3

TERMINOLOGY

Aboutness, Representation, and Metasemantics

We didn't refute Alfred's content-scepticism, but at least we got him to agree to explore strategies for attributing content to AI. That's the goal of the rest of this book. As we pointed out to Alfred, the most salient prima facie argument for doing that is that AI is presented to us, by its producers, as having content. AI systems are presented as *saying* things, as *making suggestions*, and sometimes *making decisions*. We, the human users, typically treat them as conversation partners and as sources of information (contentful information, that is).

There should be no disagreement about the fact AI systems are often (and arguably typically) presented to end-users in this way. There are indefinitely many illustrations of this from academic work to advertisements. Here's a tiny collection of the kinds of claims we have in mind (all emphasis ours). From academic papers:

> SmartBot can fall into false or impossible *beliefs*. For example, SmartBot can *believe* one of its cards has no valid value as all possible cards are inconsistent with the observed play according to SmartBot's convention. (Bard et al 2019)[1]

[1] https://arxiv.org/pdf/1902.00506v1.pdf.

Making AI Intelligible: Philosophical Foundations. Herman Cappelen and Joshua Dever, Oxford University Press (2021).
 DOI: 10.1093/oso/9780192894724.003.0003

> Robots must *know how* to be gentle when they need to interact with fragile objects, or when the robot itself is prone to wear and tear.
>
> (Huang et al. 2019)

> For a given set of desired performance measures, i.e. cycle time, work-in progress, and utilisation of three different testers, the neural network *suggests* a suitable design of scheduling rules, and the number of each type of tester needed to achieve management's goal. (Alam et al. 2004)

> It is W [weights linking nodes] that constitutes what the network *knows* and determines how it will respond to any arbitrary input from the environment. (Tam 1991)

From more technical computer science blogs:

> Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to *recognize* patterns. They *interpret* sensory data through a kind of machine perception, labeling or clustering raw input. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or time series, must be translated…

> A neural network is a corrective feedback loop, rewarding weights that support its correct *guesses*, and punishing weights that lead it to err. (Nicholson, skymind.ai)[2]

From a more or less general interest computer science blog:

> DeepXmas: AI *knows* if you are naughty or nice…

> This AI home security system can use deep-learning and *figure out* when kids are making messes, or doing things that need action. This type of technology could save a life in the future (e.g. kid choking on a blind cord). ("We don't want the AI to just become a person or kid detector, we want it to understand *naughtiness*.")[3]

[2] https://skymind.ai/wiki/neural-network.

[3] https://towardsdatascience.com/deepxmas-ai-knows-if-you-are-naughty-or-nice-2bd00b2ad3d2?gi=f35896d7ff04.

Loose Talk, Hyperbole, or 'Derived Intentionality'?

Recall from the previous chapter that Alfred dismissed all of this as *loose talk* or *hyperbole*. He thinks it is useful, but fundamentally misleading jargon. Maybe it is not just misleading, but also false. According to the sceptic, AI can't, literally speaking, perform speech acts, nor can AI think, believe, or have any of the mental states that we humans have. The sceptic says we find it natural to say things like:

> The calculator *says* that 87x9= 783,

but calculators don't really say anything. Our theory of what it is to perform the speech act of saying need not account for calculator-speech. That ordinary speech is filled with false anthropocentric descriptions of calculators shouldn't mislead philosophers. The same goes for talk about the kinds of advanced AI we focus on in this book.

It is very important to emphasize that we don't take ourselves to have refuted that kind of view. Our goal is a more modest one: as a counterbalance to that dismissive view, we will consider some first steps towards making (some) content attributions true. The best way to work out an alternative to the no-content view is actually to try to work out some of the details of an alternative.

Here is a reason for being a bit interested in our effort. Start by asking why we take content talk more seriously for people than we do for calculators. It's because people act in lots of complicated ways that the content attributions help make sense of, while the calculator doesn't really have complex actions (a purely physical account of what's going on with circuits in the calculator lets us understand what we want to understand). But the AI systems that

we are considering also act in very complicated ways that aren't illuminated by looking at what's going on with the microstructure. It is particularly interesting that their complicated forms of behaviour aren't the same as the human complicated forms of behaviour. This is what we need to engage in: what we below will call *anthropocentric abstraction*.

None of this amounts to a conclusive proof that certain forms of AI can perform certain forms of speech acts. However, note that if, after reading this book, you end up finding our efforts unconvincing, you'll do so because you have philosophical arguments against what we say. You have in effect used theories about the metaphysics and methodology of content attribution to help you understand AI and our interactions with AI. That's support for one of our central messages: there should be more interaction between theories about the metaphysics of content and theories of AI.

Aboutness and Representation

Compare two things. On the one hand, a sock—say, the left sock worn by Alan Turing when he started writing his famous paper on computability—and on the other, this sentence:

(1) The Eiffel Tower is in Paris

The sock is not about anything. It just exists. It can be worn, washed, and mended, but it doesn't represent anything. The sentence 'The Eiffel Tower is in Paris', on the other hand, *is* about something—it exhibits what philosophers imaginatively call 'aboutness'. English speakers have an easy time identifying what it

is about: most obviously, it is about Paris and the Eiffel Tower. It's about those two objects because 'Paris' is the name of Paris and 'the Eiffel Tower' is the name of the Eiffel Tower. However, it is not just about those two objects. It is also about the latter being located in the former. On one view, for example, there's something called a fact or a situation, *that the Eiffel Tower is in Paris*, that the sentence represents.[4]

A striking fact about (1) is that by virtue of its representational properties, it is the kind of thing that can be true or false.[5] If the Eiffel Tower is in Paris, then (1) is true. If the Eiffel Tower is not in Paris, then (1) is false. As it happens, it has the property of being true. Another way to put this: (1) represents the world as being a certain way. If the world is that way, then (1) is true. If the world is not that way, then (1) is false. The world is that way, so it is true. The sock, on the other hand, cannot be true or false. It doesn't represent the world as being any way at all.

The phenomenon of aboutness is so familiar to us (and so central to our lives) that it is easy to forget or overlook how amazing it is. We just used the word 'Paris', sitting in Oslo, and somehow that word manages to 'reach' all the way to a physical structure 1,555 km away (and does so without a passport or a plane ticket). Somehow the word 'Paris' connects with Paris. Not only can aboutness cross space in a seemingly mysterious way, it can also cross time. The expression 'Emperor Kanmu' denotes a Japanese emperor who lived more than 1,000 years ago. Just having read the previous sentence, you, our reader, can now use the expression 'Emperor Kanmu' to *talk about Kanmu*.

[4] For introductory work on the metaphysics of facts, see Armstrong (1997) and Mulligan (2007).

[5] We sidestep issues as to what the most fundamental 'truth-bearers' are because they aren't immediately relevant.

A sentence like (1) is an artefact. It consists of objects that we have constructed, i.e. words. But it is not only artefacts that have aboutness. We humans can believe, think, hope, fear, expect, conjecture, etc. In so doing, our minds are directed at features of the world in much the same way sentences of a language are. Just as you can say, in English, that the Eiffel Tower is in Paris, so you think, or believe, or hope, or fear that the Eiffel Tower is in Paris. In the nineteenth century, the philosopher Franz Brentano (Brentano 1874) introduced (although cognates, meaning similar things, had already existed in, for example, Latin) the term 'intentionality' to denote this ability of the human mind to represent. In the current literature, the terms 'intentionality', 'representation', and 'aboutness' are often used interchangeably (though in some theoretical contexts they are distinguished). In what follows we'll for the most part use 'representation', sometimes 'aboutness', and leave 'intentionality' behind.

AI, Metasemantics, and the Philosophy of Mind

There is a vast literature spanning many subdisciplines of philosophy that attempts to give an account of what representation amounts to and how it comes about. For more than 100 years, philosophers have been concerned with questions such as:

- By virtue of what can a *sentence of English*, say (1) above, be about the Eiffel Tower?
- By virtue of what is *the thought that the Eiffel Tower* is in Paris about Paris?
- What is the connection between the answer to those two questions?

When applied to language, these kinds of questions are often described as parts of *metasemantics*.[6] When the focus is on the intentionality of the human mind, the relevant literature is often classified as philosophy of mind.

To put this book into perspective, it's worth noting that most of the contributions to the philosophy of AI have drawn on work done in the philosophy of mind. It has not been based on work done in the philosophy of language (or the intersection of philosophy of language and philosophy of mind), and not paid any heed to the externalist tradition in the philosophy of language that is the theoretical foundation both of this work and of much of the most important work in twentieth-century philosophy of language and mind. It is hard to prove a negative, but the reader could look at the bibliographies of recent overview works—of Bringsjord and Govindarajulu (2018) on the philosophy of AI, or—perhaps more pertinently—Bruckner (2019) for the philosophy of deep learning. We see literally no works of philosophy of language there. The same thing applies to an extensive review of social sciences (including philosophy) on explainability and AI (Miller 2018). The closest engagement we've found with philosophy of language is in the monograph Floridi (2011), but in that work there is no Burge, no Kripke, no Putnam, no externalism. In the same vein, the otherwise excellent book *How to Build a Brain* (Eliasmith 2013) is

[6] See, for example, Brentano (1911) and Crane (1998) for people who take aboutness to be fundamental to theory of mind, the language-of-thought theorists like Fodor (1975), and the teleosemanticists like Dretske (1980) and Millikan (1984). Classic works on the representational properties of language include Russell (1905), Strawson (1950), Kripke (1980) (discussed at length below), Donnellan (1966), and Evans (1982). More recent work that shows these issues remain live concerns includes Recanati (2012) and Hawthorne and Manley (2012), and a textbook introduction that brings one right up to date on active issues in the field is Cappelen and Dever (2018).

written as if the externalist tradition doesn't exist.[7] Other recent work approaches the topic from a functionalist perspective (López-Rubio 2018); from a Kantian perspective (Schubbach forthcoming); from a teleosemanticist perspective (Shea 2018) or using sui generis theoretical tools (Floridi, the book just mentioned); through the lens of more venerable philosophical conceptions of abstraction as found among the British empiricists and their followers (Buckner 2018); or from the perspective of modern compositional semantic theories (Nedft 2020). In all this we find no mention of the externalist tradition, a strange gap in the literature we're aiming here to fill.

[7] There is of course significant work on the extended mind hypothesis (Clark and Chalmers 1998) and while this is a form of externalism, it is not one based in the Kripke, Burge, Putnam tradition. There is also work on what is called 'embodied embedded cognition', and while this could also be called a form of externalism, it is entirely different from the tradition we are relying on here.