7

APPLICATION

Predication and Commitment

In the previous two chapters, we've sketched metasemantic stories that specify the facts in virtue of which SmartCredit refers to Lucie and the property *high risk* in its output tokening of 'Lucie is high risk'. However, getting metasemantic stories that ground the meanings of these component parts of SmartCredit's output is arguably insufficient. We also need to explain why SmartCredit, in combining the expressions 'Lucie' and 'is high risk' in its output, doesn't just mention a person and a property, but also predicates the property of the person. In short: can SmartCredit (and other AI systems) produce or entertain or express full propositions?

In the philosophical literature, there is a very venerable literature about the nature of propositions. Our focus in what follows will not be about whether propositions are abstract entities, structured, etc. Our concern is primarily with the phenomenon called variously *entertaining* or *expressing* a proposition by predicating a property of an object. We are interested in predication qua an important semantic phenomenon, not qua a solution to the problem of the unity of propositions or the nature of propositions.

We'll pause here for a moment and recall the unrefuted sceptic from Chapter 2: according to him, what we are doing now is

Making AI Intelligible: Philosophical Foundations. Herman Cappelen and Joshua Dever, Oxford University Press (2021).
© Herman Cappelen and Joshua Dever. DOI: 10.1093/oso/9780192894724.003.0007

adding mistakes on top of our earlier mistakes. The earlier mistakes, according to the sceptic, were to look for genuine reference in the output of AI. We hope that a sceptic reading the previous chapters would react a bit concessively: maybe, they might say, there's a case to be made for reference to the property of being high risk and to Lucie, along the lines suggested. However, that need not be accompanied by conceding that the further step of finding a complete proposition (*that Lucie is high risk*) in the AI output is reasonable, nor agreeing that we should attribute to the system a commitment to that propositional content.

As before, our response is this: while we're not unsympathetic to the sceptic's position, we think the best way to test it is to try to see if an account of propositional content can be found. This is particularly *important* because there are hardly any efforts to do so. If you end up convinced that what we are about to propose is a potential way forward, then that's progress. If you find our effort entirely unconvincing, then it's a bit of additional support for the sceptic.

Predication: Brief Introduction to the Act Theoretic View

To understand the full range of AI content, we need to think that SmartCredit can not only denote Lucie and the property *high risk*, but also *predicate* the property of being high risk to Lucie. In order to create a model of how that can happen, we need to understand the act of predication. Here we encounter the same kind of dilemma we've faced throughout this book: there are very many theories of predication. Our book is brief and we cannot comprehensively engage with that entire literature. What we will do instead is explore these issues against the background of a particular theory: the act theoretic view. An obvious drawback of this strategy is that if you, our reader, is adamantly opposed to this view, what follows will at best be of conditional interest to you. We hope, however, that the general strategy (of de-anthropocentrizing a view of predication) will be of use to you, even if the way we implement it isn't.

The view we will work with for the sake of argument is found most recently in the work of Scott Soames and Peter Hanks (we'll mostly be focusing on their respective monographs, both from 2015, but see also Soames 2019, and, for a slightly larger overview of the logical space of contemporary theories in this vein, Soames et al. 2014). According to Soames and Hanks, propositions are act types. They don't have intrinsic representational properties. Token acts (of the relevant type) are the original bearers of representation, truth conditions, and truth value. This is a change from traditional ways of thinking about propositions and representational objects more generally. They've traditionally been thought of as entities that somehow had an existence independent of the act of grasping them. The Fregean picture was of the mind as reaching out to—or grasping—entities that had existence independently of the act of grasping. The act type view reverses that picture: the primary explanatory element is the act of predication. Propositions understood as abstract objects, are act types instantiated by those token acts.

What is the act of predicating? Soames takes this to be a primitive notion. It picks out something we can easily recognize. For example, to think that *Lucie is high risk* involves denoting high risk, denoting Lucie, and then predicating the former of the latter. This token act of predicating belongs to a type. That type is the *proposition that Lucie is high risk*. The proposition that *Lucie is high risk* (i.e. the act *type*) has representational properties derivatively: it's the *token* act of that type that is true or false.

Is it an objection to Soames's view that he doesn't offer us an analysis of predication? He thinks not, and we agree. As he writes:

> One might ask what we mean by 'predication'—what, in effect, the analysis of predication is. Although it is unclear that an informative answer can be given to this question, it is equally unclear that this is anything to worry about. Some logical and semantic notions—like negation—are primitive. Since this elementary point typically doesn't provoke hand-wringing, it is hard to see why the primitiveness of predication should. (Soames 2015: 30)

Although primitivism about predication is defensible, it's worth considering at least one opposing view which tries to say more. The essence of predication, according to Hanks, is the ability to sort things into groups:

> Acts of predication are acts of sorting things into groups. When you predicate a property of an object you sort that object with other objects in virtue of their similarity with respect to the property. To predicate the property of being green of something is to sort that thing with other green things. This act of sorting can be done behaviorally, for example by picking the object up and putting it with other green things, or it can be done in thought, by mentally grouping the object with other green things, or in speech, by saying that it is green. (Hanks 2015: 64)

So understood, the *ability* to predicate, i.e. categorize, is a basic biological function that human beings share with the rest of the animal kingdom. Hanks approvingly cites Susan Gelman, who says:

> [A]ll organisms form categories: even mealworms have category-based preferences, and higher-order animals such as pigeons or octopi can display quite sophisticated categorical judgments.
>
> (Gelman 2003: 11)

Hanks uses the example of sniffer wasps (and bees) to illustrate how basic this ability is. Sniffer wasps can be trained to do various things, for example, detect landmines (and also various narcotic substances). In so doing, the wasp, according to Hanks, predicates the property of smelling like—for example—TNT to various objects. The act of predication, on this view, is the act of flying to those things.[1]

In the rest of this chapter, we explore whether a de-anthropocentrized version of the act theoretic view can be applied to AI systems. If that can be done, there is some evidence that the final step is achievable: Not only can SmartCredit denote the property of being high risk and Lucie, it can also predicate the former of the latter.

Turning to AI and Disentangling Three Different Questions

In order to understand predication in AI, we'll disentangle three difference questions that are not always separated in the literature:

Q1: What is it in a sentence that means predication? There are several candidates for what that may be:

[1] Absent the ability to use language to characterize its mental act, there is an irresolvable indeterminacy in the content of the wasp's judgement. Nothing fixes *what* the wasp judges or believes.

(i) the concatenation of subject and predicate;
(ii) the space between subject and predicate; and
(iii) the root note in a tree having subject and predicate as daughter nodes.

The choice of syntactic object will depend on your background syntax and assumptions about how syntax, at the most basic level, interfaces with semantics. We won't take a stand on that here (for some helpful discussion, see King 2007 especially 33–6). What is important is that there's some syntactic feature that means or expresses predication. The question of how that syntactic feature ended up expressing predication is different from the question of what predication is, just as the question of how names—for example—ended up referring is different from the question of what reference is.

Q2: Which mental act is the mental act of predication? According to the act theoretic view, there is some kind of mental act that humans (and other animals, e.g. wasps) can perform. That act is, in part, constitutive of propositions. Since this act will be implemented in different ways in different animals, we can ask: how do we identify the particular mental state that is the act of predication? Moreover, since it's plausible that syntactic predication expressed in language somehow represents the mental act of predication we're assuming, having a grasp of what the latter is like will help us understand what syntactic predication expresses.

Q3: The metasemantics of predication: Having distinguished (1) and (2), we can distinguish two metasemantic questions:

(a) Why does this bit of syntax (i.e. the act of concatenation) mean predication? How did that part of syntax end up

having that meaning? (Analogous to how we can ask: how does a name end up having the referent that it does?)
(b) Why does this mental act mean predication? That might seem like a surprising question, but the underlying motivation for it is this: it is not just the intrinsic features of that act which makes it an act of predication. We'll see below that it is, in part, the functional role of that act—and we'll suggest that functional role can be spelled out in a teleofunctional way.

Note that both of these lattermost questions are questions in metasemantics that pattern with the metasemantic questions we have discussed about 'high risk' and 'Lucie' in Chapters 5 and 6. The three questions are closely connected. For Soames or Hanks, the answer to question (1) is that the relevant bit of syntax roughly expresses the performance of the mental act. The answer to (1) therefore partly depends on the answer to (2). However, the answer to (1) is not fully derivative of the answer to (2) because in order to answer (1), we need two things. Firstly, we need an answer to (3a), i.e. an account of why concatenation (or whatever) is hooked up to this mental act, rather than to something else. We also need an answer to (3b), i.e. a metasemantic explanation for why the mental act means what it does.

The Metasemantics of Predication: A Teleofunctionalist Hypothesis

At this point we are going to add a new theory to the mix: teleofunctionalism. We do this for three reasons. First, it seems to us natural to think that this is, at least in part, a motivating thought behind the act theoretic view. Second, we want to use this as yet another illustration of how externalist theories can be helpful in giving an account of interpretable AI. Third, in de-anthropocentrizing the notion of predication we find in contemporary philosophy of language, we will need to find some way of identifying predication independently of its realization by human mental states, and functionalist theories in general are well-placed to do this.

As we interpret the act theoretic view, it identifies a mental act, *A*, that we perform. What makes *A* an act of predication? Well, it's because the characteristic function of that act is to give rise to states of belief/judgement in us, which then give rise to characteristic kinds of behaviour. There's a familiar functionalist story sitting in the background here—to know what something represents, think about what kind of representational content would make sense of that something as a mediator between its characteristic inputs and outputs. It's then *teleo*functional because what matters is not the actual input/output performance, but what it's intended/designed/evolved to do. We can summarize this as the TF-Hypothesis:

> *TF-Hypothesis*: A mental act is the act of predication because of its teleofunctional role in giving rise to judgements that guide action.

The basic idea is that no mental act can be the act of predication in isolation from the function it performs. The relevant function is that of giving rise to judgements that then guide action. If a mental act doesn't give rise to judgements that play a role in guiding action, it would not be the act of predication. In the case of the sniffer wasp, the act of flying to a location is followed by acts of trying to extract sugar (that's how they are trained:

the TNT is laced with sugar and so the sniffer wasps are conditioned to fly to objects that smell like TNT). In the case of humans, the actions we perform are infinitely richer and impossible to make precise (and the claim we are making doesn't require that it can be).

In sum, the first part of the proposal is this: (i) a mental act cannot be the act of predication in isolation from function; (ii) the relevant function is that of giving rise to judgements that guide action; and (iii) that we perform acts with this function can be explained along teleofunctional lines. This gives us a way of identifying predication of the sort that Soames and Hanks are concerned with without committing ourselves to any particular architecture that implements it.

Some Background: Teleosemantics and Teleofunctional Role

To understand our appeal to teleofunctional role, it will be helpful to say a few words about teleosemantic theories more generally. The basic thought is that the very fact that we (and, as we're arguing, AI systems) have content, as well as the particular contents we (and they) have is to be explained in terms of the idea of *function*. Paradigmatically, for teleosemanticists, these functions are biological. Thus, for example, the function of the heart is to pump blood; that's what it's there for, where in turn this notion of a particular thing being *there for* a particular function is to be cashed out in terms of evolutionary history—of natural selection. We evolved hearts because they are efficient ways to help spread oxygen and nutrients around the body (roughly).

The basic idea behind teleosemantics is that we treat representations just like we treat any biological adornment. Why does a cat represent birds? Well, why does a cat have whiskers? The answer to the latter question—roughly, we aren't vets—is that whiskers help cats navigate tight spaces. Having whiskers confers an advantage on cats, an advantage that in their evolutionary prehistory made them more apt at navigating their environment than similar felids which lacked whiskers. The whiskered cats successfully reproduced more, and so whiskers were selected for.

The same applies to contents. A cat that has no capacity to represent birds is a poor cat. It will miss out on opportunities for food, and so is less likely to thrive and reproduce, and so less likely to produce other cats. Cats that can represent birds will be well-fed and attractive mates, and more likely to produce other cats, which will be more likely than not themselves to be able to represent birds.

There are many teleosemantic theories, and many objections to teleosemantic theories, and while it's beyond both our aims here and our ability to decide between them, it will be useful to consider briefly some options and live questions which are relevant for this book. So, we might wonder what sort of representations we can attribute on the basis of teleosemantic reasoning. We considered attributing content like that associated with the full sentence 'there is a bird there' to our cat, but can we also attribute sub-sentential contents, such as representations of CATS, or—more abstractly and difficulty—BEING THERE? Theorists like David Papineau (presented in, for example, 1987, and more recently defending against some objections in 2001) think that our content attributions should be 'top-down', concentrating on representations of full beliefs and desires primarily as opposed to their component concepts (if the latter exist at all). Others—and this will arguably be particularly relevant for us—point out that this theory struggles with creatures that lack evolutionary history, or with creatures whose evolutionary history didn't involve the things they represent. It's good that we don't attribute iPad thoughts to cats, because they evidently don't care about them. We do greatly care about iPads, but iPads don't really figure in the history—on the evolutionary timescale—of our species. Davidson's famous Swampman example (1987) is of a creature molecule by molecule identical to us created by some—say—quantum mechanical fluke which, as such, has no evolutionary history (because it isn't the child of a child of a child . . . whose lineage is shaped by natural selection). To such a creature, it seems, we can't attribute any selected-for biological function, and thus no content—an arguably bad result.

Some wonder whether behaviour and evolutionary history are sufficient to give us *determinate* representations. Thus Fodor (1990) complains that considering functions will run into extensionality problems familiar from the philosophy of content in general. If we want to attribute to a frog the concept FLY because it behaves as if it has the concept, should not we equally want to attribute to it the concept of SMALL, DARK, FLYING THING? After all—at least if we stipulate that all and only flies are small, dark flying things—the two functions seem identical from a biological point of point. Content, the objection goes, is indeterminate in a way that is unattractive.

Let us emphasize: this is very much scratching the surface of a gigantic debate. There are many sophisticated accounts out there and an ongoing research program concerned with dealing with issues like the above.[2] But, for our purposes, it doesn't matter. Just as we were happy to take the basic Kripkean picture, deanthropocentrize a bit, and see how far we could get, so the foregoing superficial survey of teleosemantics suffices for our purposes.

We use some of the ideas behind basic teleosemantics in a new way: to give an account of how a certain state can mean predication and also to give an account of what predication is.

Predication in AI

For AI, we can raise questions analogous to Q1–Q3. Start with the assumption that there is some aspect of the AI output that expresses predication. There is an initial question how to identify that external aspect of AI predication. Here are some options:

- If the output is linguistic in form, then maybe the answer here is the same as the answer in the normal linguistic case.
- However, not all AI outputs need be in linguistic form: If AlphaZero just moves pieces on the board (maybe it's connected to a magnetic system that lets it move the pieces), we can ask: what aspect of its output counts as the predication of 'good to move to A4' to the queen? One possible answer appeals to teleofunctionalism: that's the designed function of that aspect of the output.

[2] Thus we haven't mentioned Ruth Millikan's seminal work (1984, 1989a,b), or Nicholas Shea's sophisticated recent account (2018). And we haven't considered the important work of Karen Neander (2006, 1991, 1996). Again, the sole reason for this is that we don't think it has immediate bearing on the points to be made in this chapter. For an overview and more references, see Neander (2018).

For our purposes, we don't need to take a stand on the correct account of how predication is realized in the system. We'll just assume it is realized in some aspect, *A*, of the output. We can then ask (on analogy with Q1): *why* does *A* mean predication?

There is also an analogy with question Q2 above: we can assume that as in the human case, there is an aspect of the machine's 'inner life' that is predication (on analogy with the 'inner' human judgement that can be expressed, e.g. linguistically). Again, it's an open (and interesting question) what this is. Some options include:

- the system's computation;
- internal contentful states of the sort that get called 'intentional internals' in the AI literature; and
- some interactive aspect of how we use the AI.

If there is some such inner state, *ST*, then the feature we called *A* above only expresses predication derivatively: it is teleofunctionally connected to *ST*. *ST*, on this view, is the 'real' act of predication. *A* expresses predication derivatively, by being teleofunctionally connected to *ST*.

AI Predication and Kinds of Teleology

Our proposal has incorporated an appeal to teleofunctionalism. One issue (of the many issues) we have not yet addressed is: *what kind of teleology are we talking about when we talk about teleofunctionalism?*

In answering this question, we can be guided by the meta-metasemantic principle in Chapter 4: knowledge-maximization. That principle guided our metasemantic theory (which in turn guided our interpretations). We can appeal to it here again in order to determine what the correct notion of teleology should be. The question then is: what kind of telos would be knowledge-maximizing for us interpreters, if we took it to be the teleofunctional role that makes a state into the state of predication?

The answer to this is far from obvious. Here are some options:

- The first place to look is at the design stage: the AI is designed by humans (or, sometimes, other AIs) and the designers will have in mind a functional role for the AI. So the simple answer is: the telos of the system is derived (maybe in some complex way) from that of its designers' intentions. It is that intended function by virtue of which some internal state (or derivatively, an external manifestation) is predication.
- Alternatively, we could treat AIs as more wasp-like. We would then ask what promotes the AI's own survival. 'Survival' in this case would need to be de-anthropocentrized—we look for whatever is the equivalent of survival for the AI system.
- A third alternative is to think about the goal, as derived from humans, not as derived from the AIs survival, but about the human-AI system as a whole. In this case, the telos is of the combination of humans and AI, not of one of those in isolation.

We are simply listing these as options. Our goal here is not to argue for a particular answer, but to show that there's a rich field of inquiry that opens up when teleofunctionalism is introduced to help us understand and interpret AI.

Why Teleofunctionalism and Not Kripke or Evans?

A reader might reasonably say: in previous chapters we appealed first to Kripke (to explain AI denotation of high risk), then to Evans (to explain AI denotation of Lucie), and then now we are suddenly using a teleofunctionalist framework to explain predication. The reader might then ask: what is going on here? Aren't these competing frameworks? How can we selectively endorse all of them?

The first part of this answer to this is that we are not endorsing any of these metasemantic frameworks. Our goal is to show how they can be developed and adopted to understanding AI. We do so by de-anthropocentrizing guided by knowledge-maximization. If one or more of these strategies are promising, that's at least the beginning of a reply to the representational sceptic, who argues that this project is not even worth pursuing. It is also an argument in favour of exploring the various externalist traditions in the philosophy of language. That tradition has not been sufficiently exploited in this domain. In short, part of the answer to 'Why teleofunctionalism?' is: just so we can talk about another tool in the externalist toolkit, and continue to develop a general 'think about the externalist relations of the AI system, not its internal computational states' theme.

More specifically, it's hard to see how any kind of tracking/anchoring story could work for predication, because it's not clear what the thing to be tracked/linked is, or what it would mean to track or link to it. Predication is the classic syncategorematic item, where you want to give meaning not directly, but via how it affects the meanings of other stuff. Conceptual role semantics is a natural thing to use for syncategorematic items (that's why 'and' has always been the best case for conceptual role semantics), and conceptual role semantics is really just a special case of functionalism.

Teleofunctional Role and Commitment (or Assertion)

So far, we have sketched an explanation of what it is about SmartCredit (internally, externally, or both) that makes it the case that SmartCredit has propositional content: that it doesn't just refer to Lucie and express the property of being high risk, but that it has the content that *Lucie is high risk*. So far, we have not explored the question of what it is about SmartCredit that makes it the case that SmartCredit *takes a stand* on that proposition: that it asserts that Lucie is high risk, or concludes that Lucie is high risk, or suggests that Jones is high risk. Here is Soames on the distinction:

> Although to entertain the proposition that o is red is to predicate redness of o, and so to represent o as red, it is not to commit oneself to o's being red. We often predicate a property of something without committing ourselves to its having the property, as when we imagine o as red, or merely visualize it as red. Hence, predication isn't inherently committing. Nevertheless, some instances of it, e.g. those involved in judging or believing, are either themselves committing, or essential to acts that are. (Soames 2019: 2)

This is a point where Soames and Hanks disagree. As Hanks puts it:

> an act of predicating greenness of something is correct just in case that thing is green. Correctness and incorrectness here are just truth and falsity...This means that the act is true just in case that thing is green. An act of predicating a property of an object is true

> or false insofar as it can satisfy or fail to satisfy the correctness conditions determined by the property. Acts of predication have truth conditions and truth-values. (Hanks 2015: 66)

For those sympathetic to Hanks's view, commitment/assertion is built into predication and so we wouldn't need a separate section on this.[3] For the sake of argument, we will tentatively assume Soames' view and see what can be added to get us from AI-predication to AI-commitment.

Theories of Assertion and Commitment for Humans and AI

The question, then, is whether we can think of ML systems as committing to a content. To answer that, we need an account of what goes into that kind of commitment. Again, there's a massive literature on this as applied to humans (for which see e.g. Brown and Cappelen 2011, or Goldberg forthcoming).

To explore the issue of whether ML systems can perform speech acts, we could proceed as we did above: we look at various theories of what it takes to perform speech acts, and then see whether ML systems satisfy those conditions. If we focus on just assertion, there are at least four categories of views:

(i) Assertions are those sayings that are governed by certain norms—the norms of assertion.

(ii) Assertions are those sayings that have certain effects.

[3] But it raises the question of how to understand embedded propositions in negation, conditionals, etc. See Hanks 2015: ch. 4 for further discussion.

(iii) Assertions are those sayings that have certain causes.
(iv) Assertions are those sayings that are accompanied by certain commitments.

Within each of these categories of views, there's a great deal of variation. For example, there are very many norm-based views and no agreement about what the relevant norms are. Here are some of the more prominent suggestions (see Cappelen 2011: 9 for this taxonomy and references):

Truth rule: One must: assert p only if p is true.
Warrant rule: One must: assert p only if one has warrant to assert p.
Knowledge rule: One must: assert p only if one knows p.
BK rule: One must: assert p only if one believes that one knows p.
RBK rule: One must: assert p only if one rationally believes that one knows that p.

Other theories of assertion construe it as an act of commitment. This view is found in a range of authors, going back to Pierce and continuing with people such as Brandom (Pierce 1934; Searle 1969; Brandom 1994). Here is a version of the view from John MacFarlane:

> (W*) In asserting that *p* at C1, one commits oneself to withdrawing the assertion (in any future context C2), if *p* is shown to be untrue relative to context of use C1 and context of assessment C2.
>
> (MacFarlane 2005: 320)

Here is a research project: for each of these, explore whether these are norms that can be followed by an ML system. Despite their differences, they all raise the more general issue: what is it to follow or obey a norm and is that the kind of thing that an ML system can do? To investigate that question, we need an account both of the nature of norms and of what it is to follow them. Our prediction is that doing so will require using many of the same strategies we used above: you'll find current theories parochial because of being too anthropocentric. Then you will need to engage in anthropocentric abstraction, and you'll find some way to create a notion of 'assertion' or 'saying' that can fit ML systems. This will have the added effect of improving normative theories of assertion or saying.

In this book we will not carry out this project, in part because one of the authors of this book is sceptical of the very category of assertion (Cappelen 2011) and the other sympathetic to the Hanks' view that predication is committal (and so the theory of predication is all we need). That said, for those who want to pursue this project, the general meta-metasemantic principle from Chapter 4 should still be of help. Applying the knowledge-maximizing principle, we should expect the speech act of assertion to be such that it is knowledge-maximizing: we should expect assertion to be the kind of thing that maximizes knowledge for the audience member (i.e. the interpreters). If you were to pursue that line, the Williamsonian view that assertion is governed by the knowledge norm is tempting. However, endorsing that view also involves accepting that there are constitutive norms of assertion. That is an additional controversial assumption, but not one we will explore further in this book (but see the references above, in particular the anthologies and handbooks, for much recent work from many different perspectives).