Getting the Picture

Robert Akerlof, Richard Holden, and Hongyi Li[*]

March 10, 2026

Abstract

Standard economic theory assumes that people have perfect information processing ability: they can derive *all* logical consequences of information *immediately*. In reality, however, people struggle to work out the consequences of information. In this paper, we offer a theory of the reasoning process—how people go about "connecting the dots." In our theory, an agent may have all of the information needed to draw a conclusion yet they still fail to see it. Our key assumption is that agents have limited "working memory." This constrains the number of pieces of information—"dots"—an agent can think about and connect *at once*. We explore how agents analyze pieces of information and work their way to a big picture—or, "narrative"—and how narratives, in turn, shape the agent's view of the parts. We show that limited working memory makes sense of why people struggle with tradeoffs, suffer from choice overload, are influenced by defaults, engage in satisficing, selectively attend to attributes, are subject to primacy effects, and may vacillate between interpretations. We apply the framework to political persuasion, showing how supplying a simple narrative can lock in an interpretation of ambiguous evidence.

JEL Classification: D01, D80, D90.

Keywords: Cognition, reasoning, perception, narratives.

[*] Akerlof: UNSW Business School and University of Warwick, r.akerlof@unsw.edu.au. Holden: UNSW Business School, richard.holden@unsw.edu.au. Li: UNSW Business School, hongyi@unsw.edu.au. We are grateful to Drew Fudenberg, Rachel Kranton, Raghav Malhotra, Stephen Morris, Sharun Mukand, Sendhil Mullainathan, Frank Schilbach, Andrei Shleifer, Richard Thaler, and DJ Thornton for helpful comments, as well as seminar participants at Harvard, MIT, Brown, Monash, and UNSW. All authors acknowledge financial support from the Manos Institute for Cognitive Economics at UNSW. Li additionally acknowledges financial support from the Australian Research Council Discovery Project DP240103257.

1 Introduction

Standard economic theory assumes that people have perfect information processing capabilities. Given a set of information, people can immediately derive all logical consequences from it. Two words are worth emphasizing here: *immediately* and *all*.

Yet there is now a vast body of evidence that people struggle with processing information. There have also been many contributions offering rationales for why people's behavior may deviate from full rationality.

With some important exceptions that we discuss below—notably the work of Bordalo, Gennaioli, Shleifer, and coauthors—these contributions can largely be characterized as involving optimization subject to constraints, such as imperfect memory, limited information, limited attention, bounded communication, or limited ability to reason about the strategies of others. But the reasoning process underlying optimization goes unexamined.

Our paper focuses on this reasoning process: how people, endowed with all the information needed to draw a conclusion, go about "connecting the dots" to determine what is optimal. To do this we adopt the perspective of the Gestalt school of psychology: that there is a fundamental difference between perceiving the parts and perceiving the whole.[1]

To give this approach purchase, we assume that agents have limited *working memory*. This constrains the number of pieces of information—"dots"—an agent can think about and connect *at once*. As a result, the agent's capacity to draw conclusions is bounded, and the speed at which they reach those conclusions is slowed. Put differently, *computation* becomes part of the agent's problem.

Understanding how people connect—or fail to connect—the dots is central to understanding economic decision-making. People are flooded with data; but to take action, they need to make sense of it. In other words, they need a big-picture view—or a "nar-

[1] In the early 20th century, a new school of psychology emerged that challenged prevailing notions regarding human perception. The so-called Gestalt theorists, such as Max Wertheimer, realized that there is a fundamental difference between perceiving the *parts* of a picture and perceiving its *whole*. As a consequence, people may struggle to see the big picture.

rative" (see Shiller (2017)). Take the 2008 financial crisis. One narrative took hold regarding rising house prices—that they were driven by fundamentals—rather than an alternative—that there was a speculative frenzy and loose lending standards. The lack of a compelling story can keep people from taking action. This may explain why defaulting people into savings plans tends to increase enrollment (see Madrian and Shea (2001) and Thaler and Benartzi (2004)).

Our goal in this paper is to develop a framework for understanding how people analyze pieces of information and reason their way to a big picture. We also aim to understand how a big picture informs an agent's view of the parts. We then show how that framework can shed new light on important economic problems ranging from the simplest choice problem to more complex economic problems like political persuasion.

Agents in our model have two types of memory: working memory (akin to a computer's RAM) and knowledge (akin to a computer's hard drive). To isolate the role of working memory, we treat knowledge as having unlimited capacity and working memory as the sole capacity constraint. A terminological note: cognitive scientists use "working memory" to refer specifically to a conscious brain system where data is temporarily stored and manipulated (see Gazzaniga et al. (2014)), distinguishing it from unconscious systems that perform similar functions.[2] We use the term more broadly, as computer scientists do, to encompass any temporary processing capacity—conscious or unconscious—that is involved in drawing conclusions from stored information.

To illustrate, consider Mr. Stitch, who is choosing between two suits. Suit 1 is made of high-quality, Loro Piana wool but has poor fit; Suit 2 fits well but the fabric is low-quality polyester. Figure 1 shows his assessments of each suit's quality and sizing, and he values one unit of quality equivalently to one unit of sizing.

Normally, we would abstract from the computation involved in such a choice. This paper makes it explicit. Mr. Stitch's problem involves four pieces of information (10, 3, 2,

[2] In cognitive science, working memory has three components: the phonological loop (for verbal information), the visuospatial sketchpad (for visual and spatial information), and the central executive (which coordinates attention and integrates information across these systems).

	Suit 1	**Suit 2**
Quality	10	2
Sizing	3	8

Figure 1

and 8) and thinking about each takes one unit of working memory. If he has four units of working memory, he can solve the problem *immediately*—he keeps everything in mind at once.

If he has only two units, he can potentially still solve the problem—but only by breaking it up into *manageable steps*. He could store the suits' qualities (10 and 2) in working memory, compute the quality difference (8), and save this to *knowledge* (long-term memory). He can do the same for sizing, and save the sizing difference (-5). These intermediate results (8 and -5) become facts he can later retrieve. Provided they take up one unit of memory each, he can load 8 and -5 and work out that he prefers Suit 1, since $8 - 5 > 0$.[3]

Thus, with two units of working memory, Mr. Stitch solves the problem *more slowly* than with four. If he has only *one unit*, the problem becomes impossible to solve. Solving the problem requires connecting dots—and connecting dots requires multiple units of working memory.

The general model we consider in the paper is one where an agent faces a choice problem and is presented with a matrix of data. We refer to the matrix as a "picture" and the elements as "pixels." The pixels might, for instance, be the characteristics of the two suits Mr. Stitch is choosing between.

We define a "feature" of the picture as some event in picture-space: for instance, "the quality difference between Suit 1 and Suit 2 is 8." We say that a feature is "big-picture" if it involves a large number of pixels (e.g. "the value of Suit 1 exceeds the value of Suit 2") rather than a small number (e.g. "Suit 1's quality is 10"). To make a choice, the agent must work out some big-picture feature—such as "the value of Suit 1 exceeds the value

[3] An alternative approach Mr. Stitch can take is to compute the value of each suit (Value $1 = 10 + 3$ and Value $2 = 2 + 8$) and then compare their values. With either approach, though, he solves the problem *more slowly* than if he has memory capacity of four.

of Suit 2.”

The agent’s time-t “knowledge” is the set of features they think apply to the picture at time t. Initially, the agent knows the values of pixels only (either all pixels or a subset). In the case where the agent knows all of the pixels, they start with complete information about the picture; however they lack a big-picture understanding—and thus still need to learn something further to make a choice.

The agent learns by loading existing knowledge into working memory and drawing conclusions. For instance, they might load “Suit 1’s quality is 10” and “Suit 2’s quality is 2,” and conclude “the quality difference between Suit 1 and Suit 2 is 8.” Each period, the conclusions the agent draws are added to their knowledge set.

The agent has limited working memory, which restricts the amount of knowledge they can load at any one time. Moreover, the agent has a limited set of *concepts* and can only store in working memory those features that are in their concept set.[4] The agent’s limited concept set makes sense, for example, of why some choice problems—such as those involving tradeoffs—are particularly hard.

We consider two variants of the model. In the first, the conclusions the agent draws are purely deductive, while in the second, the agent extrapolates. In the version with extrapolation, the agent adds features to knowledge when they “fit the data” sufficiently well—in the sense that there are relatively few alternative explanations. For instance, Mr. Stitch might extrapolate from “the quality difference between Suit 1 and Suit 2 is 8” to “the value of Suit 1 exceeds the value of Suit 2.”

Extrapolation makes sense of a variety of phenomena—such as satisficing behavior (in the sense of Simon (1955)), selective attention to attributes, and choice overload. Choice overload reflects the idea that as a choice problem gets more complex—for instance, Mr. Stitch has more suit options or suits have more characteristics—it becomes harder to extrapolate to a conclusion about which option is best.

When an agent extrapolates, the interpretation they arrive at may depend on the path

[4] Cognitive scientists and neuroscientists have shown that the brain encodes complex information in terms of concepts to reduce cognitive load (see, for example, Baddeley and Hitch (1974)).

their thinking takes—and which narratives take hold early. In cognitive psychology, this idea is referred to as "perceptual rivalry" and there are many famous examples—such as the rabbit-duck illusion (Figure 2), which can be seen as a rabbit facing right or a duck facing left.

We illustrate perceptual rivalry with the example of a voter, Mr. Judge, who is evaluating a politician about whom there is mixed evidence. Mr. Judge is trying to decide whether the politician is a "good egg" or a "bad egg"—and they can reach either conclusion depending upon what they pay attention to initially. With this example, we also show how persuasion works in our framework. For instance, the politician can influence the voter by supplying them with a "simple" narrative.

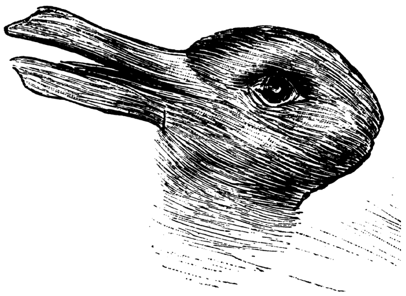


Figure 2: The Rabbit-Duck Illusion

Related Literature

There is, of course, a vast literature in economics on bounded rationality, starting with the work of Simon (1955) who pointed to "limits on computational capacity" as an important constraint on "actual human choice." The highly influential line of work on heuristics and biases, initiated by Kahneman and Tversky, suggests that people rely on mental shortcuts to make decisions due to cognitive limitations, though these shortcuts can sometimes lead to systematic errors (see Kahneman and Tversky (1974) and Thaler (1980)).

A growing body of experimental work demonstrates that many behavioral phenomena classically attributed to nonstandard preferences are in fact driven by computational constraints (see, e.g., Oprea (2024), Enke and Graeber (2023), and Ambuehl et al. (2022)).

Enke and Graeber (2023) also develop a theoretical framework in which cognitive noise compresses judgments toward a default. These papers show that when decision problems are computationally demanding, subjects exhibit patterns resembling loss aversion, probability weighting, and other classic anomalies—even in settings where preferences play no role or where the anomalies disappear when computational demands are reduced. This is very much in accord with the perspective we take.

The recent literature on bounded rationality can largely be characterized as involving optimization subject to constraints. Constraints that have been considered include imperfect memory (Mullainathan (2002) and Wilson (2014)), limited information (Sims (2003)), limited attention to decision-relevant variables (Gabaix (2014)), limited precision in communication (Cremer et al. (2007)), limited ability to reason about other agents' strategies (Stahl and Wilson (1994), Crawford and Iriberri (2007)), or bounded communication (Ellison and Holden (2014)). Our focus, by contrast, is on "connecting the dots"—that is, the process by which people try, successfully or unsuccessfully, to work out what is optimal.

An important exception is the strand of work by Bordalo, Gennaioli, Shleifer, and coauthors (hereafter BGS et al.) which seeks to "get inside people's heads" and explore the consequences for decision-making (Bordalo et al. (2023, 2012, 2013a,b, 2016, 2020, 2024); Mullainathan et al. (2008)). One of their central ideas is that people *categorize* situations, grouping similar cases together and applying the same model of reasoning within categories. In doing so, they may transfer information from a situation where it is useful to one where it is not—which can explain the presence of framing effects. They also argue that agents tend to notice features that are prominent, contrasting, or surprising when comparing situations within the same category (for instance, an item that is particularly expensive compared to similar items the decision-maker recalls). A second key idea is that decision-makers only attend to some of the many features of a decision problem. Thus, the aspects of the problem that are salient play a critical role in determining what people choose.

The approach of BGS et al. is highly complementary to ours. Indeed, in Section 3, we show that our framework also generates selective attention to attributes. Moreover, “features” in our model are similar to “categories” in BGS et al. We see our contribution in relation to BGS et al. as both providing a general characterization of how agents make sense of data and making explicit the computation involved. Moreover, we highlight the role of limited working memory and show how it drives a variety of behavioral phenomena.

Our paper also relates to a literature on narratives and mental models (e.g. Bénabou et al. (2018); Eliaz and Spiegler (2020); Gibbons et al. (2021); Schwartzstein and Sunderam (2021); Wojtowicz (2024)).[5] Relative to this literature, we adopt a distinct definition: we think of narratives as understandings of the big picture.

Our paper draws on the insights of a large empirical and theoretical literature in cognitive psychology. One closely-related strand concerns “chunking” (see Miller (1956) and Chase and Simon (1973)), which demonstrates people’s ability to compress multiple strands of information (such as the digits of a phone number) into less memory-intensive chunks (see Section 2.3 for further discussion). We also relate to a theoretical literature on “minimum description length” (see Chater (1996), Feldman (2000), and Chater and Vitányi (2003)), which argues that people favor explanations that are encoded simply. Our framework offers a foundation for this intuition: simple concepts occupy less working memory, making them easier to incorporate into the reasoning process. Another influential literature (see Tenenbaum et al. (2006) and Oaksford and Chater (2007)) casts induction as Bayesian inference guided by fixed, hard-wired priors—often taking the form of hierarchical Bayesian networks. By contrast, in our framework, the agent needs to learn

[5] Schwartzstein and Sunderam (2021) conceive of an agent as choosing a mental model from an available set that is potentially influenced by a persuader. Eliaz and Spiegler (2020) conceive of narratives as causal interpretations of events (specifically, directed acyclic graphs). Gibbons et al. (2021) conceive of narratives in terms of categorizations and examine how a group’s shared narrative affects their capacity to cooperate. Bénabou et al. (2018) model the process by which narratives disseminate. Perhaps most closely related is Wojtowicz (2024), who considers a different mechanism that can lead to path dependence: the speed at which new data arrives. In Wojtowicz (2024), an agent repeatedly evaluates their existing model against “nearby” alternatives, and switches whenever the alternative has better fit. If data arrives too quickly, path dependence may arise and the agent may not converge to the maximum-likelihood model.

what model they are in without priors to guide them.[6]

Our work also draws on the rich literature in computer science on computation and complexity. In particular, we build on insights from database theory (see Abiteboul et al. (1995)), where a central question is how to store data and extract useful conclusions from it efficiently. The distinction in that literature between data storage and data retrieval parallels our distinction between knowledge and working memory: in both settings, the bottleneck lies not in what is stored but in what can be brought to bear at any one time.

Roadmap

The paper proceeds as follows. Section 2 presents a version of the model in which the agent is purely deductive and shows how working memory constraints limit the agent's ability to draw conclusions. Section 3 considers the possibility that the agent extrapolates as well, and explains the emergence of choice "anomalies" such as satisficing and choice overload. Section 4 shows how the model can be applied to understand perceptual rivalry, where the initial focus of an agent's attention determines which of several conclusions they ultimately draw. Section 5 examines the role narratives play in the agent's reasoning process and explores the role they play in persuasion. Section 6 briefly discusses two natural ways of extending the model. Section 7 concludes. Proofs are in the appendix.

2 Deduction

2.1 Model

An agent is presented with a matrix of data P:

[6] Most work in psychology (like economics) emphasizes the outputs of inductive processes—the conclusions people ultimately draw—whereas we examine how people build up partial insights step by step.

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1N} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{M1} & p_{M2} & p_{M3} & \cdots & p_{MN} \end{bmatrix}$$

with elements p_{xy} drawn from a finite domain Ω: $p_{xy} \in \Omega$. We refer to matrices of this type as "pictures" and the elements as "pixels." We will sometimes use Q to denote a generic picture and q_{xy} to denote a pixel of Q. Let $\mathcal{P}$ denote the set of all pictures.

Picture P might, for example, be the data presented to Mr. Stitch regarding suit quality and sizing (see Figure 3), where 10 and 3 represent the quality and sizing of Suit 1, and 2 and 8 represent the quality and sizing of Suit 2. Alternatively, P might be a visual picture—for instance, one where the pixels are "black" and "white" ($\Omega = \{\text{black}, \text{white}\}$).

$$P = \begin{bmatrix} 10 & 2 \\ 3 & 8 \end{bmatrix}$$

Figure 3: Picture in Mr. Stitch Example

Features

We refer to any proper, non-empty subset f of $\mathcal{P}$ as a *feature* and we say that f *applies* to P if $P \in f$. Let $\mathcal{F}$ denote the set of all features.[7]

Returning to Mr. Stitch, an example of a feature would be:

$$f_1 = \{Q \in \mathcal{P} : q_{11} + q_{21} \geq q_{12} + q_{22}\},$$

which represents the event "the value of Suit 1 is weakly greater than the value of Suit 2," or equivalently, "Suit 1 is an optimal choice"—hence the label f_1.

We say that pixel p_{xy} supports a feature if the feature is informative about that pixel. We define a feature's "size" as the number of pixels in its support.

[7] Features are akin to events in probability theory. Just as an event E "occurs" if the state lies in set E, a feature f "applies to a picture" if the picture lies in set f.

Definition 1. *We say that p_{xy}* supports *feature f (or, equivalently, that p_{xy} is in the support $s(f)$ of feature f) if, for some realization of the other pixels, knowing that f applies to P narrows the set of values p_{xy} could take.*

It is easy to show that the support of feature f_1 is the set of all pixels of P. We will refer to features with many pixels in the support as "big-picture" features or "narratives."

Knowledge and Concepts

The agent's *knowledge* is the set of features they think apply to P. We denote the agent's knowledge at time t by K_t. The agent may add to knowledge over time, in discrete steps.

Not all features are knowable. We denote by $C \subset \mathcal{F}$ the set of knowable features. We refer to features $f \in C$ as *concepts* and set C as the agent's *conceptualization*. For now, we treat the agent's conceptualization as exogenous, however later, we will discuss how it might be endogenized.

The idea that people find it difficult to conceptualize certain features of data accords with a large body of research in cognitive psychology. For instance, it has been shown that people have trouble identifying faces when they are shown upside-down—a phenomenon known as the face inversion effect (Farah et al., 1998; Yin, 1969).

The Agent's Initial Knowledge

We assume that, at the outset, the agent knows the value of every pixel of P, but they do not know bigger-picture features.

To formalize this, we refer to a feature that specifies the value of a particular pixel, and nothing else, as a *pixel value*: that is, a feature of the form $\{Q \in \mathcal{P} : q_{xy} = \omega\}$. We denote the set of all pixel values by F_{pixels} and the set of pixel values that apply to P by $F_{pixels}(P)$. In a slight abuse of terminology, we will typically refer to $F_{pixels}(P)$ simply as the "pixels of P." We assume that the agent can conceptualize all possible pixel values: $F_{pixels} \subseteq C$. This reflects the idea that the "raw data" (i.e. pixels) that the agent is shown is easy to conceptualize.

The agent's initial knowledge is then given by $K_0 = F_{pixels}(P)$. This means, for example, that if Mr. Stitch is shown the picture in Figure 3, he would know at the outset each pixel's value (10, 3, 2, and 8) but no bigger-picture features, such as whether "the value of Suit 1 is weakly greater than the value of Suit 2."

Learning

The agent learns about the picture through a sequence of deductions. We denote the agent's thoughts at time t by T_t. A thought is composed of two elements: $T_t = \{\phi_t, W_t\}$. We refer to the first element $\phi_t \in C$ as the "candidate concept" and the second element $W_t \subseteq K_t$ as the agent's "working memory."

ϕ_t is a feature the agent is considering adding to knowledge and W_t consists of features the agent already knows apply to the picture ($W_t \subseteq K_t$). The agent adds ϕ_t to knowledge if it is deducible from the features in working memory. To give an example, suppose the agent loads into working memory "the quality of Suit 1 is 10" and "the quality of Suit 2 is 2." The agent would add "Suit 1's relative quality is +8" to knowledge if this was the candidate concept.

Formally, ϕ_t is deducible if $\cap_{f \in W_t} f \subseteq \phi_t$. The agent's knowledge at time $t+1$ is $K_{t+1} = K_t \cup \{\phi_t\}$ if ϕ_t is deducible and $K_{t+1} = K_t$ otherwise.

Limited Working Memory

Importantly, the agent has limited working memory capacity. The agent can load at most L features into working memory: $|W_t| \leq L$. As a result, the agent can only use a limited amount of their knowledge at any one time.

The Agent's Problem

Initially, our focus will be on determining *which features* of the picture the agent is able to learn—and *how long it would take* the agent to learn those features.

However, we have in mind that the agent faces a choice problem and needs to figure out the problem's "decision-relevant features." In Mr. Stitch's case, a decision-relevant feature is “the value of Suit 1 is weakly greater than the value of Suit 2.” Learning this feature tells Mr. Stitch that it is optimal to choose Suit 1.

More generally, think of the agent as facing a choice problem where they must choose an action a from a set A. They have a utility function $u : A \times \mathcal{P} \to \mathbb{R}$ that depends upon their choice of action and which picture they face. We think of decision-relevant features as those which tell the agent which action is optimal. For instance,

$$f_a = \{Q \in \mathcal{P} : u(a, Q) \geq u(a', Q) \text{ for all } a' \in A\}$$

represents the event “a is an optimal action.” We think of the agent as ready to *choose* action a when they deduce feature f_a.

2.2 What is deducible, and when?

We will now examine which concepts an agent can deduce from a picture and how many reasoning steps deductions require. Two forces matter throughout: the agent’s working-memory capacity L and the agent’s conceptualization C. We begin with a definition and then present two propositions that isolate the role of each force.

Definition 2. *Let $C(P)$ denote all of the concepts that are applicable to P.*

1. *Concept $f \in C(P)$ is deducible in τ steps ($f \in D_\tau$) if there exists a sequence of thoughts $T_0, ..., T_{\tau-1}$ such that $f \in K_\tau$.*
2. *Concept $f \in C(P)$ is deducible ($f \in D$) if it is deducible in τ steps for some finite τ.*

Working-Memory Limits

The following proposition isolates the role of working-memory capacity.

Proposition 1.

1. *If $L \geq M \times N$, the agent can deduce any applicable concept in a single step: $D_1 = C(P)$.*

2. *If $L = 1$, the agent cannot deduce anything new about P: every deducible concept $f \in D$ satisfies $p \subseteq f$ for some pixel value $p \in K_0$.*

3. *If $1 < L < M \times N$, then there exist conceptualizations for which:*

   (i) Some, but not all, of the applicable concepts can be deduced: $K_0 \subset D \subset C(P)$.

   (ii) Some concepts are only deducible in multiple steps: $D_1 \subset D_2$.

Part (1) follows from the observation that when $L \geq M \times N$, the agent can load all pixels into working memory at once, which suffices to deduce any concept that applies to the picture. Part (2) reflects the fact that substantive deduction requires integrating multiple pieces of information (i.e. "connecting dots"). With only one unit of working memory, no integration is possible. Part (3) is the substantive intermediate case: with limited but nontrivial working memory, the agent can deduce some applicable concepts but not others, and some deductions require multi-step reasoning chains.

An Illustration of Multi-Step Reasoning: Mr. Stitch

To understand part (3) of the proposition better, consider the problem faced by Mr. Stitch. As before, let

$$f_1 = \{Q \in \mathcal{P} : q_{11} + q_{21} \geq q_{12} + q_{22}\},$$

denote the event "Suit 1 is weakly better than Suit 2."

Suppose f_1 is a concept ($f_1 \in C$). Per part 1 of Proposition 1, if $L \geq 4$, Mr. Stitch can deduce f_1 in a single step by putting his knowledge of all four pixels into working memory. If $L < 4$, he cannot deduce f_1 in a single step. However, he may still be able to do so in multiple steps by making use of additional concepts.

For example, suppose $L = 2$. Let

$$g_1 = \{Q \in \mathcal{P} : q_{11} - q_{12} = 8\}, \qquad g_2 = \{Q \in \mathcal{P} : q_{21} - q_{22} = -5\},$$

so that g_1 represents "Suit 1's relative quality is +8" and g_2 represents "Suit 1's relative sizing is -5." If g_1 and g_2 are concepts, it is possible to use them to deduce f_1 in three steps:

1. Mr. Stitch loads the top two pixels (10 and 2) into working memory and deduces g_1.
2. Mr. Stitch loads the bottom two pixels (3 and 8) into working memory and deduces g_2.
3. Mr. Stitch loads g_1 and g_2 into working memory and deduces f_1.

At each step, Mr. Stitch remains within his working-memory budget ($L = 2$) since he stores only two concepts at a time.[8]

This example illustrates the value of multi-step reasoning. Multi-step reasoning allows Mr. Stitch to operate within his working-memory constraints by reformulating the problem in terms of concepts that are more economical than pixels for the task at hand.

Dominant Options

Let us compare Mr. Stitch's original problem with a second problem in which Suit 1 dominates Suit 2 on both quality and sizing (see Figure 4). Intuitively, this second problem is easier, since no tradeoff is present. A large experimental literature documents that choices involving attribute tradeoffs are more difficult, requiring greater cognitive effort and exhibiting greater sensitivity to framing and context, whereas dominance relations simplify choice (see especially Tversky (1972)).

	Suit 1	**Suit 2**
Quality	8	4
Sizing	5	2

Figure 4: An easier choice problem

[8] Note that g_1 and g_2 are not the only intermediate concepts that Mr. Stitch could use. For instance, he could also use concepts g_1' and g_2' that represent the values of $q_{11} + q_{21}$ and $q_{12} + q_{22}$: g_1' is the set where $q_{11} + q_{21} = 13$ and g_2' is the set where $q_{12} + q_{22} = 10$.

The model accounts for this empirical regularity in the following way. In dominance problems, the agent can rely on *coarse, attribute-by-attribute comparisons*, rather than on exact quantitative computations.

Formally, let

$$h_1 = \{Q \in \mathcal{P} : q_{11} - q_{12} \geq 0\}, \qquad h_2 = \{Q \in \mathcal{P} : q_{21} - q_{22} \geq 0\},$$

so that h_1 represents "Suit 1 is better quality than Suit 2," and h_2 represents "Suit 1 has better sizing than Suit 2." With a working-memory constraint of $L = 2$, Mr. Stitch can deduce f_1 in three steps with h_1 and h_2 as intermediate concepts:

1. Mr. Stitch loads the top two pixels (8 and 4) into working memory and deduces h_1.
2. Mr. Stitch loads the bottom two pixels (5 and 2) into working memory and deduces h_2.
3. Mr. Stitch loads h_1 and h_2 into working memory and deduces f_1.

Thus, dominance permits a deduction that relies only on coarse intermediate concepts.

By contrast, in tradeoff problems such as the earlier example, such coarse features are insufficient: knowing only whether Suit 1 has better quality and whether Suit 1 has better sizing does not determine which suit has better quality-plus-sizing. Instead, the agent must represent finer information—either by holding more pixels simultaneously in working memory or by constructing more particular intermediate concepts, such as g_1 and g_2.

This explanation for why tradeoff problems are harder implicitly relies on an assumption about the agent's conceptualization: namely, that coarse features such as h_1 and h_2 are more likely to be part of the conceptualization than finer-grained features such as g_1 and g_2. As we discuss later when we endogenize the agent's conceptualization C, this assumption is well founded. Coarse, attribute-by-attribute comparisons like h_1 and h_2 are simple, broadly applicable, and reusable across many choice problems, whereas features encoding exact quantitative relationships are more specialized and problem-specific. When the agent faces limits or costs in maintaining concepts, it is therefore natural that the conceptualization favors features that generate repeated savings in working memory. As a result, dominance-based problems are more likely to align with the agent's existing intermediate concepts, while tradeoff problems more often require concepts that are unavailable or costly to construct. This asymmetry makes dominance problems systematically easier to solve, even holding fixed the agent's working-memory capacity.

Conceptualization Ability

Proposition 1 focused on the role of working-memory capacity, taking the agent's conceptualization as given. The examples above, however, already illustrate that the agent's conceptualization plays an independent and crucial role in deduction. We now turn to the role played by the agent's conceptualization. Let $D(C)$ denote the set of deducible concepts under conceptualization C, and let $k(f, C)$ denote the minimal number of steps required to deduce f given C. We obtain the following proposition.

Proposition 2. *Given a picture P, if $1 < L < M \times N$:*

1. *Expanding the set of concepts reduces the number of steps required to deduce any previously deducible concept. That is, for any C and $C' \supset C$, and any $f \in D(C)$, $k(f, C') \leq k(f, C)$, with strict inequality for some P,f, C, and C'.*

2. *Expanding the set of concepts increases the set of concepts from the original set that are deducible. That is, for any C and $C' \supset C$, $D(C') \cap C \supseteq D(C)$, with strict inclusion for some P, C and C'.*

3. *If the set of concepts contains all features that apply to P, then every applicable concept is deducible: $D(C) = C(P)$.*

Part (1) of the proposition formalizes the idea that richer conceptualizations speed deduction. Intuitively, even if a concept is deducible with conceptualization C, a richer conceptualization C' may allow the agent to reach that concept using fewer intermediate steps—for example, by replacing a long chain of pixel-level inferences with a shorter chain involving higher-level features.

Part (2) says that expanding the conceptualization makes new deductions possible. Some concepts may be out of reach simply because the agent lacks the right intermediate concepts to bridge between pixels and the target. Adding concepts can create these missing links, allowing the agent to deduce concepts that were previously inaccessible.

Part (3) describes an extreme case: if the agent has a sufficiently rich conceptualization—so rich that it contains all features of the picture—then every concept is deducible through some sequence of deductions (provided $L \geq 2$). There may still be features, however, that are only deducible in a large number of steps.

Taken together, Propositions 1 and 2 show that bounded rationality arises from two distinct sources: limits on working memory and conceptualization limits. Finally, these results highlight an important sense in which working memory and conceptualization act as substitutes. With sufficient working memory, the agent can deduce any concept directly from pixels, even in the absence of intermediate concepts. Conversely, with a sufficiently rich conceptualization, even a very small working-memory capacity can suffice to deduce features. Intermediate concepts allow the agent to accomplish, over multiple steps, what would otherwise be done in a single step with many pixels loaded simultaneously.

2.3 Chunking

An important finding in cognitive psychology is that memory is limited not so much by the sheer volume of information we are trying to remember as by how that information is organized. This idea was pioneered by Miller (1956), who introduced the term *chunking* to describe the process of binding small, individual pieces of information into a smaller number of meaningful units, or "chunks," that can be stored more efficiently in memory. A canonical example is the telephone number: the same digits are far easier to remember when organized into familiar chunks (e.g. 512–3482) than when presented as an unstructured sequence (e.g. 5–1–2–3–4–8–2).

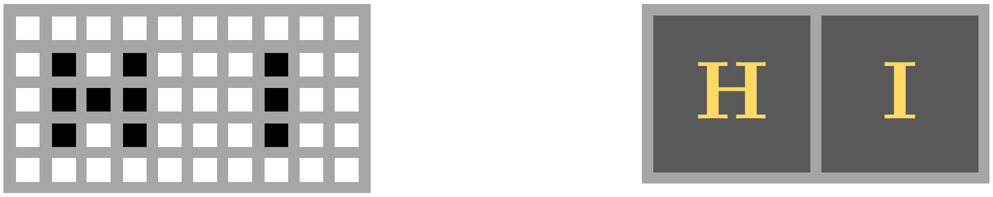
A more striking illustration of chunking comes from chess. In a well-known experiment, Chase and Simon (1973) showed expert and novice chess players a board position for five seconds and then asked them to reconstruct it from memory.[9] When the position was taken from an actual game, experts dramatically outperformed novices: on average, experts correctly recalled the locations of about 16 out of 20 pieces, while novices recalled only about 4. When the position was random, however, this advantage disappeared. Both groups performed poorly, recalling only about 3 pieces on average. The standard interpretation is that experts learn to chunk familiar chess configurations, which allows them to store and retrieve complex positions efficiently. Random configurations do not afford such chunking, leaving experts no better off than novices.

These findings map cleanly onto our model. In our model, working memory is constrained by the number of concepts an agent can load simultaneously (L), not by the amount of raw data those concepts encode. Chunking corresponds to the availability of concepts that combine multiple lower-level concepts into one concept that occupies a single unit of working memory. Importantly, such chunks may be lossless: they may encode exactly the same information as the underlying features, but in a more economical form.

To make this point concrete, consider Figure 5. Panel (a) shows a 5×10 matrix of pixels that can be thought of as analogous to the positions of pieces on a game board. Suppose that one agent—analogous to an expert—has concepts "H" and "I" which respectively pin down the left and right 5×5 submatrices of the picture, while another agent—analogous to a novice—does not. The first agent can load the entire board in working memory using just these two concepts, as illustrated in Panel (b); whereas the second agent might be forced to load all fifty pixels as distinct concepts.

From this perspective, expert performance reflects differences in conceptualization rather than differences in working-memory capacity. Expert chess players possess con-

[9] Chase and Simon (1973) built on earlier work on chunking in chess by De Groot (1965).



(a) Game Board

(b) Game Board - Chunked

Figure 5: Chunking a Game Board

cepts corresponding to common board configurations, which allows them to represent complex positions compactly. Novices lack these concepts and must instead rely on less efficient representations. Random positions fail to activate experts' stored concepts, so they lose their advantage over novices.

2.4 Endogenizing the Agent's Conceptualization

For simplicity, much of our analysis treats the agent's conceptualization C as exogenous. This allows us, for instance, to isolate the role of conceptualization in generating boundedly rational behavior. At the same time, C is clearly endogenous. Expert chess players, for example, possess concepts—such as common board configurations—that novices lack. A satisfactory account of reasoning limits must therefore explain why some features are conceptualized while others are not.

Our approach is to assume that the agent has a limited capacity to conceptualize. Specifically, there is a bound on the number of features the agent can include in C: $|C| \leq \Lambda$, where Λ is the agent's concept limit.[10] We might then ask: given the choice problem the agent faces, which features are most valuable to include in C? In particular, which concepts help the agent deduce the decision-relevant features of the problem, or allow those deductions to be made more quickly?

It would be inappropriate, in our view, to think of the agent as literally choosing C to maximize an objective function. Our focus, of course, is on a boundedly rational agent for whom such optimization would be difficult. We see it as more natural to think of the

[10] Alternatively, one could assume a cost of adding features to C.

agent's conceptualization C as evolving over time, according to some algorithm. Such an algorithm might gradually push C in the direction of optimality, however. Thus, characterizing the optimal C may shed light on the agent's actual C.

Mr. Stitch revisited

Consider why, under limited conceptual capacity, Mr. Stitch may find it easy to choose between suits when one dominates the other, but difficult to choose when the problem involves a tradeoff. For simplicity, focus on choice problems $P \in \mathcal{P}_1$ in which Suit 1 is weakly preferred to Suit 2, so that Mr. Stitch is able to choose if and only if he can deduce f_1 ("Suit 1 is better than Suit 2").

Suppose again that Mr. Stitch has two units of working memory ($L = 2$). Pixels and the decision-relevant features f_1 and f_2 ("Suit i is better") are concepts, and in addition he can accommodate exactly two further concepts. We refer to these additional concepts as his *intermediate concepts*. We ask: which two intermediate concepts are best, given this constraint?

When limited to only two intermediate concepts, Mr. Stitch cannot deduce f_1 for every $P \in \mathcal{P}_1$. The set of choice problems in $\mathcal{P}_1$ is simply too broad for any fixed pair of intermediate concepts to deal with all cases.

To build intuition, consider some natural candidates. The concepts

$$h_1 = \{Q \in \mathcal{P} : q_{11} - q_{12} \geq 0\}, \qquad h_2 = \{Q \in \mathcal{P} : q_{21} - q_{22} \geq 0\}$$

encode whether Suit 1 has better quality and better sizing. As we saw before, these concepts allow Mr. Stitch to choose correctly whenever Suit 1 dominates Suit 2. However, they fail in tradeoff cases where Suit 1 is worse on one dimension but sufficiently better on the other—for example, when its relative sizing is -5 but its relative quality is 8, so that f_1 applies even though h_1 does not.

One might try to repair this by introducing thresholds. For instance, let

$$h_1' = \{Q \in \mathcal{P} : q_{11} - q_{12} \geq 5\}, \qquad h_2' = \{Q \in \mathcal{P} : q_{21} - q_{22} \geq -5\}.$$

These concepts allow Mr. Stitch to choose correctly in some non-dominant cases, such as the one just described, where Suit 1 has a large advantage on one attribute and a bounded disadvantage on the other. But they perform poorly elsewhere: for example, they fail when Suit 1 has small advantages in both quality and sizing, a case in which the original coarse comparisons h_1 and h_2 would be effective. Any alternative choice of two thresholds exhibits a similar tradeoff. Strengthening the thresholds improves performance in some regions while worsening it in others.

Since no pair of intermediate concepts allows Mr. Stitch to deduce f_1 in every instance, a natural objective is to choose intermediate concepts that *maximize* the number of pictures $P \in \mathcal{P}_1$ for which f_1 is deducible. Under this criterion, h_1 and h_2 are optimal. Intuitively, the features "Suit 1 has better quality" and "Suit 1 has better sizing" apply broadly and can be reused across many choice problems. They guarantee correct choice in all dominance problems, and no other pair of intermediate concepts can strictly expand the set of solvable pictures without sacrificing performance elsewhere.

Boolean decomposition: Ms. Lodge

To understand more generally how to identify intermediate concepts that work well, it is helpful to consider an additional example. This second example illustrates how the problem of selecting intermediate concepts is closely related to a classic problem in computer science: designing efficient computational circuits through the decomposition of Boolean functions (see Wegener (1987)).

Consider the problem of Ms. Lodge who observes three binary attributes of a house $p_i \in \{0, 1\}$: whether it is urban (p_1), is close to cafes (p_2), and has a garden (p_3). She chooses $a \in \{0, 1\}$, whether to rent the house, and prefers to rent it if and only if either (i) it is urban and close to cafés, or (ii) it is suburban and has a garden. Like Mr. Stitch, Ms.

Lodge has working-memory capacity of two ($L = 2$).

Let $F(p_1, p_2, p_3)$ be the Boolean function that takes value 1 if renting is optimal and 0 otherwise. There is a direct correspondence between F and the decision-relevant feature f_1 ("it is optimal to rent"): f_1 applies to P if and only if $F(p_1, p_2, p_3) = 1$. The function F can be written as

$$F(p_1, p_2, p_3) = (p_1 \wedge p_2) \vee (\neg p_1 \wedge p_3),$$

where $\wedge$ denotes "and," $\vee$ denotes "or," and $\neg$ denotes "not." This representation shows that F can be decomposed as

$$F = G_1 \vee G_2,$$

with $G_1(p_1, p_2) = p_1 \wedge p_2$ and $G_2(p_1, p_3) = \neg p_1 \wedge p_3$. Each G_i corresponds naturally to an intermediate feature:

$$g_1 = \{Q \in \mathcal{P} : G_1(q_1, q_2) = 1\}, \qquad g_2 = \{Q \in \mathcal{P} : G_2(q_1, q_3) = 1\}.$$

These features—"good urban house" (an urban house close to cafes) and "good suburban house" (a suburban house with a garden)—are exactly the intermediate concepts Ms. Lodge needs to deduce f_1. Each depends on only two pixels and is therefore deducible within her working-memory constraint. Moreover, by loading g_1 and g_2 into working memory, she can deduce f_1, since f_1 applies if and only if either g_1 or g_2 applies.

This example illustrates a general point. Any decision problem induces a Boolean function F of the underlying pixels that encodes which actions are optimal in which states. From this perspective, identifying useful intermediate concepts is equivalent to identifying decompositions of F that are well adapted to the agent's cognitive constraints. This viewpoint places the problem squarely alongside a class of problems that has been studied extensively in computer science, particularly in the context of circuit design, where the central objective is to decompose complex Boolean functions into simpler components. While we do not pursue this connection further here, the analogy provides a disciplined way of thinking about conceptualization under constraints and points to a rich set of tools for analyzing which intermediate concepts are effective.

3 Extrapolation

Consider again Mr. Stitch's original problem, where Suit 1 has better quality and Suit 2 has better sizing. In Section 2, we showed that if Mr. Stitch has working memory of two ($L = 2$) and lacks appropriate intermediate concepts, he is unable to *deduce* that "Suit 1 is better than Suit 2." All the information needed to reach the conclusion is available, but his deductive abilities are not up to the task. This raises a natural question: in such circumstances, is there some other means by which Mr. Stitch might reach a conclusion?

In this section we consider the possibility that the agent supplements deduction with *extrapolation*. Whereas deduction requires that a candidate concept (ϕ_t) be logically implied by what is in working memory (W_t), extrapolation allows the agent to move from the information they currently have loaded to a conclusion that is merely *suggested* by that information. For example, Mr. Stitch might temporarily focus on the quality dimension alone, loading his knowledge of Suit 1's and Suit 2's quality into working memory, and jump to the conclusion that "Suit 1 is better overall" because it is much better on the attribute currently in view. This conclusion is not deductively warranted, but it may be compelling enough that Mr. Stitch is willing to treat it as if it were true.

At a general level, extrapolation serves two distinct purposes. First, there may be conclusions that are out of reach for an agent even though the raw data are fully observed. In such cases, extrapolation is a way of "filling in" the parts of the reasoning chain the agent cannot explicitly execute. Second, there may be conclusions that are not deductively attainable because the agent simply does not have enough information. Here, too, an agent may wish to extrapolate from what they do see to a broader pattern that is only imperfectly supported by the data.

To illustrate the second purpose, consider an investor deciding whether to buy shares in a new technology firm. The fundamental state of the world is rich: it includes the firm's true long-run profitability, the quality of its management team, the durability of its competitive advantage, and so on. The investor observes only a handful of signals—perhaps a recent earnings announcement, a few news articles about the firm's latest product, and the fact that several peers have invested. From these limited data points, the investor may extrapolate to a narrative such as "this is a high-growth firm poised to dominate its market" or, alternatively, "this is a hype-driven bubble stock." Each narrative goes well beyond what is logically implied by the sparse information available, but each may nevertheless guide choice in the absence of full deductive justification.

Extrapolation expands the set of conclusions an agent can reach, but it comes at the cost that the agent's conclusions may be incorrect.

3.1 Model

To capture extrapolation, instead of assuming that at time t the agent adds a candidate concept ϕ_t to knowledge only when it is *deducible* from the features currently in working memory W_t, we allow the agent to add ϕ_t whenever it is a sufficiently good *extrapolation* from W_t. Concretely, the agent adds ϕ_t to knowledge if its *fit* with W_t, denoted $\Psi(\phi_t; W_t)$, weakly exceeds a threshold $\alpha \in (0,1]$.

While we are not wedded to any particular functional form, we adopt the following fit function.

Let $\delta(S) \subseteq \mathcal{P}$ denote the set of pictures consistent with all features in S, i.e. $\delta(S) = \bigcap_{f \in S} f$, and define $\psi(S) = \log|\mathcal{P}| - \log|\delta(S)|$ as a measure of how strongly the feature set S restricts the feasible pictures. The fit of ϕ_t with working memory W_t is then

$$\Psi(\phi_t; W_t) = 1 - \frac{\psi\big(W_t \cup \{\phi_t\}\big) - \psi(W_t)}{\psi(\{\phi_t\})}.$$

Intuitively, $\psi\big(W_t \cup \{\phi_t\}\big) - \psi(W_t)$ is the *incremental restriction* induced by adding ϕ_t on top of W_t, while $\psi(\{\phi_t\})$ is the restriction induced by ϕ_t on its own. Thus $\Psi(\phi_t; W_t)$

is a normalized measure of how “redundant” ϕ_t is given what is already in working memory.[11]

This functional form has several attractive properties:

1. Fit is higher when ϕ_t holds in a larger fraction of pictures consistent with W_t, i.e. when $\psi(W_t \cup \{\phi_t\}) - \psi(W_t)$ is small.

2. Fit attains its upper bound of one ($\Psi(\phi_t; W_t) = 1$) if and only if ϕ_t is *deducible* from W_t (equivalently, $\delta(W_t) \subseteq \delta(\{\phi_t\})$), because then $\delta\big(W_t \cup \{\phi_t\}\big) = \delta(W_t)$.

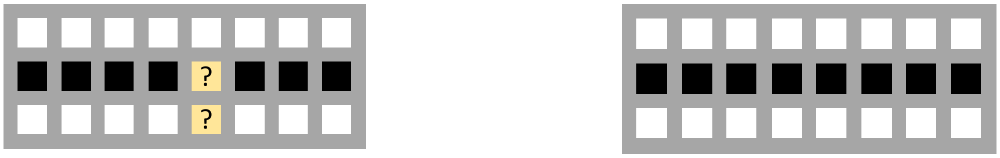
3. Fit equals $-\infty$ if W_t and ϕ_t are incompatible ($\delta\big(W_t \cup \{\phi_t\}\big) = \varnothing$).

4. Fit equals zero when W_t and ϕ_t are independent (W_t tells us nothing about ϕ_t); for instance, if $W_t = \varnothing$.

Finally, note that when $\alpha = 1$, the agent adds ϕ_t if and only if it is deducible from W_t. Thus $\alpha = 1$ nests the baseline (deduction-only) model.

The Agent’s Initial Knowledge

A second change we make is that we allow some pixels to be unknown to the agent: specifically, $K_0 \subsetneq F_{pixels}(P)$.



(a) Initial knowledge.

(b) Horizontal-line feature (ϕ_t).

Figure 6: Example of Incomplete Initial Knowledge

Figure 6 gives an example where the agent does not initially know two of P’s pixels. Based on the pixels the agent does know, they might extrapolate to a feature like the

[11] More precisely: under the uniform distribution over pictures, $\log|\delta(S)|$ is the Boltzmann entropy of the posterior support given S, and $\psi(S) = \log|\mathcal{P}| - \log|\delta(S)|$ is the information gained by imposing S. Then $\psi\big(W_t \cup \{\phi_t\}\big) - \psi(W_t) = \log|\delta(W_t)| - \log\Big|\delta\big(W_t \cup \{\phi_t\}\big)\Big|$ is the information gain from adding ϕ_t conditional on W_t, and $\Psi(\phi_t; W_t)$ normalizes this conditional gain by the unconditional informativeness of ϕ_t.

horizontal-line feature (ϕ_t). In doing so, the agent forms a view about the values of the unobserved pixels—illustrating how extrapolation can serve as a basis for *prediction*.

3.2 Mr. Stitch and Extrapolation

Let us now examine how Mr. Stitch chooses between suits using extrapolation. We assume that suit characteristics take integer values between 1 and 10 ($\Omega = \{1, 2, ..., 10\}$). We also assume that Mr. Stitch has working memory capacity of two ($L = 2$) and no intermediate concepts—so that he is incapable of choosing using deduction alone.

	Suit 1	**Suit 2**
Quality	10	2
Sizing	3	8

Figure 7: Mr. Stitch's Problem

We consider two distinct ways in which Mr. Stitch might extrapolate.

Selective attention to attributes

One possibility is that Mr. Stitch might focus on one attribute of the problem, such as suit quality. That is, he attends to the pixels "10" and "2." This is analogous to the selective attention to attributes such as price or quality in the work of BGS et al.

If Mr. Stitch stores the quality data in working memory and considers whether "Suit 1 is better" (feature f_1), the fit is:

$$\Psi(f_1; W_t) = 1 - \frac{\psi(W_t \cup \{f_1\}) - \psi(W_t)}{\psi(\{f_1\})}$$

$$= 1 - \frac{(\log 10^4 - \log 99) - (\log 10^4 - \log 10^2)}{\log 10^4 - \log(1/2)(10^4 + 670)} \approx 0.98.$$

This is because $|\mathcal{P}| = 10^4$ (each pixel can take ten values, so there are 10^4 possible pictures); $|\cap_{f \in W_t} f| = 10^2$ (only two of the four pixels are pinned down by W_t); $|\cap_{f \in W_t \cup \{f_1\}} f| = 99$ (of the 10^2 possible pictures given W_t, only one has the property that Suit 2 is strictly preferred to Suit 1); and $|\cap_{f\in\{f_1\}}| = (1/2)(10^4) + (1/2)(670) = 5335$ (this is the number of pictures where Suit 1 is weakly preferred to Suit 2). Intuitively, fit is close to one because, given Suit 1's large quality advantage, there is only a single case where Suit 2 is strictly preferred.

Thus, provided Mr. Stitch's fit threshold $\alpha \leq 0.98$, he is able to extrapolate to "Suit 1 is better" based on the quality data.[12]

Can Mr. Stitch extrapolate to "Suit 2 is better" based on the same data? A similar analysis shows that the fit of "Suit 2 is better" with the quality data is -4.58. Intuitively, fit is negative because the quality data constitutes evidence *against*—rather than for—the proposition that "Suit 2 is better." Thus, this extrapolation is not possible.

More generally, the fit of the proposition "Suit i is better" depends upon the quality of Suit i relative to Suit j (see Table 1). As Suit i's relative quality falls, the fit of "Suit i is better" falls. Thus, for a given extrapolation threshold α, Suit i's relative quality must exceed a certain threshold for the extrapolation "Suit i is better" to be feasible.[13]

Suit i's relative quality	Fit of "i is better"	Fit of "j is better"
9	1	-6.33
8	0.98	-4.58
7	0.95	-3.48
6	0.90	-2.66
5	0.83	-2.02
4	0.74	-1.48
3	0.62	-1.03
2	0.48	-0.62
1	0.29	-0.27
0	0.05	0.05

Table 1: Fit as a function of Suit i's relative quality

Table 1 also shows that, as Suit i's relative quality falls, the fit of "Suit j is better" rises.

[12] Here, Mr. Stitch selects an option based on an attribute. In a slightly richer version of the problem, we can also consider rejection of an option based on an attribute. This simply requires that Mr. Stitch has concepts of the form "Suits $i \in S$ are strictly worse than Suits $i \notin S$." Tversky (1972) argue that sequential elimination is a common practice in complex choice settings.

[13] Notice that when the quality difference is zero, both f_1 and f_2 have positive fit (0.05). The reason is that f_i corresponds to a *weak* preference for Suit i, not a strong preference. Consequently, the intersection of f_1 and f_2 is non-empty.

Thus, if Suit i's relative quality is sufficiently low, it is possible to extrapolate to "Suit j is better."

Mr. Stitch might alternatively focus on the second attribute (sizing), where Suit 2 has an advantage of five. In this case, the fit of "Suit 2 is better" is 0.83—which we can see from Table 1, since Mr. Stitch's problem is symmetric with respect to quality and sizing.

Thus, if Mr. Stitch has a high extrapolation threshold ($0.83 < \alpha < 0.98$), he can only extrapolate to "Suit 1 is better." However, if Mr. Stitch's extrapolation threshold is lower ($\alpha < 0.83$), he can extrapolate to either "Suit 1 is better" or "Suit 2 is better" depending upon whether he focuses on quality or sizing.

When Mr. Stitch has a low extrapolation threshold, then, there is some ambiguity to what he will choose (it depends upon what he focuses on). This aligns with an experimental literature documenting a random element in people's choices. For instance, Agranov and Ortoleva (2017) presented subjects with the same choice questions repeated multiple times and found that a large majority of subjects exhibited stochastic choice, selecting different options across repetitions. A follow-up survey by Agranov and Ortoleva (2022) documented that randomization appears across a wide range of domains, including objective lotteries, ambiguous gambles, social preferences, and time preferences.

Satisficing

Rather than focusing on one attribute, Mr. Stitch might focus on one option. For instance, he might focus on Suit 1 (pixels "10" and "3"). In this case, the fit of "Suit 1 is better" is:

$$\Psi(f_1; W_t) = 1 - \frac{\psi(W_t \cup \{f_1\}) - \psi(W_t)}{\psi(\{f_1\})}$$

$$= 1 - \frac{(\log 10^4 - \log 72) - (\log 10^4 - \log 10^2)}{\log 10^4 - \log(1/2)(10^4 + 670)} \approx 0.48.$$

The values of $\psi(W_t)$ and $\psi(\{f_1\})$ are the same as before. However, $\psi(W_t \cup \{f_1\}) = \log 10^4 - \log 72$, where 72 is the number of Suit 2 quality-sizing combinations for which Suit 1 is weakly preferred.

More generally, as shown in Table 2, the fit of "Suit i is better" is increasing in the value of Suit i (i.e. Suit i's quality-plus-sizing). Thus, for a given extrapolation threshold α, Suit i's value must exceed a certain level for the extrapolation "Suit i is better" to be feasible.

Value of i	Fit of "i is better"	Fit of "j is better"	Value of i	Fit of "i is better"	Fit of "j is better"
20	1	-6.33	10	-0.27	0.29
19	0.98	-4.58	9	-0.62	0.48
18	0.95	-3.48	8	-1.03	0.62
17	0.90	-2.66	7	-1.48	0.74
16	0.83	-2.02	6	-2.02	0.83
15	0.74	-1.48	5	-2.66	0.90
14	0.62	-1.03	4	-3.48	0.95
13	0.48	-0.62	3	-4.58	0.98
12	0.29	-0.27	2	-6.33	1
11	0.05	0.05			

Table 2: Fit as a function of Suit i's value

This type of extrapolation can be thought of as *satisficing* in the sense of Simon (1955). Mr. Stitch is willing to choose Suit i only if its value exceeds a level where it is "good enough." Intuitively, if Suit i is sufficiently good, Mr. Stitch can extrapolate to the conclusion "Suit i is better."

Table 2 also shows that, as Suit i's value falls, the fit of "Suit j is better" rises. Thus, if Suit i's value is *below* a threshold, the agent is able to extrapolate to "Suit j is better."

3.3 Struggling to Choose

In order to make a decision, Mr. Stitch needs a narrative about which option is best (f_1 or f_2). Absent such a narrative, he struggles to choose.[14]

[14] The challenge of decision-making in the absence of a guiding narrative is well-supported across disciplines. The work of David Tuckett and coauthors on conviction narratives emphasizes the role of storytelling in shaping confidence, suggesting that people construct narratives to navigate uncertainty and make complex choices with conviction (see Johnson et al. (2023)). Without such narratives, they argue, decision-making becomes fraught with doubt and hesitation. Shafir et al. (1993)'s notion of "reason-based choice" similarly highlights how a lack of coherent rationale complicates choice, as individuals struggle when they cannot find a unifying reason or story. Donald Davidson's theory of action is foundational here, proposing that intentional action depends on having reasons that make sense within a larger interpretive framework (see Davidson (1963)). Cognitive load theory (Sweller (1988)) further suggests that without an overarching narrative to integrate disparate information, cognitive burden increases, height-

Several factors affect whether it is difficult to construct such a narrative. One factor is working memory capacity (L). For example, Mr. Stitch can deduce which suit is better if he has memory capacity of four, but he may not be able to deduce—or even extrapolate—which suit is better with memory capacity of two.

A second factor is Mr. Stitch's conceptualization (C). A richer conceptualization aids deduction and extrapolation. For example, we saw earlier that with working memory capacity of two, Mr. Stitch cannot deduce which suit is better if he has no intermediate concepts; however, he can do so with the right ones.

We now discuss two further factors: the complexity of the decision problem and the similarity of the options.

Choice Complexity

The empirical literature on choice overload suggests that people find it difficult to choose when facing complex problems. For instance, Iyengar and Lepper (2000)'s classic "jam study" shows that consumers purchase more jam when presented with fewer options (six) than when offered a larger assortment (twenty-four).

Consistent with this evidence, it is harder for Mr. Stitch to choose when his problem is more "complex"—meaning that suits have more attributes or there are more suit options. Figure 8 gives one example, in which suits have a third attribute (price).

	Suit 1	**Suit 2**
Quality	10	2
Sizing	3	8
Color	4	6

Figure 8

Consider what happens as we increase the number of attributes, under the assumption that Suit i's value is the sum of its attribute values. Recall that if Mr. Stitch focuses on

ening the risk of analysis paralysis (Baumeister et al. (1998)). Additionally, Ricoeur (1992) and McAdams (1993) underscore the role of narrative in constructing a cohesive self, while Chang (2017)'s work on "hard choices" illustrates the difficulty of making decisions in the absence of clear evaluative criteria.

quality (pixels "10" and "2") and there are two attributes, the fit of "Suit 1 is better" is 0.98. As shown in Table 3, fit steadily decreases as the number of attributes rises—falling to 0.75 with five attributes. Thus, Mr. Stitch might be able to choose Suit 1 based on selective attention to quality when there are two attributes, but not when the problem is more complex.

# of attributes	# of suits	Fit
2	2	0.98
3	2	0.88
4	2	0.81
5	2	0.75

Table 3: Fit of "Suit 1 is better" when data on quality is in WM

A similar exercise shows how fit changes with the number of suits (assuming two attributes). Recall that when there are two suits and Mr. Stitch focuses on Suit 1 (pixels "10" and "3"), the fit of "Suit 1 is better" is 0.48. Table 4 shows that fit falls as the number of suits increases—declining to 0.26 with five suits.

# of attributes	# of suits	Fit
2	2	0.48
2	3	0.36
2	4	0.30
2	5	0.26

Table 4: Fit of "Suit 1 is better" when data on Suit 1 is in WM

Choice Similarity

An empirical literature also suggests that people find it hard to choose when options are similar. For instance, when options are close in attractiveness, people engage in costly deferral even though doing so is irrational (e.g. Tversky and Shafir (1992) and Dhar (1997)).[15]

Consistent with these findings, Mr. Stitch finds it hard to choose when options are similar. To illustrate, consider the two choice problems in Figure 9. Panel (a) shows one

[15] In Tversky and Shafir (1992)'s experiment, a rational agent might be more inclined to defer when the overall attractiveness of options declines, but not when options become closer in relative attractiveness. Nonetheless, they find that relative attractiveness matters.

choice problem, and panel (b) shows a second in which Suit 1 is unchanged but Suit 2 is more attractive—and more similar to Suit 1.

	Suit 1	**Suit 2**
Quality	7	2
Sizing	5	5

(a)

	Suit 1	**Suit 2**
Quality	7	6
Sizing	5	5

(b)

Figure 9

Mr. Stitch finds it easier to choose in case (a) than case (b). To see this, suppose he has an extrapolation threshold $\alpha = 3/4$. In choice problem (a), by focusing on quality, Mr. Stitch can choose Suit 1. Moreover, this is the only conclusion he can reach, and selective attention to quality is the only route to reaching a conclusion.[16] In choice problem (b), because the quality difference shrinks to just one, Mr. Stitch can no longer choose between the suits.

Dealing with difficult choices

What happens when agents struggle to choose? One possibility is that they may go with a default—in effect, *not* making a choice. This may help explain why defaults are so influential in complex contexts, such as retirement-plan selection (see Madrian and Shea (2001) and Thaler and Benartzi (2004)).

A second possibility is that an agent may gradually lower their fit threshold α (i.e., they become more willing to extrapolate) until they reach a point where making a choice is feasible.[17] This is consistent with a large experimental literature showing that difficult choices take longer to make (e.g. Payne et al. (1993)).

[16] If Mr. Stitch focuses on quality, the fit of “Suit 1 is better” is 0.83, which exceeds his threshold. If he instead focuses on sizing, he cannot reach a conclusion, since the fit of ”Suit 1 is better” is only 0.05. Focusing on Suit 1 (value 13) yields a fit of 0.48, and focusing on Suit 2 (value 7) yields a fit of 0.74—neither of which exceeds his threshold.

[17] It is easy to show that an agent can make a choice if the fit threshold is sufficiently low.

4 Perceptual Rivalry

A striking property of extrapolation is that the same evidence can be seen in multiple—potentially contradictory—ways. When an agent extrapolates beyond the data, the interpretation they arrive at may depend on the path their thinking takes—which features come to mind first, and which narratives take hold early. In this section, we explore this possibility and its implications.

To fix ideas, consider a voter, Mr. Judge, who is evaluating a politician about whom he has mixed evidence—some good and some bad. We will show that, depending on how Mr. Judge processes this evidence, he may come to see the politician as a "good egg" or as a "bad egg." Both interpretations are somewhat consistent with the data, but once one takes hold, it can be difficult to dislodge.

The idea that the same information can give rise to multiple, competing interpretations is well established in cognitive psychology, where it is known as perceptual rivalry (see Leopold and Logothetis (1999)). A classic illustration is the rabbit–duck illusion (Figure 2): the same drawing can be seen as a rabbit or as a duck, but it is difficult to see both at once. Once one interpretation takes hold, it tends to resist displacement by the other. Other well-known examples include the Necker cube, in which the same wireframe drawing can be perceived with either face in front, and Rubin's vase, which alternates between a vase and two facing profiles.

4.1 Mr. Judge

Suppose a voter, Mr. Judge, is deciding whether to vote for a politician. The politician is described by a vector of n characteristics:

$$P = \begin{bmatrix} p_1 & p_2 & p_3 & \cdots & p_n \end{bmatrix},$$

where $p_i \in \{\text{good}, \text{bad}\}$ and $n > 4$. The politician is a "good egg" if there are at most two bad characteristics and a "bad egg" if there are at most two good characteristics. Notice that the politician cannot be both a good egg and a bad egg. Mr. Judge finds it optimal to vote for the politician if he is a good egg and for an alternative candidate if he is a bad egg.

Mr. Judge has working memory capacity of two ($L = 2$), an extrapolation threshold $\alpha = 1/2$, and only sees four characteristics of the politician—p_1, p_2, p_3, and p_4. In other words, Mr. Judge's initial knowledge set K_0 is a subset of the pixels of P. From this limited data, let us consider how Mr. Judge extrapolates to a conclusion about whether the politician is a good or bad egg. For simplicity, suppose the only concepts Mr. Judge has besides pixels are "good egg" and "bad egg."

Since Mr. Judge's working-memory capacity is two, all he can do at time 1 is evaluate "good egg" or "bad egg" against two of the politician's characteristics/pixels (the p_i's). Table 5 shows, for different values of n, the fit of "good egg" when evaluated against two good pixels or two bad pixels. The table shows that, when $n \leq 7$, the fit of "good egg" exceeds $\alpha = 1/2$ if it is evaluated against two good pixels. Thus, when $n \leq 7$, Mr. Judge can extrapolate to "good egg" by considering the politician's two good characteristics.

	Fit("good egg"|2 good)	Fit("good egg"|2 bad)
$n = 5$	0.81	-2
$n = 6$	0.65	-1.60
$n = 7$	0.53	-1.33
$n = 8$	0.45	-1.15
$n = 9$	0.38	-1.01
$n = 10$	0.33	-0.91

Table 5: Fit Thresholds for Mr. Judge

Given the symmetry of the problem, Fit("good egg"|2 good)=Fit("bad egg"|2 bad). Thus, if Mr. Judge is able to extrapolate to "good egg" off of the politician's two good characteristics, he is *also* able to extrapolate to "bad egg" off of the politician's two bad characteristics.

This demonstrates the first point we wish to make: when an agent extrapolates, there may be competing—and mutually inconsistent—interpretations they can reach, depending on where the thought process begins. The same mixed evidence can yield opposite conclusions.

Path Dependence

While we have shown that there are two views Mr. Judge can reach of the politician—good egg or bad egg—we have not yet shown path dependence in Mr. Judge's thinking. Path dependence would mean that, if Mr. Judge initially decides the politician is a good egg, it subsequently becomes hard to see him as a bad egg (or vice-versa).

To get path dependence, it turns out we need one additional modeling ingredient: we need to make an assumption about the *salience* of features. Cognitive psychologists argue that, once we have one way of seeing a picture, that interpretation is highly salient (see Leopold and Logothetis (1999)). It is hard, for instance, to keep "rabbit" out of mind once this is one's interpretation of the rabbit-duck picture. We formalize this idea below, and show that it generates path dependence.

To illustrate the idea, suppose that "good egg" or "bad egg," once learned, are salient. That is, it is hard to keep these features out of working memory. Then, if Mr. Judge decides the politician is a "good egg" at time 1, this will be salient at time 2: it will be one of the features in working memory. With "good egg" in working memory, Mr. Judge is unable to extrapolate to "bad egg" at time 2: since a "good egg" cannot be a "bad egg," the fit of "bad egg" is $-\infty$. Intuitively, the politician's bad characteristics are dismissed as the small foibles of an otherwise good egg. In Section 4.2, we make this argument precise.

4.2 Salience and Path Dependence

We turn now to incorporating salience into our model. We will proceed as follows. First, we define the "power" of a feature. Then, we make an assumption about the relationship between a feature's power and its salience. Finally, we show that, under this assumption, an agent's thinking can exhibit path dependence.

Power

We define the "power" of feature f on region R of the picture based on how much feature f narrows the set of possibilities on region R.

To formalize this notion, call any nonempty subset of pixel locations $S \subseteq \{1, \ldots, M\} \times \{1, \ldots, N\}$ a *region*. We say two pictures are equivalent on region R if they are identical at every pixel in R; accordingly, R partitions $\mathcal{P}$ into a set of equivalence classes which we call the *subpictures of* $\mathcal{P}$ *on* R and denote as $\mathcal{R}$.

Given a set of features F, let $\delta_R(F) \subseteq \mathcal{R}$ be the set of subpictures of $\mathcal{P}$ on R that are consistent (on R) with every feature in F: that is, $\delta_R(F)$ consists of all $P_R \in \mathcal{R}$ where for each feature $f \in F$, the set $P_R \cap f$ is nonempty.

Definition 3. *We define the power of a set of features F on region R as follows:*

$$\Gamma_R(F) \ \equiv \ 1 - \frac{\log |\delta_R(F)|}{\log |\mathcal{R}|}.$$

Thus $\Gamma_R(F) \approx 0$ means that F rules out very little on R, while $\Gamma_R(F) = 1$ means that F pins down the pixels in R completely (there is only one feasible assignment on R).[18]

Application to Mr. Judge

Let us return to our example and calculate the power of "good egg" on the region consisting of all pixels. Let us also compare the power of "good egg" with the power of two pixels or one pixel.

Applying the definition, we see that the power of "good egg" is:

$$1 - \frac{\log(\binom{n}{2} + n + 1)}{\log(2^n)},$$

since, ex ante, there are 2^n politician types (i.e. 2^n possible pictures) and $\binom{n}{2} + n + 1$ possible types conditional on knowing he is a "good egg."

[18] One may think of the power of a set of features F as the reduction in entropy on region R relative to complete ignorance.

The power of k pixels is:

$$1 - \frac{\log(2^{n-k})}{\log(2^n)} = \frac{k}{n},$$

since there are 2^{n-k} possible politician types conditional on knowing the value of k pixels.

Table 6 compares the power of "good egg" against the power of two pixels and one pixel. Intuitively, "good egg" is more powerful than two pixels or one pixel in pinning down the politician's type when the overall number of politician characteristics is large ($n \geq 7$).

n	Power of "good egg"	Power of two pixels	Power of one pixel
5	0.20	0.40	0.20
6	0.26	0.33	0.17
7	0.31	0.29	0.14
8	0.35	0.25	0.13
9	0.39	0.22	0.11
10	0.42	0.20	0.10

Table 6: Power Calculations

Power and Salience

We make the assumption that features are more salient—more likely to enter working memory—when they are more powerful on the support of the feature being evaluated (ϕ_t). Specifically, we assume the following.

Assumption 1. *Suppose, $f \in K_t \setminus \{\phi_t\}$ is weakly more powerful on the support of ϕ_t than $F \subset K_t \setminus \{\phi_t\}$. If the agent loads every element of F into working memory (W_t), then they must also load f.*

Under Assumption 1, the agent will substitute a set of features in working memory with an equivalent but simpler "chunked" feature (e.g., "H" in place of the pixels representing "H").[19] More generally, prioritizing powerful features—as the agent does under Assumption 1—helps them economize on working memory.

[19] Note that the agent could have the unchunked features in working memory—but only if they have the simpler, chunked feature as well.

Establishing Path Dependence

We are now in a position to return to the Mr. Judge example and verify the informal argument above regarding the possibility of path dependence. We will show that, under Assumption 1, Mr. Judge exhibits path-dependent thinking when $n = 7$.

Recall that, when $n \leq 7$, Mr. Judge can extrapolate at time 1 to either "good egg" or "bad egg" (depending upon whether he focuses on the politician's good characteristics or bad characteristics). Moreover, when $n \geq 7$, "good egg" and "bad egg" are more powerful features than two pixels or one pixel on the region consisting of all pixels.

Suppose then that $n = 7$ and that Mr. Judge reaches the conclusion "good egg" at time 1. Suppose, at time 2, he then considers "bad egg" ($\phi_t =$ "bad egg"). The support of "bad egg" is all pixels of P. Thus, "good egg" has greater power than two pixels or one pixel. Per Assumption 1, Mr. Judge cannot load the two bad pixels into working memory at time 2—nor can he load one bad pixel without also loading "good egg." Consequently, it is impossible to extrapolate to "bad egg" at time 2.

By a symmetric argument, if Mr. Judge reaches the conclusion "bad egg" at time 1, "good egg" will be rejected at time 2.

This proves, by example, the proposition stated below.

Proposition 3. *Under Assumption 1, the agent's thinking may exhibit "path dependence." That is, for some (α, L, P, C, K_0), there exist features f and f' and finite thought sequences (i) $T_0, ..., T_t$ and (ii) $T'_0, ..., T'_t$ such that:*

- *Under sequence (i), $f \in K_\tau$ and $f' \notin K_\tau$ for all $\tau > t$ for any subsequent sequence,*
- *Under sequence (ii), $f' \in K_\tau$ and $f \notin K_\tau$ for all $\tau > t$ for any subsequent sequence.*

Primacy effects

One immediate implication of path dependence is that the order in which information is revealed matters. Consider a variant of the Mr. Judge example in which the politician's two good characteristics are revealed first, followed by the bad ones, or vice-versa (i.e.

the pixels are revealed at different times rather than all at once). Whichever information is revealed first is likely to determine which narrative locks in—"good egg" or "bad egg."

This connects to a well-known finding in psychology: primacy effects. In a classic study, Asch (1946) presented subjects with identical lists of personality traits in different orders. One group read that a person was "intelligent, industrious, impulsive, critical, stubborn, envious," while another read the same traits in the reverse order. Asch found that subjects who saw the positive traits first formed more favorable impressions—even though the informational content was identical. Anderson (1965) documented a similar pattern, finding that a trait's influence on impressions decreased with its ordinal position.

4.3 Inconsistencies and Deletion of Features

In the Mr. Judge example, something odd can happen: both "good egg" and "bad egg" can be in knowledge, even though they are inconsistent. For example, if $n = 5$, Mr. Judge can extrapolate to "good egg" at time 1 off of the good pixels. He can then extrapolate to "bad egg" at time 2 off of the bad pixels (since "good egg" is less powerful than two bad pixels when $n = 5$).

This illustrates a more general point: the possibility of *inconsistencies* in the agent's thinking. We say that there is an inconsistency in the agent's thinking if there is no picture consistent with all of the features in K_t: $\cap_{f \in K_t} f = \emptyset$.

It seems natural that the agent would try to eliminate such inconsistencies. In this section, we allow the agent to resolve inconsistencies by deleting features as well as adding them. We first augment the model to allow for deletion, then examine the extent to which deletion resolves inconsistency, and finally show that deletion can give rise to *cycles* in the agent's thinking.

Deletion

Let us suppose that the agent not only adds features to knowledge when fit exceeds a threshold; they also delete features from knowledge when fit is below a threshold.

Specifically, we divide features into three categories: (1) "axiomatic" knowledge (K_0), (2) learned features ($K_t - K_0$), and (3) features outside of knowledge. At time t, the agent may evaluate a concept ϕ_t outside of knowledge or one that has been learned. If the feature is outside of knowledge ($\phi_t \notin K_t$), the agent adds the feature to knowledge if fit exceeds α: $\Psi(\phi_t; W_t) \geq \alpha$. If the feature is a learned feature ($\phi_t \in K_t - K_0$), the agent deletes the feature from knowledge if fit is below β: $\Psi(\phi_t; W_t) \leq \beta$. We assume that $\alpha > \beta$ and that β may be positive or negative. In contrast to learned features, axiomatic knowledge cannot be deleted.

Eliminating Inconsistencies

Deletion makes it possible for the agent to eliminate some inconsistencies from their thinking, but not necessarily all.

We say that there is an "inconsistency of order k" if there is a set of k features in knowledge, $F \subseteq K_t$, that are inconsistent ($\cap_{f \in F} f = \emptyset$) but that removing any feature from F eliminates inconsistency (for all $f' \in F$, $\cap_{f \in F - \{f'\}} f \neq \emptyset$). If K_t is inconsistent and has n elements, it must have an inconsistency of order k for some $k \leq n$.

To illustrate, the pair of features "good egg" and "bad egg" are an inconsistency of order two: deleting either "good egg" or "bad egg" eliminates inconsistency. The following, by contrast, is an order-three inconsistency: (i) $x > 0$, (ii) $y > 0$, and (iii) $x + y = 0$. Any pair of statements are consistent, but the three statements taken together are inconsistent.

Proposition 4 shows that the ability to delete features makes it possible to resolve inconsistencies of order $k \leq L + 1$—however, not necessarily inconsistencies of higher order.

Proposition 4. *If an inconsistency of order $k \leq L + 1$ exists at time t, there is a thought T_t that eliminates the inconsistency at time $t + 1$. However, for any $L \geq 1$ and any $k > L + 1$, for some $(\alpha, \beta, P, C, K_0)$ and finite thought sequence $T_0, ..., T_{t-1}$, an inconsistency of order k exists at time t that no subsequent thought sequence can eliminate.*

Intuitively, the agent's working-memory capacity allows them to detect and eliminate some inconsistencies. If there is an inconsistency F of order $k \leq L+1$, the agent can evaluate one of the features $f \in F$ against the remaining features in F. Given the inconsistency, the fit of f with $F - \{f\}$ is $-\infty$, which means f will be deleted from knowledge—resolving the inconsistency. However, the agent cannot easily detect inconsistencies of higher order and these inconsistencies may remain part of their thinking.

Cycling

The ability to delete features from knowledge can lead to cycling in thinking.

To illustrate, consider what happens in the Mr. Judge example when features can be deleted. When $n = 7$, there are two stable interpretations of the politician ("good egg" and "bad egg") if $\beta < -1.33$. However, Mr. Judge cycles between "good egg" and "bad egg" when $\beta > -1.33$.

To see why cycling occurs when $\beta > -1.33$, suppose Mr. Judge extrapolates to "good egg" at time 1 by focusing on the politician's two good characteristics. At time 2, he re-evaluates "good egg," only this time he focuses on the politician's two bad characteristics. Table 5 shows that the fit of "good egg" with the two bad characteristics is -1.33, so Mr. Judge deletes "good egg" (given that $\beta > -1.33$). Having deleted "good egg," at time 3 he can extrapolate to "bad egg" by focusing on the bad characteristics. By symmetry, he can then delete "bad egg" by focusing on the good characteristics and extrapolate back to "good egg." Thus, Mr. Judge oscillates between rival interpretations, never settling on one.

This proves, by example, the proposition stated below.

Proposition 5. *Under Assumption 1, cycling is possible. That is, for some $(\alpha, \beta, L, P, C, K_0)$, there exist features f and f' and a working memory sequence $T_0, T_1, ...$ such that:*

- $f \in K_{t_1}$ *and* $f' \notin K_{t_1}$,
- $f' \in K_{t_2}$ *and* $f \notin K_{t_2}$,
- $f \in K_{t_3}$ *and* $f' \notin K_{t_3}$,

for some t_1, t_2, t_3 *with* $t_1 < t_2 < t_3$.

Indeed, cycling between rival interpretations is one of the most robust findings in the literature on perceptual rivalry. When subjects view ambiguous figures such as the Necker cube or the rabbit–duck illusion for an extended period, they typically report that perception alternates between the two interpretations (Leopold and Logothetis (1999)). Analogous cycling has been documented in the context of judgment and decision-making. For instance, Rothman et al. (2017) find that individuals frequently alternate between opposing positions rather than maintaining a stable view—a pattern they term "vacillation."

Cycling and choice

One might ask what our framework says about how people make choices when their thinking can cycle. While other views are possible, the position we take is that if the agent has a choice to make at time t and they have a feature f_a in knowledge at time t (where f_a denotes "action a is optimal"), they choose action a. The agent is convinced at time t that action a is optimal—so they go with it. The agent might later change their mind and decide that action a is not optimal. This, in our view, would be associated with a feeling of regret—in the spirit of Loomes and Sugden (1982).

5 The Role of Narratives

In this section, we examine the role narratives play in the agent's reasoning process more closely.

5.1 Bottom-up and Top-down Thinking

Narratives, we argue, play a dual role. On the one hand, they may be the end product of reasoning—the big-picture understanding that the agent is working toward. On the other hand, narratives are also a powerful input into further reasoning. Once an agent has arrived at a big-picture understanding, they can use it to draw conclusions about parts of the picture that they have not yet examined or that they could not make sense of previously.

A simple example illustrates both directions. Consider a Wheel of Fortune contestant presented with the puzzle “_OOD _OR_ING.” Starting from the visible letters, the contestant may extrapolate to a big-picture interpretation—“GOOD MORNING.” The contestant may then use this narrative to supply the missing letters—G, M, and N—filling in details that the local information alone could not reveal.

These two directions of reasoning connect naturally to a distinction that is central in cognitive psychology: between “bottom-up” and “top-down” thinking (see Marr (2010); Neisser (2014)). Bottom-up thinking involves building interpretations incrementally from local concepts toward global ones. In our framework, this corresponds to constructing narratives—assembling pixel-level information into big-picture statements. Top-down thinking reverses this direction: higher-level representations are used to interpret or infer lower-level information. In our framework, this corresponds to deploying narratives—using an existing big-picture understanding to learn about smaller-picture features of the picture.

A variety of experiments demonstrate the role of top-down processing (see Gilbert and Li (2013)). For instance, objects are identified more easily and more rapidly when they appear in consistent contexts—such as a toaster in a kitchen rather than a toaster on a beach (see Biederman et al. (1982); Oliva and Torralba (2007); Palmer (1975)). These effects arise even when the local visual information defining the object is unchanged.

We formally define bottom-up and top-down thinking in our model as follows.

Definition 4. *Suppose the agent learns concept f at time t. Let $s(f)$ be the support of concept f. We say that:*

- *The extrapolation involves “bottom-up thinking” if, for all concepts $g \in W_t$, $s(g) \subset s(f)$.*
- *The extrapolation involves “top-down thinking” if there is a concept $g \in W_t$ with $s(f) \subset s(g)$, and the agent does not learn f if g is omitted from W_t.*

The following proposition shows that both forms of thinking are essential: there are concepts that can only be learned through bottom-up reasoning, and others that can only be learned through top-down reasoning.

Proposition 6. *There exists picture P, conceptualization C, parameter L, and concepts f and f' such that:*

(a) In any sequence of thoughts where f is extrapolated, the extrapolation of f involves bottom-up thinking.

(b) In any sequence of thoughts where f' is extrapolated, the extrapolation of f' involves top-down thinking.

Part (a) is relatively straightforward. Since the agent's initial knowledge consists entirely of pixels, the agent's first reasoning steps must involve moving from the small picture toward a more global understanding. Some concepts can only be reached through this kind of bottom-up process.

What is less immediate is why some extrapolations require top-down thinking. Starting from pixels, using a larger concept to infer a smaller one may appear indirect. The key reason is that bottom-up paths to a target concept may be unavailable. In particular, there may be no simple sequence of intermediate concepts that allows the agent to build up to a desired concept f while respecting the working-memory constraint. In such cases, the agent may instead extrapolate to a related big-picture concept f' that is easier to construct bottom-up and that summarizes the picture in a compact way. Once f' is known, the agent can then engage in top-down thinking, using this narrative to infer finer-grained concepts that would otherwise be inaccessible.[20]

Seen this way, top-down thinking is not a substitute for bottom-up reasoning but a complement to it. Narratives provide a compact, memory-efficient way of representing

[20] Even when the agent is purely deductive, there are concepts that can only be learned using top-down thinking. The bottom-up path to deducing a feature f may be blocked if there are no appropriate intermediate concepts. Even if the direct path is blocked, the agent may be able to deduce a big-picture feature f' in a bottom-up fashion and then use f' to deduce f. For further details, see the proof of Proposition 6.

the picture, and using them as part of the thought process can help the agent learn additional things.

5.2 Supplying Narratives

An important way of influencing others is suggesting narratives to them. This is a common practice in politics and marketing, among other domains. Political campaigns routinely craft interpretive frames that tell voters how to understand events and candidates (Chong and Druckman, 2007; Entman, 1993): a tax increase becomes either "investing in our future" or "government overreach," depending on who supplies the narrative. Similarly, brands use storytelling to shape how consumers perceive products and form attachments (Escalas, 2004; Green and Brock, 2000)—tap water is nearly free, yet a narrative about purity, nature, and wellness can persuade consumers to pay a thousandfold markup for bottled water— only a modestly different product.

Formally, what we have in mind is that a persuader suggests a narrative ϕ_t to a target agent along with some supportive facts (W_t). Thus, supplying a narrative is akin to determining what the agent thinks at time t ($T_t = (\phi_t, W_t)$). Returning to the Mr. Judge example, the politician might want to suggest the "good egg" narrative to Mr. Judge and give the two good pixels as supporting evidence. By contrast, the opposing politician might want to suggest the "bad egg" narrative and give the two bad pixels as evidence.[21] If Mr. Judge's thinking exhibits path dependence, whichever politician manages to supply their narrative first captures Mr. Judge's vote.

This illustrates a general point: controlling the narrative—and, in particular, being the first to establish one—is widely seen as crucial in politics and beyond. Politicians invest heavily in framing the narratives around their campaigns and around controversies,

[21] Observe that this type of persuasion differs from the canonical economic model of persuasion, in which a principal influences an agent only by choosing what information to disclose (e.g. Kamenica and Gentzkow (2011)). In disclosure-based models, the mapping from signals to beliefs is fixed: the principal maximizes by choosing an information structure given the agent's fixed inference rule. Supplying a narrative, by contrast, can change the inference rule itself, by shifting which high-level structures the agent uses to interpret signals. In short, classical persuasion changes what the agent sees; narrative supply can change how the agent interprets what they see.

because the initial frame shapes the lens through which voters interpret all subsequent information (Entman, 1993; Lakoff, 2014). This dynamic helps explain why political actors sometimes go so far as to release damaging information about themselves preemptively, a tactic known as “stealing thunder” (Arpan and Roskos-Ewoldsen, 2005). The goal is not to harm one’s own standing but to deny opponents the chance to break the story and thereby set its interpretive frame. In short, the race to supply a narrative first is fierce precisely because early narratives can lock in an interpretive frame that proves difficult to dislodge.

Simple Narratives

All politicians understand the importance of keeping narratives simple. Take Ronald Reagan’s famous line “Government is not the solution to our problem, government is the problem” or Franklin Roosevelt’s pronouncement “The only thing to fear is fear itself,” which resonate even today. As the political consultant Frank Luntz puts it: “the most memorable political language is rarely longer than a sentence.”[22]

In terms of the model, we can think of a simple narrative as one that is a concept for the agent (or the set of agents) the persuader is targeting. Supplying a narrative is only persuasive if the agent can conceptualize it—and therefore think about it.

One corollary is that simple, *false* narratives can be persuasive over complex, *true* ones. Suppose, for instance, Mr. Judge can conceptualize “good egg” but is unable to conceptualize “bad egg”; and suppose the politician is a bad egg—not a good egg. Mr. Judge will be convinced by the “good egg” narrative—even though it is false—because it is simple.[23]

The Connection to Anchoring Effects

Our framework also offers a natural way to think about anchoring effects—the well-documented tendency for an initial value to disproportionately influence subsequent

[22] Luntz (2007), p. 7.

[23] In line with this idea, Jonathan Swift famously wrote: “Falsehood flies, and the Truth comes limping after it.” A modern version of this adage is Brandolini’s (2013) law: “the amount of energy needed to refute bullshit is an order of magnitude bigger than to produce it.”

judgments (Tversky and Kahneman, 1974). In the classic demonstration, subjects who first considered whether the percentage of African nations in the United Nations was above or below a randomly generated number subsequently gave estimates biased toward that number. Anchoring effects are remarkably robust: they arise across a wide range of domains, including purchase decisions, legal sentencing, real-estate appraisals, and negotiations (Furnham and Boo, 2011).

In our framework, we can think of many anchoring effects as instances of narrative supply. The anchor provides a starting narrative that shapes how the agent interprets the details of the picture—and the decision the agent ultimately makes.

6 Discussion

Here, we briefly discuss two natural ways of extending the model.

6.1 Probabilistic Features

Throughout the paper, we assumed that the features of the picture the agent learns are *deterministic*. However, there are many settings of interest where it makes sense to think about features that are *probabilistic*.

To illustrate, suppose P represents a sequence of n coin flips:

$$P = \begin{bmatrix} p_1 & p_2 & p_3 & \cdots & p_n \end{bmatrix},$$

where $p_i \in \{\text{heads}, \text{tails}\}$; and suppose the agent observes the first $n-1$ flips.

There are a variety of deterministic features the agent might learn about P. For instance, they might learn that m of the first $n-1$ flips are heads. However, these deterministic features are not particularly useful for predicting the nth flip. Really, what the agent would like to do is extrapolate from the first $n-1$ flips to a probabilistic feature of the form: "coin flips are equally likely to come up heads or tails, and flips are conditionally independent." Let us call this the "fair coin" feature.

It is easy to extend our framework so that the agent can extrapolate to probabilistic features as well as deterministic ones. Moreover, doing so captures the classic distinction between *risk* and *Knightian uncertainty*.

If the only features of P that the agent knows are the pixels (i.e. the first $n-1$ flips), they are in a state of *uncertainty* about the nth flip. The agent has no way of thinking about whether it will come up heads or tails—no way of even assigning odds. On the other hand, if the agent has learned the "fair coin" feature, the problem becomes one involving *risk*, in which the agent has a clear prior. In this sense, the distinction between risk and uncertainty is not a primitive of the environment but an emergent property of the agent's reasoning: it depends on whether the agent has managed to extrapolate from the data to an appropriate probabilistic feature.

While we do not explore the implications of probabilistic features further here, we believe they offer a promising avenue for understanding behavior under uncertainty—including well-known puzzles such as the Ellsberg paradox.

6.2 From Concepts to Codes

Throughout the paper, we assumed that the agent has a set of concepts C, that they are able to think about features in that set, and that they are unable to think about features outside that set. An alternative approach we might take is to think of the agent as *encoding* features. The idea that cognition operates over internal codes has deep roots across several disciplines. In neuroscience, a large literature studies how populations of neurons encode stimuli—with sparse, distributed codes emerging as a leading account of how sensory systems efficiently represent high-dimensional environments (Barlow, 2001; Olshausen and Field, 1996). In AI, a major theme of recent research is learning disentangled representations—internal codes in which distinct dimensions correspond to distinct generative factors of the data (Bengio et al., 2013).

Specifically, suppose the agent has a code $C : B \to \mathcal{F}$ that maps the set B of finite-length binary strings to features $\mathcal{F}$. Think of the agent, then, as storing a set of binary strings $b_1, ..., b_n$ in working memory—each representing some feature—subject to a constraint on total length:

$$\sum_{i=1}^{n} length(b_i) \leq L,$$

where $length(b_i)$ denotes the length of string b_i and L denotes the agent's working memory capacity. Under this formulation, features—rather than taking up a fixed amount of working-memory space (normalized to one) or an infinite amount of space—take up more or less space depending upon the length of their associated binary strings.

Assuming there is some regularity to the underlying picture-space, it can be advantageous to the agent to use a "compositional code" to represent features, where the codewords for features are built out of reusable parts. To illustrate, features f_1, f_2, and f_3 might be represented by the following strings:

$$f_1 : \underbrace{11}_{\mathit{square}} \quad \underbrace{000}_{\mathit{position}\ 1} \quad \underbrace{10}_{\mathit{size}\ 1}, \qquad f_2 : \underbrace{11}_{\mathit{square}} \quad \underbrace{111}_{\mathit{position}\ 2} \quad \underbrace{01}_{\mathit{size}\ 2}, \qquad f_3 : \underbrace{00}_{\mathit{circle}} \quad \underbrace{000}_{\mathit{position}\ 1} \quad \underbrace{01}_{\mathit{size}\ 2}.$$

A compositional code makes sense of why an agent might think not just in terms of features (i.e. a particular square of a particular size in a particular location) but in terms of objects (squares) and operators (position, size, rotation). There is a natural analog to the concept of object-oriented programming and, more broadly, to modular and compositional design principles in computer science, where complex systems are built by combining and reusing a small set of well-defined components (Gamma et al., 1994). In cognitive science, this idea connects to the language-of-thought hypothesis, which holds that mental representations have a compositional structure—built from reusable primitives combined according to systematic rules—that explains the productivity and systematicity of human thought (Fodor, 1975; Fodor and Pylyshyn, 1988).

Moving from concepts to codes is useful because it formalizes what it means for an agent to have a *model*. Suppose that, in the agent's code, there are two binary strings, b_1 and b_2, that both map to feature f. We can think of b_1 and b_2 as two distinct models of f.

To illustrate, consider Figure 10. The feature depicted could be represented by a binary string b_1 with associated meaning "a square, rotated 45 degrees to the right" or a binary string b_2 with associated meaning "a square, rotated 45 degrees to the *left*." These strings describe the same data—but they do so *differently*.

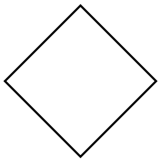


Figure 10: A square rotated 45 degrees left, or 45 degrees right.

Consider a second example. The string of numbers 3, 5, 7 might be described as "consecutive odd numbers" or as "consecutive *prime* numbers." Once we move out of sample, it becomes clear why the choice of model matters. Under the first model, the predicted next number is 9, while under the second model, it is 11.

Notice that, if the agent has two models to choose from—b_1 and b_2— and the former has a shorter string, there is an advantage to using the former since it conserves working memory. In this sense, the agent tends to gravitate towards models of the data that are compact and parsimonious. This closely relates to the literature on minimum description length in cognitive psychology which argues that people tend to favor explanations that are encoded simply (see Chater (1996), Feldman (2000), Chater and Vitányi (2003)).

7 Conclusion

The Gestalt psychologists' observation that seeing the parts is different from seeing the whole has become a foundational idea in cognitive psychology and neuroscience. Integrating information, and developing an understanding of what it means, is seen as one of the key challenges our brains face. Computer science and information theory are also based on the idea that processing information is not straightforward. Economics, by contrast, typically treats the information processing problem as a black box.

This paper builds a framework that helps connect economics with these other disciplines' perspective on information processing. We assume that agents have unlimited ability to store information—unlimited hard-drive space, so to speak. However, they have limited working memory (akin to RAM). Working memory is the place where agents integrate information and develop a big-picture understanding of what it means. We show that limited working memory bounds agents' ability to draw conclusions.

We also explore, in this context, the use of extrapolation by agents. On the one hand, extrapolation allows agents to reach further than they can with deduction alone, potentially reaching superior conclusions. On the other hand, extrapolation introduces the possibility of drawing incorrect conclusions.

This paper leaves a number of unanswered questions. There are aspects of the reasoning process that require further exploration. How, for instance, are the contents of the agent's working memory determined? The work of BGS et al. suggests that there are bottom-up processes at work (i.e. the salience of information) while the literature on rational inattention emphasizes the role of top-down processes (i.e. people make decisions regarding what deserves attention).

Our focus in this paper has primarily been on the role of limited working memory. However, the paper also brings to light the importance of limited cognitive ability (which we can think of as the agent's clock speed at performing deductions/extrapolations). When cognitive ability is limited, it matters not only *what* the agent is able to conclude but also *how quickly* they are able to reach conclusions.

Appendix

A.1 Proofs

Proof of Proposition 1.
Part 1: for any applicable concept $\phi \in C(P)$, the agent can set $W_0 = K_0$ (the $M \times N$ pixel values that apply to P) and $\phi_0 = \phi$. Since $L \geq M \times N$, the entire set K_0 fits in working memory. Moreover, $\cap_{f \in W_0} f = \{P\}$, and $P \in \phi$ (because ϕ applies to P), so $\{P\} \subseteq \phi$. Hence ϕ is deducible in a single step. Since ϕ was arbitrary, $D_1 = C(P)$.

Part 2: if $L = 1$, then $|W_t| \leq 1$ in every period t. If $W_t = \emptyset$, then $\cap_{f \in W_t} f = \mathcal{P}$, and the only ϕ with $\mathcal{P} \subseteq \phi$ would be $\phi = \mathcal{P}$, which is not a feature (features are proper subsets of $\mathcal{P}$). If $W_t = \{g\}$ for some $g \in K_t$, then ϕ_t is deducible only if $g \subseteq \phi_t$. Since the superset relation is transitive, iterating over multiple time periods yields: any $f \in D$ satisfies $p \subseteq f$ for some pixel value $p \in K_0$. That is, every deducible concept is logically implied by a single pixel value. In particular, the agent cannot integrate information from multiple pixels: no concept whose truth depends on two or more pixel values jointly can be deduced. If the conceptualization contains no proper supersets of pixel values (other than the pixel values themselves), then $D = K_0$.

Part 3: consider the following example. P is a 1×7 picture where every pixel equals 0, with $\Omega = \{0, 1\}$. Working memory capacity is $L = 2$. The conceptualization C consists of all features with support of size 1, 2, or 3, together with the single feature $f_{\text{full}} = \{P\}$.

Part 3(i): We show that f_{full} (support size 7) is applicable but not deducible, so $D \subset C(P)$.

Since f_{full} is the only concept with support size greater than 3, every feature available to load into working memory has support of size at most 3. At each step, the agent loads two such features, which jointly constrain at most $3 + 3 = 6$ pixels. Since $f_{\text{full}} = \{P\}$ requires all 7 pixels to equal 0, and the intersection of any two loadable features leaves at least one pixel unconstrained, f_{full} cannot be deduced regardless of the number of intermediate steps.

Moreover, not all applicable concepts are in K_0 (for instance, the feature $\{Q : q_1 = q_2 = 0\}$ applies to P, is a concept, and is not a pixel value), so $K_0 \subset D$.

Part 3(ii): We show that some concepts require multiple steps: $D_1 \subset D_2$.

Consider the concept $g = \{Q : q_1 = q_2 = q_3 = 0\}$, which has support size 3. This concept is not deducible in a single step, because loading two pixel values (e.g. p_1 and p_2) constrains only two pixels, and $\{Q : q_1 = q_2 = 0\} \not\subseteq g$ (since q_3 is unconstrained). However, g is deducible in two steps:

1. Load p_1 and p_2 into working memory; deduce $h_{12} = \{Q : q_1 = q_2 = 0\}$.
2. Load h_{12} and p_3 into working memory; deduce g (since $h_{12} \cap p_3 = \{Q : q_1 = q_2 = q_3 = 0\} \subseteq g$).

Thus $g \in D_2 \setminus D_1$. □

Proof of Proposition 2.

1. Any thought sequence that is feasible under C is also feasible under C'; hence $k(f, C') \leq k(f, C)$ and $D(C') \cap C \supseteq D(C)$.

   Suppose $L = 3$. Consider a 5×1 picture P consisting of white pixels $p_1, \ldots, p_5$. Define the following features. f_{12} specifies that p_1 and p_2 are white. f_{34} specifies that p_3 and p_4 are white. f_{123} specifies that p_1, p_2, p_3 are white. $f_{12345} = \{P\}$ specifies that all pixels are white. Let $C = \{f_{12}, f_{34}, f_{12345}\}$ and $C' = \{f_{12}, f_{34}, f_{123}, f_{12345}\}$. Note that $k(f_{12345}, C') = 2$: given C', the agent can deduce f_{123} from pixels p_1, p_2, p_3 in one step, then deduce f_{12345} from f_{123}, p_4, and p_5 in an additional step. On the other hand, $k(f_{12345}, C) = 3$: given C, to deduce f_{12345} the agent must load features whose intersection pins down all five pixels; but after one step the agent knows at most one of f_{12} or f_{34}, and combining either with two additional pixels covers only four of five pixels. The agent must therefore deduce both f_{12} and f_{34} (two steps) and then combine f_{12}, f_{34}, and p_5 (a third step).

2. Now suppose $L = 2$. Consider a 4×1 picture P consisting of white pixels $p_1, \ldots, p_4$. Let f_{12} specify that p_1 and p_2 are white, f_{34} specify that p_3 and p_4 are white, and $f_{1234} = \{P\}$ specify that all pixels are white. Let $C = \{f_{12}, f_{1234}\}$ and $C' = \{f_{12}, f_{34}, f_{1234}\}$. It is immediate that f_{1234} can be deduced under C' (via f_{12}, then f_{34}, then f_{1234} from f_{12} and f_{34}) but not under C (since the only non-pixel concept available as an intermediate is f_{12}, and f_{12} combined with any single pixel covers at most three of four pixels).

3. Let the pixel values in P be $p_1, \ldots, p_{MN}$. Consider features $f_1, \ldots, f_{MN}$ where $f_i = p_1 \cap \cdots \cap p_i$. Each f_i applies to P and is therefore in C. Note that given $L \geq 2$, f_{i+1} can be deduced from f_i and p_{i+1}; so by induction $f_{MN} = \{P\}$ is deducible. Then for any $g \in C(P)$, since $P \in g$ we have $\{P\} \subseteq g$; loading f_{MN} into working memory therefore suffices to deduce g. Hence $D(C) = C(P)$.

□

Proof of Proposition 3. A proof by example is provided in the main text (pages 38-38). □

Proof of Proposition 4. Suppose an inconsistency of order $k \leq L + 1$, consisting of a set of features $F = \{f_1, \ldots, f_k\}$ with $\cap_{f \in F} f = \emptyset$, exists at time t. Since $K_0 = F_{pixels}(P)$ consists of pixel values that all apply to the true picture P, the features in K_0 are mutually consistent. Therefore, at least one element of F must be a learned feature (i.e. in $K_t \setminus K_0$); relabel if necessary so that $f_k \in K_t \setminus K_0$. Consider the thought $T_t = (\phi_t, W_t)$ where $W_t = \{f_1, \ldots, f_{k-1}\}$ and $\phi_t = f_k$. Since $|W_t| = k - 1 \leq L$, this fits within the working-memory constraint. Since F is inconsistent, $\delta(W_t \cup \{\phi_t\}) = \emptyset$, so the fit of ϕ_t with W_t is $-\infty \leq \beta$; thus T_t leads to f_k being deleted, restoring the consistency of the remaining features.

Suppose $k > L + 1$. Consider a picture P consisting of $k - 1$ pixels, each taking value a, with $\Omega = \{a, b, c\}$. We will specify α later. Consider the following set F of k features:

$$f_i = \{Q : \text{pixel } i \text{ equals } a\} \quad \text{for } i \in \{1, \dots, k-1\}$$
$$f_0 = \{Q : \text{exactly one of pixels } 1, \dots, k-1 \text{ does not equal } a\};$$

and suppose the only non-pixel concept in the conceptualization is f_0. Note that F is an inconsistency of order k: the features in F are jointly inconsistent (since $f_1, \dots, f_{k-1}$ force all pixels to equal a, contradicting f_0), but removing any single feature restores consistency.

Using base-2 logarithms throughout, one can check that the fit of f_0 with working memory W_t containing any $\ell \leq L$ pixel values (all specifying "a") is

$$\Psi(f_0; W_t) = 1 - \frac{(k-1-\ell)\log_2 3 - 1 - \log_2(k-1-\ell)}{(k-1)\log_2 3 - 1 - \log_2(k-1)},$$

which is strictly increasing in ℓ. (Here $\psi(\{f_0\}) = \log_2 3^{k-1} - \log_2 2(k-1) = (k-1)\log_2 3 - 1 - \log_2(k-1)$, since there are $2(k-1)$ pictures satisfying f_0—for each of the $k-1$ pixels, there are two non-a values it could take; and similarly $\psi(W_t \cup \{f_0\}) - \psi(W_t) = (k-1-\ell)\log_2 3 - 1 - \log_2(k-1-\ell)$, since conditioning on ℓ pixels equaling a leaves $k-1-\ell$ pixels that could deviate, each in two ways.)

Pick α so that

$$\alpha < 1 - \frac{(k-1-L)\log_2 3 - 1 - \log_2(k-1-L)}{(k-1)\log_2 3 - 1 - \log_2(k-1)};$$

then by loading exactly L pixel features into memory, the agent can extrapolate to feature f_0. It follows that the thought $(\phi_0, W_0) = (f_0, \{f_1, \dots, f_L\})$ leads to the inconsistency F existing in knowledge.

Further, pick $\beta < 0$. The fit of f_0 evaluated against any $\ell > 0$ pixel values is strictly positive (since $n \log_2 3 - 1 - \log_2 n$ is strictly increasing in n for $n \geq 1$, and $k-1-\ell < k-1$ for $\ell > 0$), and the fit with $\ell = 0$ equals zero. Since $\beta < 0$, the feature f_0 can never be deleted. Moreover, the pixel features $f_1, \dots, f_{k-1}$ are axiomatic ($\in K_0$) and cannot be deleted. It follows that the inconsistency F, once it exists, can never be eliminated. □

Proof of Proposition 5. A proof by example is provided in the main text (page 41). □

Proof of Proposition 6. Consider the following example. Suppose $L = 3$ and suppose $\alpha = 1/2$. Suppose $\Omega = \{\text{white}, \text{black}\}$. Picture P is a 3×7 matrix with all white pixels. The only concepts (other than the pixel features) are:

- f^i is the feature "the top two pixels in each of columns $1, \dots, i$ are all white".
- f^P is the feature "all pixels in P are white".
- h is the feature "the seven pixels in the bottom row are all white".

We claim that h can only be learned if f^P is in working memory. To see why, notice that other than the pixels, no other concept involves any of the pixels of h. Consequently, adding any concept other than pixel values or f^P to working memory does not improve fit with h. It follows that, excluding f^P, the best possible fit with h involves adding three of the pixels from the bottom row to working memory. Evaluating h in this way, the fit is

$$1 - \frac{\psi(W_t \cup \{\phi_t\}) - \psi(W_t)}{\psi(\{\phi_t\})} = 1 - \frac{7-3}{7} = \frac{3}{7};$$

as the fit is below the threshold $\alpha = 1/2$, the concept h cannot be learned from pixels alone; and thus cannot be learned from bottom-up thinking.

Consider, on the other hand, the following multi-step thought process. In the first step, the agent loads the top two pixels from the first column into working memory, and extrapolates to f^1. In the next six steps, for each step $i \in \{2, \ldots, 7\}$, the agent loads f^{i-1} and the top two pixels from column i into working memory, and extrapolates to f^i. In the eighth step, the agent loads f^7 into working memory and extrapolates to f^P. In the ninth and final step, the agent loads f^P into working memory and extrapolates to h. We can check that each extrapolative step is successful. In particular, in the first seven steps and the ninth step, the fit of the candidate concept is exactly one. In the eighth step, the fit of f^P given $W_t = \{f^7\}$ is

$$1 - \frac{\psi(W_t \cup \{\phi_t\}) - \psi(W_t)}{\psi(\{\phi_t\})} = 1 - \frac{21-14}{21} = \frac{2}{3};$$

so fit in each step lies above the threshold $\alpha = 1/2$, and this thought process successfully leads to the extrapolation of h. Given that h is extrapolated from f^P where $s(f^P) \supset s(h)$, this extrapolation involves top-down thinking.

Finally, note that all concepts other than f^P itself involve a strict subset of the pixels of P; thus the extrapolation of f^P must involve bottom-up thinking. □

References

Abiteboul, Serge, Richard Hull, and Victor Vianu, *Foundations of databases*, Vol. 8, Addison-Wesley Reading, 1995.

Agranov, Marina and Pietro Ortoleva, "Stochastic Choice and Preferences for Randomization," *Journal of Political Economy*, 2017, *125* (1), 40–68.

_ **and** _ , "Revealed Preferences for Randomization: An Overview," *AEA Papers and Proceedings*, 2022, *112*, 426–430.

Ambuehl, Sandro, B Douglas Bernheim, and Annamaria Lusardi, "Evaluating deliberative competence: A simple method with an application to financial choice," *American Economic Review*, 2022, *112* (11), 3584–3626.

Anderson, Norman H, "Primacy effects in personality impression formation using a generalized order effect paradigm.," *Journal of personality and social psychology*, 1965, *2* (1), 1.

Arpan, Laura M and David R Roskos-Ewoldsen, "Stealing thunder: Analysis of the effects of proactive disclosure of crisis information," *Public Relations Review*, 2005, *31* (3), 425–433.

Asch, Solomon E, "Forming impressions of personality.," *The journal of abnormal and social psychology*, 1946, *41* (3), 258.

Baddeley, A D and G Hitch, "Working Memory," in G H Bower, ed., *The psychology of learning and motivation: Advances in research and theory*, Vol. 8, New York: Academic Press, 1974, pp. 47–89.

Barlow, Horace, "Redundancy Reduction Revisited," *Network: Computation in Neural Systems*, 2001, *12* (3), 241–253.

Baumeister, RF, E Bratslavsky, M Muraven, and DM Tice, "Ego depletion: is the active self a limited resource?," *Journal of Personality and Social Psychology*, 1998, *74* (5), 1252–1265.

Bénabou, Roland, Armin Falk, and Jean Tirole, "Narratives, imperatives, and moral reasoning," *NBER Working Paper 24798*, 2018.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent, "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, *35* (8), 1798–1828.

Biederman, Irving, Robert J. Mezzanotte, and Janice C. Rabinowitz, "Scene perception: Detecting and judging objects undergoing relational violations," *Cognitive Psychology*, 1982, *14* (2), 143–177.

Bordalo, Pedro, John J Conlon, Nicola Gennaioli, Spencer Y Kwon, and Andrei Shleifer, "Memory and probability," *Quarterly Journal of Economics*, 2023, *138* (1), 265–311.

_, Nicola Gennaioli, and Andrei Shleifer, "Salience theory of choice under risk," *Quarterly Journal of Economics*, 2012, *127* (3), 1243–1285.

_ , _ , and _ , "Salience and asset prices," *American Economic Review*, 2013, *103* (3), 623–628.

_ , _ , and _ , "Salience and consumer choice," *Journal of Political Economy*, 2013, *121* (5), 803–843.

_ , _ , and _ , "Competition for attention," *Review of Economic Studies*, 2016, *83* (2), 481–513.

_ , _ , and _ , "Memory, attention, and choice," *Quarterly Journal of Economics*, 2020, *135* (3), 1399–1442.

_ , _ , Giacomo Lanzani, and Andrei Shleifer, "A Cognitive Theory of Reasoning and Choice," 2024.

Chang, Ruth, "Hard choices," *Journal of the American Philosophical Association*, 2017, *3* (1), 1–21.

Chase, William G and Herbert A Simon, "Perception in chess," *Cognitive Psychology*, 1973, *4* (1), 55–81.

Chater, Nick, "Reconciling simplicity and likelihood principles in perceptual organization.," *Psychological review*, 1996, *103* (3), 566.

_ and Paul Vitányi, "Simplicity: a unifying principle in cognitive science?," *Trends in cognitive sciences*, 2003, *7* (1), 19–22.

Chong, Dennis and James N Druckman, "Framing theory," *Annu. Rev. Polit. Sci.*, 2007, *10* (1), 103–126.

Crawford, Vincent P. and Nagore Iriberri, "Level-K Auctions: Can a Nonequilibrium Model of Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions?," *Econometrica*, 2007, *75* (6), 1721–1770.

Cremer, Jacques, Luis Garicano, and Andrea Prat, "Language and the Theory of the Firm," *Quarterly Journal of Economics*, 2007, *122* (1), 373–407.

Davidson, Donald, "Actions, reasons, and causes," *Journal of Philosophy*, 1963, *60*, 685–700.

De Groot, A D, *Thought and choice in chess*, Mouton, 1965.

Dhar, Ravi, "Consumer preference for a no-choice option," *Journal of Consumer Research*, 1997, *24* (2), 215–231.

Eliaz, Kfir and Ran Spiegler, "A model of competing narratives," *American Economic Review*, 2020, *110* (12), 3786–3816.

Ellison, Glenn and Richard Holden, "A theory of rule development," *The Journal of Law, Economics, & Organization*, 2014, *30* (4), 649–682.

Enke, Benjamin and Thomas Graeber, "Cognitive uncertainty," *Quarterly Journal of Economics*, 2023, *138* (4), 2021–2067.

Entman, Robert M, "Framing: Toward clarification of a fractured paradigm," *Journal of communication*, 1993, *43* (4), 51–58.

Escalas, Jennifer Edson, "Narrative processing: Building consumer connections to brands," *Journal of consumer psychology*, 2004, *14* (1-2), 168–180.

Farah, Martha J, Kevin D Wilson, Maxwell Drain, and James N Tanaka, "What is" special" about face perception?," *Psychological review*, 1998, *105* (3), 482.

Feldman, Jacob, "Minimization of Boolean complexity in human concept learning," *Nature*, 2000, *407* (6804), 630–633.

Fodor, Jerry A., *The Language of Thought*, Harvard University Press, 1975.

Fodor, Jerry A and Zenon W Pylyshyn, "Connectionism and cognitive architecture: A critical analysis," *Cognition*, 1988, *28* (1-2), 3–71.

Furnham, Adrian and Hua Chu Boo, "A literature review of the anchoring effect," *The journal of socio-economics*, 2011, *40* (1), 35–42.

Gabaix, Xavier, "A sparsity-based model of bounded rationality," *Quarterly Journal of Economics*, 2014, *129* (4), 1661–1710.

Gamma, Erich, Richard Helm, Ralph Johnson, and John Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software*, Addison-Wesley, 1994.

Gazzaniga, Michael S., Richard B. Ivry, and George R. Mangun, "Cognitive neuroscience: the biology of the mind," 2014.

Gibbons, Robert, Marco LiCalzi, and Massimo Warglien, "What situation is this? Shared frames and collective performance," *Strategy Science*, 2021, *6* (2), 124–140.

Gilbert, Charles D and Wu Li, "Top-down influences on visual processing," *Nature Reviews Neuroscience*, 2013, *14* (5), 350–363.

Green, Melanie C and Timothy C Brock, "The role of transportation in the persuasiveness of public narratives.," *Journal of personality and social psychology*, 2000, *79* (5), 701.

Iyengar, Sheena S and Mark R Lepper, "When choice is demotivating: Can one desire too much of a good thing?," *Journal of Personality and Social Psychology*, 2000, *79* (6), 995.

Johnson, Samuel GB, Avri Bilovich, and David Tuckett, "Conviction narrative theory: A theory of choice under radical uncertainty," *Behavioral and Brain Sciences*, 2023, *46*, 1–26.

Kamenica, Emir and Matthew Gentzkow, "Bayesian persuasion," *American Economic Review*, 2011, *101* (6), 2590–2615.

Lakoff, George, *The all new don't think of an elephant!: Know your values and frame the debate*, Chelsea Green Publishing, 2014.

Leopold, David A and Nikos K Logothetis, "Multistable phenomena: changing views in perception," *Trends in cognitive sciences*, 1999, *3* (7), 254–264.

Loomes, Graham and Robert Sugden, "Regret theory: An alternative theory of rational choice under uncertainty," *The economic journal*, 1982, *92* (368), 805–824.

Luntz, Frank, *Words That Work: It's Not What You Say, It's What People Hear*, Hachette UK, 2007.

Madrian, Brigitte C and Dennis F Shea, "The power of suggestion: Inertia in 401 (k) participation and savings behavior," *Quarterly Journal of Economics*, 2001, *116* (4), 1149–1187.

Marr, David, *Vision: A computational investigation into the human representation and processing of visual information*, MIT press, 2010.

McAdams, Dan P, *The stories we live by: Personal myths and the making of the self*, William Morrow, 1993.

Miller, George A, "The magical number seven, plus or minus two: Some limits on our capacity for processing information.," *Psychological Review*, 1956, *63* (2), 81.

Mullainathan, Sendhil, "A memory-based model of bounded rationality," *Quarterly Journal of Economics*, 2002, *117* (3), 735–774.

_ , Joshua Schwartzstein, and Andrei Shleifer, "Coarse thinking and persuasion," *Quarterly Journal of Economics*, 2008, *123* (2), 577–619.

Neisser, Ulric, *Cognitive psychology: Classic edition*, Psychology press, 2014.

Oaksford, Mike and Nick Chater, *Bayesian rationality: The probabilistic approach to human reasoning*, Oxford University Press, 2007.

Oliva, Aude and Antonio Torralba, "The role of context in object recognition," *Trends in Cognitive Sciences*, 2007, *11* (12), 520–527.

Olshausen, Bruno A. and David J. Field, "Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images," *Nature*, 1996, *381* (6583), 607–609.

Oprea, Ryan, "Decisions under risk are decisions under complexity," *American Economic Review*, 2024, *114* (12), 3789–3811.

Palmer, Stephen E., "The effects of contextual scenes on the identification of objects," *Memory & Cognition*, 1975, *3* (5), 519–526.

Payne, John W, James R Bettman, and Eric J Johnson, *The Adaptive Decision Maker*, Cambridge University Press, 1993.

Ricoeur, Paul, *Oneself as another*, The University of Chicago Press, 1992.

Rothman, Naomi B, Michael G Pratt, Laura Rees, and Timothy J Vogus, "Understanding the dual nature of ambivalence: Why and when ambivalence leads to good and bad outcomes," *Academy of Management Annals*, 2017, *11* (1), 33–72.

Schwartzstein, Joshua and Adi Sunderam, "Using models to persuade," *American Economic Review*, 2021, *111* (1), 276–323.

Shafir, Eldar, Itamar Simonson, and Amos Tversky, "Reason-based choice," *Cognition*, 1993, *49* (1-2), 11–36.

Shiller, Robert J, "Narrative economics," *American Economic Review*, 2017, *107* (4), 967–1004.

Simon, Herbert A, "A behavioral model of rational choice," *Quarterly Journal of Economics*, 1955, pp. 99–118.

Sims, Christopher A, "Implications of rational inattention," *Journal of Monetary Economics*, 2003, *50* (3), 665–690.

Stahl, Dale O. and Paul W. Wilson, "Experimental evidence on players' models of other players," *Journal of Economic Behavior and Organizations*, 1994, *25* (3), 309–327.

Sweller, John, "Cognitive load during problem solving: Effects on learning," *Cognitive Science*, 1988, *12* (2), 257–285.

Tenenbaum, Joshua B, Thomas L Griffiths, and Charles Kemp, "Theory-based Bayesian models of inductive learning and reasoning," *Trends in cognitive sciences*, 2006, *10* (7), 309–318.

Thaler, Richard H and Shlomo Benartzi, "Save more tomorrow™: Using behavioral economics to increase employee saving," *Journal of Political Economy*, 2004, *112* (S1), S164–S187.

Tversky, Amos, "Elimination by aspects: A theory of choice.," *Psychological review*, 1972, *79* (4), 281.

_ **and Daniel Kahneman**, "Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty.," *science*, 1974, *185* (4157), 1124–1131.

_ **and Eldar Shafir**, "Choice under conflict: The dynamics of deferred decision," *Psychological Science*, 1992, *3* (6), 358–361.

Wegener, Ingo, *The complexity of Boolean functions*, John Wiley & Sons, Inc., 1987.

Wilson, Andrea, "Bounded memory and biases in information processing," *Econometrica*, 2014, *82* (6), 2257–2294.

Wojtowicz, Zachary, "Model Diversity and Dynamic Belief Formation," *Working paper*, 2024.

Yin, Robert K, "Looking at upside-down faces.," *Journal of experimental psychology*, 1969, *81* (1), 141.