawing from a weak proposal distribution, relative to problem size, means that many hypotheses have to be tested before finding good ones. Since the good hypotheses are preferred over the bad hypotheses in decisions, they are more likely to be returned. In many cases however, the best hypothesis that is found will still have relatively low probability on the posterior, and so low probability hypotheses will still be attested subjects' responses.

To make this all concrete, imagine the limiting case of not using *any* evidence to inform the proposal distribution. As we noted earlier, in this case the best you can do is often just to sample from the prior and accept or reject hypotheses depending on whether they are consistent with evidence ('rejection sampling', see page 11). In the limit, samples from this algorithm converge to the posterior distribution. When the observed evidence is surprising (i.e. low absolute probability) however, very many samples have to be drawn from the prior in order to get one that is consistent with the evidence. Since for real world inference most evidence is surprising, a lot of computational work will have to be done to get even a single accepted sample. If participants in experiments such as the category learning studies were doing rejection sampling, then they might only have the computational resources to deliver a single sample that passes this demanding check. Decisions made over this one sample approximation to the posterior will deliver exact probability matching behavior.

This particular explanation of probability matching-like behavior is probably too neat for a couple of reasons. For one, I don't think it's likely that cognition is actually rejection sampling – the method is just far too inefficient. For another, we shouldn't be too confident that people are exactly probability matching. Many experiments involve fitting a prior to subjects' responses, which can make it difficult to diagnose exact probability matching from nearby behaviors. The experiments that have looked at cognitive inference with *independent* estimates of the priors have found evidence for both exact probability matching (Denison et al. 2013) and in other cases for response strategies that are slightly lower variance (although still far from optimal, e.g. Mozer et al. 2008).[78]

While rejection sampling account of probability matching behavior isn't exactly right, it illustrates a more general account of the more general behavior. When a sampling-based inference algorithm draws samples from a more diffuse proposal distribution, more work will have to be done in order to find hypotheses that score reasonably well on the posterior. When the computational budget is limited, the system will not reliably find the best hypotheses. However, because it can recognize better hypotheses when they are sampled, the system will be more likely to deliver better hypotheses than weaker ones. This is just the behavior we see in canonical cases of cognitive inference.

[78] Mozer et al. show that in the particular case of prediction about familiar topics subjects' behavior is roughly what you'd expect if they were drawing two samples from the posterior and making a decision on that basis.

The final bit of behavioral evidence I want to discuss here are the exceptions to the rule – cases of slow(er) perceptual inference and fast(er) cognitive inference. The amortization hypothesis suggests that there will be instances of slower perceptual inference when our EIPDs are frustrated by a lack of cues in the image or adversarial arrangements of cues. We'll see a possible example of this in the next section. The hypothesis also suggests that there may be special instances of cognition where cognition has extensive exposure to a circumscribed domain and so becomes faster and more accurate for much the same reasons that typical perception is fast and accurate. Cases of expertise may be like this. People who have extensive exposure to circumscribed domains, from chess to medical diagnosis, tend to be able to find better solutions more quickly than novices who may know the same information (e.g. the rules of chess) but lack the experience. Such speed and accuracy may be due to the mind's ability to learn concentrated EIPDs with extensive exposure to a cognitive domain. In this way, amortization can offer a unified account of both the *Speed Difference* between typical perception and cognition and the exceptional cases of faster cognition and slower perception.[79]

VI. Neural Signatures of Amortization

In this section I argue that the *Perceptual Amortization Hypothesis* draws convergent evidence from recent findings in computational neuroscience.

Computational neuroscience aims to develop quantitative models of what different brain areas are doing and how that contributes to intelligent behavior. One source of evidence that a particular model captures some aspect of the representations and computations going on in the brain is if the model can predict biological neural activity when the brain is performing a similar task to the model, say, processing a visual stimulus. When it comes to the computational neuroscience of vision, the most successful models – as judged by this metric – are Deep Convolutional Neural Networks (DCNNs). A *neural network* is a set of nodes and directed edges each of which take on real numbers, called *activations* for the nodes and *weights* for the edges. The activation of a given node is a function of the activations of the nodes leading into it, multiplied by the weights of those edges, and passed

[79] For a research program showing further evidence of amortization in cognition, see Dasgupta et al. (2020).

through an 'activation function'. A network is said to be 'deep' if it has many layers of nodes and 'convolutional' if the weights satisfy some further constraints that intuitively require different nodes in the same layer to compute the same function.

DNNs, convolutional and otherwise, can be 'trained' to compute interesting functions. We might train a DNN to do image classification by first encoding an image into the activations at the first layer, passing those activations through the network to get a readout – a pattern of activations in the final layer that we can treat as a classification of the object in the image – and then tweaking the weights in the network if the output differs from what we wanted. DNNs trained in roughly this way can learn very complex and useful functions. They might, for example, learn a series of visual features which are diagnostic of category membership. To do this, they generally require lots of data (e.g. millions of images) on the same or very similar problems (e.g. classifying the objects in those images).

As I mentioned, DCNNs have proved the best models to date of the online computations involved in perception, at least as measured by the capacity to predict neural data. Artificial neurons in DCNNs trained to perform visual tasks like object classification can be mapped to populations of biological neurons (e.g. in the visual cortex of non-human primates), such that the activations of the artificial neurons in response to images predict the activations of the corresponding biological neurons when the animal is shown the same image. This neural predictivity can be seen both quantitatively – explaining about 50% of the variance on average – and qualitatively – in the sense that the models spontaneously reproduce aspects of visual processing like the visual hierarchy, with earlier layers of the model best predicting activation in anatomically earlier visual areas and later layers best predicting later visual areas (Yamins et al. 2014, Güçlü & van Gerven 2015, Khaligh-Razavi & Kriegeskorte 2014). The resulting predictions generalize far outside the training images and even outside of naturalistic input. Stimuli that are engineered to excite artificial neurons outside of their typical range produce atypically large activations in the corresponding biological neurons (Bashivan et al. 2019, see Figure 4b for example features). This neural predictivity appears to be a byproduct of the model being trained to perform the visual task – models that do better at classification accuracy are also better neural predictors (Yamins & Dicarlo 2016). So far this line of work has focused on vision, but there are early

signs that some of these results may generalize to other modalities. DNNs trained to localize sounds predict neural activity in auditory cortex (Kell et al. 2018).

While results like these are impressive, the current generation of DNNs also leave a lot to be desired as models of human perception. For one, current DCNNs exhibit striking behavioral differences with people, most famously showing susceptibility to 'adversarial examples' (images that the network confidently classifies incorrectly but that would never fool a human being) and exhibiting biases to learn classifications on the basis of texture and hyper-local geometry, rather than the global shape bias characteristic of human classification (See e.g. Firestone 2020 for a theoretical discussion, and Bowers et al. 2023 for catalog of these effects).

Another shortcoming of the current models has to do with neural predictivity itself. The current models that are so successful at neural prediction for most images fail dramatically for a subset of images that are classified more slowly by primate vision (Kar et al. 2019).

To unpack this a little: of the images that primate vision classifies correctly, some are solved more slowly than others. This slower solution time shows up both in the reaction times of monkeys performing the classification task and in the time it takes for the image to be decodable from the relevant areas of visual cortex.[80] Images can be binned according to this 'solution time.' Most images are fast resolving, in the sense the object category can be decoded in around 100ms, while other images are slower resolving, decodable only after 150, 200, or 250ms. The neural predictivity of DCNN models is impressively high for the fast resolving images, but rapidly falls off for slower resolving images, and trending toward chance performance for the slowest (see Figure 6).

By the same logic that was used to argue that DCNNs really do capture something about the representations and computations used in temporally early vision, the representations and computations that are involved in temporally later vision seem to be missing from these models.

[80] This is a restricted, linear decoding, rather than a flexible non-linear decoding. All of the decoding and mapping results mentioned in this section use linear mappings.

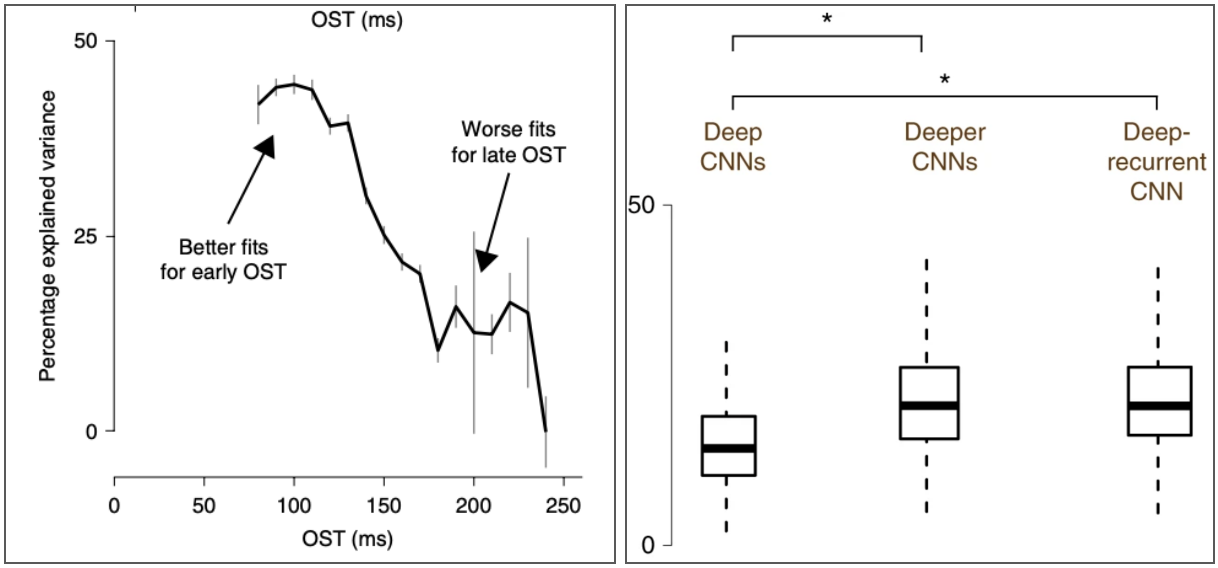


Figure 6. *Neural predictivity of DNN models trends rapidly downwards for slower resolving images (Left). OST is 'object solution time,' the time for the image classification label to be decodable from the relevant areas of macaque visual cortex. Adding layers or recurrence to DNNs delivers better predictivity, but nothing like the earlier successes (Right) (Graphs from Kar et al. 2019).*
[Figure 6]

A natural thought is that what's missing from current DCNNs is recurrent processing. Indeed, we know based on pharmacological interventions that recurrent processing in the brain is a big part of how these slow resolving images get resolved (Kar et al. 2021). Adding recurrence to DCNN models helps a little, but intriguingly, doesn't remotely return neural predictivity to the highs seen for temporally early vision (see Figure 6). This suggests that what recurrence is *doing* in biological vision is different from what it is doing in DCNN models. The upshot of this is that while DCNN models appear to capture something real and important about temporally early perceptual processing, they appear to be much less good models of what happens after the first 100ms or so.

I'd like to explain these findings by way of two theses that tie them to what we've been discussing. The first, which I think is uncontroversial, is that DCNNs of this kind are realizers of Experience-Informed Proposal Distributions. This means that evidence of DCNN-like computations in temporally early vision is evidence of EIPD-implementing computations in early vision, and so

convergent evidence for the amortization hypothesis. The second thesis – which goes beyond the current argument – is that the fits to neural data drop-off for slower resolving images because these are images where the original DCNN-implemented proposal offers a less good initial interpretation of the scene, requiring more explicitly inferential procedures of hypothesis checking to take over. This results in both longer processing times (slow perception) and in representations and computations that differ from those that go on in a DCNN.

Start with the first thesis, that DCNNs are realizers of EIPDs. This should be clear from the training regime of DCNNs and related networks.[81] In order to realize an EIPD a system must use information detectable in the image to concentrate probability mass over more likely hypotheses. The features it uses to do so should be acquired by way of extensive exposure to the domain.[82] This is just what these networks do. At the start of training, when the weights of the network have been randomly initialized, the network outputs random blends of activation patterns in response to input images. These outputs can be, and often are, normalized and treated as relatively flat probability distributions. By vast numbers of exposures to the domain, combined with their update procedure, the networks learn to detect features in the image that are diagnostic of various hypotheses. The features they learn are well understood, if odd, amalgamations of textures and local geometry (see Figure 7). Formally, these are discriminative models – they learn to approximate P(H|E) or a decision over it. This corresponds to learning to concentrate probability mass over more likely hypotheses.[83] In this way, trained discriminative DNNs are realizers of EIPDs.

Note that this is not the controversial claim that these systems are just memorizing the training data (e.g. Marcus 2020). Everything I've said is consistent with DNNs learning features that generalize far outside of their training distribution. Future iterations of these models may even move in the direction of learning more interpretable features (although it's not clear that would make them better

[81] In particular, of 'discriminative' networks (more on this below).

[82] This could correspond to exposure over ontogenetic or phylogenetic time; I won't commit one way or the other.

[83] Assuming the posterior is more peaked than the prior; a safe assumption in this case. Whether DNNs are incentivized to learn the posterior or the maximum of the posterior depends on technical details of their training. Cross entropy against a delta distribution and L2 loss are proper scoring rules and incentivize learning the true P(H|E). L1 loss incentivises placing all probability mass on the most likely hypothesis, argmaxP(H|E).

models of temporally early vision – there is some evidence that textures and hyper-local geometry are just the features that temporally early human vision uses, e.g. Evans & Treisman 2005[84]).

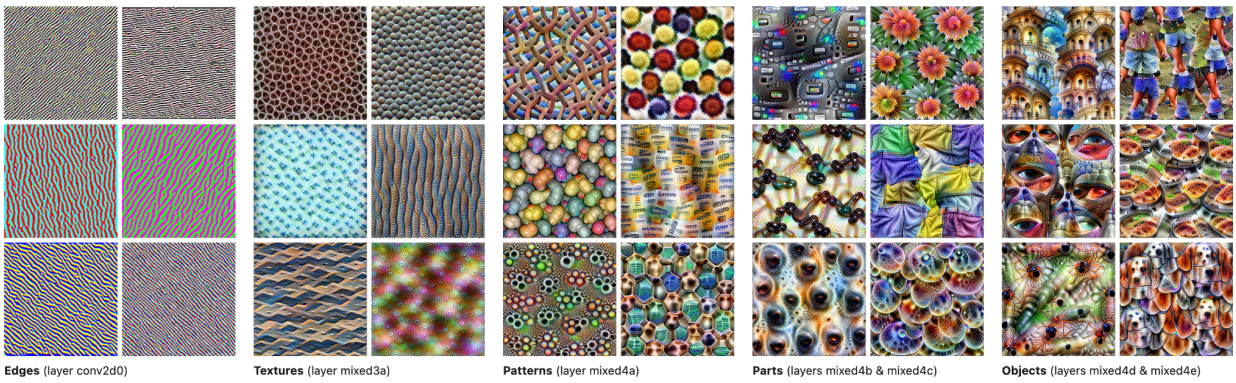


Figure 7. *Visualizations of the features at several layers of a DCNN. These features are optimized for discriminating between images of objects of various categories (Olah et al. 2017).*

[Figure 7]

So discriminative DNNs are realizers of EIPDs. Evidence of computational similarity between temporally early vision and DCNNs then is evidence that temporally early vision involves EIPDs. This is convergent neural evidence for the *Perceptual Amortization Hypothesis*.

What about the drop off in fits to neural data? Well, there are many things that could be responsible for this, but I want to explore one possibility because it's related to what we've been discussing here. This is that slow resolving images are slow because they are images where the original proposal from the EIPD is wide of the mark, perhaps due to degenerate or adversarial cues in the image. In these cases, later inferential processes of a different nature kick in. These involve making changes to the hypotheses originally prioritized by the EIPDs and explicitly checking these new hypotheses against the incoming data. The result is longer processing times, because more hypotheses have to be checked, and greater *dissimilarity* between biological neural processes and DCNNs ,

[84] ... the features we are referring to need not be simple hardwired physical features but may be learned features of intermediate complexity (cf. Ullman, Vidal-Naquet, & Sali, 2002) that characterize a target category (e.g., beaks or open wings for birds, smooth shapes and metallic textures for vehicles). (Evans & Treisman 2004, p.1477)

because of the change in both representations and computations. This dissimilarity increases the longer these late stage processes run for.

Concretely, what these late stage inferential processes might involve is making adjustments to the 3D scene proposed by the EIPD and then comparing the image predicted by the new hypothesis to incoming data in order to score that hypothesis. This can be done repeatedly to test novel 3D scene representations in a sampling procedure not unlike what we saw in cognition in Section V. For example, the EIPD might propose that vision is looking at a panther in a tree in such-and-such a position. This hypothesis about the 3D scene can be used to generate a prediction – the 2D image that such a scene is likely to produce – which can be compared to the observed image to assess the hypothesis' likelihood. When the predictions stop improving or some other criterion is met, vision settles on that hypothesis.[85]

While admittedly speculative, there are reasons to take this idea seriously. For one, these slow resolving images are images that primate vision gets right, but that DCNN models get wrong. Whatever these unmodeled processes are, they seem to be delivering improved classification performance. Sampling-based inference fits the bill.[86] Second, there's ample neural evidence that prediction plays a role in online perception – evidence that's usually marshaled to argue for predictive coding views (see Howhy 2013 for review). The scoring of sampled hypotheses involves making predictions about the incoming sensory data in just this way. Third, there is evidence that in cases where perception oscillates between different percepts because neither is strongly preferred (an extreme case of slowly resolving percepts) these temporal dynamics are well modeled by sampling-based inference (e.g. Gershman et al. 2012). Finally, when DCNNs are trained to amortize inference in a model capable of sampling-based inference, the resulting networks deliver better fits to neural data than do networks which are trained on classification tasks alone, without the support of a sampling-based model (Yildirim et al. 2020). Each of these give us some reason to believe that temporally late visual processing is implementing a distinct, sampling and prediction-based approach to inference, compared to the discriminative procedures of temporally early vision.

[85] Or, in some artificial cases like binocular rivalry, cycles forever.

[86] See e.g. Kulkarni et al. 2015 for an implementation of this idea and comparison to neural baseline.

If this second thesis is right, it helps to unify some distinct traditions in vision science. Some traditions emphasize discriminative models, including both DCNNs and the feature detection models of classic vision science (see Figure 4). Others emphasize 'generative' models, or those that can generate images in virtue of representing a joint probability over hypotheses and evidence, P(H,E). These include many predictive coding models, bayesian models more generally, and certain other neural networks. The current view is one on which both of these traditions get something deeply right about perception. Explaining both the speed as well as the dynamics and robustness of vision may require both a discriminative and a generative model working in tandem.

To sum up then, the success of CNNs at predicting neural activity for temporally early vision suggests that primate vision has access to a function realizing strong EIPDs, offering convergent evidence for the *Perceptual Amortization Hypothesis*. That a particular model should deliver impressive neural fits to fast resolving images and progressively worse fits to slower resolving images is mysterious until we imagine that the model in question is a model only of the first part of vision – the proposal distribution. Imagining how EIPDs fit in with the rest of visual inference helps us make sense of these otherwise puzzling findings.

VII. Conclusion

We started this paper with the question 'why is seeing fast and thinking slow?'. I've argued that the reason has to do with the way in which perceptual problems are mutually informative about one another's solutions, allowing perception to leverage large amounts of prior exposure to the domain to anticipate the results of computationally challenging inference. This allows the costs of inference to be prepaid, significantly reducing the time that must be spent solving any individual instance. In contrast to perception, cognitive inference problems are much more diverse. People have much less exposure to the full breadth of cognitive inferences, relative to that diversity. This limits the role that amortizing inference can play in typical cognition (although in the special cases where people do have extensive exposure to a circumscribed domain, the strategy can be leveraged to some degree). I showed how this *Perceptual Amortization Hypothesis* could explain the speed of perception, as well as a raft of other data,

including perception's relative accuracy, the particular pattern of errors seen in cognitive inference, and otherwise puzzling findings in computational neuroscience. I'll end here with some thoughts on how this fits into larger discussions about perception and cognition and the interplay between AI and cognitive architecture.

One set of consequences has to do with the use of speed for identifying ambiguous mental processes as perception or cognition. The current investigation brings with it both good news and bad news for these arguments. The good news is that typical cases of perception really do appear to be faster than typical cases of cognition and we have a computational framework that suggests that this pattern is non-accidental and likely to carry over to as of yet unstudied instances of both.[87] The bad news is that the theory predicts that some atypical cases of cognition may be fast for much the same reason that perception is. Cognitive inference will be fast when the inference is relatively circumscribed and people have extensive experience performing it. Unfortunately, that set plausibly includes many of the cases where speed has been invoked to argue that a given process is perceptual. This includes arguments for concepts, confidences, and demographic features in perception. These inferences may ultimately turn out to be perceptual, but showing that this is the case requires controlling for expert cognition. We may need new ideas about how to do this.[88]

Another set of consequences has to do with the interplay of AI and cognitive architecture. If the current view is right, then there are deep computational differences between the kinds of inferential operations that happen quickly in the human mind and those that take more time. The speed of perception is explained by the fact that relatively more of what goes on in perception is done by these fast operations. The slower processes still play a role, even in perception, in allowing for recovery when experience-informed proposal distributions are frustrated by degraded or adversarial cues in an image. Relatively more of the work in cognition is done by these slow processes. That has some downsides – cognition is slow, relatively inaccurate, and limited to small problems – but also some upsides – cognition can engage in sustained reasoning in domains with which it has extremely little exposure. AI of late has seen unprecedented progress across a wide variety of perceptual and cognitive tasks, from

[87] We can also add tests for accuracy to these methods as a related diagnostic feature.
[88] We might start by first assessing whether expert cognition can ever be *as fast* as perception.

object classification, to game play, to delivering fluent prose. Debatably, however, a lot of this success has been limited to things that people do quickly. If there is a deep computational difference between the things that we do quickly and those that take more time, then we should be cautious about extrapolating progress in AI on the former into the future. As we look to automate more of the tasks that people solve slowly, we may need a broader toolkit.

Chapter 3: An Architecture For Central Cognition

Abstract: People are able to navigate a world of tremendous complexity. We keep track of many things that touch our lives, from the aesthetic preferences of a partner to the evidence for scientific theories. When we see something on the news we can recognize consequences for disparate parts of our lives, from our financial decisions to the well being of a close friend. We put that information to use in plans that succeed often enough to shape the world around us. How do we do this? One of the chief challenges in answering this is taming the massive computational costs that arise as we approach the problem. (The challenge is sufficiently daunting that many thinkers have concluded doing so is impossible.) This paper attempts to sketch an answer to this question. I first discuss the computational costs that arise in the course of straightforward attempts to reproduce human cognition and the impossibility arguments that appreciation of those costs have inspired. I argue that certain methods in contemporary AI get us part of the way towards a solution, overcoming key barriers. These methods represent an important part of an account of human cognition, but fall short in key ways. I present one way that these methods could be used as part of a larger system capable of more fully addressing the challenge.[89]

[89] Many thanks to EJ Green, Josh Tenenbaum, Jack Spencer, Alex Byrne, Ned Block, and Laurie Paul for comments on earlier drafts of this paper, and to Lionel Wong, Alex Lew, and Tan Zhi-Xuan for extensive discussion.

I. Introduction: An Effective, Efficient Procedure

Imagine you're trying to figure out whether Bob is in his office. How do you do this? Epistemologists have a normative answer to this question. For a Bayesian, you'd start with your prior degree of belief and then update on any evidence that bears on the proposition – maybe the fact that it's a weekday or that the university is in session. The result is a posterior that represents your all-things-considered belief that Bob is in his office. On non-Bayesian views, the right norms might be consistency checking new beliefs against old beliefs to deliver the most conservative possible update, or drawing an inference to the best explanation.[90] One thing that all of these have in common is that they are holistic. The posterior, the best explanation, and the most conservative update are all sensitive to what else the reasoner believes. This sensitivity is very fine-grained. A few beliefs can radically reconfigure the space of what is relevant to think about. Right now, my beliefs about what is going on in Ukraine is not relevant to whether I think Bob is in his office. But if I believe, or found new evidence, that Bob had been recruited by the CIA as their slavic culture expert, then what is relevant to think about when fixing my beliefs could turn on a dime. This kind of holism makes normative operations for fixing belief extremely computationally demanding. The number of different possibilities that need to be considered is enormous – encompassing all possible ways the world could be that would make it more or less likely that Bob is in his office. The same applies if we want to find the *best* explanation or the *most* conservative belief update. Under very weak assumptions about what it takes to consider a possibility from a computational perspective, these operations demands astronomically many computational steps (see Brooke-Wilson 2023). The same applies for normative operations for planning and decision making.

A natural thought is that we don't do this, we consider just a few relevant possibilities and connections at any given point. Since we don't consider all the possibilities, the computational costs of doing so are irrelevant. But how do we determine what is relevant to consider on a given occasion? Since it is the above holistic and context-sensitive process that determines what is relevant, computing relevance threatens to be computationally infeasible as well. What's more, people are reasonably good at determining relevance. Good enough that we get by in a world of tremendous complexity. We can form reasonable beliefs about the things that affect our lives, from the whereabouts of a colleague to the plausibility of scientific hypotheses. We form plans to navigate an open ended world, from running errands to managing an investigation. Matching our

[90] If this is different from Bayesian updating, cf. debates in the philosophy of science (Van Fraasen 1989, Lipton 2004, Weisberg 2009, Henderson 2014)

performance in these areas is the holy grail of the field of artificial intelligence. What computable, computationally tractable process underwrites our ability to do this?

This paper examines the computational challenge the mind faces in solving such open world problems, why tools from contemporary AI get us part of the way (but not all of the way) there, and lays out a possible route way forward. Section 2 explores why approximating normative processes of belief fixation is computationally demanding and presents classic arguments that it is impossible to do within a computational framework. Section 3 draws on methods in contemporary AI to offer a counterexample to these impossibility arguments and some positive evidence that part of the problem can be solved. Section 4 presents some evidence that these same methods get us only part of the way toward an efficient procedure; missing critical properties of normative operations that human cognition possesses. Section 5 presents a positive view of human domain general cognition, drawing on a broader set of resources from AI. On this view, the mind uses certain processes to efficiently construct small models, tailored to task demands, and others to approximate normative operations in those models.

II. Intractability and the Computational Theory of Mind

In this section, I reconstruct classic arguments for the thesis that domain-general cognition incurs astronomical computational costs, of the kind that should make us skeptical that such a thing could exist. There is a weaker and a stronger version of this thesis. The weaker version holds only that *exact* domain general cognition – realized by computing exact versions of normative operations like Bayesian inference over a large body of beliefs – is computationally intractable. The more ambitious version of these arguments attempt to show a stronger thesis, that even *approximating* these operations reasonably well is intractable. While I ultimately reject the stronger thesis and the arguments for it, they give us a sense of why meeting the tractability challenge is hard and what is needed to do it.

2.1 Computational Intractability of Domain General Cognition

To inquire whether domain general cognition is tractable, we need to lay some groundwork. As I'll be thinking about it, computational tractability applies to *computational procedures*. A computational procedure is a series of finitely many applications of a finite set of mechanically specifiable operations, without recourse to

anything outside of these operations.[91] When a computational procedure involves too many operations to be feasible, we say that that procedure is computationally *intractable*. There is some relativity built into this definition (Feasible for whom? With what resources?), but many important classes of procedures require genuinely astronomical numbers of operations when applied to real world sized problems, obviating this relativity. If a procedure is intractable in this sense, it follows that the brain is not doing it. While there is significant uncertainty about the computational resources of the brain, it does not have astronomical resources. Computational *procedures* can be used to solve computational *problems*. Computational problems are a set of inputs, a set of outputs, a set of ordinal or metric structure over those outputs, and a mapping from inputs to a given structure over outputs that defines how well the problem has been solved.[92] A computational problem is intractable when the most efficient computational procedure for solving it is intractable.

Why should we think that the normative operations that have been proposed as theories of cognition are intractable? It's helpful to start here by considering the exact computation of normative operations for reasoning and planning, before attempting to generalize the tractability concern to approximate versions of these. Consider what's necessary for the exact computation of, say, Bayesian inference. When I consider whether Bob is in his office, evaluating this proposition normatively to deliver my all-things-considered belief requires considering the space of things that could have kept Bob at home today, could have caused him to come in and then leave, reasons he might be traveling, etc. Evaluating one factor, like the possibility of him traveling, involves considering numerous further questions: Does he have local or distant family? Is today a special occasion? How often does he visit out-of-town friends? Is he attending a conference? Each question leads to more considerations, creating a 'combinatorial explosion' of possibilities. The complexity is a result of the interdependencies among beliefs and, at its limits, closely tracks the number of *concepts* the agent possess. The more concepts an agent has, the more states of affairs she can represent. The thoroughgoing holism of normative operations means that each of these states of affairs must be considered to determine whether it should be updated in light of new information. The number of states of affairs to be considered is an exponential function of the number of logically independent concepts. If I am considering, for example, how the items in a box are arranged, that number of possibilities is given by the number of objects and the number of atomic properties (properties that do not entail one another, e.g. color, shape, location) that can be freely composed. Each addition

[91] This last condition can be weakened in targeted ways, asking for example, what would be tractably computable if the system had answers to particular kinds of intractable queries, say, by asking an oracle.

[92] The metric or ordinal structure reflects the fact that answers to computational problems are not always right or wrong, but are often better or worse than one another. Better procedures are those that deliver better performance on a problem.

of a new property or new object increases the number of possibilities exponentially. (If I start with one item which has one of two shapes and one of two colors, there are 2^2 or 4 possibilities. If I add that it could have one of two sizes, the number of possibilities is 2^3, or 8, and so on.) Under weak assumptions about the computational cost of evaluating a possibility, the result is computational costs that scale exponentially with the number of logically independent concepts (Consideration of space precludes a more thorough discussion here, but see Brooke-Wilson 2023 for more detail). The end result are computational costs that are truly astronomical, outnumbering even for toy cases of belief updating the number of particles in the observable universe. The same concerns apply to normative operations for decision making. In order to know what I should do now, I need to consider both how I take the world to be and the many ways my actions might affect it, entailing exponential costs to make decisions *exactly* according to normative standards such as Expected Utility Decision Theory.

We'll talk about approximation strategies for most of the rest of the paper, but it is worth noting that many approximation strategies still face this problem. Take 'satisficing', for example, the strategy of setting the goal of finding a solution that's 'good enough' rather than the best one. Satisficing was proposed by Herbert Simon (1956), in early appreciation of how the challenge of solving seemingly intractable problems shaped the mind. An agent that satisfices eschews the goal of finding the expected utility maximizing decision (or the best possible explanation, or the most conservative belief update) and settles instead for any solution that delivers some threshold of utility (explanatory virtues, conservatism, etc.). Abandoning the ambition of finding the best possible solutions to our cognitive problems is doubtless necessary to explain tractability, but besides naming a necessary condition, satisficing actually buys us very little. Assuming we're not setting the bar for 'good enough' trivially low (in which case it would be hard to square with human performance in reasoning and planning) solutions that are good enough are still going to be a negligible portion of the space of possible plans, explanations, or updates. One can start to appreciate this by imagining sampling, at random, from the space of possible plans – randomly composing actions and evaluating them for whether they accomplish various goals. For non-trivial goals, one could sample in this way until the end of time and never find a plan that's 'good enough'. The upshot of all this is that even once we've reconciled ourselves to only ever approximating normative operations over large domains, it is far from obvious how to do that within a computational framework, or even whether it can be done at all.

This challenge, often referred to as the 'combinatorial explosion', has been a central challenge for AI since its inception. The failure to appreciate the severity of the problem is often seen as the reason for the disconnect between the exuberance of early AI (many of the pioneers of the field made predictions about when

human-level AI and other accomplishments would be achieved which in hindsight seem to have been wildly optimistic) and the field's subsequent disappointments. The infamous Lighthill report (1973), written by mathematician James Lighthill, which highlighted the failure of AI research to live up to its promises and is often considered a cause of a decade long cessation in funding for the field (the first 'AI Winter'), reads,

> [This report] single[s] out one rather general cause for the disappointments that have been experienced: failure to recognise the implications of the combinatorial explosion. This is a general obstacle to the construction of a self-organizing system on a large knowledge base...

The combinatorial explosion is more than just a practical problem for the field of AI. It's been the basis for skeptical arguments for some surprising conclusions. These include arguments that domain general cognition is impossible, that human-level or 'strong' AI is impossible, and arguments that the computational theory of mind – the view that mental processes are computational processes and the basis for our science of mind – must be false.

Arguing for the first of these conclusions, that the mind must break down into parts dedicated to processing only select kinds of information, Tooby & Cosmides (19940 write that 'combinatorial explosion paralyzes any system that is truly domain-general' (91). This requires that the mind must be implemented via distinct modules, dedicated to at best approximating normative operations over limited domains (one module for our reasoning about other minds, another for our reasoning about physics, and so on). Carruthers (2007) draws a similar conclusion, arguing that '...cognition must be organized into networks of distinct computational systems, whose internal processes are appropriately *frugal'* (53, emphasis in original).

Dreyfus (1972) turns the issues related to the combinatorial explosion into an argument against the possibility of human-level AI. Interspersed with examples from the AI problems of the day and an attempt to develop his own taxonomy of problems, Dreyfus writes,

> [Many problems are] in principle reproducible [by computation] but in fact intractable. As the number of elements increases, the number of transformations required grows exponentially with the number of elements involved (205)

> [This] difference in degree between simple and complex systems turns out in practice, however, to be a difference in kind – exponential growth becoming a serious problem… (207)

> [As a result, certain] human forms of 'information processing' cannot be reproduced in any program. (208)

If human-level AI is impossible, it follows that the computational theory of mind, which holds that the mind is a machine, must also be false. Jerry Fodor, who made his career as a stalwart defender of the project of cognitive science, ultimately came to believe that the inability of computational systems to deal with holistic and global operations falsified the computational theory of mind. In characteristically blunt prose, Fodor (2000) writes that '…the computational theory of mental processes doesn't work for abductive inferences' (41). Because of this, '… sooner or later, we will *all* have to give up on the Turing story as a general account of how the mind works…' (47, emphasis in original).

These are surprising conclusions. They include a very unfamiliar picture of a set of processes that should be deeply familiar to us, the impossibility of the ambitions of a whole field, and the felicity of the foundational assumption of our science of mind. Such conclusions require strong arguments. At minimum, they require more than just appreciating the challenge of the combinatorial explosion, but some positive reasons to believe that the challenge can't be effectively overcome. Dreyfus and Fodor attempt to offer such arguments by establishing the impossibility of tractably computing *relevance*.

2.2 Relevance & The Impossibility of Approximation

So, why think that domain general operations can't be *approximated*? Start with what it means for one operation to approximate another. For present purposes, we'll say that one operation approximates another if the first has an input-output function that is close to that of the second according to some reasonable measure. For example, one function might provide a reasonable approximation to another if the two agree for most problem instances, or for most problem instances they're likely to encounter, or if the outputs of the approximating function are within a certain error of the approximated function. The relevant measure can also be composition of these – for example, the quality of an approximation may be the average disagreement between the two functions, weighted by the probability of encountering an input. We're interested in whether the global application of normative operations can be approximated in this sense.

One way to attempt to approximate an operation that is rendered intractable by the size of its domain is to focus computation on just those elements that make an outsized contribution to the result. This will be a good strategy for approximation when instances are such that a few relevant entries largely determine the result. Call problem instances that have this property *sparse*. As an example of a sparsity, consider multiplying two lists of numbers, each number on the first list with each number of the second list, and then summing the result. Performing this operation requires a number of computational steps roughly the size of the first list multiplied by the size of the second. If, however, both lists contain mostly zeros, then we can save ourselves significant computational work by just leaving those zero entries out. Here we gain considerable computational savings by taking advantage of the *sparsity* of these lists.

Many theorists have noted that a similar kind of sparsity plausibly applies to normative operations for belief fixation and decision making. In most cases of planning or reasoning, only a small proportion of the things I have beliefs about are relevant. When I'm considering whether Bob is likely to be in his office, my beliefs about events in Ukraine are generally not relevant – considering them or not makes no difference to my judgment. It's only in very special instances where such facts become relevant. A natural thought is that, if the mind could limit the domain of its computation to just those considerations that are relevant on any given occasion, then this could dramatically reduce the computational costs of approximating a global application of an operation – perhaps enough to tame intractability.

Skeptics about unified central cognition are well aware of this. Their case for intractability relies on arguments that this apparent route to tractable approximation is illusory. Critically, they point out that what is relevant to a given instance of a normative operation is highly *context sensitive*, in the sense that fine-grained differences in the problem instance entail very different relevant domains. This makes it very hard to determine what is relevant to a given problem instance. Fodor writes that 'because of the context sensitivity of many parameters of quotidian abductive inferences, there is typically no way to delimit a priori the considerations that may be relevant to assessing them' (Fodor 2000, p. 37). Dreyfus expresses a similar idea, writing that 'there do not seem to be any words or objects which are always relevant and always have the same significance…' (Dreyfus 1972, p. 201).

This yields the following argument –

Intractability of Relevance:

P1. Tractable approximation of global operations requires tractable, general-purpose ways of determining relevance

P2. The context sensitivity of relevance means that there are no tractable, general purpose ways of determining relevance

C. Global operations are intractable to approximate.

Why accept Premise 2? The argument skeptics offer is one of exclusion. They argue that none of the broad classes of methods on offer for determining relevance really solves the problem. A first divide is between methods which are themselves inferential – computing the relevance of individual candidate consideration based on their relationship to the background beliefs and the task at hand – and those that are heuristic – using some property as a proxy for relevance. If the process that determines relevance is itself inferential, then that won't explain tractability – it just passes the buck on to a higher level, where the tractability of this higher inferential process must be explained. On the other hand, if the process is heuristic, it is either impracticable or also secretly passes the buck, either to a higher level heuristic or some inferential process that had merely been made imperspicuous. This suggests that the problem is insoluble.

The first of these claims is straightforward enough. If it would be intractable to approximate inference over some large domain, it is intractable to inferentially compute what is relevant – this would require just the kind of normative operation over a massive domain that we were hoping to avoid. Things get interesting in evaluating the second. Why believe that heuristic solutions to relevance are unavailable? Given that relevance facts are acutely sensitive to background beliefs, such that we should not expect any small set of properties to be highly indicative of relevance in all cases, the heuristic solution must either involve many heuristics picking up on different properties in different circumstances, or a small set of highly general heuristics which are sensitive to many different properties. But if we take the first route and have many heuristics, how does the mind determine which to apply on any given occasion? Fodor writes that making this determination would seem to be just another abductive inference –

> Perhaps, then, real cognition in real heads achieves an appearance of abductive success by local approximations to global processes; and perhaps the problem of calculating these approximations is solved heuristically, case by case. Such a proposal would be entirely compatible with the idea that cognition is computation… The prima facie objection to this suggestion is that it is circular if the

> inferences that are required to figure out which local heuristic to employ are themselves often abductive. Which there's every reason to think that they often are. (Fodor 2000, p. 42)

Determining which heuristic to apply by some inferential process really would just knock the tractability can down the road. A natural follow on thought is that perhaps solves this problem heuristically as well, creating a hierarchy of heuristics. Dreyfus points out that this threatens a regress (note that Dreyfus uses the language of 'situations' and 'contexts' instead of heuristics, language reminiscent of the 'frames' of the discussion in AI at the time, but the meaning is the same) –

> This need for prior organization [to determine relevance] reappears in AI as the need for a hierarchy of contexts in which a higher or broader context is used to determine the relevance and significance of elements in a narrower or lower context... But if each context can only be recognized in terms of features selected as relevant and interpreted in terms of a broader context, the AI worker is faced with a regress of contexts. (Dreyfus 1972, p. 200-201)

Just how powerful a hierarchy of heuristics can be depends on a lot of details of the problem. But Dreyfus is right here to police against magical reasoning. We should not be satisfied with merely invoking more heuristics to determine which heuristics to apply to a given situation without some account of where and how the buck stops – without this it's all too easy to brush a deep problem under the rug of implementation details.

If marshaling many heuristics is infeasible, the alternative solution is to have a small set of highly general heuristics. Fodor points out that there do not seem to be any good proposals for what a general purpose heuristic guide to relevance might look like. Even highly general heuristics like, 'if I'm reasoning about things in Cambridge, only consider things in Cambridge' simultaneously rule out to many things (Bob's connection to Ukraine, if I believe those exist) and let's in too many (considering everything else in Cambridge still entails a wildly intractable problem). One possibility that has been seriously offered is to attempt to defer relevance to prior learning. This is the 'Sleeping Dogs' strategy, proposed by AI and robotics researcher Drew McDermott. The thought is that the mind might determine relevance by deferring to the past. The strategy recommends that we 'consider those things that were relevant the last time you faced a similar problem and nothing else.' But this proposal faces a serious problem. There are many natural, true descriptions of events in your past, each of which

groups different past events with the current problem instance. How do you know under which description to group past and present? Fodor writes,

> 'Just do what you did last time.' But what did I do last time? Was it that I tiptoed past a sleeping dog? Or was it that I tiptoed past a sleeping brown dog? Or that I tiptoed past a sleeping canine pet of Farmer Jones's? Or that I tiptoed past a creature that Farmer Jones had thoughtfully sedated in order to enable me to tiptoe safely past it? It could well be that these are all descriptions that I think are true of what happened last time. So, the question I'm faced with is: Which of these descriptions is relevant to deciding what I ought to do this time? (Fodor 1987, p. 119)

The problem then is that the description under which you would like to group past events and present circumstances are those descriptions that determine an events relevance profile. Depending on one's background beliefs, it may be relevant that your interaction was with a dog, Farmer Jones's dog, this dog, or a particular kind of dog. But determining which descriptions are relevant to relevance would seem to be just the problem of determining relevance in a new guise!

The upshot of all of this is that it is hard to see how the mind might tractably solve the relevance problem. Proposed solutions seem to just pass the buck along. Skeptics conclude on this basis that relevance cannot be tractably computed, opening the gates to many revisionary theses about how minds work and what is possible for machines. In the next section, we'll look at what is wrong with these arguments and how relevance might, after all, be tractably computable.

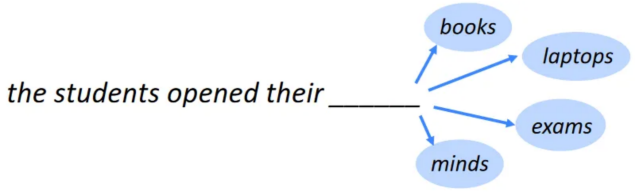
III. Relevance Function

So far we've seen an argument that approximating global operations is intractable. Essential to the argument was the premise that relevance could not be tractably computed (that is, there could be no tractable 'relevance function'). In this section, I draw on a class of contemporary AI systems, Large Language Models, to make the case that tractable relevance functions are indeed possible. If this is right, then the above intractability argument is unsound and tractable relevance functions may well be a resource that an account of cognition can appeal to.

Start with some terms. *Neural networks* are sets of nodes and weights that propagate activations to perform computations in a way loosely analogous to the way the brain works (cf. LeCun et al. 2015). A *Large*

Language Model (LLMs) is a neural network that has been trained on vast amounts of text (generally hundreds of billions or trillions of words) to perform 'next word prediction,' or predicting the next word on the basis of those that came before it.[93] Training is done using the goal of 'masked prediction'. This means that the model is presented with a string of words drawn from the internet with the final word hidden from the model or 'masked'. The model is then tasked with predicting the masked word based on those that it has seen. This prediction is compared to the actual word appearing in the text and the model updated based on the difference between its prediction and the observed word. For example, a model that saw the words 'the students opened their ___' might generate a range of guesses from 'books' to 'laptops', 'exams', or 'minds'. If the word that actually appeared is 'laptops' then the model is updated to be more likely to produce that completion in similar contexts in the future.

Next Word Prediction:



Lopardo (2019)

Models trained in this way learn to generate plausible sounding text for many different contexts. Feed them some input text and they'll complete it in a reasonable way.[94] This makes LLMs extremely general in their applications. Models that are good at next word prediction can be channeled to solve many other tasks that can be cast as next word prediction problems in the right linguistic context. Question answering is one example – often, the most plausible text following a question, like 'what is the capital of France', will be the answer to that question, 'Paris', so question answering can be reduced to an instance of next word prediction. Similarly for machine translation. Feeding an LLM a set of pairs of sentences in a base and target language puts the model in the linguistic context of translation. If the model is then fed a new sentence in the base language, the model will often complete the passage with a translation of that sentence into the target language.

[93] Some models also take into account the words on either side of a missing word. This distinction won't be important here.
[94] Especially when steered by further interventions that we won't get into. See Ziegler et al. (2019) & Bubeck et al. (2023) for further details.

Machine Translation:

1a. The quick brown fox jumps over the lazy dog.

1.b. Der schnelle braune Fuchs springt über den faulen Hund.

2a. Tom eagerly packed his suitcase for the trip ahead.

2b. _____

LLMs have shown themselves to be capable of drawing simple inferences from text, telling jokes, and composing fictional stories to various specifications, among many other tasks. The success of models like this has important implications throughout AI and beyond. What's interesting for our purposes is these models's capacity for *relevance*. Consider the following 'natural language abduction task'. Here, the model is fed two observations with a non-obvious connection between them, for example, 'It was a gorgeous day outside', and 'she asked her neighbor for a jump-start'. The model's task is to provide a hypothesis that connects the observations to one another, e.g. 'Mary decided to drive to the beach, but her car would not start due to a dead battery' (Bhagavatula et al. 2020). Success at this task requires drawing on concepts not available in the observations – in this case, drawing on concepts that could tie the gorgeous day to the need for a jump-start. This makes determining relevance a subtask of this task. Successful performance demonstrates a capacity for relevance.

Natural Language Abduction Task:

Obs1: It was a gorgeous day outside.

Obs2: She asked her neighbor for a jump-start.

Hyp: ___

(Mary decided to drive to the beach, but her car would not start due to a dead battery.)

LLMs can perform this task reasonably well. The best models tested about 80% of the time, as judged by human raters (Allen Institute Leaderboard).[95] The models that have so far been tested are older models (12B parameters), and newer models are very likely to match human performance on this task.[96] This suggests that

[95] Allen Institute for AI https://leaderboard.allenai.org/.

[96] As judged by inter-rater agreement human performance on this task is about 94%.

some LLMs are capable of efficiently computing the kind of relevance that's needed to do well on this task. And this kind of relevance shares much with the kind of relevance that is at issue in the arguments above: The concepts aren't mentioned in the task description but must be brought in by the model; in principle the concepts can come from anywhere provided enough creativity to make the story fit; and the definition of relevance depends on a fine-grained way on the observation statements and the plausible causal connections between them.[97]

A capacity for relevance is echoed in informal experience with these models. The models can riff on just about anything. They draw connections between arbitrary domains and concepts. At times they'll make subtle reasoning errors or blatantly contradict themselves, but they are always *on topic* – bringing in things that are either relevant or plausibly relevant. Ask a good LLM 'what are some ways that the war in Ukraine could impact whether an MIT philosophy professor is in his office?' and it'll produce a strong list of candidate possibilities, including everything from personal involvement in the conflict to campus disruptions due to protest. Tell it you've ruled out the initial possibilities and they'll take this into account, generating a whole new set of relevant considerations. (Anecdotally, people tend to run out of steam before generating as many relevant possibilities as a capable LLM – I encourage the reader to try for themself before looking at the LLM generated answers in Figure 1.) The models are also context sensitive in many of the ways that Fodor and Dreyfus emphasize. Add details to the prompt – like that the professor has been a lifelong pacifist, or simply mention that it's Noam Chomsky – and these details change the relevant candidate possibilities that are offered. The models can even generate reasonably plausible rankings of the probabilities attached to these possibilities (See Figure 1).[98]

[97] There are, of course, other ways in which the task might differ from the kind of relevance at issue above – for example, if the problem is very under-constrained with just two observations, and this may make coming up with just one or a few relevant concepts to establish a single causal link is too easy a relevance problem to be a good proxy for the capacity to generate relevant considerations in the wild. That said, when it comes to relevance, having more constraints means more to go on (LLMs are known to do better when they have more text to constrain their generations), so adding more constraints might make the problem easier.

[98] Note that throughout this paper I will use GPT-4 to illustrate various 'successes of LLMs'. It should be noted that GPT-4 is a very particular LLM, trained in a particular way (using additional training generated specifically to support reasoning abilities and from human feedback). Empirically, GPT-4 performs well on many tasks where other LLMs flounder. As such, these examples should be taken as existence proofs about what *some* LLMs can do, rather than as evidence for claims about what *typical* LLMs can do.

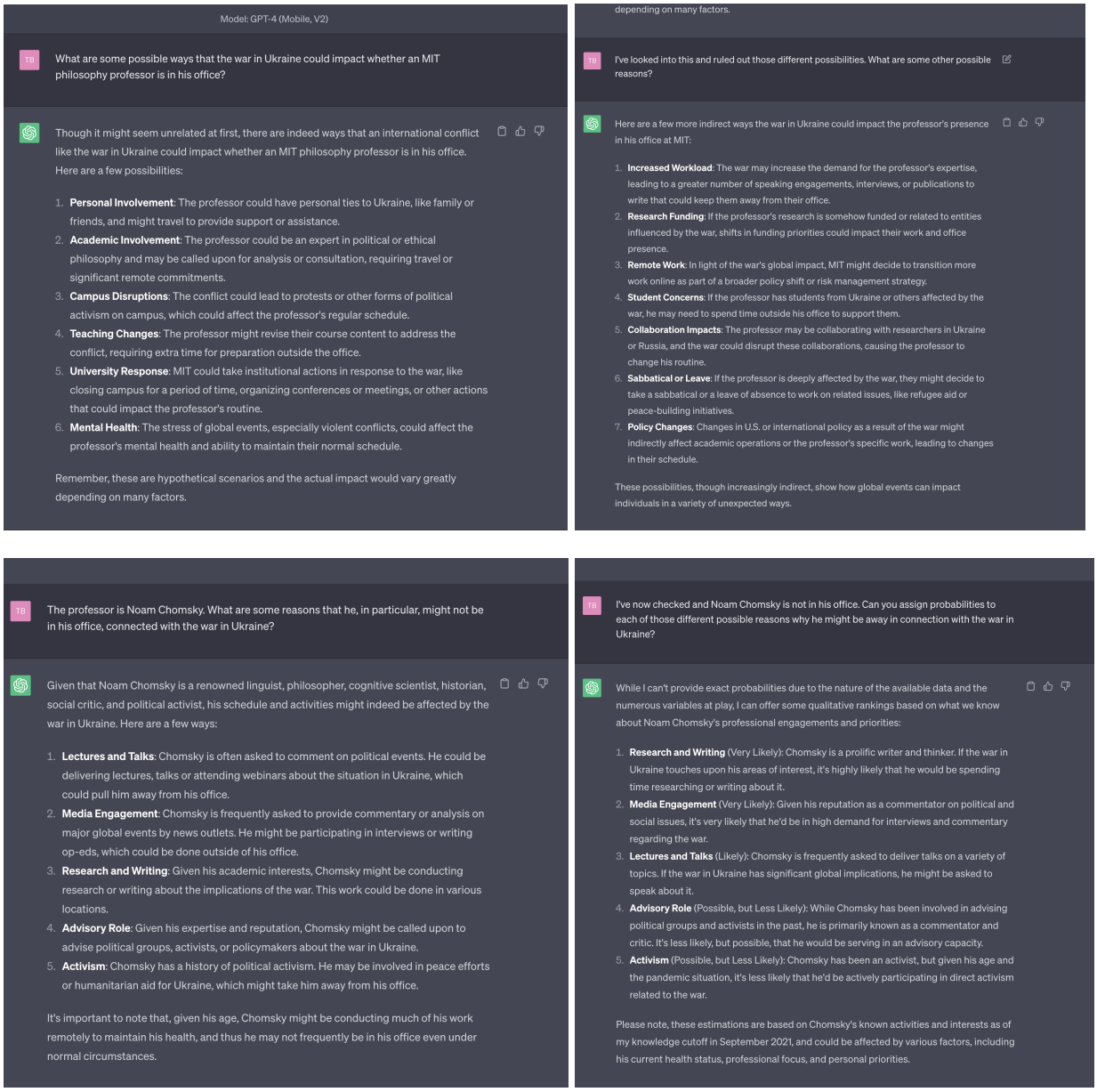


Figure 1: Several examples of a language model asked to generate considerations relevant to our target problem, delivering plausible connections between arbitrary concepts like whether an MIT philosophy professor is in their office and the events unfolding in Ukraine.[99]

Studies and examples like these suggest that, whatever their other weaknesses may be, LLMs are capable of tractably computing relevance in many cases. Whether this is the right kind of relevance is a subtle question,

[99] Generations from GPT-4 in June, 2023. (GPT-4 is a proprietary system behind an API and subject to periodic updates. As such, the system's behavior may change over time.)

but it's enough to cast significant doubt on the critical premises of the skeptical arguments we've been examining. In particular, it allows us to reject *premise 2* – LLMs represent a method for determining relevance without regress and without passing the buck on to inference at a higher level.

Where did the reasoning in defense of Premise 2 go wrong then? The answer is instructive. LLMs represent a variant of the 'Sleeping Dogs' heuristic. Recall that sleeping dogs heuristic says, 'consider as relevant anything that was relevant last time you were in a similar situation'. This is hard to apply in practice because there are many true descriptions of the things you've done previously, and determining which things under which description should be treated as guides to relevance seems to require nothing short of inference about which things are relevant and why. Skeptics like Fodor and Dreyfus were right to have these concerns. What they overlooked is that the hard work of determining which *descriptions* are guides to relevance can also be offloaded to learning, in addition to the actual considerations associated with events that fall under those descriptions. Provided there is enough data and time to learn from it, a system can learn these descriptions or similarities by starting with a very large set of possible descriptions and then making small, soft updates that lead over time to strong connections between certain classes of events and concepts.[100] Cast in the language of the discussion above, the basic strategy here is to start off relatively indifferent between a large set of possible descriptions which might be proxies for relevance profiles – relatively indifferent, that is, between whether what is relevant is to let sleeping dogs lie, let farmer Jones's sleeping dogs lie, etc. – and then make small updates in response to experience. If the fact that it was farmer Jones's dogs in particular turns out to be important, that is likely to come out in many experiences with sleeping dogs.[101] A large body of data and significant amount of training are enough to offload relevance from reasoning to learning in this way. And once learned, a relevance function of this kind can be run cheaply. Provided various conditions are met enough offline learning can compensate for a lot of online reasoning (see Chapter 2 of this dissertation for more discussion).

By learning a tractable, general purpose relevance function, LLMs represent an existence proof that undermines intractability arguments. There is, of course, room for new versions of relevance skepticism to emerge. We do not yet have an existence proof that a system relying on LLM-like relevance function could do as

[100] Provided there are some means for preferring some predicates over others, least New Riddle type concerns prevent any learning from happening at all. Clearly neural networks, people, and any other system that does in fact learn has some means of preferring some predicates consistent with the data over others.

[101] Note that since experiences with sleeping dogs can be both rare and costly if you make the wrong move, the hope is that this general strategy applies to more abstract concepts and scenarios, allowing for a degree of 'generalization' (extending what a system has learned to cases it hasn't encountered before). Here again LLMs offer reason for optimism – while their training data is vast, the vastness of possible questions is greater. Their ability to answer so many questions strongly suggests they learned *some* generalizable relevance knowledge.

well as people do, only a counterexample to an argument that such a thing could not exist. The current discussion could even point the way towards more subtle versions of skepticism. Our route to denying premise 2 required offloading the work of online inference about relevance to prior learning over large amounts of data. LLMs famously require a lot more linguistic data than people get – by one estimate about *10,000 times* more words than a typical 13 year old is exposed to (Wardstadt & Bowman 2022). What's more, the particular kind of data that LLMs consume – an internet's worth of stories and discussion of various topics – could be uniquely useful for learning relevance relations between concepts. A new form of skepticism could argue that if a system can rely on exposure to this much data, then relevance can be computed tractably, but for more humanly realistic amounts of data this is not possible.

Such a line of attack touches on important issues. Figuring out how (or whether) relevance could be tractably computed subject to more of the constraints people face represents a fruitful line of inquiry. Relevant questions here are things like – What kinds of data are available to the human learner to learn relevance? And what is the minimum that's needed for a machine to do so? Building a new skeptical case against unified cognition, strong AI, or the computational theory of mind would require answers to these questions. The case would not be easy to make. One relevant consideration is that people have many sources of data that LLMs lack, including large amounts of information from perception and from the ability to intervene on the world. Such sources of data may more than compensate for the paltry linguistic data that people have in comparison to LLMs. Another relevant consideration is that people may start out part of the way there when it comes to learning relevance. Many developmental psychologists believe that infants start life with a range of concepts, including concepts for agents, objects, cause and effect. When it comes to learning what is relevant to what, starting out with an ontology like this is windfall. LLMs in contrast must learn everything *de novo*. Considerations like these suggest that, while there is certainly room for a new version of the old skeptical challenge, the case would really have to be made. As things stand, we have an existence proof that relevance can be tractably computed, and no positive case that people's minds couldn't exploit a similar strategy to do so.

The upshot of this is that, for arguably the first time, we have a serious candidate method for the tractable computation of relevance. This opens the door to serious investigation into the structure of cognition. With a relevance function in hand, we can begin to imagine what an architecture for cognition might look like. LLMs themselves represent a natural starting point here. If LLMs are capable of tractably computing relevance, could they offer a full account of human domain general cognition? In the next section, I argue that they could not.

IV. Weaknesses of LLMs

We've seen that LLMs offer an existence proof that relevance can be tractably computed. A natural question then is whether LLMs, and nearby deep learning systems trained on different kinds of data, offer a full account of domain-general cognition. I'll call the hypothesis that they do the 'Pure Deep Learning Hypothesis'[102]. In this section, I'll say why I think we should be dissatisfied with this hypothesis. In doing this, I'll focus on LLMs as the most successful instances of deep learning systems (considered as candidate models of human cognition), but I'll focus on properties that LLMs share with other deep learning systems. The key phenomena I'll highlight are the patterns of co-occurrence of reasoning abilities in people vs. LLMs. While LLMs are quite capable of reproducing human-like behavior on many tasks, the fine-grained way in which their abilities pattern is strikingly different from what we see in people. This difference suggests a deeper difference in how people and LLMs think, with LLMs relying frequently on heuristic solutions where people exhibit deeper reasoning abilities. This motivates the search for an architecture of human cognition that can model these deeper reasoning abilities.

A range of case studies show that human and LLM reasoning abilities pattern differently. I'll look at just a few here – starting with simple cases which make diagnosing errors easy and progressing to instances of everyday reasoning, where we see similar errors. Start with multiplication. Most schoolchildren learn a simple algorithm, longform multiplication, for solving multiplication problems. This algorithm has children break down a multiplication problem potentially involving two large numbers into a series of smaller multiplicaiton problems and then a sum of the results (see figure N for details). After learning the algorithm, children (and adults) may make mistakes computing the numbers – especially mistakes like forgetting to carry or simple arithmetical errors – but the algorithm can in principle be applied to numbers of arbitrary size. Multiplying two five digit numbers, for example, is not a difficult task.

LLMs like those in the GPT series can also solve some multiplication problems, but their abilities show a strikingly different pattern. GPT-4, for example, shows near ceiling performance (close to 100%) for multiplying 2 by 2 digit numbers (e.g. 72 x 89), but performance drops off *precipitously* as the numbers get bigger. The model 92% accuracy on 3-digit-by-3 digit problems, to 4% accuracy on 4x4 digit problems. When asked to multiply 5 digit by 5 digit numbers, GPT-4 is near floor performance (near 0% accuracy, see Figure N). This held regardless of whether the model was asked to write out its reasoning (akin to using pen and paper) or

[102] Sometimes called the 'Scaling Hypothesis' (Gwern 2020) – the idea being that a connectionist system that matches the human brain for scale (100 trillion parameters, rather than the hundreds of billions or trillion parameters of today's models) would be a good model of human cognition.

to generate answers outright. It held when the model was given further examples of multiplication problems, as question-answer pairs or with full trains of reasoning demonstrating the longform multiplication algorithm. It held even after the models had been extensively 'fine-tuned', i.e. subject to further training, on tens of thousands of multiplication problems (Dziri et al. 2023; Choi 2023).

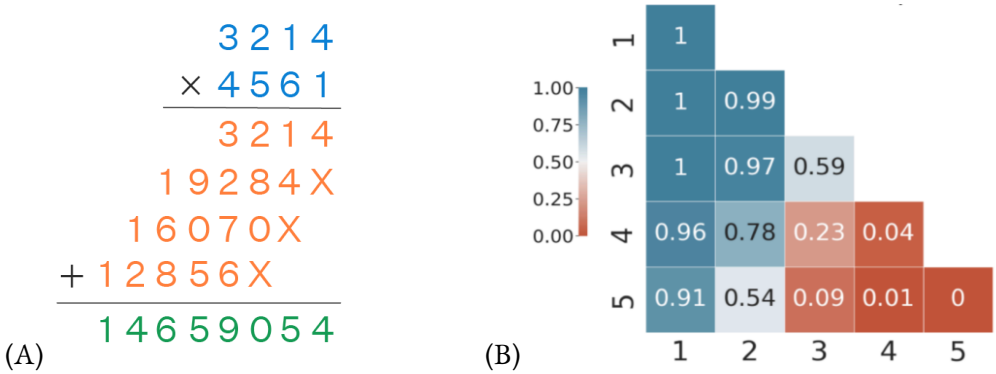


Figure 2: (A) An instance of longform multiplication of the kind schoolchildren learn. GPT-class struggle to learn this algorithm or any other solution to multiplication. (B) Pattern of errors seen in GPT-4 by number of digits in the input – performance on 1-digit-by-1-digit problems (top left) is at ceiling, while performance at 5-digit-by-5-digit problems is at floor. (Graph reprinted from Dziri et al. 2023).

Why does GPT's performance on multiplication drop off so precipitously as problems get larger? And why does extensive further training not change this pattern? One possible explanation for this behavior is if the model were solving these problems by making extensive use of memory, without learning a corresponding general solution to fall back on when memory turns up blank. A memory-based solution would exhibit just this striking pattern of failures. If we consider just the 2-digit-by-2-digit multiplication problems, there are about 6,000 of them. This is small enough that a model like GPT-4 which is trained on trillions of words has likely seen most of them. By the time we get to 5-digit-by-5-digit problems, there are over 8 billion such problems, and it becomes unlikely that a model will have seen more than a small fraction of them. As the size of the problems increases exponentially, LLM performance drops precipitously.

There are, of course, many different ways that memory might be used to solve these problems. One possibility is memorizing question-answer pairs, the way a student might cheat on a test, but this is not the only possibility. Other memory-based strategies might involve memorizing intermediate results of the longform multiplication algorithm, or memorizing statistical dependencies between individual digits in the input numbers

and individual digits in the output numbers. There's some evidence that this last is what models are actually doing. For example, they tend to get some digits right and some wrong for large problems (Figure N). Those they get right tend to be digits like the first and last digit which are statistically easiest to predict from the input digits, e.g. because the last digit depends only on two digits of the input values, rather than many (Dziri et al. 2023).

This pattern of errors, suggestive of LLMs struggling to learn principled solutions to problems, can be seen in other domains as well. Take planning. Ida Momennejad and colleagues at Microsoft Research looked at a large set of state of the art LLMs and assessed them for their ability to extract a structured map of an environment from a passage of text and then to plan a route over that map. For example, they'd feed the model a vignette describing how a series of rooms are connected by hallways 'you enter a room, room 1, and walk through an open door to room 2,' and then ask the model to either reproduce the underlying graphical structure or to plan over that graph, for example, 'find a route to room 7.' Models were generally able to reproduce the graphical structure, outputting triples of the form {room, opendoor, room}, but struggled to plan effectively over that structure (see Figure N). The best performing model, GPT-4, succeeded about 30-40% of the time on the hardest graphs, with other models performing considerably worse. The hardest graphs had 15 and 21 nodes. When it came time to plan, LLMs would frequently hallucinate edges, proposing moves between unconnected rooms, they would get caught in loops, revisiting the same rooms repeatedly while trying to navigate elsewhere, and they'd miss obvious paths, especially when those paths were not explicitly stated in the original vignette, but instead implied by the network structure (See Figure N). For comparison, people who can navigate Cambridge and Somerville are able to plan on a graph with several hundred nodes,[103] while formal studies of human planning suggest that people are capable planners (for example, they are better modeled by optimal models than by heuristic models, Callaway et al. 2022). This pattern of errors suggests that the capacity to recover the structure of an environment and then plan over it is comparatively lacking in LLMs.[104]

[103] The number of intersections in these towns: https://dataverse.harvard.edu/dataverse/osmnx-street-networks

[104] Similar patterns of errors are seen when models are asked to plan in other domains – rearranging small numbers of blocks (e.g. Valmeekam et al. 2023).

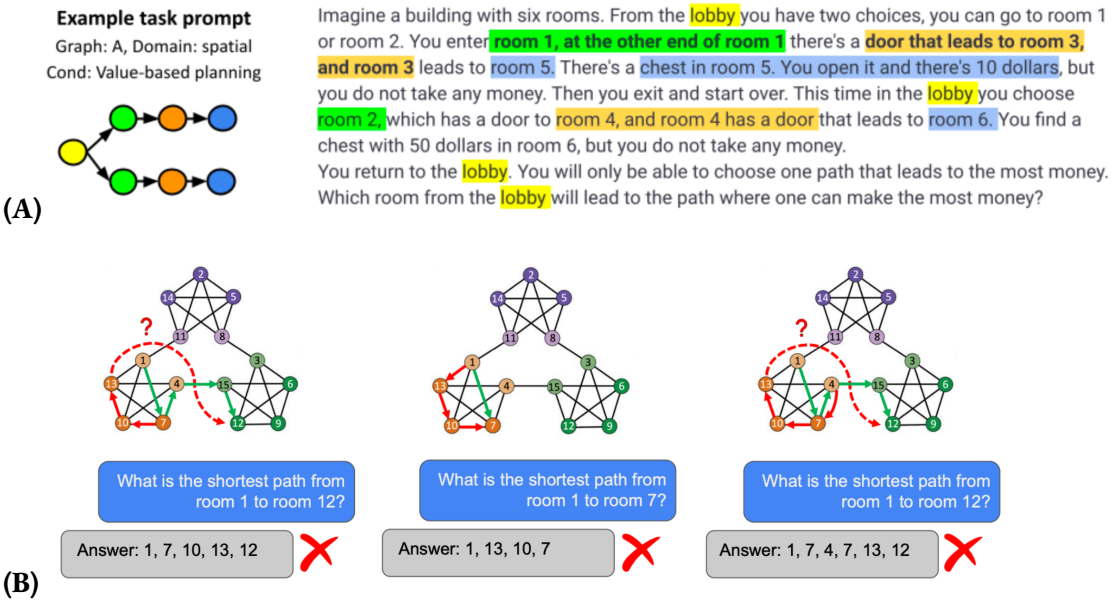


Figure 3: (A) Example prompt given to LLMs and the underlying graph. Models were tasked with various planning problems over the resulting graph, including planning a traversal and planning a detour when one path was broken. (B) Models struggled with these planning tasks, hallucinating edges, missing obvious paths, and getting caught in loops. (Both reprinted from Momennejad et al. 2023).

A final example of such failures comes from Theory of Mind (or ToM). ToM is the ability to infer other people's mental states from their actions. People have a profound capacity for ToM. They infer rich information about agents' desires from very simple actions, such as how an agent explores a space (Baker et al. 2017). They infer fine-grained, quantitative relations between the agent's costs and benefits based on the paths agents take (Jara-Ettinger et al. 2016). And they accurately infer agent's goals even when the agent's actions are relatively inefficient ways to achieve those goals, ostensibly by reasoning about agent's plans, which are often imperfect (Zhi-Xuan et al. 2020). Signs of some of these capacities, e.g. assessments of preferences, costs, and benefits, are seen even in very young infants looking time behavior (Liu et al. 2019).

LLMs appear to struggle to reason about other minds. False belief tasks offer one case study here. One part of ToM reasoning is the ability to keep track of other agents' beliefs, including when the content of those beliefs differs from the way we know the world to be – i.e. when agents have false beliefs. Classic psychological tests probe this ability by presenting people (generally young kids) with vignettes in which a character acquires a false belief about the world and then asking people questions about what that agent will do, where the answer depends on the content of their beliefs. If kids are able to keep track of the agent's beliefs as distinct from their

own representation of the environment, then they'll accurately predict the agent's ineffective actions. So, for example, kids might be given a story in which two characters are in a room, the first places an object in a box before leaving, and the second character moves the object to a second box before the first returns. Kids are then asked where the first character will look for the object upon returning. Being able to solve the task requires children keep track of the mental state of the first character as distinct from the state of the environment.

Recently, researchers have been exploring whether LLMs can keep track of false beliefs by giving them the same tests. Intriguingly, GPT-4 appears to be able to solve these tests, showing near perfect performance on vignettes with this classic structure (substituting the names of characters and objects so as to avoid a direct match to previously published studies, which may have been in the training data of large models) (Kosinski 2023, Bubeck et al. 2023). Similar to the multiplication example above, however, stepping outside of familiar problem instances reveals shortcomings. While GPT-4 is at near ceiling accuracy for vignettes with the same underlying fact pattern as those used in classic psychology experiments, it stumbles badly when tested on new vignettes, which differ in critical respects that change the final answer. So, for example, if the vignettes are changed so that the second character moves the object 'onto' the second box, rather than 'into' it, or if both boxes are transparent, the model still suggests that the second character will look in the original location (i.e accuracy drops from ceiling to floor) (Ullman 2023, Shapira et al. 2023). On the face of it, this suggests that what the model had actually learned was a pattern of facts and answers wedded to the original vignette type, rather than a generalized capacity to track other minds.

This is not the only explanation available for these failures. Another possibility is that GPT-4 doesn't understand the significance of transparency for the agent's epistemic state, perhaps because the model is trained only on text. This explanation, while tempting, sits poorly with the observation that GPT-4 seems to have an intimate understanding of transparency in other cases. Asked whether a person who is hungry and walks into a room with a transparent (or opaque) box containing their favorite food, GPT-4 gives answers that sensibly turn on the person's visual awareness of the food (see Figure N). Since these vignettes are almost certainly in the model's training data (as a staple of developmental psychology papers) a natural explanation for these failures is that the model has learned to give the answers suggested by the original vignettes, rather than the adjusted versions. That is, the model exploits a memory-based strategy for tracking mental states in these vignettes, rather than reasoning about how minds work.

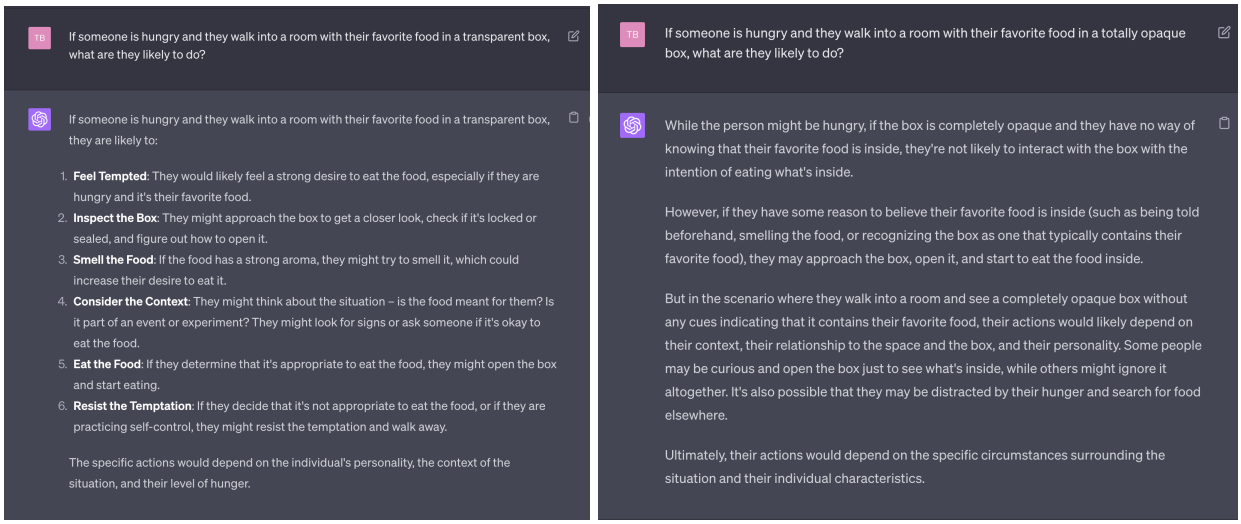


Figure 4: Other questions posed to GPT-4 suggest a nuanced understanding of how transparency affects the accessibility of information, suggesting that a lack of this kind information is not responsible for its failures on the altered ToM vignettes from (Ullman 2023).[105]

These are three cases where models fail to find general solutions to their problems where people succeed. These failures with big models are reminiscent of failures well-documented in smaller models (see McCoy et al. 2019) and consonant with the kinds of memory-based solutions that have been revealed by careful mechanistic investigations into these models (for example, documenting that models learn skip-trigram statistics in order to deliver on in-context next word prediction, see Olsson et al. 2022). Results like these and others suggest that models of this class are finding superficial, albeit sometimes extremely *subtle*, strategies for solving the problems posed to them, in lieu of the more general strategies that underwrite the human ability to, say, reason about other minds or perform multiplication.

This discussion is, of course, far from definitive. Future evidence could turn this conclusion on its head. But the current evidence suggests that, despite their successes at tasks like relevance, LLMs fall short of offering us an architecture capable of the kind of cognition seen in people. This motivates the search for a new architecture, one that can combine the ability to tractably compute relevance with more principled ways to reason about the resulting contents.

[105] Generations from GPT-4 in July, 2023

V. Bespoke Model Construction

We've seen that LLMs are capable relevance functions, but also some evidence that they struggle to learn to reason as consistently as people. Is there a way to have both relevance and reliable reasoning? This section lays out a high-level proposal. The basic idea is to use a dedicated relevance function to deliver a set of considerations relevant to some task. Once specified, these concepts can be used to build a model. Approximate normative operations of the kind that have long been used to model thought – bayesian updating, consistency checking, algorithmic planning – can then be computed in this model. In principle, the operations within the model can be kept tractable by keeping the model small (cf. Brooke-Wilson 2023). Such a system could reason over anything, although only a small number of things at once. The hope of such a system would be that reasoning over the most relevant things could approximate reasoning over everything.

5.1 Model Construction Overview

We can spell this out in stages to make the idea more concrete. We start out with a problem representation, whether a description of the problem in natural language or some language of thought. The problem can be anything that people solve in cognition (as opposed to perception or motor control). Reasoning and planning problems will be canonical cases here.[106] The problem representation is first (1) fed into a relevance function – perhaps implemented by an LLM or something else. The relevance function outputs a set of representations. These are couched in a language of thought with the structure to support familiar approximations to normative operations, for example, using probabilistic operators to capture degrees of belief, or graphical primitives to express graphs. As its name suggests, a good relevance function should output those representations which are *relevant*, in a sense to be made precise, to the task at hand. In the second stage (2), these representations serve as the raw materials for a process of *model construction*, during which a formal model of the problem domain is synthesized. Because the model is written in a language of thought, it can support familiar operations – computations for Bayesian updating, consistency checking, or classical planning algorithms. In the third stage (3), these operations are computed over the constructed model, resulting in

[106] That is, high-level planning problems – problems that abstract away from motor control. Planning your errands for the day counts as cognition, the precise sequence of movements involved in grasping is part of motor control (Mylopoulos 2021).

candidate answers to the task – e.g. probabilistic estimates over variables (i.e. credences) and executable plans. A fourth and final stage (4) involves checking these outputs against the task demands that started the process, assessing whether the end result of computation in the model answered the question at hand (if the original problem was a question) or whether the output plans deliver the desired goal (in the case of a planning problem). I'll say more about each of these stages in turn. Note however, that they are only meant to offer a proof of concept. Toward the end of the section, I'll flag some of the ways in which the reality might be considerably more complex.

To present this idea in more detail, consider the example that started this paper. We are trying to determine whether Bob is likely in his office. On the current view, the goal of our cognitive operations when confronted with this task is to come up with a model meeting the following specifications – it should support a normative operation of belief fixation, and it should include the body of considerations (i.e. the subset of my evidence and linking hypotheses) that bear most directly on the question at hand. The first step here is to query a relevance function. If this is an LLM or similar model, the query could be the natural language question 'Is Bob in his office?'. On other occasions, the input to the relevance function could be a sentence in the language of thought, or a vector from another neural model. At this stage, all that's needed is a representation of the question that is sufficient for determining relevance when fed into a relevance function.

The next step cares more about the details of the representation. Here the relevance function outputs a set of representations which can be composed into a structured model, of the kind that will support familiar operations like those discussed above. The relevance function should also output a particular operation that will be computed in the resulting model – for example, bayesian updating or prediction if the problem is a reasoning problem, or planning operations if the problem is a planning problem. This is like choosing which tools are appropriate to solving the problem (see Figure 6). There are many candidates for such structured representations. Different formats will be applicable for different kinds of tasks. A natural class of representations to reach for is code. Code, like thought, is highly flexible. Much like thought, code can represent everything from a list of beliefs, to a map, a 3D scene, or a physical model capable of simulation. Model construction might involve any of these. Sticking with our motivating example, when wondering whether Bob is in his office the relevant considerations might include: the fact that Bob is often in his office on days when he teaches; that if Bob is out of town then he's not in his office; that Agustin's presence makes Bob's more likely

because they're friends; and that Bob is often at the APA when it is in session. These will serve as the basis for model construction in the next stage.[107]

A natural question is what counts as a 'relevant' consideration. The goal of the entire model construction process is to build a model such that approximate normative operations computed over that model will best approximate the intractable application of the corresponding normatively ideal operations over the totality of the agent's beliefs. Representations are *relevant* insofar as they improve this approximation and *irrelevant* insofar as they make little difference. This is a graded concept of relevance and one that is highly relative (a particular consideration can be relevant in certain conditions, say, if considered on its own, but not in others, for example, if other considerations screen off its impact). [This idea can be expressed precisely. For example, if the goal of the process is to build a model such that approximate bayesian inference in the model best approximates the posterior you would get if you could compute exact bayesian inference over the totality of your evidence,[108] then the *decrement* to the divergence between small model and ideal model posteriors delivered by an edit to the small model gives an exact quantitative measure of relevance. For a planning problem, where the objective is to maximize the expected utility of planning in the model on the expectation defined by your total evidence, the *difference* in the expected utility of the resulting plan that results from adding a consideration to the small model gives a precise measure of relevance to planning. It is a challenging, and deep, question how the learnable components of such a system can be trained to approximate this intractable objective. (I return to this question in the next section.)]

Returning to our process of model construction, the second stage involves the actual synthesis of a model using the outputs of the relevance function as raw materials. The relevance function may be involved here as well. On different versions of the architecture the relevance function may output a simple list of building blocks, a full model, or be called recursively to build a model piecemeal. If we imagine the simplest case, where the relevance function outputs a model, then this process is much like asking an LLM to output functioning code. If instead the model is built piecemeal, this is like asking an LLM to navigate a tree of edits to a program resulting in functioning code. Where the relevance function can't be relied on, say, because the code to be generated is too novel or sophisticated, many other search techniques are available. These include stochastic search techniques, which make random edits biased by various heuristics, in effect taking a random walk in the

[107] Note that which relevant considerations are output by the relevance function (and so come to mind) on any particular occasion will be a function of what description of the problem is input to the relevance function. This means that slightly different 'framings' of the problem might result in different outputs (i.e. different things 'coming to mind').
[108] Or some suitably idealized abstraction over your evidence, if your evidence does not deliver a coherent model.

direction of a successful program, and enumerative techniques, which explore a portion of the space of possible programs exhaustively, exploiting clever ways to compress the space to make the search tractable. Combinations of these methods (neural, stochastic, and enumerative) have proven effective at managing large search spaces and continue to be explored for the particular case of code and model synthesis (see Ellis et al. 2020, Wong et al. 2023). The result of this stage is, in the case of belief fixation, a model in which connections between what is known (our evidence, observed or recalled from memory) and what we would like to know (the question at issue) are explicitly represented. The resulting model might include, for example, a variable of interest, *Bob_in_his_office,* evidence variables such as *Day_of_the_week* and *Term_in_session*, and the probabilistic or causal connections between them. A model of this kind supports Bayesian updating, delivering an estimate of whether, based on the evidence deemed relevant, Bob is likely to be in his office.

Finally, there is a stage on which the results of such reasoning and planning are checked against the world, ensuring that our models remain moored in reality and offering a learning signal for the parts of this process that are learned, such as the relevance function. What 'checking' looks like will depend on the question that started this process. In the simple cases of belief fixation, we will at times get feedback from the world – Go check: Is Bob in his office? If your mental model assigned high probability to 'yes' and he's not there, you might have overlooked something. That disconnect provides a signal that can be used to improve your relevance function. The same goes for simple cases of planning – are you able to find an effective plan for the goal in the model you've constructed? Does acting on that plan in the world produce the desired effects? In many other cases, however, checking directly against the world may not be possible. In these cases, we may check through further reflection – drawing on memory or further reasoning. How does this verdict about Bob square with what I remember about MIT's class schedule, or what I can deduce about Agustin's whereabouts? When the stakes are low or checking is hard, checking may be skipped altogether. When checks are failed – e.g. we discover a conflict between a current inference and a remembered fact – that conflict may be fed back into the relevance function to begin a new episode of reasoning aimed at reconciling the conflicting data. Thus conflicts in reasoning can be resolved with more reasoning.

Here I've presented this process as a series of four discrete stages, but this is merely meant as a proof of concept. More plausible versions of the view are likely to differ in how the different stages trade-off between one another. For example, while the above explication suggests a single call to the relevance function to define a set of concepts and beliefs relevant to the task, it seems more likely there would be repeated calls during model construction, interleaved with various kinds of checking. This is because partially constructed models and the

checks they pass or fail can supply important context for determining relevance. If a model in a partial state of construction is missing a concept that can causally tie two variables together, for example, information like this can be an important input into a relevance function (as in the abduction task discussed in Section III). Similarly, there may be good reasons to interleave the stages of model construction and model checking. Some checks will be applicable to partially constructed models, and whether partial models pass or fail such checks can provide critical guidance to the process of construction.[109] The space of ways to flesh out an architecture of this kind is vast. And a plausible account of human cognition would doubtless involve considerable complexity. What each of these potential realizations have in common is that they approach the problem of general purpose cognition by focusing computational power on computing over small bespoke models built using a large body of domain general information.

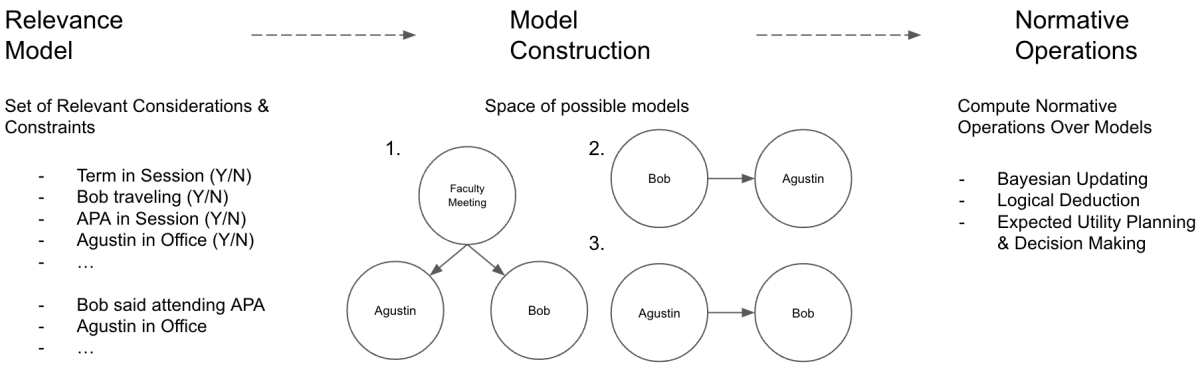


Figure 5: First three steps of the architecture. In the first, a relevance function takes in a task specification and outputs a set of relevance considerations. Next, model construction takes in these considerations and outputs a model or models. Finally, familiar computational operations for reasoning and planning are computed in the resulting model(s). The second and third steps involve operations that are intractable when computed over a large domain, but here are computed only over a domain circumscribed by the relevance function. (Note that plausible versions of this view will be less serial in their operation.)

[109] Checks that could apply to a partial model include type checks or the application of a value function (see Silver et al. 2016). Knowing how a model fails to pass a check can provide information to a relevance function – today's LLMs can make simple edits to some code in light of the bugs it produces when run, albeit with weak reliability.

How does an architecture of this type improve reasoning? Start with the multiplication discussed above. Here the basic idea is simple: Taking a pure learning approach that attempts to learn multiplication from nothing more than pairs of numbers (or series of numbers in the case of training with intermediate values) is a hard learning problem. There are many, many ways that numbers can be associated that partially overlap with multiplication. Picking multiplication from among the bunch is challenging without some bias. In contrast, even a very flexible model can learn to associate word problems with the calls to a calculator, if a calculator is among a reasonably small number of tools it has access to. Assuming a model has access to the right primitives, learning to *build* a calculator may also be an easier learning problem than learning to multiply from scratch. This seems to be the case with today's LLMs. While GPT-4 struggles profoundly with multiplication when asked to do it itself, it has no problem making the syntactically appropriate calls to a calculator. It's even able to *write the code* to build a calculator, using python's high-level arithmetical primitives, when prompted (see Figure 6). This is a very simple version of the idea above. Instead of learning to reason, which represents a very hard problem, a Model Construction architecture of the kind on offer here, learns to build and use the kinds of models that natively support reasoning in principled ways.[110]

[110] Note that GPT has probably presumably seen the code for a calculator many times in its training data, so this is not a statement about GPT's generalization abilities, just about the relative ease of learning to piece together the functions for a calculator vs. learning to calculate. Even if GPT has memorized the code for a calculator, it shows that this memorization task is easier than memorizing a general solution to multiplication in the weights of the network.

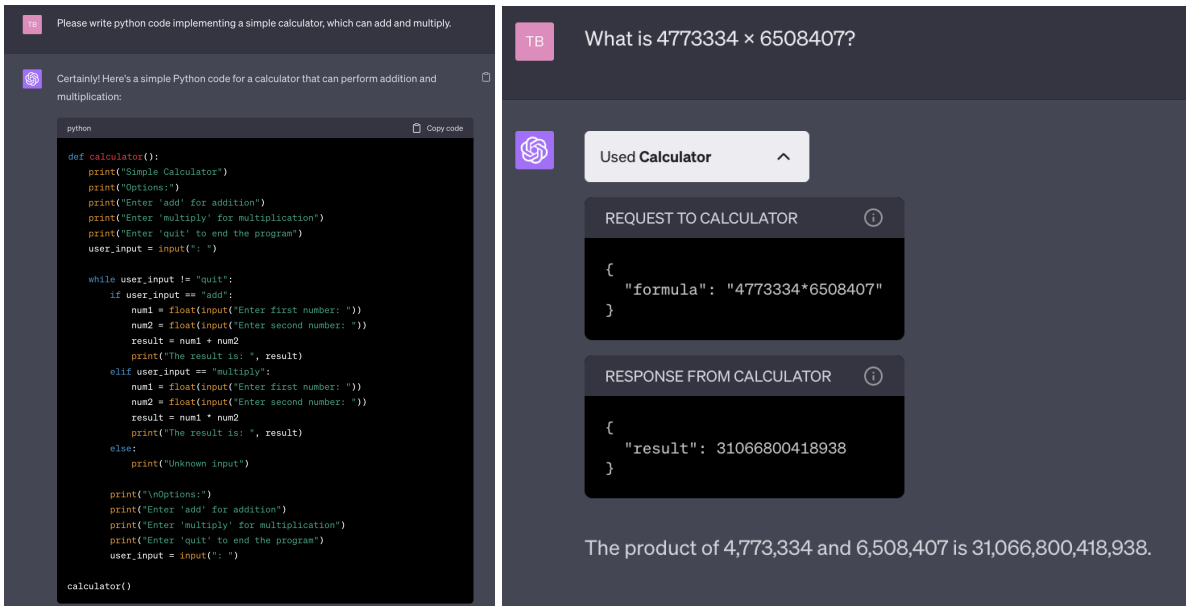


Figure 6 While GPT struggles to learn a general algorithm to multiply two numbers, with near zero accuracy for 5x5 digit multiplication, it can learn to generate the code for a functioning calculator using arithmetical primitives and to make appropriate calls to a calculator for much larger multiplication problems.[111]

We can apply a similar approach to the ToM case. Computational models representing agents, their beliefs, desires, and actions, can reproduce feats of reasoning, as evinced by much modeling work in cognitive science (discussed in Section 3). On their own, these models are limited in their explanatory power. We can't assume that people have just the models required to reason about a given experimental setup in their minds when they walk into the laboratory and that makes it difficult to know what to make of the fit between people and a given cognitive model. But we *can* explain people's systematic behavior on ToM tasks with an architecture that can synthesize small models when faced with a novel task. If cognition works this way, then the synthesized models in the mind could deliver the behavior captured by the scientific models by supporting the very same kind of reasoning. We can see one way this idea could go in current work exploring the possibility of using LLMs to translate between natural language and a probabilistic language of thought. Wong, Grand, and colleagues used an LLM to translate between naturalistic dialogue and code for several tasks, including a simple ToM task. By translating successive sentences like 'people can either bike or walk' and 'Alex loves sushi, but hates pizza' into lines of code in a high-level programming language, their system slowly built up a model that could support ToM style reasoning (see Figure 8). Once constructed, a model of this kind can support systematic performance

[111] Generations from GPT-4 in July, 2023

on a broad set of tasks, including inference to an agents' goals, predictions about their actions, and counterfactuals about what they would do in different environments. Each of these types of 'queries', or questions posed to the model, can be computed in one and the same model using different familiar operations.[112] The ability to edit and generate models on the fly expands this coverage further. Armed with the concept of an agent, for example, such an architecture could reason systematically in situations involving different numbers of agents – always deploying one and the same concept. Of course, such a model will also exhibit some failures of systematicity, when different environments that call for the same representations are nevertheless represented differently, because of a necessarily imperfect relevance function. Far from being a shortcoming, this could give an architecture of this kind a way to explain some of the failures of systematic reasoning seen in people, such as framing effects, where the way a problem is posed has undue influence on the answers people arrive at.[113]

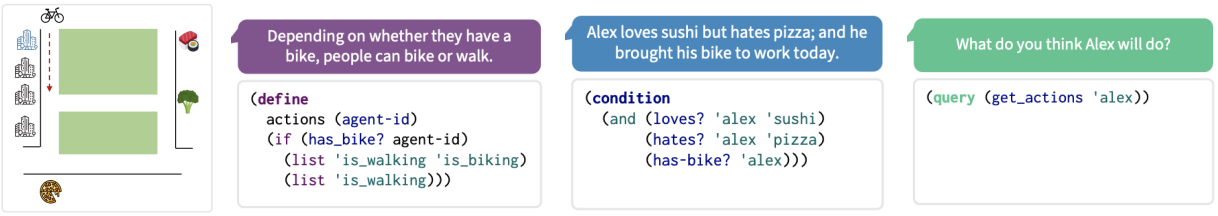


Figure 7: A sketch of how a model might go from a perceptual or natural language representation of a scenario to a structured language of thought representation that can support familiar approximations of normative operations like bayesian inference or classical planning. Fed an image of an environment, calls to an LLM are made to produce code defining a simple set of actions, agent preferences, and a code based translation of questions to ask of the model 'What do you think Alex will do?' (from Wong, Grand, et al. 2023).

There is nothing special about theory of mind here. The same methods can apply to arbitrary domains. This generality comes from two properties of the model. The first is the broad coverage delivered by LLMs and similar models. Serving as relevance functions, these systems are extremely expressive, and can map highly unstructured data of arbitrary format to an equally arbitrary space of possible outputs. The second property is the expressivity of code, which can support arbitrary computations and data types. Jointly, these properties mean that an architecture of this type is unlimited in important respects (I'll discuss some of its important limits

[112] Approximate Bayesian inference, prediction, and counterfactual evaluation, respectively.
[113] See Dasgupta et al. (2017) for some early exploration of a sympathetic account of framing effects.

in the next section). This means that an architecture of this type can help explain the success of models is the many areas of human cognition where such models have been applied, intuitive physics (Battaglia et al. 2013, Allen et al. 2020), concept learning (Goodman et al 2011, Piantadosi et al. 2016), planning (Ho et al. 2022), and beyond.

The availability of an architecture like this reshapes important debates about what computational models teach us in computational cognitive science. On the face of it, computational models using familiar approximations to normative operations have been shown to provide strong qualitative and quantitative fits to human behavior across a range of domains. There has been a lively debate about what lessons we should take from this. A realist perspective holds that the computational processes described by the model are implemented in the course of human cognition, while an antirealist perspective views the processes of human cognition as merely input-output equivalent to those described by the model. A problem for the realist perspective is explaining how the human mind could possess the huge diversity of small models needed to realize the familiar operations described by the models of computational cognitive science. The architecture offered here makes the realist position on this issue tenable. It expands our theory of cognitive operations to include not just the operations reproduced by a cognitive model, but also those performed by the cognitive modeler. A general capacity for model construction explains how, wherever we seem to look, we find modelable operations.

VI. Conclusion

We started this paper wondering how domain general cognition is accomplished despite the massive computational costs suggested by models of thinking. These costs have led many theorists to skepticism about the possibility of domain general cognition consistent with the computational theory of mind. AI has faced similar practical and theoretical challenges, and similar skepticism about the tenability of its ultimate goals. Answering this challenge in both cognitive science and AI requires a way to approximate normative operations over large bodies of belief. In this paper, I've attempted to answer these in-principle challenges and sketch a path forward. I've argued that a class of methods in contemporary AI, large language models, get us part of the way there, by answering skeptical arguments to the effect that tractably computing relevance is impossible. Empirically, such systems tractably compute many aspects of relevance. These systems answer longstanding skeptical arguments by offloading much of the work of reasoning about relevance to prior learning – using large amounts of data and weak assumptions to learn highly context-sensitive relevance relations. The success of such systems opens up, arguably for the first time, the space to propose and evaluate domain general cognitive

architectures. LLMs offer a natural candidate architecture. Despite their initial success with relevance however, LLMs do not seem to capture important aspects of human reasoning. Current evidence suggests that these models generally fail to learn normative operations that people manifestly do. This can be seen in domains as diverse as arithmetic, planning, and theory of mind. Such failures motivate the search for a new architecture for human-like cognition. I've offered one alternative view, designed to exploit the powerful relevance function capabilities of LLM-like systems while making space for approximations to normative operations. On this view, LLM-like systems are used to tractably compute the relevance of various considerations which serve as the basis for the construction of small models in which normative operations can be tractably approximated. Finally, I showed how this view links up with the large body of empirically validated models from computational cognitive science. Positing an architecture which can synthesize these models in response to task demands offers a route to a realist explanation of the empirical successes of Bayesian computational cognitive science.

If correct, this view of how the mind works has consequences for several other areas of philosophy, including epistemology, decision theory, and philosophy of AI. Start with epistemology. Traditional epistemology asks the question, what does my evidence support? In real human reasoning however, there is another epistemological question that is always prior to this one – namely, what is relevant to think about? Only once the mind has determined what is relevant to consider can we get to evaluating the bearing of any evidence on our prior beliefs. How the mind determines relevance is liable to have profound impacts on what beliefs we come to. The consequences of the relevance strategy adopted are liable to ramify throughout one's belief system, as subsequent updates build on one another. The same is true for our values. Practical reasoning depends on determining which of the very many things we value is relevant to consider on a given occasion. What we consider will have significant impacts on what we decide. There is a great deal of work to be done in epistemology and the philosophy of action to tease out the normative principles that govern relevance judgements and the epistemic and practical implications of failures to follow these norms.

Commitments about cognitive architecture also have consequences for decision theory. Traditional decision theory aims to model how agents ought to make decisions. Classically, an agent is modeled with all of her beliefs and their entailments in view when making decisions. Fragmented decision theory has been developed as an alternative. It attempts to build a decision theory with more realistic assumptions about the limits on which of her epistemic commitments an agent can 'see' at any one time. People are generally limited to a small subset of their beliefs and small subset of their entailments. For such fragmented agents, an essential part of decision theory is deciding how to move between fragments – bringing different beliefs and entailments into

view. Fragmented decision theory is clearly tackling an important problem. It is limited however, by a lack of a theory of what 'fragments' actually amount to and what kinds of 'moves' between fragments are realistically available to an agent. The cognitive architecture developed here makes a start on answering these questions. Fragments are mental models, and the 'moves' between them are a diverse set, including computing in a model, building a new model, editing an existing model, and building a model to reason about another, among others Each of which will produce very different kinds of transitions between fragments. A normative theory that strives for psychological realism (motivated by the idea that ought implies can) should think about the norms that govern such transitions. (In doing so, decision theory can lay the normative foundations for the construction of cognitive architectures.)

Finally, there are implications of this view for the philosophy of AI. Right now the hope in much of the field is that scaling the current models to the size of the human brain will deliver human-level intelligence. The current line of thought suggests that this might be wrongheaded. Unaided connectionist models struggle with things like approximating normative operations beyond toy domains (as we move, for example from 2-by-2 digit multiplication to 5-by-5, or from planning on a small graph to planning on a larger graph). Such operations are difficult to acquire in a data-driven way because, as problems get larger, the number of data points necessary to pin down the relevant function increases rapidly. Based on the current evidence, we can venture that genuinely big problems, like reasoning and planning on the scale of many hours and coordinated steps, could prove particularly challenging for pure connectionist models. If these problems turn out to be an important roadblock in the way forward for AI, a turn toward more human-like bespoke model construction may be prudent. Such systems need not have very much built into them – a few familiar operations for reasoning and planning and the basic architecture of model construction – but these basic pieces may provide significant leverage on problems that are difficult to tackle with data and parameters alone.

References

Acuna, Daniel E., Max Berniker, Hugo L. Fernandes, and Konrad P. Kording. 2015. "Using Psychophysics to Ask If the Brain Samples or Maximizes." *Journal of Vision* 15 (3): 1–16. https://doi.org/10.1167/15.3.7.

Adams, Wendy J. 2007. "A Common Light-Prior for Visual Search , Shape , and Re Fl Ectance Judgments." *Journal of Vision*. https://doi.org/10.1167/7.11.11.Introduction.

Agus, Trevor R., Clara Suied, Simon J. Thorpe, and Daniel Pressnitzer. 2012. "Fast Recognition of Musical Sounds Based on Timbre." *The Journal of the Acoustical Society of America* 131 (5): 4124–33. https://doi.org/10.1121/1.3701865.

Alais, David, and David Burr. 2004. "The Ventriloquist Effect Results from Near-Optimal Bimodal Integration." *Current Biology* 14 (3): 257–62. https://doi.org/10.1016/j.cub.2004.01.029.

Allen, Kelsey R, Kevin A Smith, and Joshua B Tenenbaum. 2020. "Rapid Trial-and-Error Learning with Mental Simulation Supports Flexible Tool Use." *The Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.XXXXXXXXXX.

Allen, Kelsey R, Kevin A Smith, and Joshua B Tenenbaum. 2019. "Rapid Trial-and-Error Learning with Mental Simulation Supports Flexible Tool Use." https://doi.org/10.1073/pnas.XXXXXXXXXX.

Amodei, Dario. 2023. "Scaling, Alignment, & AGI in 2 Years."

Apple. 2015. "Supplier List 2015." https://www.aicd.com.au/content/dam/aicd/pdf/news-media/glc/2015/Apple_Supplier_List_2015.pdf.

Arora, Sanjeev, and Boaz Barak. 2007. *Computational Complexity: A Modern Approach*. Cambridge University Press.

Baker, Chris L., Julian Jara-Ettinger, Rebecca Saxe, and Joshua B. Tenenbaum. 2017. "Rational Quantitative Attribution of Beliefs, Desires and Percepts in Human Mentalizing." *Nature Human Behaviour* 1 (March): 0064. https://doi.org/10.1038/s41562-017-0064.

Baker, Nicholas, and Philip J Kellman. 2018. "Abstract Shape Representation in Human Visual Perception." *Journal of Experimental Psychology. General* 147 (9): 1295—1308. https://doi.org/10.1037/xge0000409.

Barragan-Jason, Gladys, Fanny Lachat, and Emmanuel J. Barbeau. 2012. "How Fast Is Famous Face Recognition?" *Frontiers in Psychology* 3 (OCT): 1–11. https://doi.org/10.3389/fpsyg.2012.00454.

Bashivan, Pouya, Kohitij Kar, and James J DiCarlo. 2019. "Neural Population Control via Deep Image Synthesis." *Science* 364 (6439): eaav9436. https://doi.org/10.1126/science.aav9436.

Battaglia, Peter W, Robert A Jacobs, and Richard N Aslin. 2003. "Bayesian Integration of Visual and Auditory Signals for Spatial Localization." *J. Opt. Soc. Am. A* 20 (7): 1391–97. https://doi.org/10.1364/JOSAA.20.001391.

Battaglia, Peter W, Daniel Kersten, and Paul R Schrater. 2011. "How Haptic Size Sensations Improve Distance Perception." *PLOS Computational Biology* 7 (6): 1–13. https://doi.org/10.1371/journal.pcbi.1002080.

Battaglia, Peter W., Jessica B. Hamrick, and Joshua B. Tenenbaum. 2013. "Simulation as an Engine of Physical Scene Understanding." *Proceedings of the National Academy of Sciences of the United States of America* 110 (45): 18327–32. https://doi.org/10.1073/pnas.1306572110.

Bear, Daniel M., Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Sun Fan-Yun, Fei-Fei Li, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, Judith E. Fan.

Bechtel, William, and Adele Abrahamsen. 2002. *Connectionism and the Mind: Parallel Processing, Dynamics, and Evolution in Networks*. Blackwell Publishing.

Beck, Jacob. 2018. "Marking the Perception–Cognition Boundary: The Criterion of Stimulus-Dependence." *Australasian Journal of Philosophy* 96 (2): 319–34.

Beierholm, Ulrik R., Konrad P. Körding, Ladan Shams, and Wei Ji Ma. 2009. "Comparing Bayesian Models for Multisensory Cue Combination without Mandatory Integration." *Advances in Neural Information Processing Systems 20 - Proceedings of the 2007 Conference*, 1–8.

Bendaña, Joseph, and Eric Mandelbaum. 2021. "The Fragmentation of Belief." In *The Fragmented Mind*, edited by D. Kinderman and A. Onofri Borgoni. Oxford University Press.

Bengio, Yoshua. 2023. "Personal and Psychological Dimensions of AI Researchers Confronting AI Catastrophic Risks." 2023.

Bhagavatula, Chandra, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. "Abductive Commonsense Reasoning," 1–18. http://arxiv.org/abs/1908.05739.

Biederman, Irving. 1987. "Recognition-by-Components: A Theory of Human Image Understanding." *Psychological Review* 94 (2): 115–47. https://doi.org/https://doi.org/10.1037/0033-295X.94.2.115.

Block, Ned. 2018. "If Perception Is Probabilistic, Why Does It Not Seem Probabilistic?" *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2017.0341.

Block, Ned. 2022. *The Border between Seeing and Thinking*. Oxford University Press.

Bloj, M. G., D. Kersten, and A. C. Hurlbert. 1999. "Perception of Three-Dimensional Shape Influences Colour Perception through Mutual Illumination." *Nature* 402 (6764): 877–79. https://doi.org/10.1038/47245.

Bogacz, Rafal, Eric Brown, Jeff Moehlis, Philip Holmes, and Jonathan D Cohen. 2006. "The Physics of Optimal Decision Making: A Formal Analysis of Models of Performance in Two-Alternative Forced-Choice Tasks." *Psychological Review*. Bogacz, Rafal: Department of Computer Science, University of Bristolu, Bristol, United Kingdom, BS8 1UB, r.bogacz@bristol.ac.uk: American Psychological Association. https://doi.org/10.1037/0033-295X.113.4.700.

Bonawitz, Elizabeth, Tomer D. Ullman, Sophie Bridgers, Alison Gopnik, and Joshua B. Tenenbaum. 2019. "Sticking to the Evidence? A Behavioral and Computational Case Study of Micro-Theory Change in the Domain of Magnetism." *Cognitive Science* 43 (8). https://doi.org/10.1111/cogs.12765.

Botvinick, Matthew, and Marc Toussaint. 2012. "Planning as Inference." *Trends in Cognitive Sciences* 16 (10): 485–88. https://doi.org/10.1016/j.tics.2012.08.006.

Brainard, David H, Philippe Longère, Peter B Delahunt, William T Freeman, James M Kraft, and Bei Xiao. 2006. "Bayesian Model of Human Color Constancy." *Journal of Vision* 6 (11): 10. https://doi.org/10.1167/6.11.10.

Brascamp, Jan W., Raymond van Ee, André J. Noest, Richard H.A.H. Jacobs, and Albert V. van den Berg. 2006. "The Time Course of Binocular Rivalry Reveals a Fundamental Role of Noise." *Journal of Vision* 6 (11): 1244–56. https://doi.org/10.1167/6.11.8.

Bresciani, Jean Pierre, Franziska Dammeier, and Marc O. Ernst. 2006. "Vision and Touch Are Automatically Integrated for the Perception of Sequences of Events." *Journal of Vision* 6 (5): 554–64. https://doi.org/10.1167/6.5.2.

Brooke-Wilson, Tyler. 2023. "How Is Perception Tractable?" *Philosophical Review*, 1–49.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei, 2020. "Language Models Are Few-Shot Learners." In *Advances in Neural Information Processing Systems*, edited by H Larochelle, M Ranzato, R Hadsell, M F Balcan, and H Lin, 33:1877–1901. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, and Scott Lundberg. 2023. "Sparks of Artificial General Intelligence: Early Experiments with Gpt-4." *ArXiv Preprint ArXiv:2303.12712*.

Carey, Susan. 2009. *The Origin of Concepts*.

Carruthers, Peter. 2007. *The Architecture of the Mind*. *The Architecture of the Mind*. Clarendon Press. https://doi.org/10.1093/acprof:oso/9780199207077.001.0001.

Chatterjee, Sourav, and Persi Diaconis. 2018. "The Sample Size Required in Importance Sampling." *Annals of Applied Probability* 28 (2): 1099–1135. https://doi.org/10.1214/17-AAP1326.

Churchland, Paul M. 1988. "Perceptual Plasticity and Theoretical Neutrality: A Reply to Jerry Fodor." *Philosophy of Science* 55 (2): 167–87.

Churchland, Paul M. 1992. *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. MIT press.

Clark, Andy. 2002. "Global Abductive Inference and Authoritative Sources , or , How Search Engines Can Save Cognitive Science." *Cognitive Science Quarterly* 2 (2): 115–40.

Clark, Andy. 2013. "Expecting the World: Perception, Prediction, and the Origins of Human Knowledge." *Journal of Philosophy* 110 (9): 469–96. https://doi.org/10.5840/jphil2013110913.

Clark, James J, and Alan L Yuille. 1990. *Data Fusion for Sensory Information Processing Systems*. Springer Science & Business Media.

Coenen, Anna, Bob Rehder, and Todd M. Gureckis. 2015. "Strategies to Intervene on Causal Systems Are Adaptively Selected." *Cognitive Psychology* 79: 102–33. https://doi.org/10.1016/j.cogpsych.2015.02.004.

Colombatto, Clara, Stefan Uddenberg, and Brian J Scholl. 2021. "The Efficiency of Demography in Face Perception." *Attention, Perception, & Psychophysics* 83 (8): 3104–17. https://doi.org/10.3758/s13414-021-02351-9.

Colonius, Hans, and Petra Arndt. 2001. "A Two-Stage Model for Visual–Auditory Interaction in Saccadic Latencies." *Perception & Psychophysics* 63 (1): 126–47.

Cosmides, Leda, and John Tooby. 1994. "Origins of Domain Specificity: The Evolution of Functional Organization." In *Mapping the Mind: Domain Specificity in Cognition and Culture*, edited by L. A. Hirschfeld and S. A. Gelman. http://books.google.com/books?hl=en&lr=&ie=UTF-8&id=n1tqi8Tux-kC&oi=fnd&pg=PA85&dq=%2522Leda+Cosmides%2522+%2522John+Tooby%2522+%2522Origins+of+domain+specificity%2522&ots=WggBYXjcKo&sig=uHeb4kmWsJqIGNlci11G-9bPcp4.

Cottier, Ben. 2023. "Trends in the Dollar Training Cost of Machine Learning Systems." *Epoch. January* 31: 2023.

Danks, David. 2014. *Unifying the Mind: Cognitive Representations as Graphical Models*. Mit Press.

Dasgupta, Ishita, and Samuel J. Gershman. 2021. "Memory as a Computational Resource." *Trends in Cognitive Sciences* 25 (3): 240–51. https://doi.org/10.1016/j.tics.2020.12.008.

Dasgupta, Ishita, Andrew K. Lampinen, Stephanie C. Y. Chan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. 2022. "Language Models Show Human-like Content Effects on Reasoning," 1–36. http://arxiv.org/abs/2207.07051.

Dasgupta, Ishita, Eric Schulz, and Samuel J Gershman. 2017. "Where Do Hypotheses Come From?" *Cognitive Psychology* 96: 1–25. https://doi.org/https://doi.org/10.1016/j.cogpsych.2017.05.001.

Dasgupta, Ishita, Eric Schulz, and Samuel J Gershman. 2017. "Where Do Hypotheses Come From?" *Cognitive Psychology* 96: 1–25. https://doi.org/https://doi.org/10.1016/j.cogpsych.2017.05.001.

Dasgupta, Ishita, Eric Schulz, Noah D Goodman, and Samuel J Gershman. 2018. "Remembrance of Inferences Past: Amortization in Human Hypothesis Generation." *Cognition* 178: 67–81. https://doi.org/https://doi.org/10.1016/j.cognition.2018.04.017.

Dasgupta, Ishita, Eric Schulz, Joshua B Tenenbaum, and Samuel J Gershman. 2020. "A Theory of Learning to Infer." *Psychological Review* 127 (3): 412.

David J. Chalmers. 2011. "A Computational Foundation for the Study of Cognition." *Journal of Cognitive Science* 12 (4): 325–59. https://doi.org/10.17791/jcs.2011.12.4.325.

Davis, Martin, George Logemann, and Donald Loveland. 1962. "A Machine Program for Theorem-Proving." *Communications of the ACM* 5 (7): 394–97.

Davis, Martin, and Hilary Putnam. 1960. "A Computing Procedure for Quantification Theory." *Journal of the ACM (JACM)* 7 (3): 201–15. https://doi.org/10.1145/321033.321034.

Denison, Stephanie, Elizabeth Bonawitz, Alison Gopnik, and Thomas L. Griffiths. 2013. "Rational Variability in Children's Causal Inferences: The Sampling Hypothesis." *Cognition* 126 (2): 285–300. https://doi.org/10.1016/j.cognition.2012.10.010.

Descartes, Rene. 2001. *Discourse on Method, Optics, Geometry, and Meteorology*. Hackett Publishing.

Dodge, Samuel, and Lina Karam. 2017. "A Study and Comparison of Human and Deep Learning Recognition Performance under Visual Distortions." *2017 26th International Conference on Computer Communications and Networks, ICCCN 2017*. https://doi.org/10.1109/ICCCN.2017.8038465.

Downey, Rodney G, and Michael R Fellows. 2013. *Fundamentals of Parameterized Complexity*. Vol. 4. Springer. https://doi.org/10.1007/978-1-4471-5559-1.

Dreyfus, Hubert L. 1972. *What Computers Can't Do: The Limits of Artificial Intelligence*.

Dziri, Nouha, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jian, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck,

Xiang Ren, Allyson Ettinger, Zaid Harchaoui, Yejin Choi, 2023. "Faith and Fate: Limits of Transformers on Compositionality," 1–37. http://arxiv.org/abs/2305.18654.

Elga, Adam, and Agustín Rayo. 2021. "Fragmentation and Logical Omniscience." *Nous* 56 (3): 716–41. https://doi.org/10.1111/nous.12381.

Ellis, Kevin, Catherine Wong, Maxwell Nye, Mathias Sable-Meyer, Luc Cary, Lucas Morales, Luke Hewitt, Armando Solar-Lezama, and Joshua B. Tenenbaum. 2020. "DreamCoder: Growing Generalizable, Interpretable Knowledge with Wake-Sleep Bayesian Program Learning," 1–22. http://arxiv.org/abs/2006.08381.

Ernst, Marc O, and Martin S Banks. 2002. "Humans Integrate Visual and Haptic Information in a Statistically Optimal Fashion." *Nature* 415 (6870): 429–33.

Evans, Karla K, and Anne Treisman. 2005. "Perception of Objects in Natural Scenes: Is It Really Attention Free?" *Journal of Experimental Psychology: Human Perception and Performance* 31 (6): 1476.

Feldman, Jacob. 2000. "Minimization of Boolean Complexity in Human Concept Learning." *Nature* 407 (6804): 630–33. https://doi.org/10.1038/35036586.

Feldman, Jacob, and Patrice D. Tremoulet. 2006. "Individuation of Visual Objects over Time." *Cognition* 99 (2): 131–65. https://doi.org/10.1016/j.cognition.2004.12.008.

Firestone, Chaz. 2020. "Performance vs. Competence in Human–Machine Comparisons." *Proceedings of the National Academy of Sciences of the United States of America* 117 (43): 26562–71. https://doi.org/10.1073/pnas.1905334117.

Firestone, Chaz, and Brian Scholl. 2016. "Seeing Stability: Intuitive Physics Automatically Guides Selective Attention." *Journal of Vision* 16 (12): 689. https://doi.org/10.1167/16.12.689.

Firestone, Chaz, and Brian Scholl. 2017. "Seeing Physics in the Blink of an Eye." *Journal of Vision* 17 (10): 203. https://doi.org/10.1167/17.10.203.

Firestone, Chaz, and Brian J Scholl. 2015. "Cognition Does Not Affect Perception: Evaluating the Evidence for Top-down Effects." *Behavioral and Brain Sciences* 39 (May). https://doi.org/10.1017/S0140525X15000965.

Flum, Jörg, and Martin Grohe. 2006. *Parameterized Complexity Theory*. Springer Science & Business Media.

Fodor, Jerry. 1987. "Modules, Frames, Fridgeons, Sleeping Dogs, and the Music of the Spheres." In *Modularity in Knowledge Representation and Natural-Language Understanding*. Cambridge: MIT Press.

Fodor, Jerry A. 1988. "A Reply to Churchland's 'Perceptual Plasticity and Theoretical Neutrality.'" *Philosophy of Science* 55 (2): 188–98.

Fodor, Jerry A. 1983. *The Modularity of Mind*.

Fodor, Jerry A. 2000. *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. *The Mind Doesn't Work That Way*.

Forster, Bettina, Cristiana Cavina-Pratesi, Salvatore M. Aglioti, and Giovanni Berlucchi. 2002. "Redundant Target Effect and Intersensory Facilitation from Visual-Tactile Interactions in Simple Reaction Time." *Experimental Brain Research* 143 (4): 480–87. https://doi.org/10.1007/s00221-002-1017-9.

Friedman, Jane. 2020. "The Epistemic and the Zetetic." *Philosophical Review* 129 (4): 501–36.

Geirhos, Robert, Heiko H. Schütt, Carlos R. Medina Temme, Matthias Bethge, Jonas Rauber, and Felix A. Wichmann. 2018. "Generalisation in Humans and Deep Neural Networks." *Advances in Neural Information Processing Systems* 2018-Decem (NeurIPS 2018): 7538–50.

Geisler, Wilson S. 2011. "Contributions of Ideal Observer Theory to Vision Research." *Vision Research* 51 (7): 771–81. https://doi.org/10.1016/j.visres.2010.09.027.

Geisler, Wilson S., and Jeffrey S. Perry. 2009. "Contour Statistics in Natural Images: Grouping across Occlusions." *Visual Neuroscience* 26 (1): 109–21. https://doi.org/10.1017/S0952523808080875.

Gershman, Samuel J, and Noah D Goodman. 2014. "Amortized Inference in Probabilistic Reasoning." *Proceedings of the 36th Annual Conference of the Cognitive Science Society (CogSci 2014)* 1: 517–22.

Gershman, Samuel J., Joshua B. Tenenbaum, Edward Vul, and Joshua B. Tenenbaum. 2012. "Multistability and Perceptual Inference." *Neural Computation* 24 (1): 1–24. https://doi.org/10.1162/NECO_a_00226.

Gershman, Samuel, Eric Horvitz, and Joshua Tenenbaum. 2015. "Computational Rationality: A Converging Paradigm for Rationality in Minds, Brains, and Machines." *Science*. papers2://publication/uuid/20A0106C-9CBA-472D-AAFB-69231964766F.

Gerstenberg, Tobias, Matthew F Peterson, Noah D Goodman, David A Lagnado, and Joshua B Tenenbaum. 2017. "Eye-Tracking Causality." *Psychological Science*.

Gibson, James J. 1950. *The Perception of the Visual World*. Oxford, England: Houghton Mifflin.

Gigerenzer, Gerd, and Daniel G. Goldstein. 2011. "Reasoning the Fast and Frugal Way: Models of Bounded Rationality." *Heuristics: The Foundations of Adaptive Behavior*. https://doi.org/10.1093/acprof:oso/9780199744282.003.0002.

Gilchrist, Alan L. 1977. "Perceived Lightness Depends on Perceived Spatial Arrangement." *Science* 195 (4274): 185–87. https://doi.org/10.1126/science.831266.

Gobet, Fernand, and Neil Charness. 2006. "Expertise in Chess." In *The Cambridge Handbook of Expertise and Expert Performance*, edited by K Anders Ericsson, Neil Charness, Paul J Feltovich, and Robert R Hoffman, 523–38. Cambridge Handbooks in Psychology. Cambridge: Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511816796.030.

Godfrey-Smith, Peter. 2003. *Theory and Reality*. *Science and Its Conceptual Foundations Series*. https://doi.org/q175 .g596.

Goldreich, Oded. 2008. *Computational Complexity: A Conceptual Perspective*. Cambridge University Press.

Goodman, Noah A., Joshua B. Tenenbaum, Jacob Feldman, and Thomas L. Griffiths. 2008. "A Rational Analysis of Rule-Based Concept Learning." *Cognitive Science* 32 (1): 108–54. https://doi.org/10.1080/03640210701802071.

Goodman, Noah D., Joshua B. Tenenbaum, and Tobias Gerstenberg. 2015. "Concepts in a Probabilistic Language of Thought." *The Conceptual Mind: New Directions in the Study of Concepts*, no. 010: 1–25. https://www.stanford.edu/~ngoodman/papers/ConceptsChapter-final.pdf.

Goodman, Noah D., Tomer D. Ullman, and Joshua B. Tenenbaum. 2011. "Learning a Theory of Causality." *Psychological Review* 118 (1): 110–19. https://doi.org/10.1037/a0021336.

Gopnik, Alison, David M. Sobel, David Danks, Clark Glymour, Laura E. Schulz, and Tamar Kushnir. 2004. "A Theory of Causal Learning in Children: Causal Maps and Bayes Nets." *Psychological Review* 111 (1): 3–32. https://doi.org/10.1037/0033-295X.111.1.3.

Green. 2020. "The Perception-Cognition Border: A Case for Architectural Division." *The Philosophical Review*.

Green, E J. 2021. “The Puzzle of Cross-Modal Shape Experience.” *Nous*.

Green, E. J., and Gabriel Oak Rabin. 2019. “Use Your Illusion: Spatial Functionalism, Vision Science, and the Case against Global Skepticism.” *Analytic Philosophy* 61 (4): 345–78. https://doi.org/10.1111/phib.12163.

Green, E.J. 2017. “On the Perception of Structure.” *Nous*.

Griffiths, Thomas L, Falk Lieder, and Noah D. Goodman. 2015. “Rational Use of Cognitive Resources: Levels of Analysis between the Computational and the Algorithmic.” *Topics in Cognitive Science* 7 (2): 217–29. https://doi.org/10.1111/tops.12142.

Güçlü, Umut, and Marcel A.J. van Gerven. 2015. “Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream.” *Journal of Neuroscience* 35 (27): 10005–14. https://doi.org/10.1523/JNEUROSCI.5023-14.2015.

Gwern. 2020. “The Scaling Hypothesis.” 2020.

Hafri, Alon, and Chaz Firestone. 2021. “The Perception of Relations.” *Trends in Cognitive Sciences* 25 (6): 475–92. https://doi.org/10.1016/j.tics.2021.01.006.

Hafri, Alon, Anna Papafragou, and John C. Trueswell. 2013. “Getting the Gist of Events: Recognition of Two-Participant Actions from Brief Displays.” *Journal of Experimental Psychology: General* 142 (3): 880–905. https://doi.org/10.1037/a0030045.

Hansen, Thorsten, Maria Olkkonen, Sebastian Walter, and Karl R. Gegenfurtner. 2006. “Memory Modulates Color Appearance.” *Nature Neuroscience* 9 (11): 1367–68. https://doi.org/10.1038/nn1794.

Harding, Glen, Julie M. Harris, and Marina Bloj. 2012. “Learning to Use Illumination Gradients as an Unambiguous Cue to Three Dimensional Shape.” *PLoS ONE* 7 (4). https://doi.org/10.1371/journal.pone.0035950.

Harnad, Stevan. 2012. "Connecting Object to Symbol in Modelling Cognition." In *Connectionism in Context*, edited by Andy Clark and Rudi Lutz, 75–90. Springer Science & Business Media.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Infernece, and Prediction*. 2nd ed. Springer.

Heitz, Richard P. 2014. "The Speed-Accuracy Tradeoff: History, Physiology, Methodology, and Behavior." *Frontiers in Neuroscience* 8 (8 JUN): 1–19. https://doi.org/10.3389/fnins.2014.00150.

Henderson, Leah. 2014. "Bayesianism and Inference to the Best Explanation." *British Journal for the Philosophy of Science* 65 (4): 687–715. https://doi.org/10.1093/bjps/axt020.

Hershenson, Maurice. 1962. "Reaction Time as a Measure of Intersensory Facilitation." *Journal of Experimental Psychology* 63 (3): 289–93.

Hinton, Geoffrey. 2023. "Full Interview: 'Godfather of Artificial Intelligence' Talks Impact and Potential of AI."

Hinton, Geoffrey E., and T. J. Sejnowski. 1986. "Learning and Relearning in Boltzmann Machines." In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*, edited by D E Rumelhart and J L McClelland.

Ho, Mark K., David Abel, Carlos G. Correa, Michael L. Littman, Jonathan D. Cohen, and Thomas L. Griffiths. 2022. "People Construct Simplified Mental Representations to Plan." *Nature* 606 (7912): 129–36. https://doi.org/10.1038/s41586-022-04743-9.

Hughes, Howard C., Patricia A. Reuter-Lorenz, George Nozawa, and Robert Fendrich. 1994. "Visual-Auditory Interactions in Sensorimotor Processing: Saccades Versus Manual

Responses." *Journal of Experimental Psychology: Human Perception and Performance* 20 (1): 131–53. https://doi.org/10.1037/0096-1523.20.1.131.

Icard, Thomas F. 2018. "Bayes, Bounds, and Rational Analysis." *Philosophy of Science* 85 (1): 79–101. https://doi.org/10.1086/694837.

Isnard, Vincent, Véronique Chastres, Isabelle Viaud-Delmon, and Clara Suied. 2019. "The Time Course of Auditory Recognition Measured with Rapid Sequences of Short Natural Sounds." *Scientific Reports* 9 (1): 1–10. https://doi.org/10.1038/s41598-019-43126-5.

Jacob, Georgin, R. T. Pramod, Harish Katti, and S. P. Arun. 2021. "Qualitative Similarities and Differences in Visual Object Representations between Brains and Deep Networks." *Nature Communications* 12 (1): 1–14. https://doi.org/10.1038/s41467-021-22078-3.

Jara-Ettinger, Julian, Hyowon Gweon, Laura E. Schulz, and Joshua B. Tenenbaum. 2016. "The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology." *Trends in Cognitive Sciences* 20 (8): 589–604. https://doi.org/10.1016/j.tics.2016.05.011.

Jenkin, Zoe. 2020. "The Epistemic Role of Core Cognition." *The Philosophical Review* 129 (2): 251–98. https://doi.org/10.1215/00318108-8012850.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. macmillan.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. "Scaling Laws for Neural Language Models." *ArXiv Preprint ArXiv:2001.08361*. https://arxiv.org/abs/2001.08361.

Kar, Kohitij, and James J DiCarlo. 2021. "Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition." *Neuron* 109 (1): 164-176.e5. https://doi.org/https://doi.org/10.1016/j.neuron.2020.09.035.

Karpathy, Andrej. 2021. "Workshop on Autonomous Driving." In *Computer Vision and Pattern Recognition CVPR*. https://www.youtube.com/watch?v=g6bOwQdCJrc.

Kell, Alexander J E, and Josh H McDermott. 2019. "Deep Neural Network Models of Sensory Systems: Windows onto the Role of Task Constraints." *Current Opinion in Neurobiology* 55: 121–32. https://doi.org/https://doi.org/10.1016/j.conb.2019.02.003.

Kell, Alexander J E, Daniel L K Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. 2018. "A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy." *Neuron* 98 (3): 630-644.e16. https://doi.org/https://doi.org/10.1016/j.neuron.2018.03.044.

Kemp, Charles, Patrick Shafto, and Joshua B. Tenenbaum. 2012. "An Integrated Account of Generalization across Objects and Features." *Cognitive Psychology* 64 (1–2): 35–73. https://doi.org/10.1016/j.cogpsych.2011.10.001.

Khaligh-Razavi, Seyed Mahdi, and Nikolaus Kriegeskorte. 2014. "Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation." *PLoS Computational Biology* 10 (11). https://doi.org/10.1371/journal.pcbi.1003915.

Kim, Junkyung, Matthew Ricci, and Thomas Serre. 2018. "Not-So-CLEVR: Learning Same-Different Relations Strains Feedforward Neural Networks." *Interface Focus* 8 (4). https://doi.org/10.1098/rsfs.2018.0011.

Kirchner, Holle, and Simon J. Thorpe. 2006. "Ultra-Rapid Object Detection with Saccadic Eye Movements: Visual Processing Speed Revisited." *Vision Research* 46 (11): 1762–76. https://doi.org/10.1016/j.visres.2005.10.002.

Knill, David C. 1998. "Discrimination of Planar Surface Slant from Texture: Human and Ideal Observers Compared." *Vision Research* 38 (11): 1683–1711. https://doi.org/10.1016/S0042-6989(97)00325-8.

Knill, David C., and Jeffrey A. Saunders. 2003. "Do Humans Optimally Integrate Stereo and Texture Information for Judgments of Surface Slant?" *Vision Research* 43 (24): 2539–58. https://doi.org/10.1016/S0042-6989(03)00458-9.

Kok, Peter, Janneke F M Jehee, and Floris P. de Lange. 2012. "Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex." *Neuron* 75 (2): 265–70. https://doi.org/10.1016/j.neuron.2012.04.034.

Koller, Daphne, and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, Massachusetts: MIT Press. https://doi.org/10.1016/j.ccl.2010.07.006.

Konrad, Koerding, and Wolpert Daniel. 2004. "Bayesian Integration in Sensorimotor Learning." *Nature* 427 (January): 1–4. papers2://publication/uuid/14D22FBF-5F45-4824-9D08-9D8C52D3B47A.

Körding, Konrad P., Ulrik Beierholm, Wei Ji Ma, Steven Quartz, Joshua B. Tenenbaum, and Ladan Shams. 2007. "Causal Inference in Multisensory Perception." *PLoS ONE* 2 (9). https://doi.org/10.1371/journal.pone.0000943.

Körding, Konrad P., Shih Pi Ku, and Daniel M. Wolpert. 2004. "Bayesian Integration in Force Estimation." *Journal of Neurophysiology* 92 (5): 3161–65. https://doi.org/10.1152/jn.00275.2004.

Körding, Konrad P., and Daniel M. Wolpert. 2006. "Bayesian Decision Theory in Sensorimotor Control." *Trends in Cognitive Sciences* 10 (7): 319–26. https://doi.org/10.1016/j.tics.2006.05.003.

Kosinski, Michal. 2023. "Theory of Mind May Have Spontaneously Emerged in Large Language Models." *ArXiv Preprint ArXiv:2302.02083*.

Kulkarni, Tejas D., Pushmeet Kohli, Joshua B. Tenenbaum, and Vikash Mansinghka. 2015. "Picture: A Probabilistic Programming Language for Scene Perception." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 07-12-June: 4390–99. https://doi.org/10.1109/CVPR.2015.7299068.

Kwisthout, Johan. 2011. "Most Probable Explanations in Bayesian Networks: Complexity and Tractability." *International Journal of Approximate Reasoning* 52 (9): 1452–69. https://doi.org/10.1016/j.ijar.2011.08.003.

Kwisthout, Johan, Todd Wareham, and Iris van Rooij. 2011. "Bayesian Intractability Is Not an Ailment That Approximation Can Cure." *Cognitive Sceince*.

Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. "Building Machines That Learn and Think like People." *Behavioral and Brain Sciences* 40: 1–72. https://doi.org/10.1017/S0140525X16001837.

Le, Tuan Anh, Atilim Gunes Baydin, and Frank Wood. 2017. "Inference Compilation and Universal Probabilistic Programming." In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, edited by Aarti Singh and Jerry Zhu, 54:1338–48. Proceedings of Machine Learning Research. PMLR. https://proceedings.mlr.press/v54/le17a.html.

LeCun, Yann. 2023. "On the Highway towards Human-Level AI, Large Language Model Is an off-Ramp." https://twitter.com/ylecun/status/1621805604900585472?t=24fvsZZLbmUBtLoYXz46Xg&s=19.

Lecun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature*. https://doi.org/10.1038/nature14539.

LeCun, Yann, and Tiernan Ray. 2022. "Meta's AI Guru LeCun: Most of Today's AI Approaches Will Never Lead to True Intelligence." *Zdnet*, 2022. https://www.zdnet.com/article/metas-ai-guru-lecun-most-of-todays-ai-approaches-will-never-lead-to-true-intelligence/.

Levy, Roger, Florencia Reali, and Thomas L Griffiths. 2009. "Modeling the Effects of Memory on Human Online Sentence Processing with Particle Filters." In *Proceedings of the 22nd Conference on Neural Information Processing Systems (NIPS)*.

Lewis, David. 1982. "Logic for Equivocators." *Nous* 16 (3): 431–41.

Lewis, Richard L, Andrew Howes, and Satinder Singh. 2014. "Computational Rationality: Linking Mechanism and Behavior through Bounded Utility Maximization." *Topics in Cognitive Science* 6 (2): 279–311. https://doi.org/10.1111/tops.12086.

Lieder, Falk, and Thomas L. Griffiths. 2019. "Resource-Rational Analysis: Understanding Human Cognition as the Optimal Use of Limited Computational Resources." *Behavioral and Brain Sciences*, 2019. https://doi.org/10.1017/S0140525X1900061X.

Lieder, Falk, Thomas L. Griffiths, and Ming Hsu. 2018. "Overrepresentation of Extreme Events in Decision Making Reflects Rational Use of Cognitive Resources." *Psychological Review* 125 (1): 1–32. https://doi.org/10.1037/rev0000074.

Lieder, Falk, Ming Hsu, and Thomas L Griffiths. 2014. "The High Availability of Extreme Events Serves Resource-Rational Decision-Making." In *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 36.

Lighthill, James, E A Feigenbaum, and P McCorduck. 1973. "Lighthill Report: Artificial Intelligence." *Science Research Council (SRC), UK*, 157.

Lipton, Peter. 2004. "Inference to the Best Explanation [1991]." London.

Little, Patrick C, and Chaz Firestone. 2021. "Physically Implied Surfaces." *Psychological Science*.

Liu, Shari, Neon B. Brooks, and Elizabeth S. Spelke. 2019. "Origins of the Concepts Cause, Cost, and Goal in Prereaching Infants." *Proceedings of the National Academy of Sciences of the United States of America* 116 (36): 17747–52. https://doi.org/10.1073/pnas.1904410116.

Lupyan, Gary. 2017. "Changing What You See by Changing What You Know: The Role of Attention." *Frontiers in Psychology* 8 (MAY): 1–15. https://doi.org/10.3389/fpsyg.2017.00553.

Ma, Wei Ji. 2019. "Bayesian Decision Models: A Primer." *Neuron* 104 (1): 164–75. https://doi.org/10.1016/j.neuron.2019.09.037.

Ma, Wei Ji. 2010. "Signal Detection Theory, Uncertainty, and Poisson-like Population Codes." *Vision Research* 50 (22): 2308–19. https://doi.org/10.1016/j.visres.2010.08.035.

MacPherson, Fiona. 2012. "Cognitive Penetration of Colour Experience: Rethinking the Issue in Light of an Indirect Mechanism." *Philosophy and Phenomenological Research* 84 (1): 24–62. https://doi.org/10.1111/j.1933-1592.2010.00481.x.

Mamassian, Pascal, and Michael S Landy. 2001. "Interaction of Visual Prior Constraints." *Vision Research* 41 (20): 2653–68. https://doi.org/https://doi.org/10.1016/S0042-6989(01)00147-X.

Mandelbaum, Eric. 2017. "Seeing and Conceptualizing: Modularity and the Shallow Contents of Perception." *Philosophy and Phenomenological Research*, 1–17. https://doi.org/10.1111/phpr.12368.

Marcus, Gary. 2020. "The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence." http://arxiv.org/abs/2002.06177.

Marotta, Jonathan J, Thomas J McKeeff, and Marlene Behrmann. 2002. "The Effects of Rotation and Inversion on Face Processing in Prosopagnosia." *Cognitive Neuropsychology* 19 (1): 31–47.

Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* MIT Press. https://doi.org/10.1016/0022-2496(83)90030-5.

Marslen-Wilson, William D. 1985. "Speech Shadowing and Speech Comprehension." *Speech Communication* 4 (1–3): 55–73. https://doi.org/10.1016/0167-6393(85)90036-6.

Marti, Sébastien, and Stanislas Dehaene. 2017. "Discrete and Continuous Mechanisms of Temporal Selection in Rapid Visual Streams." *Nature Communications* 8 (1). https://doi.org/10.1038/s41467-017-02079-x.

McCoy, R. Thomas, Junghyun Min, and Tal Linzen. 2020. "BERTs of a Feather Do Not Generalize Together: Large Variability in Generalization across Models with Similar Test Set Performance," 217–27. https://doi.org/10.18653/v1/2020.blackboxnlp-1.21.

McCoy, R. Thomas, Ellie Pavlick, and Tal Linzen. 2019. "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference," no. 1. http://arxiv.org/abs/1902.01007.

Meng, Yuan, Neil R Bramley, and Fei Xu. 2018. "Children's Causal Interventions Combine Discrimination and Confirmation." In *Cognitive Science Society*, 762–67.

Michotte, Albert, G O Thine, and Geneviève Crabbé. 1964. *Les Complements Amodaux Des Structures Perceptives*. Publications universitaires.

Moreno-Bote, R., D. C. Knill, and A. Pouget. 2011. "Bayesian Sampling in Visual Perception." *Proceedings of the National Academy of Sciences* 108 (30): 12491–96. https://doi.org/10.1073/pnas.1101430108.

Moreno-Bote, Rubén, Asya Shpiro, John Rinzel, and Nava Rubin. 2010. "Alternation Rate in Perceptual Bistability Is Maximal at and Symmetric around Equi-Dominance." *Journal of Vision* 10 (11): 1–18. https://doi.org/10.1167/10.11.1.

Morgenstern, Yaniv, Richard F. Murray, and Laurence R. Harris. 2011. "The Human Visual System's Assumption That Light Comes from above Is Weak." *Proceedings of the National Academy of Sciences of the United States of America* 108 (30): 12551–53. https://doi.org/10.1073/pnas.1100794108.

Morrison, John. 2016. "Perceptual Confidence." *Analytic Philosophy* 57 (1): 15–48. http://www.columbia.edu/~jrm2182/Morrison.Perceptual.Confidence.pdf.

Mozer, Michael C., Harold Pashler, and Hadjar Homaei. 2008. "Optimal Predictions in Everyday Cognition: The Wisdom of Individuals or Crowds?" *Cognitive Science* 32 (7): 1133–47. https://doi.org/10.1080/03640210802353016.

Murray, Richard F., Khushbu Patel, and Alan Yee. 2015. "Posterior Probability Matching and Human Perceptual Decision Making." *PLoS Computational Biology* 11 (6): 1–16. https://doi.org/10.1371/journal.pcbi.1004342.

Mylopoulos, Myrto. 2021. "The Modularity of the Motor System." *Philosophical Explorations*, 376–93.

Nanay, Bence. 2018. "The Importance of Amodal Completion in Everyday Perception." *I-Perception* 9 (4). https://doi.org/10.1177/2041669518788887.

Nussenbaum, Kate, Alexandra O. Cohen, Zachary J. Davis, David J. Halpern, Todd M. Gureckis, and Catherine A. Hartley. 2020. "Causal Information-Seeking Strategies Change Across Childhood and Adolescence." *Cognitive Science* 44 (9). https://doi.org/10.1111/cogs.12888.

Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert. 2017. "Feature Visualization." *Distill* 2 (11): e7.

Olkkonen, Maria, Thorsten Hansen, and Karl R Gegenfurtner. 2008. "Color Appearance of Familiar Objects: Effects of Object Shape, Texture, and Illumination Changes." *Journal of Vision* 8 (5): 1–16. https://doi.org/10.1167/8.5.13.

Olsson, Catherine, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, Chris Olah, 2022. "In-Context Learning and Induction Heads."

Papineau, David. 2021. *The Metaphysics of Sensory Experience*. *The Metaphysics of Sensory Experience*. https://doi.org/10.1093/oso/9780198862390.001.0001.

Park, Jaewoo, and Murali Haran. 2018. "Bayesian Inference in the Presence of Intractable Normalizing Functions." *Journal of the American Statistical Association* 113 (523): 1372–90. https://doi.org/10.1080/01621459.2018.1448824.

Phillips, Ben. 2019. "The Shifting Border Between Perception and Cognition." *Nous* 53 (2): 316–46. https://doi.org/10.1111/nous.12218.

Piantadosi, Steven T., Joshua B. Tenenbaum, and Noah D. Goodman. 2016. "The Logical Primitives of Thought: Empirical Foundations for Compositional Cognitive Models." *Psychological Review* 123 (4): 392–424. https://doi.org/10.1037/a0039980.

Piantadosi, Steven T., Joshua B. Tenenbaum, and Noah D. Goodman. 2012. "Bootstrapping in a Language of Thought: A Formal Model of Numerical Concept Learning." *Cognition* 123 (2): 199–217. https://doi.org/10.1016/j.cognition.2011.11.005.

Pinker, S. 1997. *How The Mind Works*. https://doi.org/10.1111/j.1749-6632.1999.tb08538.x.

Pylyshyn, Zenon. 1999. "Is Vision Continuous with Cognition? The Case for Cognitive Impenetrability of Visual Perception." *Behavioural and Brain Sciences* 22: 341–423.

Quilty-Dunn, Jake. 2019. "Unconscious Perception and Phenomenal Coherence." *Analysis* 79 (3): 461–69. https://doi.org/10.1093/analys/any022.

Quilty-Dunn, Jake. 2019. "Attention and Encapsulation." *Mind and Language*. https://doi.org/10.1111/mila.12242.

Rahnev, Dobromir, and Rachel N Denison. 2018. "Suboptimality in Perceptual Decision Making." *Behavioral and Brain Sciences* 41: 1–43. https://doi.org/10.1101/060194.

Ramachandran, V. S. 1988. "Perception of Shape from Shading." *Nature*. https://doi.org/10.1016/0002-9394(88)90349-2.

Rescorla, Michael. 2015. "Bayesian Perceptual Psychology." In *Oxford Handbook of Perceptual Psychology*, 193–99.

Riesenhuber, Maximilian, and Tomaso Poggio. 1999. "Hierarchical Models of Object Recognition in Cortex." *Nature Neuroscience* 2 (11): 1019–25.

Rock, Irvin. 1983. *The Logic of Perception*. MIT Press.

Rousselet, Guillaume A., Marc J.M. Macé, and Michèle Fabre-Thorpe. 2003. "Is It an Animal? Is It a Human Face? Fast Processing in Upright and Inverted Natural Scenes." *Journal of Vision* 3 (6): 440–55. https://doi.org/10.1167/3.6.5.

Scholl, Brian J., and Patrice D. Tremoulet. 2000. "Perceptual Causality and Animacy." *Trends in Cognitive Sciences*.

Schulz, Laura. 2012. "Finding New Facts; Thinking New Thoughts." In *Rational Constructivism in Cognitive Development*, edited by Tamar Kushnir and Fei Xu, 269–94. Elsevier.

Schulz, Laura E., and Alison Gopnik. 2004. "Causal Learning Across Domains." *Developmental Psychology* 40 (2): 162–76. https://doi.org/10.1037/0012-1649.40.2.162.

Schwitzgebel, Eric. 2008. "The Unreliability of Naive Introspection." *Philosophical Review* 117 (2): 245–73.

Serre, Thomas. 2019. "Deep Learning: The Good, the Bad, and the Ugly." *Annual Review of Vision Science* 5: 399–426. https://doi.org/10.1146/annurev-vision-091718-014951.

Sexton, Nicholas J, and Bradley C Love. 2022. "Reassessing Hierarchical Correspondences between Brain and Deep Networks through Direct Interface." *Science Advances* 8 (28): eabm2219. https://doi.org/10.1126/sciadv.abm2219.

Shams, Ladan, Wei Ji Ma, and Ulrik Beierholm. 2005. "Sound-Induced Flash Illusion as an Optimal Percept." *NeuroReport* 16 (17): 1923–27. https://doi.org/10.1097/01.wnr.0000187634.68504.bb.

Shannon, Claude E. 1961. *The Shannon Centennial: 1100100 Years of Bits*.

Shapira, Natalie, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. "Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models." http://arxiv.org/abs/2305.14763.

Shea, Nicholas. 2023. "Moving beyond Content-Specific Computation in Artificial Neural Networks." *Mind and Language* 38 (1): 156–77. https://doi.org/10.1111/mila.12387.

Siegel, Susanna. 2017. "How Is Wishful Seeing like Wishful Thinking?" *Philosophy and Phenomenological Research*.

Siegel, Susanna. 2012. "Cognitive Penetrability and Perceptual Justification." *Nous* 46 (2): 201–22. https://doi.org/10.1111/j.1468-0068.2010.00786.x.

Silins, Nicholas. 2016. "Cognitive Penetration and the Epistemology of Perception." *Philosophy Compass* 11 (1): 24–42. https://doi.org/10.1111/phc3.12292.

Silver, David, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, and Marc Lanctot. 2016. "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature* 529 (7587): 484–89.

Simon, Herbert Alexander. 1965. *The Shape of Automation for Men and Management*. Vol. 13. Harper & Row New York.

Simon, Herbert Alexander. 1997. *Models of Bounded Rationality: Empirically Grounded Economic Reason*. Vol. 3. MIT press.

Sipser, Michael. 2013. *Introduction to the Theory of Computation*. Cengage Learning.

Smith, Kevin A, Lingjie Mei, Shunyu Yao, and Jiajun Wu. 2019. "Modeling Expectation Violation in Intuitive Physics with Coarse Probabilistic Object Representations," no. NeurIPS: 1–11.

Sokal, A. 1997. "Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms." In *Functional Integration*, 131–92. Springer. https://doi.org/10.1007/978-1-4899-0319-8_6.

Spelke, Elizabeth S., and Katherine D. Kinzler. 2007. "Core Knowledge." *Developmental Science* 10 (1): 89–96. https://doi.org/10.1111/j.1467-7687.2007.00569.x.

Stalnaker, Robert. 1984. *Inquiry*. MIT Press.

Stone, Mervyn. 1960. "Models for Choice-Reaction Time." *Psychometrika* 25 (3): 251–60. https://doi.org/10.1007/BF02289729.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2020. "Energy and Policy Considerations for Deep Learning in NLP." In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 3645–50. https://doi.org/10.18653/v1/p19-1355.

Sundareswara, Rashmi, and Paul R. Schrater. 2008. "Perceptual Multistability Predicted by Search Model for Bayesian Decisions." *Journal of Vision* 8 (5): 1–19. https://doi.org/10.1167/8.5.12.

Szymanik, Jakub, and Rineke Verbrugge. 2018. "Tractability and the Computational Mind." In *The Routledge Handbook of the Computational Mind*, 339–54. Routledge.

Tang, Hanlin, Calin Buia, Radhika Madhavan, Nathan E Crone, Joseph R Madsen, William S Anderson, and Gabriel Kreiman. 2014. "Spatiotemporal Dynamics Underlying Object Completion in Human Ventral Visual Cortex." *Neuron* 83 (3): 736–48. https://doi.org/https://doi.org/10.1016/j.neuron.2014.06.017.

Tenenbaum, JCoshua B, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. 2011. "How to Grow a Mind: Statistics, Structure, and Abstraction." *Science* 331 (6022): 1279–85. https://doi.org/10.1126/science.1192788.

Tenenbaum, Joshua B, and Thomas L Griffiths. 2006. "Optimal Predictions in Everyday Cognition." *Psychological Science* 17 (9): 767–73.

Tessler, Michael Henry, and Noah D. Goodman. 2019. "The Language of Generalization." *Psychological Review* 126 (3): 395–436. https://doi.org/10.1037/rev0000142.

Thorpe, Simon, Denis Fize, and Catherine Marlot. 1996. "Speed of Processing in the Human Visual System." *Nature* 381 (6582): 520–22. https://doi.org/10.1038/381520a0.

Tokunaga, Rumi, and Alexander D. Logvinenko. 2010. "Material and Lighting Hues of Object Colour." *Ophthalmic and Physiological Optics* 30 (5): 611–17. https://doi.org/10.1111/j.1475-1313.2010.00733.x.

Tooby, John, and Leda Cosmides. 1992. "The Psychological Foundations of Culture." In *The Adapted Mind: Evolutionary Psychology and the Generation of Culture.*, 19–136. New York, NY, US: Oxford University Press.

Turing, Alan. 1936. "On Computable Numbers, With an Application to the Entscheidungsproblem." *Proceedings of the London Mathematical Society*. http://draperg.cis.byuh.edu/archive/winter2014/cs320/Turing_Paper_1936.pdf.

Ullman, Tomer. 2023. "Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks." http://arxiv.org/abs/2302.08399.

Ullman, Tomer D., Noah D. Goodman, and Joshua B. Tenenbaum. 2012. "Theory Learning as Stochastic Search in the Language of Thought." *Cognitive Development* 27 (4): 455–80. https://doi.org/10.1016/j.cogdev.2012.07.005.

Ullman, Tomer D., Elizabeth Spelke, Peter Battaglia, and Joshua B. Tenenbaum. 2017. "Mind Games: Game Engines as an Architecture for Intuitive Physics." *Trends in Cognitive Sciences* 21 (9): 649–65. https://doi.org/10.1016/j.tics.2017.05.012.

Ullman, Tomer D., Andreas Stuhlmüller, Noah D. Goodman, and Joshua B. Tenenbaum. 2018. "Learning Physical Parameters from Dynamic Scenes." *Cognitive Psychology* 104: 57–82. https://doi.org/10.1016/j.cogpsych.2017.05.006.

Beers, Robert J. Van, Anne C. Sittig, and Jan J. Denier Van Der Gon. 1999. "Integration of Proprioceptive and Visual Position-Information: An Experimentally Supported Model." *Journal of Neurophysiology* 81 (3): 1355–64. https://doi.org/10.1152/jn.1999.81.3.1355.

Fraassen, Bas C Van. 1989. *Laws and Symmetry*. Clarendon Press.

Leeuwen, Neil Van, and Tania Lombrozo. 2023. "The Puzzle of Belief." *Cognitive Science* 47 (2): e13245.

Opstal, John Van. 2016. "Chapter 13 - Multisensory Integration." In *The Auditory System and Human Sound-Localization Behavior*. Academic Press.

Rooij, Iris van. 2008. "The Tractable Cognition Thesis." *Cognitive Science* 32 (6): 939–84. https://doi.org/10.1080/03640210801897856.

Villalobos, Pablo, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, and Marius Hobbhahn. n.d. "Will We Run out of Data ? An Analysis of the Limits of Scaling Datasets in Machine Learning."

Vul, Edward, Noah Goodman, Thomas L. Griffiths, and Joshua B. Tenenbaum. 2014. "One and Done? Optimal Decisions from Very Few Samples." *Cognitive Science* 38 (4): 599–637. https://doi.org/10.1111/cogs.12101.

Wang, Chun, and Gongjun Xu. 2015. "A Mixture Hierarchical Model for Response Times and Response Accuracy." *British Journal of Mathematical and Statistical Psychology* 68 (3): 456–77. https://doi.org/10.1111/bmsp.12054.

Warstadt, Alex, and Samuel R. Bowman. 2022. "What Artificial Neural Networks Can Tell Us about Human Language Acquisition." https://doi.org/10.1201/9781003205388-2.

Weisberg, Jonathan. 2009. "Locating IBE in the Bayesian Framework." *Synthese* 167 (1): 125–43.

Weiss, Yair, Eero P. Simoncelli, and Edward H. Adelson. 2002. “Motion Illusions as Optimal Percepts.” *Nature Neuroscience* 5 (6): 598–604. https://doi.org/10.1038/nn858.

Wilder, John D., Wendy J. Adams, and Richard F. Murray. 2019. “Shape from Shading under Inconsistent Illumination.” *Journal of Vision* 19 (6): 1–15. https://doi.org/10.1167/19.6.2.

Witkin, Andrew P, and Jay M Tenenbaum. 1983. “On the Role of Structure in Vision.” In *Notes and Reports in Computer Science and Applied Mathematics*, edited by Jacob Beck, Barbara Hope, and Azriel B T - Human and Machine Vision Rosenfeld, 481–543. Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-084320-6.50022-0.

Wolpert, David H, and William G Macready. 1997. “No Free Lunch Theorems for Optimization.” *IEEE Transactions on Evolutionary Computation* 1 (1): 67–82. https://doi.org/10.1109/4235.585893.

Wong, Lionel, Gabriel Grand, Alexander K. Lew, Noah D. Goodman, Vikash K. Mansinghka, Jacob Andreas, and Joshua B. Tenenbaum. 2023. “From Word Models to World Models: Translating from Natural Language to the Probabilistic Language of Thought,” 1–94. http://arxiv.org/abs/2306.12672.

Wozny, David R, Ulrik R Beierholm, and Ladan Shams. 2010. “Probability Matching as a Computational Strategy Used in Perception.” *PLOS Computational Biology* 6 (8): 1–7. https://doi.org/10.1371/journal.pcbi.1000871.

Wu, Jiajun, Ilker Yildirim, J.J. Lim, W.T. Freeman, and J.B. Tenenbaum. 2015. “Galileo : Perceiving Physical Object Properties by Integrating a Physics Engine with Deep Learning.” *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, 1–9.

Xu, Fei, and J.B. Joshua B Tenenbaum. 2007. “Word Learning as Bayesian Inference.” *Psychological Review* 114 (2): 245. https://doi.org/10.1037/0033-295X.114.2.245.

Yalcin, Seth. 2018. “Belief as Question-Sensitive.” *Philosophy and Phenomenological Research* 97 (1): 23–47. https://doi.org/10.1111/phpr.12330.

Yamins, Daniel L.K., and James J. DiCarlo. 2016. “Using Goal-Driven Deep Learning Models to Understand Sensory Cortex.” *Nature Neuroscience*. https://doi.org/10.1038/nn.4244.

Yamins, Daniel L.K., Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J. DiCarlo. 2014. “Performance-Optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (23): 8619–24. https://doi.org/10.1073/pnas.1403112111.

Yildirim, Ilker, Mario Belledonne, Winrich Freiwald, Joshua Tenenbaum, Mario Belledonne, Winrich Freiwald, and Joshua Tenenbaum. 2018. “Efficient Inverse Graphics in Biological Face Processing.” *Science Advances* 6 (10): 282798. https://doi.org/10.1126/sciadv.aax5979.

Zhang, Shizhuo Dylan, Curt Tigges, Stella Biderman, Maxim Raginsky, and Talia Ringer. 2023. “Can Transformers Learn to Solve Problems Recursively?” http://arxiv.org/abs/2305.14699.

Zhi-Xuan, Tan, Jordyn L. Mann, Tom Silver, Joshua B. Tenenbaum, and Vikash K. Mansinghka. 2020. “Online Bayesian Goal Inference for Boundedly-Rational Planning Agents.” *ArXiv*, no. NeurIPS.

Ziegler, Daniel M, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. “Fine-Tuning Language Models from Human Preferences.” *ArXiv Preprint ArXiv:1909.08593*.