# 6

# APPLICATION

*Names and the Mental Files Framework*

## Does SmartCredit Use Names?

In the previous chapter, we outlined a proposal for how a de-anthropocentrized version of Kripke's theory could explain how SmartCredit could represent the property of being high risk. In the sentence we started with 'Lucie is high risk', that property is attributed to a person, Lucie. We now turn to the question of how SmartCredit can refer to Lucie.

An initial observation: we cannot assume that the lexical item 'Lucie' plays a significant role in SmartCredit's neural network. What happens is this: some information about Lucie is initially fed into the system. This will include various financial and demographic information. The system will then 'collect' more information. If we idealize a bit, we can think of the system as ending up with a potentially enormous collection, C, of information. What SmartCredit is then programmed to do is assess whether an individual that has all the properties in C is high or low risk. In the case we have been considering, the conclusion is that the individual is high risk. If this is the right description, SmartCredit is different from regular English speakers in that proper names play a

*Making AI Intelligible: Philosophical Foundations.* Herman Cappelen and Joshua Dever, Oxford University Press (2021).
 DOI: 10.1093/oso/9780192894724.003.0006

rather limited role in its computational structure. Property clusters play a central role. What we assume when we translate the output into the sentence: 'Lucie is high risk' is that Lucie satisfies a certain property cluster. We can think of the output as in effect being of the form:

> Something satisfying C is high risk.

Then there's a background assumption, that a particular person, Lucie, satisfies C. What happens at the output point is that this assumption is added and we present the output as if it's directly about Lucie, i.e. as:

> Lucie is high risk.

If the story we've just told is correct, that presentation of the output is tendentious because it relies on an implicit assumption: that the person we refer to with the name 'Lucie' is someone who satisfies C.

One advantage of this picture is that, if correct, it means that we don't need to add an account of how SmartCredit can have competency with names. What we need is an account of how it can use the predicates that are components of C (and we gave an account of that in the previous chapter), and then an account of predication (which we will give in the next chapter).

A disturbing feature of this account is that the real output of SmartCredit, i.e. *that something satisfying C is high risk*, could become impossible to understand and track. The property cluster that C is an abbreviation for will potentially track properties and interconnections between properties that we cannot express. Initially, the input we give SmartCredit might be tractable for us, but as it starts

collecting information from varied sources, the resulting complexity will, in many cases, be too complex for a human mind, even if we did have the terminology to express it.

If the output is something we cannot grasp, that makes the assumption that Lucie satisfies the conditions in C one we cannot fully grasp (or understand). If we can't grasp it, we can't assess it. In other words, we are then building in a tacit assumption that we're incapable of assessing. Since C will contain an enormous amount of information, we can safely assume that it will frequently happen that some of the information doesn't apply to the individual to whom we take it to apply, in this case Lucie. We then face questions about how to treat the output of SmartCredit when some, but not all or most, of the information in C fails to apply to the individual we interpret the output to be about.

In other words, if SmartCredit has no capacity for representing in some way analogous to how we represent using proper names, then we face both communicative and epistemic obstacles in our engagement with it.

## The Mental Files Framework to the Rescue?

We just outlined some problems for a certain model of how AI represents Lucie. Anyone familiar with the last 120 years of philosophy of language will recognize that analogues of these problems have been extensively discussed. One of the fundamental divides in the literature is between theories according to which names are clusters of descriptions and theories according to which they are not. In the above paragraph, we in effect

first proposed a descriptivist view and then raised some problems for it.

Rather than rehearse that entire debate here, we will do as we did in the section of predicates: use one of the standard theories and see if it can be applied to an AI system like SmartCredit. The framework we will appeal to is the so-called mental file framework. Early work on this includes Lockwood (1971), Perry (1980), and Evans (1982). Recanati (2012) is a recent and comprehensive presentation and defence of the view. In what follows, we will rely in large part on Recanati's view.

At the core of the theory is the idea that human cognition centrally involves clusters of properties. In saying that they are clusters, we mean that the properties are presented as co-instantiated. Such clusters are, in this literature, called mental files. Using the metaphor of files, we can talk of the files as consisting of properties, where that is to say that these properties are co-instantiated. Here is Aiden Gray's useful summary of the view:

> The role of a file is to collect and store information derived from a single object. Files are temporally enduring—an agent maintains a file over time, adding new information derived from the same object. A thinker can employ a file to think and reason about the referent of the file. (Gray 2016: 348)

Three important observations about the properties contained in a file:

1. **A negative thesis**: The object the file is about is not the object that has all the properties in the file. In Evans' terminology: the denotation of a file is not achieved through fit.

2. **A positive thesis**: Reference to an object is determined by various external relations. We'll say more about these below; they can be epistemic, historical, and causal.
3. **A corollary**: The file for a person, say Lucie, can fail to describe her: As long as the external connection holds between Lucie and the file, the file is about Lucie even if *all* the information contained in it fails to apply to Lucie.

According to the mental file framework, we should think of proper names as associated with mental files. For each name, a different mental file. The files can evolve over time (as we gain and lose information), can be combined, and can sometimes be divided.

Two important questions arise for this kind of view. First, what does it take for a mental file to be associated with a name: e.g. under what conditions is a file the Lucie-file? The leading theorists are quite vague on this. Again, Gray gives a good description:

> One sometimes sees the claim that a particular file is 'labeled' with a name—for example, Recanati (2012, pg. 190) and Lockwood (1971, pg. 208). This is, at best, a placeholder for an account of the connection between a file and a name, and, at worst, a misleading metaphor.

We think Gray is right that the mental file framework is deeply metaphorical and that these metaphors are both integral to the attractiveness of the view and potentially misleading. Our heads contain no real files, no labels, and no filing cabinets. Insofar as the theory trades heavily on these metaphors, it's in danger of misleading us. But for now, we'll put those concerns aside. We'll focus on the good parts of the theory and use them to help us understand SmartCredit.

The second question that is important for our purposes is how to understand the relation between a file and its referent. We know from 1 and 2 above that it's not through fit, but through some kind of external relation. To assess the theory, we will need more details. Recanati's favoured term for the relation is 'epistemically rewarding'. He says:

> The characteristic feature of the relations on which mental files are based, and which determine their reference, is that they are epistemically rewarding […] They enable the subject to gain information from the objects to which he stands in these relations.
>
> (Recanati 2012: 35)

Perception of an object is a paradigm of an epistemically rewarding relation. That, however, obviously cannot be the full story about names because we can talk about Cicero using 'Cicero' despite never smelling or touching or hearing or seeing him. So if we use 'Cicero' to talk about Cicero in virtue of having a file that stands in an epistemically rewarding relation to Cicero, then such relations must include, for example, information gained through testimony.

## Epistemically Rewarding Relations for Neural Networks?

To extend Recanati's framework AI, we need to de-anthropocentrize the notion of an epistemically rewarding relation. We have ideas about what would constitute such relations for humans and, as we have seen, perception is often presented as a paradigm. Independently of considerations having to do with AI, we know

that this is too anthropocentric: surely there can be creatures that refer using, say, 'Lucie' to refer to Lucie, just as we do, but don't perceive the world the way we do or, more generally, gain knowledge about the world in ways completely different from us. Maybe they rely entirely on telepathy. Maybe they gain knowledge in ways we have never thought of or cannot fully understand. It would be parochial to stipulate that such creatures cannot use 'Lucie' to refer to Lucie.

The kinds of AI that we are discussing are, to a significant extent, alien. In particular, the algorithms governing neural networks will do things we don't or can't fully understand. As a result, the epistemically rewarding relation will look different from the human paradigms.

What we need to do is familiar by now: the models developed by mental file theorists must be de-anthropocentrized and applied to AI systems like SmartCredit. The resulting theory should satisfy the two core elements of our positive theory:

- It will be externalist (because the relation between a file and the object the file is about is external to the speaker).
- It will be developed by abstracting away from anthropocentric components of existing theories.

Note that when Recanati and other mental file theorists describe epistemically rewarding relations that human speakers rely on, they do so in rather abstract terms. They don't, for example, go into the details of how particular perceptual systems like smell or touch work (and end up being epistemically rewarding). They are not even particularly clear on just what counts as 'rewarding' or 'epistemic'. It's just assumed that, say, perception is a paradigm of

something that's an epistemically rewarding relation. Similarly, it is not to be expected that our theory goes into great detail of all the various epistemically rewarding and reference determining relations that AIs stand in to objects. What we have to say here will be at a fairly high level of abstraction.

The first step is to appeal to the meta-metasemantic principle of knowledge maximization from Chapter 4. According to this principle, the relevant epistemically rewarding relation should be knowledge maximizing. The most natural way to do that is to build knowledge maximization directly into the account of 'epistemically rewarding'. Gray cites the following passage from Williamson:

> A causal relation to an object (property, relation, . . .) is a channel for reference to it only if it is a channel for the acquisition of knowledge about the object (property, relation).
>
> (Williamson 2007, 264, ellipses in original)

The more detailed story then is an answer to the following question: which specific such relation maximizes the interpreter's knowledge if used to determine reference? The answer to this will to a large extent vary between AI systems. It will depend on the details of how the system works, how it was created, and how it is used.

In saying this, we don't mean to be defeatist about a general theory. Some of the material from the previous chapter is directly relevant here. In Chapter 5, we developed an account of anchoring for predicates that relied on the idea that a property anchors *a training process*. In many cases, elements of the training process will also be important in understanding the epistemically rewarding relation to an object that determines reference of singular terms. Here it is natural to appeal to some of the central notions developed by

Gareth Evans (one of the early proponents of mental file theory). According to Evans, the denotation of the use of a term T is fixed by what he called the 'dominant source' of the information associated with T. This was the way Evans developed Kripke's causal theory. He agreed with Kripke that reference is not fixed by descriptive fit, i.e. not fixed by what the associated descriptions pick out. However, he diverged from Kripke in giving the associated descriptions a reference-determining role: not through fit, but through an external causal relation. The notion of a dominant causal source is now at the centre of the theory and will need further elaboration. However, for our purposes, we will simply use this schematic idea and apply it to AI. The schematic idea is this:

> *Denotation through dominant causal source:* A system can denote an object that is the dominant causal source of a set of information given as input in the training stage.

If we return to our simple case of SmartCredit and Lucie, the initial stage involves information being fed into the system. Call the conjunction of that, C. Think of C as a mental file. There will be an object that is the dominant causal source of C. If things go well, this will be Lucie. If so, SmartCredit can refer to Lucie through Lucie being the dominant causal source of C. If so, it is correct to describe the output of the system as being of the form: *Lucie is high risk* (and not just: 'Someone satisfying C is high risk').

## Case Studies, Complications, and Reference Shifts

There are several concerns about Evans' theory. Central among these is that we need more clarity in what counts as a dominant

source of a body of information. That it is dominance that determines reference means that not all of the associated information needs to have the referent as a source: there can be misinformation mixed in that doesn't have the referent as a source. Evans is also clear that dominance is not simply a matter of quantity: it is not a matter of a simplistic counting information and then locating the source of the majority. Some of the information is more heavily weighted than others. Evans is also clear that over time, dominance can shift. He illustrates this with the example of 'Turnip' (1973: 306). The example involves a youth, A, with the nickname 'Turnip'. He leaves his village at an early age. Many years later, a different person, B, settles in the village. The old villagers believe that A has returned and refers to B using 'Turnip'. At first, the dominant causal source associated with 'Turnip' will be the A. Over time, however, this can change: the dominant source of information can be shifted from A to B. In Evans' example, it's easy to see how that can happen: as villagers see more and more of B, the file associated with 'Turnip' will gradually fill up with more significant information that has B as its source. The information that has A has its source will gradually fade into relative insignificance. The possibility of a name shifting referent over time was one of the central motivations for Evans' theory. Evans argued that Kripke's theory couldn't account for such reference shifts and that his alternative could do so easily.

We mention this because the kind of flexibility that Evans' version of the mental file theory provides can be useful for interpreting AI. First, it should be possible for the initial data to contain what we would naturally classify as *misinformation* about A. One way this could happen: some of the descriptive material, D, in the A-file has another object, B, as its source. It correctly characterizes

B, but fails to describe A. As long as B isn't the dominant source of the information in the file, the file as a whole can have A as its denotation. It simply contains the misinformation that A is D. Moreover, gradual shifts can happen as follows: the information we feed the system can initially have Lucie as its dominant source. As information is added over time, the dominant source can shift to another person. This can happen in two ways:

(i) What counts as dominant information can shift even if the total amount of information doesn't shift. This can happen even as the informational content of a file in an AI system is stable.
(ii) Information can be added to the file. This can change the dominant causal source and that again can result in reference shift for the file as a whole.

Whether (i) is an option will depend on how dominance is understood. As we see it, dominance is unlikely to be understood independently of human interests. What counts as dominance is not an objective feature that can be read off the world independently of what interpreters care about. If that is right, then a gradual change in what we care about can result in a change in what the file refers to. Note that these issues about how to understand dominance comprise a meta-metasemantic question. Our guiding principle is knowledge-maximization and it tells us that dominance should be construed in a knowledge-maximizing way.

Many cases will not fit cleanly into either of (i) or (ii) above. Consider the following: suppose we have an AI system that is set up to make assessments of economic health. Initially, it is fed information about the economic situation in the US. It starts

giving outputs of the form 'The economic outlook is G'. Since the US is the dominant source of the information the system is operating on, we should interpret this to mean that the economic outlook in the US is G. Doing so is knowledge-maximizing. Now suppose that over time, the system starts to focus more specifically on the input of economic data from California (it is still using the entire data set, but changes its focus to California). This could happen for several reasons: maybe that turns out to be the data that's most predictive or that its algorithms can do the most with. It keeps producing outputs of the form 'Economy is doing great' or 'Economy is doing poorly'. Now we ask: *which* economy does it refer to, the US, or just California, or something else?

The general form of an answer to that is guided by the general principle that it refers to whatever place it's in an epistemically rewarding relation to. Our knowledge-maximizing principle tells us to construe 'epistemically rewarding' as a relation that is knowledge maximizing for us as interpreters/users. However, just saying that leaves us with a range of possible outcomes. Here is one possible outcome: it might turn out that its predictions are more accurate about California, although it does well enough for the US as a whole. If so, there are competing considerations for how to understand 'maximizing'. It will matter what we care the most about: precise knowledge or maximally general knowledge? The answer to that, again, might depend on what we are going to do with this information and how we use it to generate further knowledge. This case shows how the metasemantics, guided by knowledge-maximization, will be sensitive to our interests and activities in ways that are hard to predict.

Next consider an AI system used by law enforcement to help determine who is guilty of various crimes. It's in effect an AI-detective and it tells about the degree of guilt of various subjects. Initially, we feed it information about Lucie's activities. After processing these, it outputs 'Innocent', if it finds no violations of any laws. If Lucie has had a couple of speeding tickets, the output is 'Guilty of minor traffic violations'. The referential issue is fairly simple since Lucie remains the dominant source of information throughout. Here is a more complicated case: we start to feed it data about a supposed mob family. The data includes a broad range of actions performed by several people over a period of time. This makes the referential question more difficult to settle. We need to decide whether it is tracking guilt of the *organization*, or guilt of *specific people in the organization*. After all, if the organization is guilty, that is the result of various crimes committed by the members of the crime family. In tracking the family, it is tracking the individuals. We can use the knowledge-maximization principle to help us adjudicate: is the epistemically rewarding relation, i.e. the knowledge-maximizing relation, one that leads to the family or to particular individuals? As in the previous case, that might depend on our interests and what we end up doing with that knowledge. It could, for example, depend on how we use it to generate new knowledge.

What both these cases are meant to illustrate is that the correct externalist interpretations can be sensitive to variations in our epistemic goals and practices. This follows quite naturally from the general principle of aiming for a metasemantics that maximizes knowledge. The measure of maximization can vary in various ways and the two cases illustrate that.

## Taking Stock

Here is what we have done:

1. We started with an outline of the mental file metasemantic framework.
2. We observed that the details of this metasemantics aren't straightforwardly applicable to AI systems—in particular, we need to abstract on the notion of an epistemically rewarding relation.
3. We took some initial steps toward de-anthropocentrizing, proposing an AI-friendly version of epistemically rewarding.
4. Finally, we outlined some choice points for that theory, using Evans' notion of a dominant source of information.

As in the case of predicates discussed in Chapter 5, this proposal is schematic, though not more so than the theories it is modelled on. It reinforces the conclusion from Chapter 4: the standard internalist approaches to AI have significant limitations. We cannot discover all the facts about the contents of the machine learning system's classifications simply by looking at the internal programming implementation of those systems. What discussion in this chapter concludes is that we need to focus in particular on the interpreters' epistemic goals and activities. What the AI system is about can vary depending on the interpreter's aims.