AI and Alien Languages

Matti Eklund
matti.eklund@filosofi.uu.se

work in progress

for Cappelen and Rachel Sterken (eds), Communicating with AI: Philosophical Perspectives

1. Introduction

In this paper, I will focus on AI systems ("AIs") as very different, or at least potentially very different, kinds of language users from what humans are. Much theorizing about language is, for natural and understandable reasons, focused on human language, primarily the natural languages we use. But when asking philosophical questions about language, we often want to consider what languages in general are, and not only consider human languages. There is some reason to think that AIs are different from us in relevant respects, so asking questions about languages used by AIs may be useful for these general questions about language.

Some of my remarks here will be very general, and have application way beyond AI. These remarks will address questions such as: what possible languages can there be, besides the languages that we humans tend to use? These general remarks may of course be relevant to AI, but only because AIs are among the putative language users that can be thought to employ languages with different semantic and metasemantic features. But other remarks, especially towards the end, will be more directly related to AI. But even those general remarks will trade only on general features of prominent language-using AIs. In brief, the relevant features are these: they are trained on very large text corpuses, and they learn language through general learning mechanisms and not through using some language-specific faculty.

It is of course possible to reasonably doubt that AIs genuinely use and understand language, as opposed merely to behaving in certain ways as if they do. They may produce outputs that are much like those of genuine language users but that is not enough to be genuine language users. But I will here by and large just set such doubts aside. I will simply assume that AI systems are genuine language users, and only ask

questions about the semantics and metaphysics of the linguistic items they consume and produce. The question of whether AI systems are genuinely use and understand language is orthogonal to the issue of how, assuming that they do, the the semantic and metasemantic facts regarding their languages relate to the semantics and metasemantics of human natural languages.

In recent work, Herman Cappelen and Josh Dever have discussed some focused on ways in which the language of AIs may be fundamentally different from human language. While they have main focusedon alien *metasemantics* and my ultimate focus will be on alien *content*, their discussion serves as a natural springboard for mine and hence I will start by discussing what they say. First I discuss metasemantics, and then I turn to my main issue here, content.[1]

2. Cappelen and Dever on metasemantics

Let me start by discussing what Cappelen and Dever say about metasemantics. The technical notion "metasemantics" has come to mean some different things in the literature, but as Cappelen and Dever discuss it, metasemantics is concerned with "what features make things mean what they do".[2]

Cappelen and Dever say the following about metasemantics as practiced:

> Philosophical work in metasemantics, because of its focus on creatures like us, has produced what we will call an *anthropocentric metasemantics*. The existing philosophical accounts of content determination are too parochial by being too focused on contingent features of human communicative/representational practices.[3]

Their point is that metasemantics as practiced is anthropocentric and needs to be "de-anthropocentrized".[4] When in their (forthcoming) they elaborate on their views on metasemantics they make the distinct point that metasemantic facts are contingent:

> Disputes between internalists and externalists are often framed as if one of these two positions is a necessary truth capturing some essential fact about the nature of

[1] Issues having to do with alien content are discussed in greater detail in Eklund (forthcoming).
[2] Cappelen and Dever (forthcoming), p. 8.
[3] Cappelen and Dever (2021), p. 69. Where I use italics, the original has boldface.
[4] Cappelen and Dever (2021), p. 69.

> content and communication. But surely this isn't right. Classic arguments for metasemantic externalism—in, for example, Kripke (1980) and Burge (1979)—start with particular contingent facts about our languages.[5]

The connection between the two passages is this. Underlying the charge that metasemantics as practiced is anthropocentric is the idea that different used languages can have different metasemantics; but then it is a contingent fact that a given metasemantic theory accurately describes why the symbols used by some given community mean what they do.

There is some reason to think that Cappelen and Dever overstate the case in this second passage. They are arguably right that "classic arguments for metasemantic externalism" do "start with particular contingent facts about our languages". And one can imagine linguistic communities whose judgments about the cases that these arguments are centered on are systematically different from our judgments, in such a way that principled application of the method of considering such judgments should lead to the conclusion that their languages work differently in relevant respects. Their word "water" is true of XYZ on Twin Earth even if our word "water" is not and the explanation is that for them the description associated with "water" determines reference in a way that it does not do so for us. But externalism can still be a necessary truth. The argument for it being so would be that in order for a creature to have thoughts and use symbols with meaning in the first place, the right relations to the environment must obtain. Once these relations do obtain and the creature has thoughts and uses symbols with meanings, internal facts about the creature might determine which thoughts and symbols have the same meanings: but it is only through relations with the external environment that any of the symbols have meaning in the first place.[6]

Even if the truth of externalism is not contingent, Cappelen and Dever's claim that metasemantic facts are contingent could still be correct. It could be that while externalism is a necessary truth, it is a contingent matter exactly which metasemantics with externalist implications is correct. In the case of the two "water"-using communities, the true stories of how meaning is determined may be systematically different, even if both stories have to incorporate some externalist element. And if AIs

[5] Cappelen and Dever (forthcoming), p. 10.
[6] Cappelen himself elsewhere seems to make this very point. See Cappelen (2018), p. 166f.

work very differently from us, the correct metasemantics for the languages they use may be different still.

Taking this picture seriously, there does appear to be a question of what makes it the case that some metasemantics is correct for some language, and another metasemantics is correct for another. The initial question which a metasemantics is meant to answer, "what features make things mean what they do?", has a language-specific higher-order counterpart:

> What features make these-and-these features make the expressions of this-and-this language mean what they do?

This latter question is language-specific for the obvious reason that it concerns the meanings of the expressions of a specific language. And it is higher-order, for it doesn't direct concern what features make things mean what they do, but its counterpart concerning what features make certain features make things mean what they do.

There is a regress looming, and for each level the new question that arises gets more complex. If we can significantly ask the higher-order counterpart of the question that a metasemantics is meant to answer, what is to stop us from asking a question of still higher-order?

> What features make thus-and-such (level 2) features make these-and-these (level 1) features make the expressions of this-and-this language mean what they do?

Cappelen and Dever themselves note that there is an endless hierarchy here but do not see this as a cause for concern. Instead they say "we endorse this endless hierarchy of theorizing", while adding, "[T]here is of course a practical limit to what we humans can process and grasp, but that isn't the limit of interesting inquiry".[7]

Cappelen and Dever seem to see only a practical problem here: we cannot actually investigate all of metasemantics, metametasemantics, metametametasemantics, etc. But if one takes seriously how they describe metasemantics (again: what features *make* things mean what they do?), a metasemantics provides an answer to how meaning is *determined*. And then the endless hierarchy corresponds to an endless chain of

[7] Cappelen and Dever (2021), p, 75.

dependence: something determines that our expressions mean what they do, something in turn determines this determination fact, and so on. And whatever to say about the possibility of such chains of dependence – many philosophers would want to say that they are impossible: there cannot be turtles all the way down – one may well be skeptical of there being an endless chain of metasemantic, metametasemantic, metametametasemantic, etc., facts.

It seems something has gone wrong. But what? Compare the following alternative, more flatfooted description of what is going on. A *general metasemantics* is a fully general account of what determines the meaning of any expression of any language. Insofar as any theory that purports to be a general metasemantics gets things wrong for any kind of expression of any language, it fails. In addition to a general metasemantics there are various specific metasemantic theories, which do not aspire to the same generality. But the relationship between a general metaemantics and a specific metasemantic theory is not one of metaphysical determination (any more than in general, the relation between a fully general claim and various related specific claims is one of such determination.)

Cappelen and Dever themselves centrally appeal to Williamson's principle of knowledge-maximization in their account of meaning. This describe Williamson as proposing this as a "meta-metasemantic principle" to the effect that "the correct metasemantics is one that maximizes knowledge for the interpretee".[8] If Cappelen and Dever are right and the knowledge-maximization principle should be regarded as a metametasemantic principle rather than as a metasemantic principle, then obviously there is reason to believe in a hierarchy of metalevels (or at least the beginnings of one). But it is unclear why it should be regarded as a metametasemantic principle, and Cappelen and Dever do not really pause on the matter. A different view is that the principle of knowledge-maximization simply is a general metasemantic principle, and any appearance that it is "metametasemantic" is just due to it being general in such a way that other claims which are regarded as metasemantic can be derived from it.

Cappelen and Dever use the example of AI to make the point that actual metasemantic theories are anthropomorphic and cannot serve as general metasemantics. But it is not clear that the example of AI is needed to make the point. Davidson's

[8] Cappelen and Dever (2021), p. 76.

Swampman example, already familiar from the literature, will do.[9] Cappelen and Dever themselves bring up this example. Swampman is a creature molecule by molecule identical to a human but with no evolutionary history, having come into existence by a fluke. If having an evolutionary history is necessary for having thoughts with content, as on some theories, then Swampman can think no thoughts with content – and that is arguably an unwanted consequence.[10]

Already the Swampman example seems to show that a metasemantic theory that heavily relies on evolutionary facts will get some possible cases wrong, even if it is extensionally correct as a theory of the metasemantics of human language. Given that the Swampman example exists, and given that Cappelen and Dever themselves take the Swampman example to be persuasive, what is the possible role and relevance of AI? One thought might be: AI is actual, whereas Swampman is merely possible. But as so often in philosophy, it is unclear why actuality matters. The principles discussed are necessarily true if true, whence merely possible cases are relevant. Sometimes merely possible cases are dialectically weak because it may be doubted that the cases are possible. But it would be odd to doubt that Swampman is metaphysically possible.

In fact, the example of Swampman has some advantages over focusing on AI. I said I would set aside doubts regarding whether Ais genuinely can think. But here such doubts are relevant Swampman seems to think thoughts with contents because of the obvious deep similarities between Swampman and us. AIs are more dissimilar from us. And what is more, in the case of AIs, one can explain away the appearance that they employ contents by noting that it is simply instrumentally useful for us to treat them as if they do. We may speak as if AIs are employing contents but the straightforward truth is that it is just convenient for us to treat them as doing so.

The general point is that even if there are content-employing AIs, and even if the metasemantics for them and the languages they employ is different from the metasemantics for human representations, already examples familiar from the literature, such as Swampman, point to the need for a more general metasemantics.

3. Alien content

[9] See Davidson (1987).
[10] Cappelen and Dever (2021), p. 127.

A distinction analogous to that between a general metasemantics and specific metasemantic theories is relevant also when we turn from metasemantics to semantics proper. It is one thing to say what the meanings of the expressions of a given language are (specific semantic theory); another to give a fully general story about meaning. The need for such a distinction becomes more plain in light of the possibility that languages may differ greatly. Already the human natural languages that are the most familiar and extensively studied contain many different kinds of constructions, and some are better understood than others. Successfully characterizing the meanings of expressions of a given human natural language, or even all human natural languages, is a far cry from characterizing meaning simpliciter. Distinguish between a general theory of meaning, seeking to characterize what meaning is, for all possible languages, and a specific theory of meaning seeking to describe the meanings of the expressions of a particular language. This is parallel to the distinction between a general metasemantics and specific metasemantic theories. Both when it comes to metasemantics and when it comes to semantics there is reason to distinguish between giving a correct account of what goes for some languages – for example human natural languages – and what goes for language generally.

There are possible perspectives given which, for principled reasons, there can be no possible languages different from the languages we speak. But the natural default view is surely that there is a vast space of possible languages, and there are languages with expressive resources of a kind not found in human natural languages. In other work, I investigate whether and to what extent there can be alien languages, and more generally, alien representations.[11] More specifically, my focus is on whether and to what extent there are languages that differ *semantically* from familiar kinds of languages, and doing so in a *structural* way (and not just in, for example, which objects and properties they have words for.

While Cappelen and Dever focus mostly on alien metasemantics, they also do discuss this sort of thing, under the heading of *alien content*. Their first remarks are these:

> …another possibility is that the crucial factor identified by the AI system cannot be described in our language. A relatively innocuous version of this would be a disease we had never detected and for which we thus had no name. Less innocuous versions

[11] Eklund (forthcoming).

> might be causal factors that don't fall nicely into our existing categories of "disease", "violence", and so on, so that we would even find it hard to work out what sort of new thing we were looking for. AI systems thus confront us with the possibility of alien content.[12]

They go on to note that in this example, "we at least know the logical category of the alien content" and "the shared logical category gives us a place to start". But "we may not always have the supporting crutch of a shared logical category".[13] And what is more,

> GPT-3 (if it means anything by its outputs) could be expressing contents without truth conditions, contents formally representable only as complex constraints on probability functions or other information measures, dynamic update rules on alien scoreboards, and so on. Again, there is no guarantee that anything "said" is something we could say or understand.[14]

What Cappelen and Dever present are three kinds of cases of what can, in some sense or other, count as alien content. First, there are the cases that in relevant ways are at bottom just like Nelson Goodman's famous "grue" example. As Goodman defines "grue", something falls under this predicate exacly if it is either green and observed before t (where t is some future time), or blue and not so observed. Intuitively, "grue" seems inherently disjunctive. (Inherently: it is not just that we happen to pick out what it stands for using a disjunction but that it seems intuitively that grueness just is disjunctive.) Call any predicate that seems similarly inherently disjunctive *grue-like*. The grue-like predicates may be so strange from our perspective that we may in practice be unable to learn them, but still they are just predicates. The expressions belong to familiar logical categories. A language differing from familiar ones just by having grue-like simple predicates would not count as alien in the sense I characterized above, since it is not thereby structurally different from familiar languages. Second, there are cases where an alien language user, for example an AI, might use subsentential expressions of different – alien – "logical categories". For example, the sentences may not contain singular terms and predicates but expressions of these alien logical categories. Third, the alien language

[12] Cappelen and Dever (forthcoming).
[13] Cappelen and Dever (forthcoming).
[14] Cappelen and Dever (forthcoming).

may not have sentences with truth conditions, but instead the sentences, or the representations most closely resembling sentences, express "constraints on probability functions", "dynamic update rules on alien scoreboards", etc. Here it is not just a matter of the alien sentences being built up from constituents of alien logical categories: the alien language does not use sentences in the ordinary sense, the things with truth conditions (in contexts), at all.

With respect to each possible kind of alienness one can ask if it is *possible* for there to be languages that are alien in the relevant sense. With respect to each kind, one can also ask whether there are *actual* examples of languages that are alien in the relevant sense. When it comes to the first kind, it seems plausible that the answers to both questions is yes. It seems rather obvious that there could be languages with grue-like predicates, including predicates that are so grue-like that we would have difficulty comprehending them. And it seems natural to speculate (whatever in the end is the truth of the matter) that some actually used human language might include simple predicates that appear to us to be grue-like in this way. But again, while Cappelen and Dever call such languages "alien" they are not alien in the sense in which I prefer to use the term. In what follows I will set aside languages that are "alien" only in this sense. I will focus only on the two other kinds of alienness. Are there possible *subsententially* alien languages – languages differing from familiar ones at the subsentential level? Are there possible *sententially* alien languages – languages differing from familiar ones by not having sentences but some sort of counterpart, not associated with truth conditions?

There are distinctions to draw beyond those already drawn. First, consider the notion of different logical categories. One way to understand what Cappelen and Dever gesture towards is that they are envisaging the possibility that *instead* of expressions of familiar logical categories, a creature could use a language whose expressions belong to alien logical categories. Another possibilty it is that a creature could use expressions of alien logical categories *in addition to* expressions of familiar logical categories. Second, consider the possibility of alternatives to sentences and truth conditions. One reaction is that contents without truth conditions should not be seen as very strange: expressivists have presented theories of actual domains of discourse according to which that is how content works there. And whether or not the expressivists are right, the possibility that they could be right does not seem so very strange. What would be more alien are contents that in some way are descriptive in much the way declarative sentences typically

are but which do not have truth conditions. The idea that the meanings of sentences should ultimately be explained or characterized in terms of something other than truth conditions is of course well known. Friends of inferentialist views and of dynamic semantics have proposed such alternative accounts. But even the friends of these alternative accounts do not reject the claim that sentences have truth conditions.

While Cappelen and Dever bring up the the possibility of AIs using alien languages, they do not pause on the question of whether the alien languages of the kinds they mention really are possible. Can there really be languages of the kind pointed to? While I do believe that the answer is yes, I think there are complications worth pausing on. Towards the end I will again focus on AIs specifically, and the reasons for taking seriously that AIs might come to use alien languages. Dever's (2020) is a searching discussion of what concepts, or meanings, there can be, but it is focused on individual concepts or words and not on the structural or general features at issue when it comes to either subsentential or sentential alienness of the kind currently at issue.

Some theorists make pronouncements that seem to assume that all possible languages in relevant respects work the way that languages familiar to us seem to do. For example, they make pronouncements to the effect that there can be no sentences without predication. Donald Davidson (2005) says,

> …if we do not understand predication, we do not understand how any sentence works, nor can we account for the structure of the simplest thought that is expressible in language. At one point there was much discussion of what was called "the unity of the proposition"; it is just this unity that a theory of predication must explain.[15]

Is Davidson right? Maybe. But I do not see that he bases his claims on anything other than consideration of how sentences of familiar languages work. It is a whole other thing to argue that all sentences of all language must employ predication. Might there be sentences with expressions belonging to alien logical categories, and might such sentences not make use of predication?[16]

[15] Davidson (2005), p. 77.
[16] Davidson's argument in his (1974) regarding the possibility of alternative conceptual schemes may be thought to be an argument for the claim quoted. I disagree. See chapter 4 of Eklund (forthcoming).

In his (1970), David Lewis remarks, "Semantics with no treatment of truth conditions is not semantics".[17] He does so in the course of criticizing the idea, seemingly defended by Katz and Postal, that one can do semantics by translating between two languages. That idea seems problematic: I can know how to translate between two languages unknown to me without understanding either. Somehow or other, semantics must relate to what one understands when one knows how to use certain expressions. But saying, with Lewis, that semantics must include a treatment of truth conditions goes beyond this. Maybe all familiar languages are such that a semantics for them must include such a treatment. But what exactly rules out that we could come across some creatures – including AIs – for which that is not so?

Compare too Lewis (1975). Lewis there famously describes two conceptions of language, languages as abstract entities and language as a social phenomenon, and proposes a way to combine the conceptions by describing under what conditions a linguistic community uses a certain language in the sense of abstract entity. What Lewis says when describing the abstract conception is this:

> What is a language? Something which assigns meanings to certain strings of types of sounds or of marks. It could therefore be a function, a set of ordered pairs of strings and meanings. The entities in the domain of the function are certain finite sequences of types of vocal sounds, or of types of inscribable marks; if σ is in the domain of a language £, let us call σ a sentence of £. The entities in the range of the function are meanings; if σ is a sentence of £, let us call £(σ) the meaning of u in £. What could a meaning of a sentence be? Something which, when combined with factual information about the world – or factual information about any possible world – yields a truth-value. It could therefore be a function from worlds to truth-values – or more simply, a set of worlds. We can say that a sentence σ is true in a language £ at a world w if and only if w belongs to the set of worlds £(σ)[18].

Truth conditions play a central role here. And they play a central role in Lewis' characterization of his synthesis of the two conceptions. In general terms, Lewis links

[17] Lewis (1970), p. 18.
[18] Lewis (1975), p. 3.

them by describing when a language in the abstract sense is used by a given population. But what is it for a language to be used by a population? Lewis says,

> My proposal is that the convention whereby a population P uses a language £ is a convention of truthfulness and trust in £. To be truthful in £ is to act in a certain way: to try never to utter any sentences of £ that are not true in £. Thus it is to avoid uttering any sentence of £ unless one believes it to be true in £. To be trusting in £ is to form beliefs in a certain way: to impute truthfulness in £ to others, and thus to tend to respond to another's utterance of any sentence of £ by coming to believe that the uttered sentence is true in £.[19]

But this account of course presupposes that the meanings of sentences – sentences generally, not just sentences of languages familiar to us – are properly characterized in terms of truth conditions (and more fundamentally, that all languages have sentences). There are some well-known issues regarding Lewis' specific conception: must not meanings be more fine-grained than truth conditions? But one can go farther and ask whether there are not alternatives to looking to truth conditions in the first place: why not truth* conditions, for some suitable alternative to truth, *truth**? One historically important example is the verificationist focus on provability. Some verificationists sought to identify truth with some kind of provability. But another line for a verificationist to take is to allow that truth is distinct from provability but still persist in giving an account of meaning in terms of the latter, and not worry about truth conditions, properly so called, at all. This is just one example of an alternative to truth. But if there is one such example there is bound to be many more.

Cappelen and Dever note that "[T]he possibility of alien content is independent of the structure of the vehicle" and that "syntactic form is no guarantee of semantic form".[20] Their point is that even of AIs produce outputs that are syntactically similar to what we produce, that does not mean that they are semantically similar. This sort of thing cuts in different ways. It can also be that AIs produce outputs that seem really strange and weird, but the outputs are different only non-semantically: the semantic form of the outputs is the same.

[19] Lewis (1975), p. 7.
[20] Cappelen and Dever (forthcoming).

Let me elaborate on this point. It is easy to concoct things that *on the face of it* are semantically alien languages. Agustín Rayo, in forthcoming work, describes a language (or "language") with strings of symbols ("names"), all of the same category, which can be interpreted to systematically correspond to sets of possible worlds, and one can imagine some sort of language user, whether human or artificial, employing these strings of symbols thereby somehow representing these possible worlds.[21] All of this seems clearly possible. It seems that Rayo successfully describes a language, that the language employs sentences with truth conditions, but the constituents of the sentences are not names or predicates in any ordinary sense: they do not refer or predicate but in other ways help to determine which set of possible worlds the sentence as a whole corresponds to. Does this mean that there are languages that are alien in the second way outlined? Not clearly. Concocting structured things that can be correlated with sets of worlds is easy. What justifies seeing these things as *representations*, and as alien representations?

There are two issues to distinguish between. One – call this *the representation issue* – concerns whether some concoction presented as an alien language amounts to a system of representations at all. For example, one can ask: Is it necessary for something to count as a system of representations that the would-be representations are associated with truth conditions? ´More importantly, when it comes to assessing Rayo's example, is it sufficient that they can be so associated? The first of these questions relates to the third kind of alienness, and the issue of whether there may be languages whose (broadly descriptive) sentences are not associated with truth conditions. (In connection with the representation issue, there is also the subsidiary question of what counts as a *linguistic* representation. Contrast: pictures, maps, diagrams,…) A different issue – call it *the alienness issue* – concerns whether the concoction really is alien. It is easy to devise a language which *appears* alien: its sentences superficially seem more like pictures than like sentences, or its symbols all belong to the same kind, or its symbols seem like nothing we would be inclined to see as symbols, or…

Rayo's purported language illustrates both issues. One can envisage doubts about whether lists correlated with sets of possible worlds really qualify as representations. And even setting aside such doubts, there remains the question of whether the language is as alien as it may seem at first glance. Compare a different example. Here is Peter Sullivan, discussing an idea from Ramey's famous article "Universals":

[21] See Rayo (forthcoming).

> [Ramsey] observed that nothing rules out propositions consisting entirely of several expressions of the same type […] He was not suggesting that we could make sense of non-sentences like 'Socrates Plato' or 'mortality senility wisdom'. Any type or category that did self-combine as those familiar ones fail to would be very different from those we employ. It would be employed in thought of a very different logical shape, and altogether alien to us.[22]

On the face of it this is a subsententially alien language (and in fact, Rayo's language, if indeed it is alien, is an instance of the kind of language Sullivan describes: its subsentential expressions are all of the same kind). But a complication is illustrated by the fact that we *can* make sense of the "non-sentences", "Socrates Plato" and "mortal senile wise". We can easily envisage a language with a convention that a sentence consisting of two names expresses that the bearers of the names stand in a given relation R; say, identity. We can easily envisage a language with a convention that a sentence consisting of a certain number of predicates expresses something like that something has the properties expressed by the predicates. "Socrates Plato" would mean that something is identical to Plato. "Mortal senile wise" would mean that something is mortal, senile and wise. Languages with these conventions would look alien. But nothing about these conventions guarantees that they are in any interesting sense alien: the differences with familiar languages are superficial. And this problematizes Sullivan's example: a language could contain sentences with expressions all of the same type and still not be alien. This is not to say that the differences will in all cases be merely superficial. But it raises the question of what decides when they are not.

There are a number of different possible reactions to the representation issue and the alienness issue. I won't here try to discuss all of them in any detail. All I want to do is to present a menu of (some of the) options. First, one might be a *superficialist* (to coin a label) and say that already minimal criteria – like, for example, matchability with sets of possible worlds – suffice for something to be a representation, and that already superficial aspects suffice for some representations to be alien.[23] There is nothing deeper that is relevant for either of these issues. Second, one might take a *metaphysics-first* line and take

[22] Sullivan (2020), p. 195f.
[23] Superficialism is similar in spirit to what Rayo (2013, pp. 13ff) calls compositionalism.

the primary issue be that of what propositions are, and then address the representation issue and the alienness issue on the basis of such considerations. Something is a language or system of representation only if some putative representations it contains express propositions. The metaphysics-first line is in itself neutral on the question of what propositions are, and with a suitably liberal view on what propositions are, the metaphysics-first view may be indistinguishable from a superficialist view. But the friend of this general idea might for example hold that all logically non-complex propositions are Russellian propositions complex entities with one or more objects and a property or a relation as constituents – and say that for some system to genuinely be a system of representations, it must be possible to assign Russellian propositions to representations of that system in the right way. On this conception, whether a language is alien depends on whether its sentences express alien kinds of propositions. But this is immediately ruled out if all propositions are Russellian. Third, one might adopt what can be called a *cognition-first* line. One might say whether something is a system of representation is determined by whether some possible agent/thinker can use it as one. And as for alienness the approach would be the following. Consider languages of the kind Sullivan considers. If any agent/thinker in any sense using such a language would do so by employing representations with familiar structure – a sentence "ab" is translated into a sentence of its language of thought of the form "aRb", then those languages would count as only superficially alien. But if the language of thought could contain sentences "ab" things would stand differently. I mention language of thought when illustrating the cognition-first line but I don't mean to imply that the cognition-first line needs to be bound up with the idea of a language of thought. There is arguably a sense in which I am using English as a system of representation even if English is not my language of thought, and even if my language of thought has a quite different structure.

I am not sure how attracted I myself am to anything like the cognition-first line, and I would not be its best advocate. But here is a way of motivating it. Consider the possibility of a language containing only sentences without structure, each corresponding to a set of possible worlds. Suppose further that the language is infinitary, and has a sentence for each set of possible worlds. Given the limitations of our minds, we humans could not speak such a language. But it can be maintained that a certain kind of higher creature might. However, it is here that a friend of the cognition-first line might demur. She might say that even given the existence of the higher creatures, there remains the question of how an unstructured sentence of their language could come to represent a

given set of possible worlds rather than another. And absent a satisfactory answer to this pressing metasemantic question, we should reject the possibility of even a higher creature using this sort of language. By contrast, it may further be speculated, a satisfactory explanation can be given of how some language-users could use some semantically structured sentences to represent sets of possible worlds. These points together would yield that while there certainly can be languages with structured sentences, there cannot be languages whose sentences are unstructured, of the kind that the higher creatures would speak.

Or that would be how a friend of the cognition-first view might reason. To stress, what I just presented is speculation. And the speculation wouldn't be something the friend of the cognition-first view is committed to: I am just presenting it to illustrate how she might reason. Even supposing the speculation is on the right track, there are further twists and turns. For example, given a language of the kind agreed to exist, it seems we can introduce semantically unstructured sentences by stipulation: "Let 'S' be a semantically unstructured sentence expressing the set of possible worlds that "snow is white" expresses". One response to this in turn is to say that such a stipulation always fails: the best one can do is to get "S" to superficially lack structure.

I am describing all this briefly, only to indicate what some relevant moves are, and my description leaves a whole host of issues unanswered. What is it for an agent/thinker to use something as a system of representation? What is it to assign propositions, Russellian or otherwise, to representations "in the right way"?

My own view is that there are alien languages of the kind described, and that hence a general theory of meaning cannot merely describe familiar kinds of languages. My belief is not based on considerations about purported examples of alien languages. Maybe all purported examples can be problematized in the way I have shown that the examples brought up can be. Instead my beliefs is based on the basic question *why not?* and consideration of, and rejection of, specific considerations to the effect that alien languages are impossible.

How do AIs relate to alien languages? AIs are different from us, and so there is a theoretical possibility that they could come to employ alien languages, if such exist. But that by itself is just a rather tenuous connection, and does not set AI systems apart from, for example, non-human animals and extraterrestrials. However, in the next section I will provide some reasons for thinking that the possibility of alien languages has a more interesting connection to AI.

4. Deep learning and theoretical linguistics

In his (2021), Gabe Dupre compares the aims of theoretical linguistics and the aims of deep learning (DL) of the kind central in AI research. His conclusion is that "we should not expect DL models to illuminate linguistic theory".[24] In brief, Dupre defends the kind of view on theoretical linguistics, conceived of as the project of "describing and explaining the properties of human language", associated with the Chomskian tradition. On the Chomskian view, "human abilities to acquire a language depend in large part on linguistically-specific innate structures". DL, by contrast, "centrally relies on the use of general (i.e., not language specific) learning mechanisms, aimed at reproducing observable data". I am attracted to the view Dupre defends, But I will not pause to defend it here.[25]

Theoretical linguistics as Dupre describes it is a branch of "cognitive psychology" and "a true theory of human language will thereby provide an account of the distinctive features of human psychology that enable us to learn and use language".[26] This might make it sound as if it is trivial that DL models should not be expected to illuminate linguistic theory: DL models are not meant to provide such an account. But as Dupre notes, it is of course a theoretical possibility that at a suitable level of abstraction, DL models function similarly to human psyches, in such a way that DL mode dols illuminate linguistic theory as characterized. However, Dupre argues that "DL systems are unlikely to complete linguistic tasks in the way that humans do, and […] we are unlikely to learn about human language through investigation of such systems".[27] Dupre relies on the traditional Chomskian distinction between competence and performance. DL systems are aimed at reproducing human performance: reproducing human linguistic behavior. But what theoretical linguistics aims at is a theory of competence, "the underlying rule-system(s) partially responsible for the acquisition and use of language".[28] Dupre argues that performance is by no means a direct reflection of competence.

The Chomskian assumptions on which Dupre relies are, as he remarks, controversial in some circles. Moreover, those opposed to the Chomskian tradition

[24] Dupre (2021), p. 617. Quoted from the abstract.
[25] Chomsky himself has, unsurprisingly, given voice to views very similar to those of Dupre. See Chomsky, Roberts and Watumull (2023), an opinion piece in the New York Times. For criticism of Chomsky's stance, see, e.g., Piantadosi (forthcoming).
[26] Dupre (2021), p. 618.
[27] Dupre (2021), p. 622.
[28] Dupre (2021), p. 626.

might, and sometimes do, take the success of these AI systems at the level of performance as evidence that one does not need to appeal to linguistically-specific innate structures. But let me not get into that. My main aim is not to determine whether Dupre is right. What I want to focus on is this. Suppose that he in fact is right. What should we then say about the semantics of the sentences produced by the relevant AI systems? One possibility is of course that the semantics of those sentences is just the same as the semantics of the corresponding human-produced sentences. We and these systems function differently internally, but the sentences we produce have the same features. This may well be the most reasonable thing to hold. But consider the following different possibility: If Dupre is right, then the sentences of DL are different from superficially similar sentences of English when it comes to linguistic structure. The sentences of DL have different structure, or perhaps, given how DL works, the sentences of DL simply do not have the kind of linguistic structure that sentences of natural languages have. The sentences produced are superficially similar but that hides significant differences when it comes to structure. Recall from earlier the example of a language with sentences like "mortal senile wise" but where what is expressed by those sentences is something utterly familiar. That is a case of a language which may look alien but at bottom is not. What we are considering now is a mirror image of this. The sentences produced look familiar, but because of differences in processing the sentences are actually alien.

Compare perhaps Rayo's language, described above. The sentences of that language have structure. But the structure is nothing like ordinary linguistic structure. The structure is more like that of descriptions of coordinates in coordinate systems. At the same time the sentences of the language have ordinary truth conditions. Maybe the sentences used by system with only general intelligence will only have structure in the sense, however it is to be described, in which the sentences of Rayo's language have structure.

To stress, it is very unclear how plausible the speculation just given voice to really is. One way to motivate it would be via the cognition-first view on structure, but the cognition-first view is itself unclear and on shaky ground. Also, it remains that what AIs produce has structure in the sense that not all strings are well-formed, and which strings are well-formed and not can be explained by appeal to differences in kind between different expressions. In general, and obviously, there is no reason why a system endowed only with general intelligence and no domain-specific learning mechanisms cannot come to learn and use the sentences of a human natural language., and use them

with the exact semantic structure they actually have. So the speculation I gave voice to faces all sorts of problems.

But even though an AI system employing deep learning can come to understand and employ the sentences of natural language, with the structure these sentences actually have, the speculation discussed suggests other possibilities. For example, if an AI system of the kind discussed develops its own language, is there good reason to suppose that this language would have the kind of structure that human natural languages do? If the systems are internally different from humans when it comes to language processing in the way Dupre describes, there is no reason to assume that a language it develops on its own has the structure that familiar human languages have.

When discussing AI metasemantics I compared Davidson's Swampman example, and suggested that Cappelen and Dever's points about anthropocentric metasemantics could have been made equally well, or better, using that example. Things stand differently with alien content. In that case it makes sense to focus on thinkers with differently structured minds, and if AIs can be said to have minds at all – something I haven't here weighed in on – they are arguably different in structure from human minds, whereas Swampman's mind would have a similar structure to human minds.

REFERENCES

Cappelen, Herman: 2018, *Fixing Language*, Oxford University Press.

Cappelen, Herman and Josh Dever: 2021, *Making AI Intelligible*, Oxford University Press.

Cappelen, Herman and Josh Dever: forthcoming, "AI with Alien Content and Alien Metasemantics", in Ernest Lepore (ed.), *Oxford Handbook of Applied Philosophy of Language*.

Chomsky, Noam, Ian Roberts and Jeffrey Watumull: 2023, "The False Promise of ChatGPT", *The New York Times*, March 8, 2023, https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html.

Davidson, Donald: 1974, "On the Very Idea of a Conceptual Scheme", *Proceedings and Addresses of the American Philosophical Association* 47: 5–20.

Davidson, Donald: 1987, "Knowing One's Own Mind", *Proceeedings and Addresses of the American Philosophical Association* 60: 441-58.

Davidson, Donald: 2005, *Truth and Predication*, Harvard University Press, Cambridge, Massachusetts.

Dever, Josh: 2020, "Preliminary Scouting Reports from the Outer Limits of Conceptual Engineering", in Alexis Burgess, Herman Cappelen and David Plunkett (eds.), *Conceptual Engineering and Conceptual Ethics*, Oxford University Press, Oxford.

Dupre, Gabe: 2021, "(What) Can Deep Learning Contribute to Theoretical Linguistics?", *Minds and Machines* 31: 617-35.

Eklund, Matti: forthcoming, *Alien Structure: Language and Reality*, Oxford University Press. (To be published June 2024.)

Lewis, David: 1970, "General Semantics", *Synthese* 22: 18–67.

Lewis, David: 1975, "Languages and Language", in Keith Gunderson (ed.), *Minnesota Studies in the Philosophy of Science,* volume VII, University of Minnesota Press.

Piantadosi, Steven: forthcoming, "Modern Language Models Refute Chomsky's Approach to Language", in Edward Gibson and Moshe Pollak (eds.), *From Fieldwork to Linguistic Theory: A Tribute to Dan Everett*, Language Science Press.

Rayo, Agustín: forthcoming, "Why I am not an Absolutist (or a First-Orderist): Reply to Menzel and Pickel".

Sullivan, Peter: 2020, "Varieties of Alien Thought", in Sofia Miguens (ed.), *The Logical Alien: Conant and His Critics*, Harvard University Press, Cambridge, Massachusetts, pp. 183-201.