4

OUR THEORY

De-Anthropocentrized Externalism

Our goal in this book is to point in what we think is the right direction for explaining the contents of AI systems, and to do some initial exploration of the territory in that direction. In this chapter, we'll set out two central claims (and a third peripheral claim) that structure our positive proposals. The first claim is directed primarily at work being done in the artificial intelligence literature. That claim contains a bit of bad news: much of the work being done on interpretability of artificial intelligence, we think, centres around an incorrect picture of how content is determined. But it also contains a bit of good news: work on the determination of content in philosophy provides a better *externalist* picture of how content is determined and sophisticated tools for developing that picture.

The second claim is directed primarily at philosophers. The second claim also starts with a bit of bad news: philosophers shouldn't get too triumphant about the special suitability of externalist theories of content determination for the AI content project. That's because consideration of some of the specific details about AI systems reveals that the externalist accounts that philosophers have developed contain crippling anthropocentric biases that make them unsuitable for use on nonhuman cases like AI systems.

Making AI Intelligible: Philosophical Foundations. Herman Cappelen and Joshua Dever, Oxford University Press (2021).
 DOI: 10.1093/oso/9780192894724.003.0004

But again there's a bit of good news: consideration of those specific details also helps point the way towards a deeper and more generalized understanding of externalism, one that gives a picture that can apply across human and nonhuman cases.

Putting the two claims together, one of the lessons is that there is room for highly productive interaction between philosophers and artificial intelligence researchers here. Both sides, we think, have been hampered by narrow perspectives. On one side, people have been approaching problems of AI content with an unnecessarily narrow picture of how contents might be determined. On the other side, people have been thinking about content determination using an unnecessarily narrow range of cases of content bearers. Each side has things the other side lacks; bringing everyone together opens up the potential for deeper and more productive work by everyone.

We'll see eventually that when the pieces are brought together, important methodological questions arise about how to carry out a research project using all of those pieces. Our third peripheral claim is then that a characteristically philosophical move will be useful: in order to get the right picture about how to develop a metasemantics of AI content, we need to think first about some *meta*-metasemantic questions.

First Claim: Content for AI Systems Should Be Explained Externalistically

Machine learning neural net programs are different from other programs in a way that matters for content. Consider a simple bit of Python code:

```
If num > 1:
    for i in range(2,num):
            if (num % i) == 0:
                    print(num," is not a prime number")
                    break
            else:
                    print(num," is a prime number")
else:
    print(num,"is not a prime number")
```

When the variable 'num' is set to 5009 and the program outputs '5009 is a prime number', there is a clear story about why that output means that 5009 is a prime number. The output means that 5009 is prime because of the programming/computational details about how that output is produced. The program produces the output '5009 is a prime number' because it tries dividing 5009 by all integers between 2 and 5008, and fails to find a nontrivial integer divisor of 5009.[1] Because not having an integer divisor is what it is to be prime, the computational production of the output is representing primeness.

As we've seen, the computational structure of machine learning neural networks makes it incredibly difficult to produce this kind of story grounding content in the programming/computational details. When we look at the vast array of internode connection strengths in the neural network of SmartCredit, or trace the computational path of Lucie's financial details as they

[1] This way of saying what the program is doing takes for granted that, for example, the programming code 'num % i' corresponds to division (mod i). A more complete explanation would explain this representational feature in terms of lower-level architecture.

percolate through that network, we don't find anything that obviously produces a helpful content-level story. This program-level obscurity then leads people to various reactions:

- Some people get tempted into scepticism about content, thinking that there's no way to tell a content-level story on the basis of such obscure computational mechanisms.
- Some people get tempted into thinking that AI systems must have wildly alien contents, representing (perhaps) massively disjunctive properties that are computationally tracked by the details of the neural network, but which can't be expressed or comprehended by humans.
- Some people think that humanly comprehensible contents can be extracted from the computational details of the neural network, but that sophisticated tools of computational intervention are needed to figure out how specific contents are grounded in specific portions of specific neural networks. (Example: the rapidly expanding body of work on feature visualization tools.)

Our first central lesson for this book is that all of the above is the wrong way to think about the problem of AI content. AI content is not a problem at the level of programming and computational detail. Instead, AI content is a problem at the level of environmental and sociological detail.

One important thought for considering the foundation of content for neural network AI systems is: all of this has happened before. This is not the first time we've encountered the problem of assigning content to systems whose computational details are enormously, and perhaps incomprehensibly, complex. That's

because all of us are neural networks of exactly that sort. And we've thought a lot about how to attribute representational content to people.

It's possible to take the computational perspective on content determination for people. To take this perspective involves thinking that what a person means by the sentences they utter, or what the contents of their beliefs are, or what features of the world they are representing in perception, are determined by the computational details of their neuroanatomy. Taken to the extreme, this sort of project leads to identifying representational functions of specific neurons, for example, 'face detection neurons'.[2] Of course, the computational approach doesn't need to be taken to this extreme—we might think that human content is grounded at some 'higher level' of computational organization. Maybe we need multiple neurons working together in the right way before they can represent anything, or maybe we need entire regions of the brain computationally organized in the right way before we get representation. But from this computational perspective, we can extract a research program: consider the computational structure of various parts of human brains at various levels of abstraction, and try to determine which of those computational structures manage to represent and what they represent.

But there's an important alternative perspective on content determination which argues that that entire research program is a mistake. *Externalist* views hold that the problem of content determination for people isn't a computational problem; it's an environmental and sociological problem. (More carefully: externalism holds that content determination isn't *uniquely* a

[2] See e.g. Axelrod et al. (2019).

computational program. While a person's internal state may do some of the work in fixing content, their external connections to their environment and society also do some of the work. So in thinking about content determination, we shouldn't look automatically and single-mindedly at computational factors.) There are many versions of externalism available, but what they have in common is a denial that a person's representational capacities are fully grounded by internal features of that person. Instead, externalist views hold that we represent the way we do in part because of features external to us. For example:

- Some of our visual experiences might be representations of faces in part because of our evolutionary history, which is a history of a social species whose survival and reproductive success depended on recognition of social cues, leading to evolutionary selection for facial recognition abilities. On this picture, it's not computational features of our neurons that make them facial detectors, it's historical and teleological features of us that make us facial detectors.
- Some of our linguistic utterances might mean what they mean in part because of the way that we are related to our larger speech community and because of what the words we use mean in that speech community. On this picture, knowing everything about the computational architecture of a particular language user could leave us far short of what's needed to assign content to that user's utterances, since we need at a minimum to know how other people are using those words as well.
- Some of our beliefs have the contents that they do because of the specific environments in which we've acquired

these beliefs. Someone who lives in an environment surrounded by water will acquire beliefs *about water*, while someone who lives in an environment surrounded by some other clear liquid will end up with beliefs about that liquid. (On this picture, looking at computational features of the brains of the believers won't reveal what their beliefs are about—the two believers in the two environments could have their different beliefs despite having the same internal computational organization.)

Those are just some quick snapshots of some externalist approaches—later in this book we'll develop some of the externalist approaches in more detail.

One of our central theses, then, is that we should pursue externalist approaches to content determination, rather than internalist computationally oriented approaches. An analogy to make the point clear:

Suppose we get interested in the question of what makes things *valuable*. So we collect various things of value: dollar bills, krone coins, gold doubloons. We start examining the samples at the microlevel, looking for the features that make them valuable. Looking at the atomic level, we don't find anything clearly value-determining. So we look at a slightly higher level, checking the chemical and molecular structures. Still nothing value-determining appears, so we go up another organization level—we look at threads in the dollar bill, or the ridges on the krone coin. Again, nothing value-determining emerges.

The problem, of course, is that this approach to the determination of value is fundamentally misguided. Value isn't the kind of feature that emerges at some specific physical level of the

description of a valuable object. No amount of probing features at the right level is going to produce useful results, because the problem isn't a vertically determined one. Explaining the determination of value needs to be done (at least in part) horizontally, by thinking about how the valuable object is related to us, our practices, and things in the environment. The determination of value is a sociological problem, not a microphysical problem.[3]

Similarly with content, we suggest, Much of the interesting work that's been done in the artificial intelligence literature on interpretability and explainability of AI has presupposed a problematic internalistic and computational perspective, and assumes that the research project needs to be centred around probing kinds of content to be found at various levels of computational organization. Externalist approaches offer much greater promise for explaining AI contents. That shouldn't be surprising. Externalist approaches, since they allow content explanation to be 'horizontal', making it an environmental and sociological question, rather than 'vertical', let us shortcut full engagement with the computational complexities and obscurities of AI systems. These are the kinds of approaches that have been most successful for 'neural network' creatures like us; it makes sense that they would also be successful for the neural networks we have created. And it's not surprising that content in general would be an externalist notion, grounded in relations of content-bearing creatures and systems to their environment—the nature and role of content,

[3] That point is compatible with the thought that the sociological *eventually* gets grounded in the microphysical in a reductionist way, but (a) it doesn't require that further thought, and (b) even if the reductionist agenda eventually works out, it remains true that the right explanatory approach to value goes through the sociological.

after all, is precisely to relate creatures and systems to their environment, so that they can encode information about the environment and properly interact with the environment.

Second Claim: Existing Externalist Accounts of Content Are Anthropocentric

Our first claim is that externalist accounts of content determination provide the right route forward. But it's not enough just to point to existing work in externalist metasemantics. Existing externalist metasemantic stories have been told as stories about people like us, but AI systems, despite some architectural similarities to people, aren't entirely like us. Our first claim started with the thought that all of this has happened before—that the content-determination problems that AI confronts us with are problems we've already encountered in thinking about our own contents. But, as Twain emphasized, history doesn't *repeat*, it *rhymes*. It's the same problem, but in a different key.

Consider a simple case (later discussion in the book will provide more sophisticated discussion of more sophisticated cases; for now, we just want proof of concept). Suppose Jones's visual perceptual experience represents the object in front of him as a snake, and we want a story about *why* snake is part of the content of that visual experience. Why does Jones's visual experience represent a snake, rather than a thin rectangular region with portions of black, red, and yellow? An externalist might at this point appeal to external features of Jones, including his evolutionary history. Jones's visual experience represents a snake because Jones is a member of a species whose visual systems evolved in an

environment containing dangerous snakes, so that having snake-recognition capacities was evolutionarily advantageous.

Obviously we aren't going to get explanations quite like that for artificial intelligence systems. Artificial intelligence systems didn't evolve—these programs aren't members of species that reproduce via offspring that mix genetic traits from two parents, they aren't at risk of being killed by predators in their environment before reproducing, they aren't subject to random mutations. More generally, AI systems have very different sorts of environmental and sociological connections than we do—these differences then create problems in taking off-the-shelf externalist tools for content determination, created as theories about human content, and applying them directly to AI systems.

So far we have been discussing externalism at a very high level of abstraction—as the general view that environmental and sociological factors can matter to content determination, and thus that content is not determined solely by the internal computational state of a content-bearer. But to have a substantive theory of AI content determination, we need to descend from that high level of abstraction and say which environmental and sociological factors matter and in what way they matter.

Philosophers have developed impressive models for how to understand the content of human language and human mental states. We have developed theories of what content is, how it can be expressed in language, and how it can be shared in communication. Those theories, however, were developed with humans as their starting point. In other words, we developed those theories by observing a specific animal, with specific biological features and evolutionary history. There are many features of our communicative patterns that are contingent on the kinds of animals we

are and the kinds of lives we lead. For example, we have mouths and ears, which make talking and listening natural. We have fingers, which make writing (and sign language) possible. And so on.

However, most people who study meaning and communication will agree that things made very differently from us can express content. If there are aliens, we might be able to communicate with them. They might be able to think things and say things to us, even if both their internal hardware and their external relation to their environment are very, very different from ours. In other words, the ability to communicate in a contentful way is *multiply realizable*: it is an ability that is not restricted to animals and certainly not to animals just like us.

These two facts (that our theories of content have been developed with humans as their starting point but beings other than humans, plausibly, can represent and communicate) are in tension, and failure to attend properly to the second fact has caused our theories to be *biased*. The bias is that most of our theories of representation are too *anthropocentric*. They are parochial because they are based on continent features of our communicative practice. These features are salient to us, but not essential to the nature of content and communication.

Philosophical work in metasemantics, because of its focus on creatures like us, has produced what we will call an **anthropocentric metasemantics**. The existing philosophical accounts of content determination are too parochial by being too focused on contingent features of human communicative/representational practices. What's needed (both for a metasemantic account that's suitable for AI systems and for a general approach to metasemantic questions that's general and robust enough to be philosophically satisfying) is a **de-anthropocentrized** metasemantics. To achieve

that de-anthropocentrization, we'll set out the idea of **anthropocentric abstraction**. In anthropocentric abstraction, we take existing externalist accounts of content determination and abstract away from these contingent and parochial features of human communication to reveal a more abstract pattern that is realizable in many kinds of creatures.

The trick with anthropocentric abstraction is that we can't simply abstract away all the details about the specifics of human engagement with environment and society. An abstracted metasemantic theory that said just that content in general (not just for creatures like us) is determined by *some kind of relation to the external environment* would be too vacuous to be useful or interesting. What's needed is to abstract just the right amount: enough to remove any undue anthropocentric bias, but not so much that we remove all content from the externalism.

Finding this abstractive sweet spot will inevitably involve careful consideration of the details of AI systems. We need to consider the points of similarity and difference between AI systems and us, so that we can see how to take externalist frameworks originally developed as tools for understanding our ability to represent the world and abstract them into tools that also explain the ability of AI systems to represent the world. We'll dive into details as we consider some specific externalist frameworks, but we'll start by noting six big-picture points of comparison:

1. **Creation**: Unlike humans, AI systems are intentionally designed and created by people who already have their own representational contents. AI systems thus give rise to special questions about how their contents relate to the contents of their creators.

2. **Limited Range**: Many AI systems have a very limited range of conceptual applications. An image recognition program might only be able deploy the contents 'cat' and 'dog', and only be able to apply those contents to photographic images. Humans, on the other hand, have a very wide range of contents that can be applied across a wide range of domains.
3. **Unclear Boundaries**: Programs, unlike people, easily break down into smaller subprograms, and easily integrate with other programs to create larger computational and functional units. Questions about what exactly has the content are thus trickier for programs than for people.
4. **Output Variability**: Some contentful AI outputs are linguistic, and at least on the face of it, these linguistic outputs have the same content as sentences in a natural language. Other contentful AI outputs are non-linguistic: AI systems can produce numerical outputs (probability distributions), moves on a game board, digital photographical images, and so on.
5. **Dedicated Integration**: AI systems typically have very specific roles that they are intended to play in our lives (assess credit risk, play chess games, etc.), and the contents they bear need to help make sense of them playing those roles. AI systems are largely single-purpose tools; we are largely many-purpose tool users; this difference between us and AI systems can matter to the details of content determination.
6. **Black Box and White Box Implementation**: Like us, AI systems have internal computational architecture that is largely black box, with computational details that are

> obscure and not revelatory of purpose or representational content. But many AI systems are in fact complicated mixtures of black box and white box components. A complex neural net might, for example, be combined with a Monte Carlo randomizing tree search algorithm whose computational implementation, purpose, and representational significance are all entirely transparent.

It is in thinking through points of difference and similarity such as these that philosophical work on metasemantics has much to learn from AI research as we look for the right abstractive sweet spot.

Third Claim: We Need Meta-Metasemantic Guidance

The problems of anthropocentric abstraction are not unique to AI systems. To attribute content to animals, we may need to engage in some anthropocentric abstraction, abstracting away from human-specific details to an approach suitable to the specific ways that other animals are embedded in their environments. The same might be true for content attribution to, for example, pictures, dance, music, and film. In all these cases we might need models that go beyond what we have developed to account for content of human beliefs and languages.

Different cases of anthropocentric abstraction will involve confronting different questions about how to abstract. We need many different metasemantic theories: a human metasemantics, explaining the specific ways in which contents of humans are grounded in specific internal features of humans and specific ways

that humans are embedded in their environments; an artificial intelligence metasemantics, explaining the different specific ways in which contents of AI systems are grounded in the different specific internal features of those systems and the different specific ways that they are embedded in their environments, and so on, for other varyingly alien bearers of content.[4]

Each of these domains will require detailed separate investigations: anthropocentric abstraction is not a unified type of theorizing. Here we focus on the problem of abstracting externalism in a way suitable for AI systems. As we'll see in the subsequent chapters, in considering how to abstract, we encounter a number of choice points. As a result, we are exploring a large logical space—a space containing multiple metasemantic theories for different content-bearing creatures and systems, and different options for how to analogize a metasemantic theory for one kind of creature to a metasemantic story for a different kind of creature. Navigating that logical space raises a methodological question: how do we decide what the right way is to abstract? Even once we are completely clear on all of the ways in which AI systems are different from and similar to us, and completely clear on what the right externalist metasemantic framework is for us, how do we decide what abstracted analogue of that framework is best for the AI systems?

Our third central claim in this book is that this methodological problem is best addressed by considering questions

[4] We don't mean to commit here to any particular way to carve up the metasemantic landscape. Maybe different kinds of AI require different kinds of metasemantics; maybe humans and some nonhuman animals all get contents determined by the same metasemantics.

of *meta-metasemantics*. Consider an explanatory hierarchy of content-related facts:

1. The **semantic** facts are facts about what contents specific content-bearing items have. It's thus a semantic fact, for example, that the word 'Aristotle' in English refers to Aristotle.
2. The **metasemantic** facts are facts that explain why the semantic facts are what they are. The semantic fact that 'Aristotle' refers to Aristotle is, for example, according to a Kripkean metasemantic approach (which we'll discuss in greater detail in Chapter 6) explained by the fact that the name 'Aristotle' is part of a causal chain of usages going back to Aristotle.
3. The **meta-metasemantic** facts are facts that explain why the metasemantic facts are the way they are. Meta-metasemantic questions are rarely explicitly addressed by philosophers. Why is the semantic fact that 'Aristotle' refers to Aristotle explained by a Kripkean causal chain metasemantics rather than by some other metasemantic account?

Answering meta-metasemantic questions, we will suggest, requires considering the theoretical role of contents. By considering what explanatory work contents and content attributions do for us, we can work out what kinds of fact could best fix semantic features so that contents can play those explanatory roles. A good meta-metasemantic framework can thus offer us the needed methodological guidance. In determining how to abstract an AI-suitable metasemantics from the existing human-targeted externalist metasemantics, we need to think about the role we

want contents to play, and then think about the details of AI systems and the role of those AI systems in our lives and our environment. From all of this we can hope to extract a specific picture of what content determination mechanisms for AI systems would be best suited to the roles that the meta-metasemantics identifies.

A Meta-Metasemantic Suggestion: Interpreter-centric Knowledge-Maximization

The field of meta-metasemantics is less well developed than metasemantics.[5] There isn't even a consensus that metasemantic theorizing should be guided by an explicit meta-metasemantics. One reason for that might be a healthy fear that it is hard to see where this will stop: why not develop the field of meta-meta-metasemantics? After all, if we need the metasemantics to guide our semantics, and we need meta-metasemantics to guide our metasemantics, why and how would this ever stop?

We recognize this as a concern to some, but we have no fear: we endorse this endless hierarchy of theorizing. There is of course a practical limit to what we humans can process and grasp, but that isn't the limit of interesting inquiry. In this book, however, we move at most three meta levels up—we will leave the explorations of higher levels to others (or to ourselves in the future). We do that with a very concrete goal in mind: to guide our theorizing about the metasemantics of AI. Moreover, we will not devote the book to arguing for our meta-metasemantic view. We will instead use a

[5] There is a bit of meta-metasemantic literature on the question of whether meta-metasemantics for externalist theories should be externalist or internalist (see e.g. Cohnitz and Haukioja 2013 for discussion and references).

proposal made by Timothy Williamson in the last chapter of *The Philosophy of Philosophy* (Williamson 2007).

Before briefly sketching that view, we should emphasize that if you have alternative theories about meta-meta-metasemantics, we encourage exploring the various ways those views will trickle down to metasemantics and that again to particular interpretations of AI output. The overall spirit of this book is to develop a framework for thinking about interpretable AI and there will be many alternative ways to fill in that framework. The use of Williamson's meta-metasemantics is just one of them.

In *The Philosophy of Philosophy*, Williamson can be read as proposing a version of the principle of charity as a meta-metasemantic principle. Roughly, Williamson's view is that the correct metasemantics is one that *maximizes knowledge for the interpretee*. Moreover, Williamson thinks this is the principle that makes externalism correct. His proposal is that a knowledge-maximization principle is the foundation of externalist metasemantics.

The argument goes as follows. Imagine a case of demonstrative misidentification: Alex is a devoted physiognomist who thinks he can tell a person's character from their face. He sees Bea, and on the basis of her appearance forms the belief he would express by saying 'She is F, G, and H'. She is none of these things: physiognomy is nonsense. But, it so happens, there is someone, somewhere (let's say New Zealand), who is F, G, and H, and has no other properties. Call that person 'Ceres'.

We can use this scenario, Williamson thinks, to shed light on our theory of reference, and in particular to draw connections between interpretation and reference. Consider the question of who Alex's utterance (or the related thought) of 'she' refers to.

Does it refer to Bea, the person in front of him, or Ceres, the person who is F, G, and H? If you aim to maximize the true beliefs you impute to someone (à la work like Davidson 1973), taking the reference to be Ceres seems like the way to go: it makes Alex's belief come out true.

But that's obviously wrong: 'a descriptive theory of reference gone mad' (2007: 263), in Williamson's words. The referent, it seems, is Bea. But how do we make that square with interpretation?

Williamson's neat idea is that what we should aim to maximize is not the interpretee's true beliefs, but their knowledge, and that doing so yields an argument for externalism. Thus, to put the matter crudely, imagine the speaker uttering the following four sentences:

- She is F.
- She is G.
- She is H.
- She is in front of me.

If we want to maximize belief, we should say that 'She' refers to Ceres, since that gets us three true beliefs, and one false belief, although it gets us no knowledge. If we want to maximize knowledge, we should say that 'she' refers to Bea, since it gets us one piece of knowledge, namely that Bea is in front of the speaker.

The reason for this is that perception is a suitable 'channel' for knowledge, whereas physiognomy isn't. But perception is of course a paradigm causal channel as well, and so Williamson, generalizing from these sorts of considerations, suggests that knowledge-maximization is a better principle of interpretation

than belief maximization because knowledge tracks causal channels in a way that true beliefs don't always. As he intriguingly suggests:

> Such examples [as the one we just considered] are of course just the analogue for demonstrative pronouns of examples Kripke and Putnam used to refute descriptive cluster theories of reference for proper names and natural kind terms. In effect, such theories are special cases of a truth-maximizing principle of charity. One fundamental error in descriptive theories of reference is to try to make true belief do the work of knowledge. (2007: 264)

Our aim here is not to get too deep into theories of interpretation.[6] But we do want to make one change to Williamson's theory: we suggest choosing a metasemantics that maximizes knowledge of the *interpreter*, not of the subject (be it a person or a machine) being interpreted. In other words, we are suggesting our metasemantic principles should be guided by a meta-metasemantic principle that tells us to pick a metasemantics that maximizes what we, the interpreters, end up knowing as a result of the interpretative enterprise.

That raises the question: *why aim for meta-metasemantic principles that tell us to maximize the interpreter's knowledge and not the interpretee?* A couple of points in reply to this: first, note that this is a question in meta-meta-metasemantics. We will not try to provide a general meta-meta-metasemantic theory here. We are not alone in not doing that. In fact, we know of no worked-out three-level metasemantic theory. Since, in a little book like this, arguments have to stop somewhere, we would be fairly comfortable simply using

[6] The most sophisticated recent work on the topic is by Robbie Williams (see 2005, 2007) and references therein. A nice overview of the space of options for charity principles is in Felman (1998).

this as a starting point. However, there's a bit more to be said. Recall our earlier point that the meta-metasemantics should be guided by considerations of what explanatory work content and content attributions are doing for us. We think it should be fairly noncontroversial that a central goal of content and content attributions to AIs is to increase our knowledge. That gives us motivation for being self-centred. It is *our* knowledge that matters, not that of the artificial systems.

What we just said leaves open the possibility that others might have other interests. The artificial systems, for example, if they have interests, might want to rely on different meta-meta-metasemantic principles when they interpret each other or us. Maybe they want metasemantic principles that maximize knowledge (or power or something else) of the artificial systems. It is also possible that our interests are aligned. That's an open question. More generally, we think the right view is one according to which *interests* of various kinds will play a central role in higher up on the meta-…-metasemantic hierarchy.

These larger questions about how to ground the various levels of metasemantic theorizing are interesting (and worthy of an entire separate monograph, since they are issues largely unexplored), but will not concern us in what follows. Instead, we will use versions of the Williamsonian principle to illustrate how meta-metasemantics can and should play an important role in the z philosophy of AI. More specifically, whenever we engage in de-anthropocentrizing (which is our way of developing a metasemantics for AI), there will be choice points. We will use knowledge-maximization as our guide when making those choices.