Sharing Our Concepts With Machines

Patrick Butlin

Forthcoming in *Erkenntnis,* please cite published version

Abstract

As AI systems become increasingly competent language users, it is an apt moment to consider what it would take for machines to understand human languages. This paper considers whether either language models such as GPT-3 or chatbots might be able to understand language, focusing on the question of whether they could possess the relevant concepts. A significant obstacle is that systems of both kinds interact with the world only through text, and thus seem ill-suited to understanding utterances concerning the concrete objects and properties which human language often describes. Language models cannot understand human languages because they perform only linguistic tasks, and therefore cannot represent such objects and properties. However, chatbots may perform tasks concerning the non-linguistic world, so they are better candidates for understanding. Chatbots can also possess the concepts necessary to understand human languages, despite their lack of perceptual contact with the world, due to the language-mediated concept-sharing described by social externalism about mental content.

1. Introduction

Babylon Health, a London-based private healthcare company, claims that their AI can 'understand and recognise the unique way that humans express their symptoms' (babylonhealth.com/ai; accessed 26 June 2020). Software that could reasonably be described in this way is not currently available to customers – at the time of writing their interactive 'symptom checker' generates a series of multiple-choice questions – but it appears that they intend to deploy a chatbot in the near future. Presumably, their intention is to offer an application which allows users to freely enter natural-language descriptions of their symptoms,

which will prompt follow-up questions, suggested diagnoses and recommendations about treatment or where to seek further help.

The grounds for optimism about the prospects of building such an application have arguably been boosted recently by some impressive results in the field of language modelling by deep neural networks. Systems using the Transformer architecture, including BERT (Devlin et al. 2019), GPT-2 (Radford et al. 2019) and GPT-3 (Brown et al. 2020), have shown the most spectacular results. GPT-3 is a huge artificial neural network, of 175 billion parameters, trained on a dataset of hundreds of billions of words of human-generated language to predict the next word from a given sentence. It shows remarkable proficiency on a range of tasks, and has been acclaimed as a significant advance on previous language models. For example, it can generate convincing, fluent imitations of news articles when prompted with titles and subtitles; human evaluators were at close to chance performance when asked to distinguish these from real articles (Brown et al. 2020). It is also notable for being able to perform new tasks after only being given a few demonstrations, without retraining.

Babylon Health's claims and GPT-3's results offer an opportunity for some applied philosophy of mind. They raise obvious questions about the capacity of AI systems to understand human language: Can GPT-3 understand language? Could a chatbot understand what we said to it? My aim in this paper is to make progress on these questions.

More specifically, I want to consider whether chatbots or language models like GPT-3 can possess the concepts that they would need to understand human languages. One requirement for successful linguistic communication is that the communicators share concepts; this claim is appealed to in arguments that concepts must have a public, shareable character (Prinz 2002, Onofri 2017). This implies that a chatbot could only understand the sentence 'I have a rash', for example, if it possessed the lexical concept RASH. Sharing concepts does not suffice for understanding even when they are associated with the same words; the pragmatic ability to make sense of utterances in context is also important, and a challenge for AI (Levesque 2014). But

possessing the right concepts is certainly one of the major elements of understanding. It is arguably what is required to make a language meaningful to a user.

This question about sharing concepts is pertinent because of an important feature which chatbots and language models have in common. This is that they interact with the world solely through text: they take only text as input, produce only text as output, and (in the case of language models) have training data composed of text alone.[1] So it is natural to wonder whether they could understand references to the everyday phenomena that we are familiar with through perceptual experience. A medical chatbot might receive 'I have a headache' or 'I have a tingling sensation in my toes' as an input. To understand these utterances it would have to share the human concepts of headaches and tingling sensations even though it had never experienced them. In comments on the Turing Test both Putnam (1981) and Davidson (1990) have argued that chatbots could not use human language meaningfully, because they lack sufficiently direct interaction with the objects and properties to which we most frequently refer.

The view which I will argue for here is that chatbots can share our concepts in the manner required to understand human languages, but language models such as GPT-3 cannot. The reason for this difference is that they have different functions. Chatbots may have functions which involve detecting and acting on non-linguistic states of affairs; for instance, a medical triage chatbot may have the function of diagnosing the user's condition and making an appropriate recommendation. In contrast, language models have purely linguistic functions, such as generating text which extends an input in a probable way. I will argue that this difference in function makes a difference to potential concept-possession in section 2, and then give a further argument for the claim that chatbots may possess concepts in section 3.

[1] A system based on GPT-3 has been developed which generates images from descriptions (called DALL-E; Ramesh et al. 2021). But my topic is those systems which use text alone as input, output and training data.

These arguments, however, will not address the problem of chatbots' perceptual impoverishment, which I will turn to in the second half of the paper (sections 4 and 5). I will argue that social externalism about conceptual content entails that chatbots could share our concepts in the way that is necessary for understanding human languages. Social externalism tells us that the content of lexical concepts is often fixed by the content of associated words, and that this works even when language users have incomplete or imperfect understanding of the relevant phenomena (Putnam 1975, Burge 1979). Chatbots are language users, and their lack of perceptual experience is a form of incomplete understanding. So they can share our concepts because they use our languages. As Millikan (2000, p. 89) puts it: 'it is possible… to have a substance concept *entirely* through the medium of language.'

Because the topic of this paper is concept-sharing, it will be important to be clear about the different ways in which concepts can vary. I will use the terms 'content', 'cognitive significance' and 'conception' to refer to three dimensions of variation. I will reserve the term 'content' for referential content; two concepts have the same content, in my sense, if and only if they have the same reference. Philosophers do not always use 'content' in this way in connection with concepts, partly due the phenomenon identified by Frege in which concepts with the same reference can make different contributions to thought. For example, Prinz (2002) writes that concepts have both 'intentional content' – that is, referential content – and more fine-grained 'cognitive content'. Prinz also suggests that concepts with the same cognitive content can have different references, such as in the case of the concepts which I and my Twin Earth counterpart express using the word 'water' (Putnam 1975). However, my preference is to use 'cognitive significance' in place of Prinz's 'cognitive content'. Two concepts have the same cognitive significance if and only if they present themselves to the thinker as 'obviously and incontrovertibly' co-referential (Schroeter 2008).

In addition to their content and cognitive significance, concepts are also associated with conceptions (Quilty-Dunn 2021). A conception is a structured body of information connected to

a concept, which might include such things as beliefs employing the concept and representations of exemplars or prototypes of the kind to which the concept refers. The conception associated with a concept can change readily, and different individuals will usually have different conceptions associated with concepts with the same reference, but this does not prevent us from understanding one another. For example, experts on Leibniz have different conceptions of Leibniz from me, because they know far more about him, but this does not prevent me from understanding them when they talk about Leibniz. This paper will test the limits of this phenomenon, because it seems that a medical triage chatbot would have a radically different conception of headaches from those who have suffered them.

2. Functions, Content and GPT-3

For a machine to understand a human language it must possess concepts which are coreferential with at least some commonly-used words in that language. It may be that these concepts also need to have the same cognitive significance as those possessed by humans, but for now we can put this issue aside and focus on relatively basic necessary conditions (I discuss it in section 5). Since we have the example of a medical triage chatbot in mind, we can consider words like 'red', 'skin' and 'rash'. To understand English a machine would need to possess concepts which are coreferential with words such as these: RED, SKIN, and RASH.[2]

There are two basic requirements for possessing a concept such as SKIN. First, a system which possesses this concept must be able to deploy representations with content which includes *skin* in cognitive processes. An animal which could think the thought *my skin is red* would meet this requirement. Second, this representational ability must take a conceptual form, the essence of which is systematicity (Evans 1982, Peacocke 1992, Camp 2009). To meet this requirement the range of contents which the system can represent must be systematically related. If the animal

[2] I follow the convention of using small caps when naming concepts.

which could think *my skin is red* could also think *this plant is green*, then to meet this requirement it would also have to be able to think *my skin is green* and *this plant is red*. I will discuss whether chatbots and language models might meet these two requirements in turn, in this and the following section.

To be able to represent kinds such as skin and rashes, and properties such as being red, a system must contain elements with functions which concern these kinds and properties (Millikan 1984, Dretske 1988). The reason for this is that representation is a essentially functional phenomenon. What it is for something to be a descriptive representation is for it to have, to a first approximation, the function of conveying or storing information about states of affairs (non-descriptive representations have other functions). The fact that descriptive representations have correctness conditions – in the form of truth or accuracy conditions – arises from this fundamental feature. For example, when the English sentence 'My skin is red' is used as a descriptive representation its function is to convey the information that my skin is red. It can perform this function when my skin is red, but not otherwise (I am using 'the information that' in the factive sense). This is why the truth-condition for this sentence is that my skin is red, and hence why its content is *my skin is red*.

A further aspect of this functional approach to representation is that the functions of elements of systems flow from the functions of the systems themselves. In particular, designed or evolved systems contain elements with the function of carrying information about states of affairs, and hence represent those states, because this allows them to condition their behaviour on those states. They therefore do so because they have functions which constitute some forms of behaviour as successful and others as unsuccessful, and because sensitivity to states of affairs facilitates successful behaviour for them. This means that what they represent depends on what is relevant for performing their system-level functions. Success or failure in behaviour can then be explained by appealing to accuracy or inaccuracy in contributing representations (Papineau 1993, Shea 2018).

Putting these ideas together gives us a way to draw inferences from the functions of whole AI systems, such as chatbots or language models, to conclusions about the contents of the representations that they might use. From the function of a whole AI system – that is, the kind of behaviour that it is supposed to produce – we can infer what information it would need to produce this behaviour, and hence what the content of its representations is likely to be. There are various ways in which such inferences are uncertain, but this method is powerful enough to support conclusions about what sorts of representations possible future chatbots might use and about what can reasonably be attributed to GPT-3.

The function of a medical triage chatbot is to conduct a conversation with its user from which it can infer a likely diagnosis, then to conclude this conversation by recommending appropriate action, perhaps informing the user of the diagnosis, and potentially taking further actions, such as alerting emergency services. By far the likeliest way for a chatbot to do this is for it to build up a representation of the user's symptoms, history and other medically-relevant features over the course of the conversation. This representation would be used in combination with stored medical knowledge to select questions to ask the user and to generate the diagnosis, recommendations, and other possible actions. The content of this representation of the user's features might include propositions such as *there is a red patch of skin on the user's calf* and *the user lives in the tropics*. The chatbot's behaviour would be likely to be successful – that is, its diagnoses accurate and its recommendations appropriate – if this representation was accurate, and unsuccessful if it was inaccurate.

We therefore have good reason to believe that future chatbots will represent some of the familiar kinds, objects and properties that are among the most common subjects of human discourse. Medical chatbots would have reasons to represent a wide range of these. Although chatbots only interact with the world through text, they may use these interactions to perform non-linguistic functions, such as diagnosing diseases.

In contrast, GPT-3 has a purely linguistic function. GPT-3 is trained to predict the next word from a given sequence, and its function is to generate likely continuations of the texts which it is provided as input. On the face of it, this suggests that it needs to represent only information about language. It must represent which words appear in its input, in which order, and enough further information about the statistics of word sequences to compute which words are likely to come next. There is also some evidence that Transformer-based language models represent syntactic properties of inputs (Rogers et al. 2020). GPT-3 therefore appears to be capable of representing the words 'skin', 'red' and 'rash' but not the worldly entities to which they refer.

A possible objection to this line of thought is that GPT-3 can answer trivia questions accurately (as can other language models; see Petroni et al. 2019). Brown and colleagues (2020) found that it achieved 64.3% accuracy on a test called TriviaQA under the zero-shot condition, in which the questions are provided without useful context. This could be used to object to the argument either on the grounds that it shows that GPT-3 has non-linguistic knowledge, or because it shows that GPT-3 can be used to perform non-linguistic functions. The first of these claims is not persuasive, however, because GPT-3's performance is better explained by the fact that likely word sequences are correlated with facts about the world. For example, suppose that GPT-3 can give the correct answer to the question 'Who was the King of England immediately before John?' (which is 'Richard I'). This could be explained either by GPT-3's representing that John succeeded Richard I, or by its representing that sentences like 'John succeeded Richard I' are common. One reason to favour the latter explanation is that GPT-3 can be expected to reproduce common misconceptions.

The argument that GPT-3 can represent the non-linguistic world because it can be used to perform non-linguistic functions, meanwhile, fails because it does not distinguish two kinds of artefact functions. What has been called the *proper* function (Millikan 1984, Preston 1998) of an artefact is the intended performance that guided its production or modification, via the mind of its maker, and hence caused it to have the properties that it possesses in its resulting form. A use

to which an artefact is put which did not influence its form is not a proper function; we might call it an *expedient* function. Since GPT-3 was trained solely to perform linguistic tasks, its subsequent use for non-linguistic tasks can only give it expedient non-linguistic functions. Representational content is fixed solely by proper functions, however, because of the distinctive role that content ascriptions play in explaining behaviour. This role involves adverting to a disposition of some part of the system which was shaped by a prior process of selection (Shea 2018).

A related argument for the claim that language models such as GPT-3 cannot understand human languages is suggested by Lake and Murphy (2020). Lake and Murphy argue that language models lack semantic knowledge on the grounds that the representations that underlie their use of language are not suitable for supporting uses such as describing salient features of the environment, forming accurate representations of the world on the basis of linguistic input, and choosing linguistic outputs so as to achieve goals. This deficiency is partly a result of language models' function, although it is also partly a result of the way in which they perform it – one could imagine a next-word-prediction system which worked in a much more human-like way. However, my critique is different from theirs. Lake and Murphy argue that language models do not represent the right information in connection with the word 'skin', in the right way, to possess (or model) semantic knowledge; my claim is that they do not employ representations which refer to the skin.

3. Concepts and Systematicity

We have seen that chatbots are likely to be able to represent the sorts of everyday phenomena necessary for understanding human languages – like skin, rashes and the colour red – and that language models such as GPT-3 are not. This means that GPT-3 is no longer a good candidate for possessing the concepts needed to understand human languages, so from this point on I will concentrate on chatbots, continuing to use the example of one used for medical triage.

Showing that chatbots can use representations with content including *skin* or *red* in computational processes does not show that they can possess the concepts SKIN or RED, because it is also possible to represent these phenomena non-conceptually. Non-conceptual content may be implicated in perception (Burge 2010, Block 2014, Neander 2017) or animal cognition (Beck 2012). So in this section I will examine whether chatbots' representations might take a conceptual form. The principal criterion for this is that they support systematic representational abilities.

The classic statement of this condition is Evans' (1982) 'Generality Constraint'. The generality constraint states that for a thinker's thoughts to take a conceptual form it must be the case that if they can think the thoughts *a is F* and *b is G*, they must also be able to think the thoughts *a is G* and *b is F*. That is, the set of representations that they are able to form in cognition must be closed under recombination of elements of their content, where this recombination accords with syntactic rules.

Camp (2009) argues that this condition can be satisfied in more or less substantial ways. A less substantial way, which she calls the 'causal counterfactual way', is for the representing system to be such that they would form any of the representations in the required range if they were subject to a corresponding stimulus. For example, an animal which is capable of forming representations with the content *this fruit is green* and *this bug is blue* might also be capable of forming *this fruit is blue*, but only if it is subject to a visual impression of a blue fruit. This way of satisfying the requirement does not entail either that the system can form the new representations without the corresponding stimulus, or that it can do anything useful with them when it has formed them. This latter point is significant. Suppose there is some behaviour which would be worthwhile for the animal given that the fruit is blue, and that the value of this behaviour could in principle be inferred by a short chain of reasoning from its background knowledge plus the information that the fruit is blue. If the animal was unable to perform this behaviour despite representing that the fruit was blue this would impugn the generality of its

thought. A more substantial way of meeting the generality constraint, according to Camp, is to be capable of forming new representations combining past elements of content in a way which is relatively independent of the current stimulus.

Camp is careful to make clear that in her view what is required for conceptual thought is systematicity, not compositionality. A representational medium exhibits compositionality if it employs representational vehicles with fixed content which can be combined and recombined to form compound representations, which inherit their content from the content of their parts. This is distinct from systematicity because systematicity is a feature of the range of representational abilities of a representation-using system, whereas compositionality is a feature of the medium that underlies these abilities. Fodor and others have argued prominently that human thought relies on a compositional representational medium, on the grounds that this would explain the apparent systematicity of our abilities (Fodor 1987, Fodor & Pylyshyn 1988). Compositionality appears to suffice for systematicity, at least of the causal counterfactual kind.

It therefore seems to be possible for chatbots to have systematic representational abilities because it is possible for them to use compositional representational media. Chatbots would be likely to benefit from using representational elements which can be combined to form a systematic range of representations, which also point to bodies of knowledge associated with their contents, as concepts are thought to (Lake & Murphy 2020, Quilty-Dunn 2021). For example, a medical triage chatbot needs to be able to represent the presence of swelling in a range of contexts (e.g. that the user's ankle is swollen, or that the swelling on their abdomen has grown rapidly) and draw on a body of knowledge about the diagnostic and prognostic significance of swelling.[3] Future chatbots are likely to rely on neural networks for major aspects of their operation, but this is not incompatible with compositionality: it has been argued that

[3] Babylon Health write that their system uses a 'Knowledge Graph' which they describe as 'one of the largest structured medical knowledge bases in the world' (babylonhealth.com/ai; accessed 29 July 2020). Danks (2014) suggests that concepts may be thought of as elements of graphical models which provide comparably structured representations of human knowledge.

compositional representation is possible within neural networks (Smolensky 1991, Greff et al. 2020), and in any case hybrid neuro-symbolic architectures are possible (e.g. Mao et al. 2019).

This argument leaves open whether chatbots could possess systematic abilities of the more substantive kind which Camp identifies, which requires relatively stimulus-independent representation. However, there is reason to believe that stimulus-independent representation would be valuable for them. In order to establish a diagnosis, or even to get a reasonably precise grasp of a user's symptoms, a medical chatbot may need to form and test hypotheses. For example, if the user says that they have a rash, then the chatbot may form the hypothesis that they have shingles, and ask questions selected to test this hypothesis. To form a hypothesis of this kind is to generate a representation of something other than the immediate input. Furthermore, an effective chatbot would not only be capable of registering a wide range of inputs from its users, including ones involving unusual combinations of elements of content, but of interpreting and acting on these inputs, such as by asking appropriate clarificatory questions. Substantive systematicity is a likely feature of future chatbots, which would therefore meet a key criterion for concept possession.

4. Social Externalism and Conceptual Content

So far we have seen that chatbots could possess concepts with content concerning non-linguistic entities such as skin, redness and rashes. To understand human languages, however, they would have to possess some of our public concepts, such as SKIN, RED and RASH. Chatbots' perceptual impoverishment seems hard to reconcile with their possession of such concepts – how could a system which never had any perceptual contact with skin have the same concept, SKIN, as a human? – but in this and the following section I will argue that this appearance is misleading. Familiar and attractive claims from current philosophy of mind entail that chatbots could possess the same concepts as we do, in the sense relevant for understanding. In this

section I will focus on content, then in section 5 I will turn to issues concerning conceptions and cognitive significance.

The main premise of my argument is social externalism about conceptual content. Content externalism is the thesis that the content of mental representations is determined partly by facts about the environment, of which the thinker may not be aware (Schroeter 2008). Social externalism is a version of content externalism which claims that the content of lexical concepts is determined partly by the meanings of the words with which they are associated (Putnam 1975, Burge 1979). According to the social externalist picture, if two people have concepts which they each associate with the same word, then their concepts are likely to have the same content, even if they have very different conceptions of the phenomena in question. This can be the case even if one or both of them has incomplete understanding of these phenomena; for example, someone with only the sketchiest of impressions of what viruses are can possess a concept with the content *virus*.

Social externalism tells us that speakers of common languages share concepts in two senses: they possess concepts with the same content, and what's more, they do so because language allows them to acquire content-matched concepts from one another. One effect of this is that communication is facilitated. A person who knows very little about viruses can communicate with an expert about them, and thus acquire knowledge about viruses – for which the ability to think about viruses is a precondition (Goldberg 2009). Another is that our use of public languages expands the range of our potential thoughts beyond those matters on which we individually have expertise, and facilitates the co-ordinated focus of our thoughts on matters of common interest (O'Madagain 2018).

I claim that if chatbots possess concepts, the content of these concepts may also be determined by the meanings of corresponding words. As social externalism has been developed, three mechanisms have been proposed to explain how word meanings can influence conceptual content, and hence how language can facilitate the sharing of concepts (in the second of the two

senses just mentioned). The reason why we should think that concept-sharing can extend to chatbots is that each of these three mechanisms would allow them to be included.

The first proposed mechanism, which was one of the key tenets of the early tradition of social externalism, is that concept-sharing happens as a result of deference to experts (Putnam 1975, Burge 1979). According to this picture, for any given lexical concept there is a group of experts, and the content of the experts' concepts is determined by their detailed and accurate conceptions of the objects or properties concerned. Those who are not experts may have intentions or dispositions to accept correction from them, and this posture of deference is what causes the content of the non-experts' concepts to track the content of those of experts. This allows non-experts to possess concepts for phenomena of which they have only very limited understanding. The content of both experts' and non-experts' concepts is determined by the intentions and dispositions which govern their use of those concepts and the associated words, but these are different in each case. In particular, the non-experts take the conditions for correct application of their concepts to be those of the corresponding words, without knowing what these are, and intend or are disposed to deploy the concepts accordingly (Goldberg 2009).

Greenberg (2014) has argued that exactly what deference amounts to, and how it works in determining content, are both underspecified. However, for our purposes what matters is what attitudes, dispositions or other properties a concept-user needs in order to count as deferring to experts, with respect to a given concept, in the manner required for the concept to be shared. The clearest criterion is that the concept-using system must be disposed to modify its use of the concept so as to track expert usage. If it learns, for instance, that experts do not take certain entities to fall under the concept, it must be disposed to stop representing those entities as falling under the concept itself. Since this phenomenon is specific to lexical concepts, the system will be so disposed if and only if it is also disposed to modify its use of the corresponding word in accordance with expert usage.

We should expect this criterion to be met by chatbots because linguistic creativity or resistance to community or expert use would be likely to impede them from good performance. In the case of medical triage chatbots, we might expect the use of technical medical concepts to be governed by the system's knowledge base, which would be updated by a special-purpose process to ensure quality. With respect to these concepts the chatbot might itself be an expert. But we can also imagine that it would show deference to human users in the way in which it used other concepts and associated words. For example, suppose a chatbot knew that the word 'tattoo' referred to an ink mark of the skin, and was consulted by a user about a rash that had apparently been caused by the use of a marker pen on their body. It might be that the chatbot would use the word 'tattoo' to refer to this mark, and be corrected by the user. The chatbot should then modify its use of this word, in accordance with what it had learnt. In this case we would have reason to say that the chatbot possessed the concept TATTOO in virtue of its deference to human users.

An alternative account of how language facilitates concept-sharing is proposed by Schroeter and Schroeter (2016). This account relies on a notion that they call 'apparent *de jure* co-reference': two token elements of thought share this relation when they seem to the thinker to be guaranteed to co-refer, in virtue of the way in which they are presented. This relation underlies identity between concepts, because what distinguishes cases in which I repeatedly use the same concept in thought (e.g. HESPERUS twice) from cases in which I use distinct concepts (e.g. HESPERUS and PHOSPHORUS) is that only in the former case can I be sure of co-reference without further thought. Building on this idea, Schroeter and Schroeter suggest that a word and a set of different thinkers' concepts can be bound together in a 'representational tradition' by being treated as *de jure* co-referential, and that concepts that belong to the same representational tradition will have the same content. The idea that *de jure* co-reference links words to concepts, and hence can transmit sameness of content from one individual to another, is particularly clear if we think about cases in which learning words scaffolds our learning about the world. When I

learnt that there is an architectural genre called 'prairie style' I was caused to acquire a concept which automatically inherited its content from the phrase.

This mechanism of concept-sharing would also work for chatbots, because their concepts would belong to common representational traditions with ours. A medical chatbot's shingles concept, for example, would be treated by it as *de jure* co-referential with the word 'shingles', except in cases where particular features of the patient's discourse indicated that they were using this word in a non-standard way.

Finally, a third account is offered by Fodor (1994) and Millikan (2000). Fodor and Millikan both advocate covariation theories of content, according to which the content of a given concept depends on its being used to track a particular object or property in the environment. Leaving aside the details of their theories, they both argue that in many cases concepts will be shared between individuals, because of our tendency to use what other people say about the environment as a guide to what it contains. Thus, for example, if my ability to identify the presence of curlews is a result of your tendency to mention it when you hear a curlew, my CURLEW concept has its content fixed by the meaning of the word 'curlew'. Fodor compares this reliance on others' expertise to the use of scientific instruments; we can use either experts or instruments to achieve reliable correlations between our thoughts and phenomena that we would not be able to recognise alone.

Once again, this mechanism will work for chatbots, because they will rely very heavily on what their interlocutors say to learn about what is going on in particular parts of the world. For example, a chatbot will be able to use a symbol that covaries reliably with headaches in its interlocutors by activating this symbol on those occasions on which the interlocutors say that they have a headache. This symbol, according to Fodor and Millikan's account, will thereby come to have the same content as the word 'headache'. The covariation will not be perfect, but Fodor and Millikan's accounts don't require this, because humans also make mistakes in

identifying entities in their environment. What matters for present purposes is that if this mechanism for concept-sharing works for humans, it will work just as well for chatbots.

Social externalism therefore gives us reason to believe that chatbots' concepts will often have the same content as those which humans use in association with the same words. This depends on chatbots' use of human languages, but it does not depend on their having a prior understanding of these languages (which would make my argument circular). Social externalism tells us that possession of lexical concepts, which is required to understand the corresponding words, is often the result of the use of those very words.

5. Conceptions and Cognitive Significance

In this section I consider whether differences in conception or cognitive significance, the two aspects of concepts other than content, would entail that chatbots could not share human concepts in the way required for understanding. In each case I appeal to arguments from the recent literature in philosophy of mind.

On the topic of conceptions, one of the central doctrines of social externalism is that it is possible to possess a concept despite having incomplete or flawed understanding of the phenomenon concerned. This would entail having a different conception from experts. One of Burge's classic cases is of a person who lacks the theoretical knowledge that arthritis affects only the joints, suspecting that it affects their thigh (1979); another is of a person who believes that sofas are religious artefacts (1986). It is a standard view among social externalists that these characters possess the concepts ARTHRITIS and SOFA. They would also take me to possess the concept HIGGS BOSON, even though I have very little knowledge of particle physics (Soames 1989). However, some theorists do claim that concept-sharing can fail where thinkers have sufficiently diminished understanding, or radically divergent conceptions (Brown 2000, Goldberg 2009).

It is therefore possible to imagine an objection which claimed that although incomplete understanding is generally compatible with concept possession, incomplete understanding of the specific kind exemplified by chatbots is not. Take the case of the concept HEADACHE. Chatbots cannot experience pain, so their understanding of headaches would be incomplete. They could not recognise headaches in themselves, unlike (presumably) all human thinkers who possess this concept. However, human headaches would have significance for medical triage chatbots, since their functions concern them, as I argued in section 2. Such chatbots might accurately represent many details about the possible causes, effects, and treatments for headaches, and might be highly capable of identifying them in others through the Fodor-Millikan route. This is the reverse of the typical human case of incomplete understanding, since it combines excellent textbook knowledge with the complete absence of 'grounding' perceptual familiarity with even related phenomena. Empiricists about concepts, such as Prinz (2002), might deny that possession of some concepts is compatible with this distinctive form of incomplete understanding.

If there are any concepts which chatbots could not possess, these would include concepts for bodily sensations and properties of perceptual experiences (or perhaps directly perceptible properties of objects), such as PAIN or RED. If these concepts were inaccessible to them, it might then follow that they could not possess closely-related concepts such as ARTHRITIS or RASH. But no argument on the lines suggested could plausibly be made that chatbots could not possess concepts of the latter kind, even though they could possess ones of the former kind. So one issue at stake here is whether there are any phenomenal concepts (Loar 1997, Papineau 2002). Phenomenal concepts are usually defined as concepts that can only be possessed by thinkers who have had conscious experiences of particular kinds, such as pain or the visual experience of the colour red. If there are phenomenal concepts, then chatbots cannot possess them; if there are none, then the fact that chatbots do not have human-like conscious perceptual experiences does not bar them from possessing the same concepts as humans.

Ball (2009) argues on social externalist grounds that there are no phenomenal concepts. He observes that if the English word 'red' expresses the concept RED, it will be possible for this concept to be shared by the mechanisms described in the previous section, provided that social externalism is true. If this is correct, then RED is not a phenomenal concept, because it can be possessed by someone who has never seen the colour red (or experienced, e.g., a red afterimage). Mary, the colour scientist in Jackson's (1982) Knowledge Argument who lives in a black-and-white room, would be able to possess the concept RED before leaving the room, thanks to her familiarity with scientific works and other English texts which discuss this colour. Phenomenal concept theorists might seek to reject this position by arguing that there are two concepts expressible by 'red', of which one is a phenomenal concept (call it RED_P), and the other is not (RED_N), and that the mechanisms of social externalism only apply to RED_N. This is the only alternative, since there must be some concept that Mary expresses by using 'red'. But this has a range of implausible consequences. For instance, suppose that before leaving the room Mary has the thought that she would express by saying that seeing red is a phenomenal state. On the phenomenal concept theorist's view, this thought must involve RED_N. But someone living in a multicoloured environment could have a thought that they would express in the same way, so the phenomenal concept theorist would have to say either that these are distinct thoughts; that the second person has two different concepts for which they would use the word 'red'; or that the second person also lacks RED_P. None of these options is attractive, particularly in the context of social externalism.

Even if Ball's argument fails, and there are phenomenal concepts, Mary would still possess (and chatbots could still possess) concepts such as RED_N. In this case there would be reason to believe that such concepts would suffice for understanding of words like 'red'. Mary is supposed to be an expert colour scientist, so presumably she would understand the sentence 'rashes are typically red', perhaps by employing her knowledge that 'red' picks out a colour with certain specific cultural associations and a dominant wavelength of 625-740nm. There are also, of

course, many people in real life with severe visual impairments and it is far from clear that they cannot understand the word 'red'. Helen Keller was deaf and blind from infancy, but she certainly understood English: she learnt to read Braille and write, earned a degree, and became an author, activist and public speaker (Stich 1983).

A similar response can be given to an alternative line of thought which might also suggest that chatbots could not understand human language due to differences in conceptions. Theorists of embodied cognition suggest that the range of concepts we are capable of possessing depends on the form of our bodies and our perceptual apparatus (Shapiro 2019). One version of this claim would be that thinkers with very different bodies could not share concepts with the same content, but social externalism provides us with an argument against this view. An alternative would be that different bodies lead to radically different conceptions, such that thinkers possess different concepts with the same content. This would be analogous to the case of RED_{P} and RED_{N}, except that the different conceptions would arise from differences in body forms rather than conscious experience. But again, even if this is true, it is doubtful whether it would prevent understanding. For example, evidence for the embodiment of concepts comes from Pulvermüller's (2005) finding that reading the word 'kick' causes activation in areas of motor cortex associated with the legs, so it might be suggested that a link with a mechanism for performing the action of kicking is an essential component of the concept KICK. But someone born without legs, such as the athlete Zion Clark, who would lack such a mechanism, could certainly understand the sentence 'Smith kicked the winning penalty'.

It might further be objected that Mary's or Keller's conception of the colour red, and Clark's conception of the action of kicking, are less different from those of most humans than a chatbot's would be. This may be so, but we still have good reasons to expect chatbots to be able to understand: they could have concepts with the same contents as ours, and significant overlaps in conception, and in general differences in conception do not seem to entail the inability to understand.

Turning to the topic of cognitive significance, the potential objection to the claim that chatbots can possess the concepts necessary to understand human languages would be that chatbots' concepts would differ from ours in this respect. Unlike in the case of conceptions, it is not obvious why one would expect the cognitive significance of chatbots' concepts to differ from ours. It is not even obvious what this claim amounts to, because cognitive significance is defined intrapersonally in the first instance. However, it does seem that failures of understanding are possible for reasons other than differences in content, which could plausibly be described in terms of cognitive significance. Loar (1976) describes a case in which two people are each aware of a third in two different ways: they are currently watching him being interviewed on television, and they also see him on the train each morning. They do not know that the man on television is the same as the man on the train. If one says 'He is a stockbroker' to the other, intending to refer to the man on television, the other may take this to be a reference to the man on the train. This would be a misunderstanding even though they are thinking of the same man.

Loar's case does not seem likely to point towards widespread or chatbot-specific problems. Prosser (2018) argues that when the use of shared words facilitates concept-sharing, the concepts thus shared will be alike in cognitive significance as well as content (although he talks of sharing 'modes of presentation' rather than cognitive significance). Prosser distinguishes between cases in which communication requires interpretation, and cases in which it is transparent. When utterances include indexicals, demonstratives, or perhaps words with common homonyms, it is necessary for hearers to interpret these words, and this generates the possibility of Loar-style cases. But otherwise we typically take it for granted that words have the same meanings in the mouths of our interlocutors as they would in ours. This is simultaneously necessary for, and made possible by, the phenomenon of concept-sharing through language. What Prosser means by calling communication 'transparent' in such cases is that we do not need to rely on interpretative premises, implicitly or explicitly, in reasoning from our interlocutors' utterances. For example, if someone says to me, 'Beech trees are native to the UK,' I can infer that beeches

have been present in the UK for thousands of years without employing a premise such as *the speaker is using 'beech trees' to refer to beech trees*. Prosser's analysis implies that if the concept-sharing described by social externalism extends to chatbots, it will have the effect that the cognitive significance of chatbots' concepts will be equivalent to that of human concepts, where these are associated with a shared word.

6. Conclusion

I have argued that language models such as GPT-3 cannot understand human languages, but that it is possible that chatbots could possess the concepts necessary to do so. Chatbots can represent the familiar objects and properties of human life – which is a crucial element in understanding languages – because they perform tasks that relate to some of these objects and properties, and must gather and store information about them in order to do so. Language models lack functions of this kind. It is also possible for chatbots to have systematic representational abilities, which would mean that they would meet the principal criterion for concept-possession, and social externalism suggests that they could share concepts with humans.

References

Ball, D. 2009. There are no phenomenal concepts. *Mind* 118(472): 935-962.

Beck, J. 2012. Do animals engage in conceptual thought? *Philosophy Compass* 7(3): 218-229.

Block, N. 2014. Seeing-as in the light of vision science. *Philosophy and Phenomenological Research* 89(3): 560-572.

Brown, J. 2000. Critical reasoning, understanding and self-knowledge. *Philosophy and Phenomenological Research* 61(3): 659-676.

Brown, T. et al. 2020. Language models are few-shot learners. *arXiv* preprint.

Burge, T. 1979. Individualism and the mental. *Midwest Studies in Philosophy* 4(1): 73-122.

Burge, T. 1986. Intellectual norms and foundations of mind. *Journal of Philosophy* 83(12): 697-720.

Burge, T. 2010. *The Origins of Objectivity*. Oxford: Oxford University Press.

Camp, E. 2009. Putting thoughts to work: Concepts, systematicity, and stimulus-independence. *Philosophy and Phenomenological Research* 78(2): 275-311.

Danks, D. 2014. *Unifying the Mind: Cognitive Representations as Graphical Models*. Cambridge, MA: MIT Press.

Davidson, D. 1990. Turing's test. In Newton-Smith and Wilkes, eds., *Modelling the Mind*. Oxford: Oxford University Press.

Devlin, J., M.-W. Chang, K. Lee & K. Toutanova. 2019. BERT: Pre-training of deep bi-directional transformers for language understanding. *arXiv* preprint.

Dretske, F. 1988. *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press.

Evans, G. 1982. *The Varieties of Reference*. Oxford: Oxford University Press.

Fodor, J. 1987. Why there still has to be a language of thought. In *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Fodor, J. 1994. *The Elm and the Expert*. Cambridge, MA: MIT Press.

Fodor, J. & Z. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition* 28(1-2): 3-71.

Greenberg, M. 2014. Troubles for content I. In Burgess and Sherman, eds., *Metasemantics: New Essays on the Foundations of Meaning*. Oxford: Oxford University Press.

Greff, K., S. van Steenkiste & J. Schmidhuber. 2020. On the binding problem in artificial neural networks. *arXiv* preprint.

Goldberg, S. 2009. Experts, semantic and epistemic. *Noûs* 43(4): 581-598.

Jackson, F. 1982. Epiphenomenal qualia. *Philosophical Quarterly* 32: 127-136.

Lake, B. & G. Murphy. 2020. Word meaning in minds and machines. *arXiv* preprint.

Levesque, H. 2014. On our best behaviour. *Artificial Intelligence* 27-35.

Loar, B. 1976. The semantics of singular terms. *Philosophical Studies* 30(6): 353-377.

Loar, B. 1997. Phenomenal states. In Block, Flanagan and Güzeldere, eds., *Consciousness: Philosophical Debates*. Cambridge, MA: MIT Press.

Mao, J., C. Gan, P. Kohli, J. Tenenbaum & J. Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words and sentences from natural supervision. In *International Conference on Learning Representations*.

Millikan, R. 1984. *Language, Thought and Other Biological Categories*. Cambridge, MA: MIT Press.

Millikan, R. 2000. *On Clear and Confused Ideas*. Cambridge, UK: Cambridge University Press.

Neander, K. 2017. *A Mark of the Mental: In Defense of Informational Teleosemantics*. Cambridge, MA: MIT Press.

O'Madagain, C. 2018. Outsourcing concepts: Deference, the extended mind, and expanding our epistemic capacity. In Carter, Clark, Kallestrup, Palermos and Pritchard, eds., *Socially Extended Epistemology*. Oxford: Oxford University Press.

Onofri, A. 2017. The publicity of thought. *Philosophical Quarterly* 68(272): 521-541.

Papineau, D. 1993. *Philosophical Naturalism*. Oxford: Basil Blackwell.

Papineau, D. 2002. *Thinking about Consciousness*. Oxford: Oxford University Press.

Peacocke, C. 1992. *A Study of Concepts*. Cambridge, MA: MIT Press.

Petroni, F. et al. 2019. Language models as knowledge bases? *arXiv* preprint.

Preston, B. 1998. Why is a wing like a spoon? A pluralist theory of function. *Journal of Philosophy* 95(5): 215-254.

Prinz, J. 2002. *Furnishing the Mind.* Cambridge, MA: MIT Press.

Prosser, S. 2018. Shared modes of presentation. *Mind and Language* 34(4): 465-482.

Pulvermüller, F. 2005. Brain mechanisms linking language and action. *Nature Reviews Neuroscience* 6: 576-582.

Putnam, H. 1975. The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science* 7: 131-193.

Putnam, H. 1981. Brains in a vat. In *Reason, Truth and History*. Cambridge: Cambridge University Press.

Quilty-Dunn, J. 2021. Polysemy and thought: Towards a generative theory of concepts. *Mind & Language* 36(1): 158-185.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei & I. Sutskever. 2019. Language models are unsupervised multi-task learners. *OpenAI Blog*.

Ramesh, A., M. Pavlov, G. Goh, S. Gray, C. Vos, A Radford, M. Chen & I. Sutskever. 2021. Zero-shot text-to-image generation. *arXiv* preprint.

Rogers, A., O. Kovaleva & A. Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *arXiv* preprint.

Shapiro, L. 2019. *Embodied Cognition*. New York: Routledge.

Schroeter, L. 2008. Why be an anti-individualist? *Philosophy and Phenomenological Research* 77(1): 105-141.

Schroeter, L. & F. Schroeter. 2016. Semantic deference versus semantic coordination. *American Philosophical Quarterly* 53(2): 193-210.

Shea, N. 2018. *Representation in Cognitive Science*. Oxford: Oxford University Press.

Smolensky, P. 1991. Connectionism, constituency and the language of thought. In Loewer & Rey, eds., *Meaning in Mind: Fodor and his Critics*. Oxford: Basil Blackwell.

Soames, S. 1989. Semantics and semantic competence. *Philosophical Perspectives* 3: 575-596.

Stich, S. 1983. *From Folk Psychology to Cognitive Science*. Cambridge, MA: MIT Press.