1

INTRODUCTION

The Goals of This Book: The Role of Philosophy in AI Research

This is a book about some aspects of the philosophical foundations of Artificial Intelligence. Philosophy is relevant to many aspects of AI and we don't mean to cover all of them.[1] Our focus is on one relatively underexplored question: Can philosophical theories of meaning, language, and content help us understand, explain, and maybe also improve AI systems? Our answer is 'Yes'. To show this, we first articulate some pressing issues about how to interpret and explain the outputs we get

[1] Thus we are not going to talk about the consequences that the new wave in AI might have for the empiricism/rationalism debate (see Buckner 2018), nor are we going to consider—much—the question of whether it is reasonable to say that what these programs do is 'learning' in anything like the sense with which we are familiar (Buckner 2019, 4.2), and we'll pass over interesting questions about what we can learn about philosophy of mind from deep learning (López-Rubio 2018). We are not going to talk about the clearly very important ethical issues involved, either the recondite ones, science-fictional ones (such as the paperclip maximizer and Roko's Basilisk (see e.g. Bostrom 2014 for some of these issues)), or the more down-to-earth issues about, for example, self-driving cars (Nyholm and Smids 2016, Lin et al. 2017), or racist and sexist bias in AI resulting from racist and sexist data sets (Zou and Schiebinger 2018). We also won't consider political consequences and implications for policy making (Floridi et al. 2018).

Making AI Intelligible: Philosophical Foundations. Herman Cappelen and Joshua Dever, Oxford University Press (2021).
 DOI: 10.1093/oso/9780192894724.003.0001

from advanced AI systems. We then use philosophical theories to answer questions like the above.

An Illustration: Lucie's Mortgage Application is Rejected

Here is a brief story to illustrate how we use certain forms of artificial intelligence and how those uses raise pressing philosophical questions:

> Lucie needs a mortgage to buy a new house. She logs onto her bank's webpage, fills in a great deal of information about herself and her financial history, and also provides account names and passwords for all of her social media accounts. She submits this to the bank. In so doing, she gives the bank permission to access her credit score. Within a few minutes, she gets a message from her bank saying that her application has been declined. It has been declined because Lucie's credit score is too low; it's 550, which is considered very poor. No human beings were directly involved in this decision. The calculation of Lucie's credit score was done by a very sophisticated form of artificial intelligence, called SmartCredit. A natural way to put it is that this AI system *says* that Lucie has a low credit score and on that basis, another part of the AI system *decides* that Lucie should not get a mortgage.

It's natural for Lucie to wonder where this number 550 came from. This is Lucie's first question:

> **Lucie's First Question**. What does the output '550' that has been assigned to me *mean*?

The bank has a ready answer to that question: the number 550 is a credit score, which represents how credit-worthy Lucie is. (Not very, unfortunately.) But being told this doesn't satisfy Lucie's

unease. On reflection, what she really wants to know is *why* the output means that. This is Lucie's second question:

> **Lucie's Second Question**: Why is the '550' that the computer displays on the screen an assessment of my credit-worthiness? What *makes* it mean that?

It's then natural for Lucie to suspect that answering this question requires understanding how SmartCredit works. What's going on under the hood that led to the number 550 being assigned to Lucie? The full story gets rather technical, but the central details can be set out briefly:

> *Simple Sketch of How a Neural Network Works*[2]
>
> SmartCredit didn't begin life as a credit scoring program. Rather, it started life as a general neural network. Its building blocks are small 'neuron' programs. Each neuron is designed to take a list of input data points and apply some mathematical function to that list to produce a new output list. Different neurons can apply different functions, and even a single neuron can change, over time, which function it applies.
>
> The neurons are then arranged into a network. That means that various neurons are interconnected, so that the output of one neuron provides part of the input to another neuron. In particular, the neurons are arranged into layers. There is a top layer of neurons—none of these neurons are connected to each other, and all of them are designed to receive input from some outside data source. Then there is a second layer. Neurons on the top layer are connected to neurons on the second layer, so that top layer neurons

[2] For a gentle and quick introduction to the computer science behind basic neural networks, see Rashid 2016. A relatively demanding article-length introduction is LeCun et al. 2015, and a canonical textbook that doesn't shirk detail and is freely available online is Goodfellow et al. 2016.

provide inputs to second layer neurons. Each top layer neuron is connected to every second layer neuron, but the connections also have variable weight. Suppose the top layer neurons T1 and T2 are connected to second layer neurons S1 and S2, but that the T1-to-S1 connection and the T2-to-S2 connections are weighted heavily while the T1-to-S2 connection and the T2-to-S1 connections are weighted lightly. Then the input to S1 will be a mixture of the T1 and T2 outputs with the T1 output dominating, while the input to S2 will be a mixture of the T1 and T2 outputs with the T2 output dominating. And just as the mathematical function applied by a given neuron can change, so can the weighting of connections between neurons.

After the second layer there is a third layer, and then a fourth, and so on. Eventually there is a bottom layer, the output of which is the final output of SmartCredit. The bottom layer of neurons is designed so that that final output is always some number between 1 and 1000.

The bank offers to show Lucie a diagram of the SmartCredit neural network. It's a complicated diagram—there are 10 levels, each containing 128 neurons. That means there are about 150,000 connections between neurons, each one labelled with some weight. And each neuron is marked with its particular mathematical transformation function, represented by a list of thousands of coefficients determining a particular linear transformation on a thousands-of-dimensions vector.

Lucie finds all of this rather unilluminating. She wonders what any of these complicated mathematical calculations has to do with why she can't get a loan for a new house. The bank continues explaining. So far, Lucie is told, none of this information about the neural network structure of SmartCredit explains why it's evaluating Lucie's creditworthiness. To learn about that, we need to consider the neural network's training history.

A bit more about how SmartCredit was created

Once the initial neural network was programmed, designers started training it. They trained it by giving it inputs of the sort that Lucie has also helpfully provided. Inputs were thus very long lists of data including demographic information (age, sex, race, residential location, and so on), financial information (bank account balances, annual income, stock holdings, income tax report contents, and so on), and an enormous body of social media data (posts liked, groups belonged to, Twitter accounts followed, and so on). In the end, all of this data is just represented as a long list of numbers. These inputs are given to the initial neural network, and some final output is produced. The programmers then evaluate that output, and give the program a score based on how acceptable its output was that measures the program's error score. If the output was a good output, the score is a low score; if the output was bad, the score is a high score. The program then responds to the score by trying to redesign its neural network to produce a lower score for the same input. There are a number of complicated mathematical methods that can be used to do the redesigning, but they all come down to making small changes in weighting and checking to see whether those small changes would have made the score lower or higher. Typically, this then means that a bunch of differential equations need to be solved. With the necessary computations done, the program adjusts its weights, and then it's ready for the next round of training.

Lucie, of course, is curious about where this scoring method came from—how do the programmers decide whether SmartCredit has done a good job in assigning a final output to input data?

The Scoring Method

The bank explains that the programmers started with a database of millions of old credit cases. Each case was a full demographic, financial, and social media history of a particular person, as well as a credit score that an old-fashioned human credit assessor had assigned to that person. SmartCredit was then trained on that data

set—over and over it was given inputs (case histories) from the data set, and its neural network output was scored against the original credit assessment. And over and over SmartCredit reweighted its own neural network trying to get its outputs more and more in line with the original credit assessments.

That's why, the bank explains, SmartCredit has the particular collections of weights and functions that it does in its neural network. With a different training set, the same underlying program could have developed different weights and ended up as a program for evaluating political affiliation, or for determining people's favourite movies, or just about anything that might reasonably be extracted from the mess of input social media data.

Lucie, though, finds all of this a bit too abstract to be very helpful. What she wants to know is why *she*, in particular, was assigned a score of *550*, in particular. None of this information about the neural architecture or the training history of SmartCredit seems to answer that question.

How all this applies to Lucie

Wanting to be helpful, the bank offers to let Lucie watch the computational details of SmartCredit's assessment of Lucie's case. First they show Lucie what the input data for her case looks like. It's a list of about 100,000 integers. The bank can tell Lucie a bit about the meaning of that list—they explain that one number represents the number of Twitter followers she has, and another number represents the number of times she has 'liked' commercial postings on Facebook, and so on.

Then they show Lucie how that initial data is processed by SmartCredit. Here things become more obscure. Lucie can watch the computations filter their way down the neural network. Each neuron receives an input list and produces an output list, and those output lists are combined using network weightings to produce inputs for subsequent neurons. Eventually, sure enough, the number '550' drops out of the bottom layer.

But Lucie feels rather unilluminated by that cascading sequence of numbers. She points to one neuron in the middle of the network and to the first number (13,483) in the output sequence of that neuron. What, she asks, does that particular number mean? What is *it* saying about Lucie's credit worthiness? This is Lucie's third question:

> **Lucie's Third Question**: How is the final meaningful state of SmartCredit (the output '550', meaning that Lucie's credit score is 550) the result of other meaningful considerations that SmartCredit is taking into account?

The bank initially insists that that question doesn't really have an answer. That particular neuron's output doesn't by itself mean anything—it's just part of a big computational procedure that holistically yields an assessment of Lucie's credit worthiness. No particular point in the network can be said to mean anything in particular—it's the network as a whole that's telling the bank something.

Lucie is understandably somewhat sceptical at this point. How, she wonders, can a bunch of mathematical transformations, none of which in particular can be tied to any meaningful assessment of her credit-worthiness, somehow all add up to saying something about whether she should get a loan? So she tries a different approach. Maybe looking at the low-level computational details of SmartCredit isn't going to be illuminating, but perhaps she can at least be told what it was in her history that SmartCredit found objectionable. Was it her low annual income that was responsible? Was it those late credit card payments in her early twenties? Or was it the fact that she follows a number of fans of French film

on Twitter? Lucie here is trying her third question again—she is still looking for other meaningful states of SmartCredit that explain its final meaningful output, but no longer insisting that those meaningful states be tied to specific low-level neuron conditions of the program.

Unfortunately, the bank doesn't have much helpful to say about this, either. It's easy enough to spot particular variables in the initial data set—the bank can show her where in the input her annual income is, and where her credit card payment history is, and where her Twitter follows are. But they don't have much to say about how SmartCredit then assesses these different factors. All they can do is point again to the cascading sequence of calculations—there are the initial numbers, and then there are millions upon millions of mathematical operations on those initial numbers, eventually dropping out a final output number. The bank explains that that huge sequence of mathematical operations is just too long and complicated to be humanly understood—there's just no point in trying to follow the details of what's going on. No one could hold all of those numbers in their head, and even if they could, it's not clear that doing so would lead to any real insight into what features of the case led to the final credit score.

Abstraction: The Relevant Features of the Systems We Will be Concerned with in This Book

Our concern is not with any particular algorithm or AI systems. It is also not with any particular way of creating a neural network. These will change over time and the cutting edge of programming today will seem dated in just a year or two. To identify what we

will be concerned with, we must first distinguish two levels at which an AI system can be characterized:

- On the one hand, it is an abstract mathematical structure. As such it exists outside space and time (it is not located anywhere, has no weight, and doesn't start existing at any particular point in time).
- However, when humans use and engage with AI, they have to engage with something that exists as a physical object, something they can see or hear or feel. This will be the **physical implementation** (or **realization**) of the abstract structure. When Lucie's application was rejected, the rejection was presented to her as a token of numbers and letters on a computer screen. These were physical phenomena, generated by silicon chips, various kinds of wires, and other physical things (many of them in different locations around the world).

This book is not about a particular set of silicon chips and wires. It is also not about any particular program construed as an abstract object. So we owe you an account of what the book is about. Here is a partial characterization of what we have in mind when we talk about 'the outputs of AI systems' in what follows:[3]

- The output (e.g. the token of '550' that occurs on a particular screen) is produced by things that are not human. The non-human status of the producer can matter in at least three ways:

First, these programs don't have the same kind of physical implementation as our brains do. They may use 'neurons', but their

[3] This is not an effort to specify necessary and sufficient conditions for being an AI system—that's not a project we think is productive or achievable.

neurons are not the same kind of things as our neurons—they differ of course physically (being non-biological), but also computationally (they don't process inputs and produce outputs in the same way as our neurons). And their neurons are massively different in number and arrangement from our neurons, and massively different in the way they dynamically respond to feedback.

Second, these programs don't have the same abilities as we do. We have emotional repertoires and sensory experiences they lack, and arguably have beliefs, desires, hopes, and fears that they also lack. On the other hand, they have computational speeds and accuracies that we lack.

Third, these programs don't have the same histories that we do. They haven't had the kind of childhoods we have had, and in particular haven't undergone the same experiences of language acquisition and learning that we have. In short, they are non-human (where we will leave the precise characterization of this somewhat vague and open-ended).

- When we look under the hood—as Lucie did in the story above—what we find is not intelligible to us. It's a black box. It will operate in ways that are too complex for us to understand. It's important to highlight right away that this particular feature doesn't distinguish it from humans: when you look under the hood of a human, what you will find is brain tissue—and at a higher level, what looks like an immensely complex neutral network. In that sense, the human mind is also a black box, but as we pointed out above, the physical material under the hood/skull is radically different.
- The systems we are concerned with are made by human programmers with their own beliefs and plans. As Lucie saw, understanding SmartCredit requires looking beyond the program itself to the way that the program was trained. But the training was done by people, who selected an initial range of data, assigned target scores to those initial training cases based on their own plans for what the program should track, and created specific dynamic methods for the program to adjust its neural network in the face of training feedback.

- The systems we are concerned with are systems that are intended to play a specific role, and are perceived as playing that role. SmartCredit isn't just some 'found artefact' that's a mysterious black box for transforming some numbers into other numbers. It's a program that occupies a specific social role: it was designed specifically to assign credit scores, and it's used by banks because it's perceived as assigning credit scores. It's treated as useful, as producing outputs that really are meaningful and helpful credit scores, and it becomes entrenched in the social role it occupies because it's perceived as useful in that way.

None of this adds up to a complete metaphysics of AI systems. That's not the aim of this book. Instead, we hope it puts readers in a position to identify at least a large range of core cases.

The Ubiquity of AI Decision-Making

SmartCredit raises concerns about what its outputs mean. But SmartCredit is only the tip of the iceberg. We are increasingly surrounded by AI systems that use neural network machine learning methods to perform various sorts of classifications. Image recognition software classifies faces for security purposes, tags photographs on social media, performs handwriting analysis, guides military drones to their targets, and identifies obstacles and street signs for self-driving cars. But AI systems of this sort aren't limited to simple classification tasks. The same underlying neural network programming methods give rise, for example, to strategic game-playing. Google's AlphaZero has famously achieved superhuman levels of performance in chess, Go, and Shogi. Other machine learning approaches have been applied to a wide variety of games, including video games such as Pac-Man, Doom, and

Minecraft.[4] Other AI systems perform variants of the kind of 'expert system' recommendation as SmartCredit. Already there are AI systems that attempt to categorize skin lesions as cancerous or not, separate spam emails and malware from useful emails, determine whether building permits should be granted and whether prisoners should receive parole, figure out whether children are being naughty or nice using video surveillance, and work out people's sexual orientations from photographs of their faces. Other AI systems use machine learning to make predictions. For example, product recommendation software attempts to extrapolate from earlier purchases to likely future purchases, and traffic software attempts to predict future locations of congestion based on earlier traffic conditions. Machine learning can also be used for data mining, in which large quantities of data are analysed to try to find new and unexpected patterns. For example, the data mining program Word2Vec extracted from a database of old scientific papers new and unexpected scientific conclusions about thermoelectric materials.

These AI systems are able to perform certain tasks at extraordinarily high levels of precision and accuracy—identifying certain patterns much more reliably, and on the basis of much noisier input, than we can, and making certain kinds of strategic decisions with much higher accuracy than we can—and both their sophistication and their number are rapidly increasing. We should expect that in the future many of our interactions with the world will be mediated by AI systems, and many of our current intellectual activities will be replaced or augmented by AI systems.

[4] See https://www.sciencenews.org/article/ai-learns-playing-video-games-starcraft-minecraft for some discussion about the state and importance of AI in gaming.

Given all that, it would be nice to know what these AI systems mean. That means we want to know two things. First, we want to know what the AI systems mean with their explicit outputs. When the legal software displays the word 'guilty', does it really *mean* that the defendant is guilty? Is guilt really what the software is tracking? Second, we want to know what contentful states the AI systems have that aren't being explicitly revealed. When AlphaZero makes a chess move, is it making it for reasons that we can understand? When SmartCredit gives Lucie a credit score of 550, is it weighing certain factors and not others?

If we can't assign contents to AI systems, and we can't know what they mean, then we can't in some important sense understand our interactions with them. If Lucie is denied a loan by SmartCredit, she wants to understand why SmartCredit denied the loan. That matters to Lucie, both practically (she'd like to know what she needs to change to have a better chance at a loan next time) and morally (understanding why helps Lucie not view her treatment as capricious). And it matters to the bank and to us. If we can't tell why SmartCredit is making the decisions that it is, then we will find it much harder to figure out when and why SmartCredit is making its occasional errors.

As AI systems take on a larger and larger role in our lives, these considerations of understanding become increasingly important. We don't want to live in a world in which we are imprisoned for reasons we can't understand, subject to invasive medical conditions for reasons we can't understand, told whom to marry and when to have children for reasons we can't understand. The use of AI systems in scientific and intellectual research won't be very productive if it can only give us results without explanations (a neural network that assures us that the ABC conjecture is true

without being able to tell us *why* it is true isn't much use). And things are even worse if such programs start announcing scientific results using categories that we're not sure we know the content of.

We are in danger, then, of finding ourselves living in an increasingly meaningless world. And as we've seen, it's a pressing danger, because if there is meaning to be found in the states and activities of these AI systems, it's not easily found by looking under the hood and considering their programming. Looking under the hood, all we see is jumbles of neurons passing around jumbles of numbers.

But at the same time, there's reason for optimism. After all, if you look under *our* hoods, you also see jumbles of neurons, this time passing around jumbles of electrical impulses. That hasn't gotten in the way of our producing meaningful outputs and having meaningful internal states. The hope then is that reflecting on how *we* manage to achieve meaning might help us understand how AI systems also achieve meaning.

However, we also want to emphasize that it's a guarded hope. Neural network programs are a little like us, but only a little. They are also very different in ways that will come out in our subsequent discussion. Both philosophy and science fiction have had an eye from time to time on the problem of communicating with and understanding aliens, but the aliens considered have never really been all that alien. In science fiction, we get the alien language in Star Trek's Darmok,[5] which turns out to be basically English with more of a literary flourish, the heptapod language of 'Story of Your Life',[6] which uses a two-dimensional syntax to

[5] See *Star Trek: The Next Generation*, season 5 episode 2.

[6] In Chiang, *Stories of Your Life And Others*, Tor Books, 2002. The book was the inspiration for the film Arrival.

present in a mildly encoded way what look like familiar contents, and the Quintans of Stanislaw Lem's 1986 novel *Fiasco*, who are profoundly culturally incomprehensible but whose occasional linguistic utterances have straightforward contents. In philosophy, consideration of alien languages either starts with the assumptions that the aliens share with us a basic cognitive architecture of beliefs, desires, reasons, and actions, or (as Davidson does) concludes that if the aliens aren't that much like us, then whatever they do simply can't count as a language.

Our point is that the aliens are already among us, and they're much more alien than our idle contemplation of aliens would have led us to suspect. Not only that, but they are *weirdly* alien—we have built our own aliens, so they are simultaneously alien and familiar. That's an exciting philosophical opportunity—our understanding of philosophical concepts becomes deeper and richer by confronting cases that take us outside our familiar territory. We want simultaneously to explore the prospect of taking what we already know about how familiar creatures like us come to have content and using that knowledge to make progress in understanding how AI systems have content, and also see what the prospects are for learning how the notions of meaning and content might need to be broadened and expanded to deal with these new cases.

The Central Questions of this Book

Philosophy can help us understand many aspects of AI. There are salient moral questions such as whether we *should* let AI play these important social roles. What are the moral and social

consequences of letting AI systems make important decisions that throughout our history have been made by humans who could be held accountable? There are also pressing questions about whether advanced AI systems could eventually make humans superfluous—this is sometimes discussed under the label 'existential risk' of AI (see Bostrom 2014). None of these is the topic of this book.

The questions we will be concerned with have to do with **how we can interpret and understand the outputs of AI systems**. They are illustrated by the questions that Lucie asked the bank in our little story above. Recall Lucie's first question:

> **Lucie's First Question**: What does the output '550' that has been assigned to me mean?

Lucie's first question is a question about how to understand a specific output of a specific program. We're not going to try to answer Lucie's question, or even to give particular tools for answering this kind of question. But we are interested in the meta-question about whether Lucie's question is a reasonable and important one. We've already observed that AI systems are frequently used *as if* questions like Lucie's made sense and had good answers—we treat these systems *as if* they are giving us specific information about the world. It's thus important to consider whether there is a sensible way to think about these programs on which questions like Lucie's first question could eventually be answered.

This perspective leads to Lucie's second question:

> **Lucie's Second Question**: Why is the '550' that the computer displays on the screen an assessment of my credit-worthiness? What *makes* it mean that?

Our central interest in this book starts with examining what kinds of answers this question could have. If the output states of AI systems do mean something, then surely there must be some *reason* they mean what they do. If we could at least figure out what those reasons are, we might be better positioned down the road to answering Lucie's first question.

The bank tried one particular method of answering Lucie's second question: they directed Lucie to the details of SmartCredit's programming. As we saw, this method wasn't obviously successful—learning all the low-level neural network details of SmartCredit's programming didn't seem to give a lot of insight into why its outputs meant something about Lucie's credit worthiness.

But that was just one method. Our central project is to emphasize that there are many other methods that are worth considering. One way to think about the project is to remember that humans, too, are content-bearing. Our outputs, like SmartCredit's outputs, at least prima facie, mean things and carry information about the world. But looking inside our skulls for an explanation of those contents isn't likely to be much more illuminating than looking inside SmartCredit's programming code was. We emphasized above that programs like SmartCredit are different from people in many important ways, and that's worth keeping in mind (and will guide much of our discussion below). But at the same time, both we and machine-learning programs like SmartCredit are systems producing outputs based on some enormously complicated and not obviously illuminating underlying computational procedure.

That fact about *us*, though, hasn't stopped us from assigning contents to people's outputs, and it hasn't stopped us from

entertaining theories about why people's outputs mean what they do. It's just forced us to consider factors other than neurocomputational implementation in answering that 'why' question. Theories about why human outputs mean what they do have appealed to mental states, to causal connections with the environment, to normative considerations of coherence and charity, to biological teleology, and to relations of social embedding. One of our central projects is then to see whether these kinds of theories can be helpfully deployed in answering Lucie's second question, and how such theories might need to be adapted to accommodate the differences between people and programs.

Lucie had a third question:

> (**Lucie's Third Question**): How is the final meaningful state of SmartCredit (the output '550' meaning that Lucie's credit score is 550) the result of other meaningful considerations that SmartCredit is taking into account?

Eventually we want a good theory of content for AI systems. A good theory of content for people needs to do more than just assign contents to the things we say—it also needs to assign contents to 'hidden' internal states of beliefs and desires, which then help make sense of, and perhaps constrain the contents of, the things we say. We should be open to the possibility that it's the same for AI systems. SmartCredit 'says' some things—it produces explicit outputs of the form of the '550' evaluation it outputs for Lucie. But in making sense of why SmartCredit's explicit outputs have the meanings that they do, we might want to attribute additional contentful states to the program—for example, we might (as Lucie does) want to be able to attribute to SmartCredit various reasons that led it to assign Lucie the credit score that it did.

On a more abstract level: AI systems produce various outputs, and we can always ask what, if anything, makes it the case that an AI system has a certain output; and AI systems produce those outputs for various reasons, and we can ask whether those reasons are contentful reasons (rather than just irreducibly complicated mathematical computations), and what, if anything, makes it the case that the reasons have the contents that they do.

The underlying facts are not in dispute: ML (machine learning) systems are (or consist of) massively complex algorithms that generate an enormous neural network with thousands or millions of interconnected 'neurons'. It is also beyond dispute that in many cases the overall structure and dynamics of that system is too complex for any human to comprehend. A burning question is now: when this system produces an output consisting of an English sentence like the examples given above, how can that output mean what those English words mean? How can we know that it tells us something about what we call creditworthiness?

'Content? That's So 1980'

A central aim in this book is to encourage increased interaction between two groups. First, AI researchers, who are producing machine learning systems of rapidly increasing sophistication, systems that look to have the potential to take on or supplement many of our ordinary processes of reasoning, deciding, planning, and sorting. And second, philosophers, who work in a rich intellectual tradition, which provides tools for thinking about content, tools directed both at determining what features of a system make it contentful (and in what ways) and at characterizing different

kinds of contents with a variety of formal tools. We want to encourage that interaction because we think that AI has a content problem—we need to be able to attribute contents to AI systems, but we're currently poorly positioned to do so.

A certain view of the history of AI research can make all of that seem like a confused retrograde step. AI researchers *tried* approaches centred around content and representation. That's what the symbolic artificial intelligence program was about, that's what led to endless projects focused on a small block world based on clear representational systems. But the wave of contemporary successes in AI has been won by *moving away* from the symbolic approaches. Neural network machine learning systems are deliberately designed *not* to start with a representational system—the whole goal is to allow data that hasn't been pre-processed into representational chunks to be filtered by neural network systems in a way that isn't mediated by representational rule systems and still produce powerful outputs. So if what we're suggesting in this book is a return to symbolic AI, and a move away from the machine learning successes, contemporary AI researchers would be understandably uninterested. (For an introduction to this sort of old-school philosophical theorizing inspired by old-school AI theorizing, see Rescorla 2015.)

But that's not what we are suggesting. Our point, in fact, is that philosophy brings to the table a collection of tools designed to find content in the wild, rather than building content into the architecture. The central problem in the philosophical study of content is this: when people go about in the world, encountering and interacting with various objects, making various sounds, having various things going on inside their heads, a bunch of contents

typically result. Some of the sounds they make have contents; some of the brain states they are in have content. Philosophical accounts of content want to explain what makes that the case: what needs to be going on so that some sounds are contentful and others are not; what needs to be going on so that some sounds mean that it's raining and other sounds mean that it's sunny.

People, of course, are the original neural network systems. So the philosophical project of content must be compatible with application to neural networks. That's because the philosophical project isn't to *build* contentful systems by setting them up with the right representational tools, but rather to *understand* the contents that we find 'in the wild'. Work in philosophy of language and formal semantics has indeed produced very sophisticated mathematical models of representation. But the philosopher's suggestion isn't that we should take those models and use them in designing good humans. We (philosophers who work on the theory of content) are not proposing that babies be pre-fitted with Montague semantics, or that axiomatic theories of meaning be taught in infancy. We just want to understand the content that certain complex systems (like people) carry, whatever the causal and historical story about how they came to carry that content.

So even if the history of AI research has made you a representation/content pessimist, we encourage you to read on. We think that intellectual engagement between philosophy and AI research has the promise of letting you have your non-content-oriented design tools and your *post facto* content attribution, too. And, we want to suggest, that's a good thing, because content plays crucial roles, and AI systems that lie wholly outside the domain of content won't give us what we want.

What This Book is **Not** About: Consciousness and Whether 'Strong AI' is Possible

There's an earlier philosophical literature on AI that we want to distance ourselves from. In influential work, John Searle (1980) distinguished between what he called Strong and Weak AI. Strong AI, according to Searle, has as a goal to create thinking agents. The aim of that research project is to create machines that *really* can think and have other cognitive states that we humans have. Searle contrasted this with the weak AI project according to which the aim was to create machines that have the *appearance* of thinking (and understanding and other cognitive states). Searle's central argument against Strong AI was the Chinese Room Argument. There's now a very big literature on the soundness of that argument (and also on how to best present the argument—for some discussion and references, see Cole 2014).

The project of this book will not engage with Searle-style arguments and we are not interested in the Strong vs Weak AI debate.

Our starting point and methodology are different from the literature in that tradition: Our goal in the first four chapters of Part I is to use contemporary theories of *semantics* and *meta-semantics* to determine whether (and how) ML systems could be interpreted. We take some of the leading theories of how language has representational properties and see what those theories have to say about ML systems. In most cases they are mixed: there's some match with what we are doing and some mismatch—and then we suggest fixes. In Chapter Four we suggest that maybe the right attitude to take is that we need to revise our meta-semantics to accommodate ML systems. Rather than use anthropocentric theories of content (i.e. theories of content based on how human

language gets content) to determine whether ML systems have content, we should revise our theories of content attribution so that ML systems can be considered representational (in effect revising what representation is, so that ML systems can be accommodated).

This strategy contrasts with the argumentative strategy exemplified by Searle's Chinese Room argument (and the tradition arising from that argument): the idea behind that strategy is to use reflections on a cluster of thought experiments to settle, once and for all, the question of whether machines can understand and have a semantics. This book doesn't engage with and only indirectly takes a stand on that form of argument.

Connection to the Explainable AI Movement

In 2018, the European Union introduced what it calls the General Data Protection Regulation. This regulation creates a 'right to explanation' and that right threatens to be incompatible with credit scores produced by neural networks (see Kaminski 2019 Goodman and Flaxman 2017, Adadi and Berrada 2018) because, as we pointed out above, many ML systems make decisions and recommendations without providing any explanation of those decisions and recommendations. Without explanations of this sort, ML systems are uninterpretable in their reasoning, and may even become uninterpretable in their results.

The burgeoning field of explainable AI (XAI) aims to create AI systems that are *interpretable* by us, that produce decisions that come with comprehensible *explanations*, that use *concepts* that we can understand, and that we can *talk to* in the way that we can

engage with other rational thinkers.[7] The opacity of ML systems especially highlights the need for artificial intelligence to be both explicable and interpretable. As Ribero et al. (2016:3–4) put it, 'if hundreds or thousands of features significantly contribute to a prediction, it is not reasonable to expect any user to comprehend why the prediction was made, even if individual weights can be inspected'. But the quest for XAI is hampered both by implementation difficulties in extracting explanations of ML system behaviour and by the more fundamental problem that it is not clear what exactly explicability and interpretability *are* or what kinds of tools allow, even in-principle, achievement of interpretability.

Doshi-Velez and Kim (2017) state the goal of interpretability as being 'to explain or present [the outputs of AI systems] in understandable terms' (2017:2) and proceed to point to real problems in modelling explanations. What they do not note is that both *understanding* and *terms/concepts* require at least as much clarification as explanation. As they point out, much work in this field relies on 'know it when you see it' conceptions of these core concepts. This book aims to show how philosophy can be used to remedy this lacuna in the literature.

The core chapters in this book aim to present proposals for how we can attribute content to AI systems. We return to the implications of this for the explainable AI movement towards the end of the book.

[7] A recent discussion in philosophy is Páez 2019. Outside of philosophy, some recent overviews of the literature are Mueller et al. 2019 and Addadi and Berrada 2018. For particular proposals about how to implement XAI, see Ribeiro et al. 2016, Doshi-Velez & Kim 2017, and Hendricks et al. 2016. For an approach that uses some ideas from philosophy to explain 'explanation', see Miller 2018.

Broad and Narrow Questions about Representation

We should note one limitation of our approach: we focus on whether the outputs of AI systems are content-bearing. In the little story about Lucie, we ask what the output '550' means. We are interested in whether that token can and should be considered contentful—in Chapter Three we put this as the question of whether there's aboutness in the AI system. This is closely connected to, but distinct from, a broader issue: Does the AI system have the ability to reason? Does it have a richer set of beliefs and so a richer set of contents? We can also ask: what is the connection between being able to represent the thought that Lucie's credit score is 550, and having a range of other thoughts about Lucie? Can a system have the ability to think only *one* thought, or does that ability by necessity come with a broader range of representational capacities? These are crucial questions that we will return to in the final chapter. Prior to that, our goal is somewhat more narrow and modest: Can we get the idea of content/representation/aboutness for AI systems off the ground at all? Are there any plausible extensions of existing meta-semantic theories that opens the door to this? Our answer is yes. In the light of that positive answer, the broader questions take prominence: how much content should be attributed? What particular content should be attributed to a particular AI system? Does SmartCredit understand 'credit worthiness', and also 'credit' and 'worthiness', and grasp the relevant compositional rule? Does understanding 'credit' involve an understanding of *money*, *borrowing*, *history*, etc? These are questions that become pressing, if the conclusions in this book are correct.

Our Interlocutor: Alfred, The Dismissive Sceptic

A character called Alfred is central to the narrative of this book. Alfred, we imagine, is someone whose job it is to make AI systems. He is very sceptical that philosophers can contribute to his work at all. In the next chapter, Alfred is having a conversation with a philosopher. Alfred argues that while he thinks talking to philosophers is a bit interesting, it is basically useless for him. According to Alfred, philosophers have nothing substantive to contribute to the development of AI.

Alfred will return at several junctions in this book. In writing this book (and thinking through these issues), Alfred has been very useful to us—we hope he is also of some interest to readers (and especially those potential readers who are entirely unconvinced that AI will profit from an injection of philosophy).

Who is This Book for?

The intended audience for this book are readers interested in starting to think how philosophy can help answer important questions about interpretable AI. They have some knowledge of philosophy, some knowledge of AI, and are interested in how to use the former to reflect on the latter.

There are some people who should not buy or read this book:

- If you are looking for a technical book that engages in great detail with the formal aspects of neural networks, then this book is not for you.

- If you're looking for a book that develops in detail a new theory about the nature of meaning, then this book is also not for you.
- If you're looking for a complete theory of interpretable AI then, unfortunately, you've also bought (or borrowed or downloaded) the wrong book.

Our goals are modest. We hope the book will help frame some important issues that we find surprisingly little literature on. AI raises very interesting philosophical questions about interpretability. This book tries to articulate some of those issues and then illustrate how current philosophical theories can be used to respond to them. In so doing, it presupposes some knowledge of philosophy, but not very much. Our hope is that it can be used even by upper-level undergraduate students and graduate students not expert in either AI or philosophy of language. We hope it will inspire others to explore these issues further. Finally, we hope it opens up a door between researchers in AI and the philosophy of language/metaphysics of content.