Minds and Machines (2022) 32:585–611
https://doi.org/10.1007/s11023-022-09609-7

ORIGINAL ARTICLE

The End of Vagueness: Technological Epistemicism, Surveillance Capitalism, and Explainable Artificial Intelligence

Alison Duncan Kerr[1] · Kevin Scharp[2]

Received: 14 May 2020 / Accepted: 2 July 2022 / Published online: 11 September 2022

Abstract

Artificial Intelligence (AI) pervades humanity in 2022, and it is notoriously difficult to understand how certain aspects of it work. There is a movement—*Explainable* Artificial Intelligence (XAI)—to develop new methods for explaining the behaviours of AI systems. We aim to highlight one important philosophical significance of XAI—it has a role to play in the elimination of vagueness. To show this, consider that the use of AI in what has been labeled *surveillance capitalism* has resulted in humans quickly gaining the capability to identify and classify most of the occasions in which languages are used. We show that the knowability of this information is incompatible with what a certain theory of vagueness—*epistemicism*—says about vagueness. We argue that one way the epistemicist could respond to this threat is to claim that this process brought about the end of vagueness. However, we suggest an alternative interpretation, namely that epistemicism is false, but there is a weaker doctrine we dub *technological epistemicism*, which is the view that vagueness is due to ignorance of linguistic usage, but the ignorance can be overcome. The idea is that knowing more of the relevant data and how to process it enables us to know the semantic values of our words and sentences with higher confidence and precision. Finally, we argue that humans are probably not going to believe what future AI algorithms tell us about the sharp boundaries of our vague words unless the AI involved can be explained in terms understandable by humans. That is, if people are going to accept that AI can tell them about the sharp boundaries of the meanings of their words, then it is going to have to be XAI.

Keywords Vagueness · Artificial intelligence · Surveillance capitalism · Epistemicism · Explainable artificial intelligence · Machine learning

✉ Kevin Scharp
ks70@st-andrews.ac.uk

Extended author information available on the last page of the article

Artificial Intelligence (AI) pervades humanity in 2022, and it is notoriously difficult to understand how certain aspects of it work. There is a movement—Explainable Artificial Intelligence (XAI)—to develop new methods for explaining the behaviours of AI systems. There are many important philosophical implications associated with XAI. We focus on the philosophical implications of XAI for vagueness. Our account, which ends with XAI, begins by noting the rise of what has been labelled *surveillance capitalism* (Wills, 2017; Zuboff, 2019).

Humans are quickly gaining the capability collect vast amounts of information about people, and this has already expanded from actions and speech to emotional states, cognitive states, and other personality aspects.[1] Many of these capabilities are powered by machine learning algorithms, which comb through these vast quantities of data and change themselves in light of processing it.[2] That is exactly what a machine learning algorithm is—it learns by changing its own parameters as it encounters more data.

The leading corporations employing the surveillance capitalist model, like Google and Facebook, collect this information about everyone so that it can be processed and sold to advertisers. Advertisers, in turn, can sell their goods or services more effectively and make greater profits by understanding and even manipulating every conceivable minute detail of their customers' lives. That is the essence of surveillance capitalism.

We are not going to challenge any of the surveillance capitalism framework presented by authors like Shoshana Zuboff, nor are we discussing its vast ethical or economic implications. Rather, we look at its *philosophical* consequences. In particular, the rise of surveillance capitalism poses a serious threat to certain theories of vagueness. Vagueness is the presence of fuzzy boundaries between categories; for example, in the spectrum of colours between red and orange, some are definitely red, some are definitely orange, and some are not definitely either. Moreover, the border between the definite and the indefinite is fuzzy as well.[3]

One prominent theory of vagueness, *epistemicism*, implies that vague words have sharp boundaries, but those boundaries are unknowable (Williamson, 1994). We argue for the following three theses:

[1] For example, see Ko (2018) for a survey on emotion recognition and Kosinski and Wang (2018) on sexual orientation detection, and see Kosinski et al. (2013) on personality trait detection.

[2] *Artificial intelligence* is the discipline devoted to creating an artificial intelligent agent, something that can do all the things a human can do, but hopefully much better. *Machine learning* is a branch of artificial intelligence that studies algorithms that display remarkable intelligent behaviour without themselves being agents. See Russell and Norvig (2020) on artificial intelligence and Alpaydin (2020) on machine learning.

[3] *Vagueness* is often defined as the presence of borderline cases of classification, but other times by reference to the *sorites paradox* reasoning (e.g., if adding or subtracting a single grain of sand does not change whether the collection is a heap, then either every collection is a heap or none are). See Dietz and Murzi (2010), Sorensen (2018), and Oms and Zardini (2019) for surveys. There has been a shocking amount of new work by leading philosophers and linguists on vagueness, including Sassoon (2013), Raffman (2014), Castroviejo et al. (2018), Bacon (2018), MacFarlane (2020a, 2020b, 2020c), Wright (2021), Fine (2020), Ripley (2021), and Salles (2021).

(1) The information being collected by surveillance capitalism enables exactly what the epistemicist theory of vagueness says is impossible. Thus, if epistemicism is true, then the rise of surveillance capitalism is the end of vagueness. In other words, the epistemicist's interpretation of surveillance capitalism should be that it is the end of vagueness. In other words, something is now possible that epistemicism entails is impossible if there is vagueness.

(2) A better interpretation of the situation is that there is vagueness but the information in question is knowable; hence, epistemicism is false. However, there is an interesting nearby doctrine that we call *technological epistemicism*, which might be true. According to technological epistemicism, vagueness is ignorance, but this ignorance can and does decrease with certain technological improvements like AI.

(3) Regardless of the truth of technological epistemicism, a key issue determining whether people trust AI to tell us about the semantic boundaries of our words will be how well the AI involved can be *explained*. That is, if we are going to accept that AI can tell us more about what we mean, then it is going to have to be *XAI*.

Overall, we think that surveillance capitalism, AI, and XAI will result in considerable elimination of vagueness, and philosophers need an account of vagueness that is compatible with these developments.

1 Epistemicism

The new digital products sold by surveillance capitalist firms, copious information about many aspects of your behaviour and that of almost everyone, was not dreamt of until recently. During the twentieth century, it was safe to assume that no one would ever be able to know that information. Theories of vagueness that developed during this time were no exception, and the assumption tended to be accepted without comment. One important theory of vagueness, epistemicism, is deeply affected by this development from surveillance capitalism.

Epistemicism is the view that vague words, like 'bald', 'heap', and 'red' have sharp boundaries (Sorensen, 2001; Williamson, 1994). For example, it implies that there *is* a particular number of grains of sand that makes a heap. One fewer, and the collection is not a heap. Likewise, for whether a given colour patch is green. For the epistemicist, no matter how indeterminate it seems, there is a fact of the matter as to whether the patch is green or not green. The epistemicist goes on to say that we do not know where this line is and furthermore, we *cannot* know (Williamson, 1994, p.185). So, while we do not know how to draw the line, there is a line to be drawn in every case of vagueness, according to the epistemicist. As such, vagueness, for the epistemicist, is a matter of what we know about language and the world rather than a feature of our language.

We should be clear about our motivations: we are not epistemicists, but we do acknowledge the intuitive appeal of the position.[4] The idea that questions like 'is this pile a heap?' and 'is this patch green?' have definite answers in every case is comforting and seems like a natural perspective. Moreover, we claim that investigating the consequences of epistemicism in the context of surveillance capitalism is illuminating because it exposes some costs of the view, but also uncovers the potential for a new version of epistemicism and a source of support for it. Because we are not concerned with conclusively establishing or refuting epistemicism, we are not going to focus on evaluating the *other* arguments for or against it.

Timothy Williamson is a prominent epistemicist who has argued that the meanings of our words are determined by something like the overall usage displayed by everyone who uses them (Williamson, 1994, pp. 130–132, 135).[5] Williamson provides a response to the most obvious objection to epistemicism: how can finite and patchy uses of linguistic expressions determine completely sharp meanings for those expressions? It seems like usage does not settle every case. Williamson's solution is simple—any cases that are not determined by usage are thereby false. He writes:

> [T]he concepts of truth and falsity are not symmetrical. The asymmetry is visible in the fundamental principles governing them, for (F) is essentially more complex than (T), by its use of negation. The epistemic theorist can see things this way: if everything is symmetrical at the level of use, then the utterance fails to be true, and is false in virtue of that failure (if it says that something is the case). In that sense, truth is primary. At the level of truth and falsity, there is no symmetry to break (Williamson, 1994, p. 208).

The symmetry Williamson is talking about pertains to cases where usage of a predicate does not determine an answer for whether it applies or does not apply to an object. Williamson's answer is that the predicate does not apply to the object in these cases because of the asymmetry between truth and falsity. Hence, the usage of the predicate together with the true/false asymmetry principle determines a meaning with a perfectly sharp boundary for the predicate.

Since we seem to know what our words mean but we do not have any inclination either way in many cases, it seems like epistemicism entails that we do not understand our own words. Williamson' reply to this objection:

> On the epistemic view, our understanding of vague terms is not partial. The measure of full understanding is not possession of a complete set of metaphysically necessary truths but complete induction into a practice. […] To know

[4] Epistemicism has been called highly *counter*intuitive by a number of philosophers, mostly because it seems difficult to understand *how* words could have acquired sharp boundaries; see Mosdell (2015) for recent discussion. We share these *metasemantic* worries, but still acknowledge the attraction of the view's *semantic* commitments.

[5] We have formulated this point in terms of collective use, but Williamson is clear that the point applies to idiolects as well. "What you mean by 'thin' does not depend solely on what you would say in your present circumstances and mood. You have no way of making each part of your use perfectly sensitive to the whole, for you have no way of surveying the whole. To imagine away this sprawling quality of your use is to imagine away its vagueness," (Williamson, 1994, p. 231).

> what a word means is to be completely inducted into a practice that does in fact determine a meaning (Williamson, 1994, p. 221).

The epistemicist view is that our linguistic practices determine where all the sharp boundaries are, but we do not need to know everything about our linguistic practices in order to know what our words mean. We just need to take part in these practices for that. Hence, for Williamson, linguistic understanding is a matter of knowing one's way around a practice rather than knowledge of a fully determinate meaning. A way to put the point is: one needs to understand enough about the global pattern of usage of an expression so that one can get right enough cases to be considered part of the linguistic community. But no one comes close to getting everything right, even though there is a right answer to any regular question like 'Is that wall green?'.

2 Williamson's Arguments for the Impossibility of Knowledge in the Borderline

The message so far: we live in the age of surveillance capitalism, and what the epistemicist assumed was impossible *seems to be possible*. In this section, we examine the arguments for the epistemicist's conclusion that knowledge is impossible in borderline cases of vagueness.

Here is an obvious objection to what has been said so far: it cannot be *the end of vagueness* because either the knowledge in question is possible or it is not. If it is possible to know the sharp borders of our meanings, then vagueness has not come to an end; rather, there never was any vagueness to begin with. Thus, the entire way of framing the discussion so far is misguided.

Reply: Epistemicism has two parts: (i) all vagueness is just ignorance of semantic features and (ii) this ignorance is impossible to overcome. What kind of possibility is at work in (ii)? It seems like it has to be something like: humans are physically incapable of doing what is required to have knowledge. One might call this *technological impossibility*—this is the kind of modality at play in claims like 'it is impossible for humans to travel to another solar system'. That seems like the most plausible and least controversial version of epistemicism, and that is the version we consider throughout. It also fits well with Williamson's remarks about what might be possible given the truth of epistemicism. He writes:

> Suppose that persons with exact physical measurements m are borderline cases for 'thin'. The epistemic theorist has no special reason to deny that a being with cognitive powers greater than any we can imagine could know of someone with exact physical measurements m whether he is thin. Who knows what such a being might know? On the epistemic view, vague utterances in borderline cases are true or false and we humans have no idea how to find out which. It is quite consistent with this view that what is a borderline case for us is not a borderline case for creatures with cognitive powers greater than any we can imagine. Equally, the epistemic theorist has no special reason to assert that such a being could know of someone with exact

> physical measurements m whether he is thin. The cognitive capacities of creatures outside the speech community are simply not to the point (Williamson, 1994, p. 212).

This passage is pretty clear that technological modality is the appropriate interpretation of Williamson's epistemicism on the impossibility of knowledge in borderline cases. It is also clear that this passage was written at a time (1994) before the dominance of AI (or even the internet!) and the impact that it has had on the cognitive capacities of "we humans".

Technological modalities change over time. For example, it was technologically impossible for Leibniz to use Microsoft Word, but not so for Bill Gates in 1975 (*before* founding Microsoft). Accordingly, for the epistemicist, the technological *im*possibility of acquiring and processing the relevant information about meanings and their borders in the past implies that there was genuine vagueness in the past, but if this *is* technologically possible *now*, then there is no genuine vagueness *now*. Hence, the rise of these technological capabilities is properly described as the end of vagueness (if epistemicism is true). We think that this is probably the most plausible route for the dedicated epistemicist to take in responding to the threat posed by surveillance capitalism.

There are four main dimensions of Williamson's argument for the technological impossibility of knowledge of borderline cases (for some example vague expression V):

1. We do not know global usage patterns of V.
2. Global usage patterns of V are variable across nearby possible worlds.
3. We do not know how global usage of V determines a meaning for V.
4. Global usage of V might depend on errors.

Williamson's argument for the impossibility of knowledge in cases of vagueness focuses on the relationship between meaning and use. He assumes for the sake of argument that the meaning of an expression supervenes on usage, and he appeals to this supervenience in his arguments. Let us consider each of these arguments in turn.

1. *We do not know global usage*. This is the obvious one to challenge on the basis of surveillance capitalism. Consider what Williamson writes about the overall pattern of usages for a word:

   > [Y]ou have no way of surveying that pattern in all its details. Since the content of the concept depends on the overall pattern, you have no way of making your use of a concept on a particular occasion perfectly sensitive to its content (Williamson, 1994, pp. 231–232). What you mean by 'thin' does not depend solely on what you would say in your present circumstances and mood. You have no way of making each part of your use perfectly sensitive to the whole, for you have no way of surveying the whole. To imagine away this sprawling quality of your use is to imagine away its vagueness (Williamson, 1994, p. 231).

There are two distinct points in each of these passages: (A) overall usage patterns are unknowable, and (B) there is no way to make any particular use cohere with the overall pattern of usage. Right now, we are only focusing on (A). It is clear that Williamson thinks *very many* uses of an expression are relevant to determining its meaning. We call this an *expansive* view of the data. If only a small portion of the use pattern was responsible for determining meaning, then this argument would not make sense. Moreover, if point (A) were to fail, it seems like (B) would fail as well, since (A) is the only reason given for (B). In other words, the only reason we cannot make our local use cohere with the global pattern of use is that we do not know the global pattern of use in detail. We investigate this issue further below.

2. *Global usage patterns vary across nearby possible worlds*. This is the primary argument Williamson uses (Williamson, 1994, pp. 216–234), and it is also the one that attracted the most attention (Caie, 2012; Gomez-Torrente, 1997; Hawthorne, 2006; Sennet, 2012). It relies on *margin of error principles*, which say that one does not know that p in a possible world w unless p is true in the worlds that are similar to w in certain ways. Note, the principle does not say that one has to have knowledge in all these nearby worlds—just that the proposition is true there. The idea is that if one's belief is false when evaluated at worlds that are similar to our world, then one's belief is true only by luck, so it does not count as knowledge. Williamson's point is that in borderline regions of vague expressions, whether some object has the property in question depends on the location of the exact border for that property. And, the location of that exact border is highly unstable across nearby possible worlds. In some worlds that are relevantly similar to ours, the exact border is slightly different. Hence, even if one could formulate a true belief about the object in question (e.g., it has the property in question), this belief would not be knowledge. It would not be knowledge because in worlds similar to ours, the belief is false.
3. *We do not know how global usage determines a meaning*. Williamson emphasizes this point several times, but never provides an argument for it. Rather, he states that we just do not know how this is supposed to work, and for all we know it is impossible to figure out. Here are a couple of places where Williamson makes the point:

   > Although meaning may supervene on use, there is no algorithm for calculating the former from the latter. Truth-conditions cannot be reduced to the statistics of assent and dissent. In particular, the line between truth and falsity is not to be equated with the line between unanimous and less than unanimous assent, or with the line between majority assent and its absence (Williamson, 1994, p. 206). The epistemic theory of vagueness makes the connection between meaning and use no harder to understand than it already is. At worst, there may be no account to be had, beyond a few vague salutary remarks. Meaning may supervene on use in an unsurveyably chaotic way (Williamson, 1994, p. 209). Even if you did know all the details of the pattern (which you could not), you would still be igno-

> rant of the manner in which they determined the content of the concept, (Williamson, 1994, p. 232).

The clear message is that one cannot just assume that knowledge of the pattern of use that determines sharp meanings will yield knowledge of meanings. One has to also know *how* that determination works.

4. *Global usage might depend on errors*. Here the problem is that words might be used in ways that do not reflect the facts about the world. Williamson's example:

> We can certainly be wrong about whether someone is thin, for we can be wrong both about the person's shape and size and about normal shapes and sizes in the relevant comparison class. These errors may be systematic; some people may characteristically look thinner or less thin than they really are, and there may be characteristic misconceptions about the prevalence of various shapes and sizes. Appeal might be made to dispositions to assent and dissent in epistemically ideal situations or given perfect information, but that is merely to swamp normal speakers of English with more measurements and statistics than they can handle. Perhaps the dispositions to assent and dissent of an epistemically ideal speaker of English would be an infallible guide to thinness, but then such a speaker might know the truth-value of 'TW is thin'. The ordinary basis for attributions of 'thin' is perceptual; such a basis is inherently fallible (Williamson, 1994, p. 207).

The point Williamson is making is that meaning is supposed to be truth conditions, but people make mistakes with words all the time. If you take all the mistakes into consideration when calculating the truth conditions, then you get the wrong truth conditions. But there is no good way of distinguishing the mistakes from the rest of the uses. So even though meaning (truth conditions) is somehow determined by use, this is not a task any human could ever understand because it would require distinguishing truth from error across all uses of the language.

3 The End of Vagueness

Our plan is to address the four major dimensions of Williamson's arguments for the impossibility of knowledge in borderline cases. Our first conclusion is that Williamson is mistaken about which aspects of use determine meaning. Williamson has an expansive view of the use data—very many of our correct uses determine the meaning of a word (maybe the mistakes do too, but given point 4 above, it seems like he might deny that). Instead, according to the linguists who study usage and formulate semantic theories for natural languages, meaning is determined by a relatively narrow range of responses, and most of what is relevant are native speakers' attitudes toward contradictions, synonymies, and entailments.

The ways in which meanings depend on usage have been enshrined in our best theories of natural language semantics for decades.[6] In particular, the sum total of evidence used to construct and adjudicate these theories constitutes the relevant facts about usage. And the semantic theories themselves, which output semantic values for all the linguistic expressions in a given language fragment, enshrine the dependence in question. That is, our best semantic theories together with an account of the evidence relevant to assessing them explain how meanings depend on facts about usage and which facts about usage are relevant. We already know this, even if we did not realize it.

The scientists that study natural languages seem to agree that *patterns* of usage are what is important. Moreover, a few key facts about usage patterns are the primary explanandum for semantic theories of natural language.[7] Insofar as these theories attribute meanings that are relevant to vagueness (and we are not going to question this assumption), we already know which aspects of usage determine meanings.

One classic statement of the basis for semantics comes from David Dowty et al.: "In constructing the semantic component of a grammar, we are attempting to account not for speakers' judgments of grammaticality, grammatical relations, etc. but for their judgments of synonymy, entailment, contradiction, and so on" (1980, p. 2). Dowty et al.'s claim is that semantic theories are answerable only to our judgments about synonymies, entailments, and contradictions—that is it. And this is from one of the most influential works in natural language semantics of all time.

For what it is worth, we prefer the view defended by Tonhauser and Matthewson (2015), which outlines exactly what an element of semantic data is, how to collect semantic data, and how these data bear evidentially on hypotheses about word and sentence meaning. Tonhauser and Matthewson think that semantic theories are answerable to more aspects of overall usage than Dowty et al., but none of these linguists think that anything close to Williamson's expansive view of the data is correct. In other words, none of these views support the epistemicist's contention that very many things people say with a word affect the meaning of that word.

A good example of how meanings depend in part on speaker's judgments about entailments comes from Donald Davidson's (1967) pioneering work on action sentences. If one treats 'he buttered the bread in the kitchen' as a three-place relation (holding between whoever he is, the bread, and the kitchen) and one treats 'he buttered the bread in the kitchen at midnight' as a four-place relation (holding between him, the bread, the kitchen, and midnight), then one cannot explain the entailment from the latter to the former. Davidson offers an account of the logical form of these sentences that treats them as being about events, and this account preserves the entailment. This advantage of Davidson's semantic theory is one reason it has gone on to be so famous and influential (Altschuler et al., 2019; Gillon, 2019).

Regardless of how to spell out the details, the contours of the right account—that is, the scientist's account—have been a foundation for natural language semantics for several generations. If Williamson or any other epistemicist is

[6] See Chierchia and McConnell-Ginet (2000).

[7] See Égré (2015) for an overview. See also Ball and Rabern (2018) for a range of perspectives.

going to dismiss or ignore what the scientists have to say about which aspects of usage determine meanings and how they do so, then there ought to be a clear reason as to why the epistemicist is right and all those scientists are wrong.

We outlined four points Williamson makes in his argument for the impossibility of knowledge in borderline cases of vagueness:

1. We do not know global usage patterns.
2. Global usage patterns are variable across nearby possible worlds.
3. We do not know how global usage patterns determines meanings.
4. Global usage might depend on errors.

Our point about the received view in linguistics on which aspects of usage determine meaning impacts three of Williamson's arguments for the impossibility of knowledge in borderline cases. It undermines Williamson's claim that we do not know how meaning supervenes on usage. It undermines his claim that the relevant usage patterns vary across nearby possible worlds. And it undermines his claim that global usage patterns include all sorts of mistakes (like thin people being called not thin).

Consider the first claim that we know how meanings supervene on usage. The relevant usage patterns determine meanings for natural language expressions in exactly the way that the linguists who construct semantic theories specify the relation between those theories and their data. That is, natural language semanticists do provide us with knowledge of the supervenience function because they give us two kinds of information—(i) the semantic theories themselves and (ii) the kinds of evidence that counts in favour or against these semantic theories. The phenomena that count as evidence for or against a semantic theory are the supervenience base for the supervenience relation in question. The outputs of the semantic theory are what supervene on this base. Of course, this knowledge of the supervenience relation that we gain from understanding semantic theories for natural languages together with their classes of evidence for and against is not perfect, but it is significant, and it shows that the technological possibility this knowledge could be improved without any obvious obstacles.

Moreover, in Williamson's arguments, the possible worlds that are similar to ours are supposed to be ones where some people's judgments about specific cases are different. Of course, these are not the features of uses that linguists think determine meaning. In order to get a possible world in which the meanings of a word is different, it would have to be a world in which native speakers make different judgments about entailments or engage in substantially different patterns of classification, or something else just as fundamental. For example, a world where people accept the sentence 'Heather broke the window' but might not accept 'the window broke' would work. This would require massive changes all over the pattern of usage, not just some slight differences on the margins. Hence, Williamson's false picture of how meanings are determined leads him to think that meanings are much more modally fragile than they really are. Hence, even if one accepts Williamson's margin of error principles, they do not undermine knowledge in borderline cases of vagueness. In all the nearby possible worlds, vague words still have the same meanings, so the sentences about borderline cases that are true in the actual world are true in the relevantly similar possible worlds as well.

Finally, once one understands which judgments by native speakers are relevant to determining the meaning of an expression, one can see that the errors Williamson mentions are irrelevant. Even if people mistake some windows for doors, they will still agree to the relevant entailments, accept the relevant synonymies, reject the relevant contradictions, and engage in the same patterns of classification. So, this point vanishes as well once the mistake about which patterns of usage are relevant is straightened out.

All that is left is Williamson's Point 1, which is undermined by the claim about surveillance capitalism collecting data on all sorts of usage patterns, including the ones that linguists think determine the meanings of our words. Remember, the technical possibility of collecting enough of the right data on linguistic usage to decrease vagueness even a small amount in only one word is enough to refute epistemicism.

We consider a series of objections to our arguments just given. *First objection*: what we have here is a clash of intuitions—those that support the epistemicist and those that support an alternative view on meaning that differs from epistemicism. There is no reason to support the latter over the former, so there is no reason to think that surveillance capitalism poses any threat at all to epistemicism.

Reply: on one side we have the epistemicist's contention that very many uses of language (or at least the correct ones) influence the completely determinate but unknown semantic features of our words. On the other side we have a consensus of scientists offering a scientific framework that has vast empirical support and predictive success. According to this framework, the semantic theories that specify semantic values of our words are not answerable to every use of language; rather, it is *patterns* of usage that matter and only certain ones at that.[8] This is obviously not a clash of intuitions, and it should be clear which side to believe.

The objector continues: one can find linguists saying things that sound like epistemicist views. For example, Christopher Kennedy writes, "As noted above, changes in our dispositions can result in changes in the extension of a vague predicate in ways that are too complicated to calculate" (2011, p. 86).

Reply: Kennedy is no epistemicist—he offers a robust objection to epistemicism just after this quotation, and one of his points is that certain changes in our dispositions can affect extensions, but the epistemicist is wrong to assume that any changes in dispositions will affect extensions. Kennedy's theory, however, does posit that there are sharp cutoffs in the semantic values of our words, which is shared with

[8] Williamson does sometimes talk about the *pattern of usage,* but he means all or most of the uses of a vague term as the quotations above demonstrate; changing even one of the uses makes a new pattern (as evidenced by his examples of other possible worlds where a vague word is used differently). When we talk about patterns of usage, *we* mean the common or central uses of a vague term as it is used by the relevant people (whatever one's view on this) over time, where a pattern of usage can retain its identity through some changes in the individual uses that make up that pattern (as a ship can retain its identity through some changes in planks).

the epistemicist. However, Kennedy does not think they are unknowable although they might have been too complicated to calculate in 2011 when this quote was published.

Second objection: meaning is not determined by usage, it is determined by *dispositions* to use. So, framing the discussion in terms of actual uses is misleading.

Reply: We agree, but it does not make any difference in Williamson's arguments or our objections. Not all dispositions to use are relevant to meaning. Your disposition to call certain windows doors does not matter at all. Your disposition to agree that if someone breaks a window then the window breaks does matter. As does your disposition to agree that one meaning of 'bank' applies to financial institutions. All the same points apply to dispositions.

Third objection: The very precise knowledge of borderline cases that Williamson argues is impossible is not the same as the knowledge that comes from semantic theories. In particular, Davidson's considerations about entailments and action sentences provide us with limited knowledge of logical forms compared to the knowledge of extensions and their borderlines, which is Williamson's focus. As such, our arguments about the patterns of usage on which meanings depend are not relevant for assessing Williamson's project.[9]

Reply: We admit that people mean a wide range of things by the term 'semantics', but what we mean by 'semantics' is the same as those who use the term to describe projects in that attribute semantic values to the linguistic items (especially sentences) in a language fragment in a roughly compositional way. The semantic values are supposed to be formal models of meanings. This is the same kind of project that Davidson was engaged in and about which the Dowty et al. quote above addressed. Remember that Davidson endorsed a Tarski-style semantic theory that attributes determinate extensions to the words in question. More recent theories like Dowty et al.'s theory and Kennedy's theory differ from what Davidson advocates in some ways but they too attribute determinate extensions to words in question, as do and all the others we referenced. Our view is that there is no distinction between the kind of knowledge Williamson says is impossible (i.e., knowledge of the determinate border) and the kind of knowledge that is provided by a specific semantic theory for a specific language fragment that makes correct predictions about specific speakers' linguistic acts. At this point, we have said enough to put the burden of proof on a defender of Williamson-style epistemicism; as far as we know, there is nothing about this point in Williamson's lengthy defense of epistemicism or those who follow him, so any sort of argument to support an objection like the one under consideration would need to be substantial enough to stand alone as a contribution to the literature.

Fourth Objection: Even if we come to know the patterns of usage that determine meaning and we know how meanings are determined by those usage patterns, it *still* will not be enough. The reason is that semantic theories output values de re, but we need de dicto descriptions of the world to satisfy the epistemicist.[10]

[9] Our thanks to an anonymous referee for this objection.

[10] [Acknowledgement redacted].

Reply: Williamson never makes this argument, but he does discuss the distinction in his chapter on vagueness in the world. There he provides a lovely illustration:

> Syntactically, the distinction between constructions de re and de dicto may be drawn for any sentence functor. In constructions de dicto, a term occurs within the scope of the functor, … In constructions de re, the term occurs outside the scope of the functor … Imagine someone who does not know that Constantinople fell in 1453; he knows only that it fell after a great siege sometime in the fifteenth century. Thus he does not know that the year Constantinople fell was before 1460. However, he does know that 1453 was before 1460. Indeed, he knows of 1453 that it was before 1460. Since 1453 is the year Constantinople fell, he knows of the year Constantinople fell that it was before 1460. Thus 'He knows of the year Constantinople fell that it was before 1460' (de re) does not entail 'He knows that the year Constantinople fell was before 1460' (de dicto) (Williamson, 1994, p. 259).

et us see how to formulate the objection exactly: imagine that the output of a semantic theory is something like: expression 'lavender' has semantic value M. One wants to know whether a certain wall is lavender or not lavender (or maybe whether it is lavender or mauve—that extra complexity can be added after considering the more basic case). That is, one wants to be able to know *that* the wall is lavender or know *that* the wall is not lavender. Nevertheless, it is clear that there is no barrier to using a semantic theory that provides de re outputs to infer de dicto knowledge. For example, the expression 'lavender' has semantic value M, and W is a wall that is a member of M; therefore, W is lavender. We are arguing that, contra the epistemicist, the use data relevant to determining sharp boundaries for vague expressions are not expansive. The fact that one might need additional information to infer de dicto results from de re predictions of a semantic theory does not matter for our purposes.

Fifth objection: Perhaps meanings do not supervene on usage patterns alone—they might depend on all sorts of mental processes and maybe other things as well.

Reply: We are not, for the sake of argument, considering this option because Williamson assumes the supervenience of meaning on use. Still, even if it turns out that this assumption is wrong, the rise of surveillance capitalism is capable of overcoming this obstacle. For example, there are algorithms that can identify your core personality traits better than your close friends simply by looking at your Facebook "likes".[11] Moreover, algorithms in the area known as *Bayesian Theory of Mind* can accurately attribute beliefs, desires, intentions, emotions, and other mental states that might be involved in determining the meanings of our words.[12] This has already been one of the most active areas of surveillance capitalism and is the basis for microtargeting advertisements and the business plan for companies like Cambridge Analytica.[13] Overall this objection takes us beyond the scope of the paper, but we have good reason to think that the same results apply even when one goes there.

[11] See Youyou et al. (2015).

[12] See Baker et al. (2011) for the founding paper of Bayesian Theory of Mind.

[13] See Ienca and Vayena (2018).

We could go on and on—there are many relevant objections the epistemicist could press, but the burden of proof at this point should be on the epistemicist to point out where our objections have gone wrong or formulate new arguments for the impossibility of knowledge in borderline cases.

So far, we have offered a number of objections against epistemicism based on surveillance capitalism and natural language semantics. Because epistemicism involves a number of interlocking mistakes, it is helpful to list the major ones.

(i) *Epistemicist mistake 1* Very many of the correct uses of a vague term are relevant to determining a semantic value for that term and hence for a determinate boundary between things to which the term applies and things to which it does not. We have argued that the right view is that only certain patterns of usage are relevant for constructing and testing theories in natural language semantics and so only certain patterns of usage are relevant for determining the semantic values of our linguistic expressions.

(ii) Epistemicist mistake 2: The correct uses that determine semantic values are unknowable. We have argued that enough of the relevant patterns of use for determining semantic values can be catalogued now with the rise of surveillance capitalism that we can make well-supported generalizations about the rest.

(iii) *Epistemicist mistake 3* The relationship between the correct uses that determine meaning and the determinate semantic value is unknowable. We have argued that we already understand how the relevant use patterns determine semantic values by way of semantic theories proposed by linguists. The correct relationship between relevant data and the attribution of semantic values by a semantic theory is one we already understand from natural language semantics. This knowledge is not perfect or impervious to update or improvement, but it contradicts what is entailed by epistemicism.

(iv) *Epistemicist mistake 4* Determinate boundaries for vague terms are unknowable. So far we have objected to Williamson's arguments for this claim, but we have not shown that the claim itself is false. We aspire in this paper to shift the burden of proof to the epistemicist, and we have done that. Many philosophical moves undermine an argument for some view so that supporters of it will feel like they need to offer some new justification or debug the old one. Nevertheless, we also think a good case can be made that this epistemicist claim is indeed false, not simply unjustified. That is a task for the next section.

4 The "Ultimate Dictionary" App

We do have reasons to think that vague terms do have more determinate boundaries than are known in many conversations and that using AI can help us come to know these boundaries better. In this section, we first propose a thought experiment that should undermine this epistemicist point directly.

We want to be careful to warn the reader to avoid confusion—the following thought experiment is not real. We are unaware of anything like this hypothetical algorithm being used in surveillance capitalism or in the literature on machine learning algorithms.[14] Before the thought experiment, we need to reflect on the roles of AI / machine learning in what we have said in order to emphasize the distinction between what is real and the thought experiment.

So far, we have mentioned that machine learning algorithms are used for various purposes by surveillance capitalist firms. These purposes include data collection, data cleaning, and data analysis.[15] *Data collection* often does not use AI, but it can, depending on the kind of data. Scraping text from a website does not require AI, but collecting text from an audio or video track often does use AI. And one project's output might be the input to another project; for example, clicked "Likes" on Facebook might be the data for a machine learning algorithm that predicts personality traits, but those personality traits might be data for another machine learning algorithm that predicts what kind of advertising would be most effective. *Data cleaning* is getting one's data sorted properly so that they can be analysed. Many of these algorithms are not AI, but some are. Finally, *data analysis* often uses AI, and with the aims of surveillance capitalism, the goal is to predict and control people's behaviour with respect to clicking links online and spending money.

Now for the thought experiment. We invite you to consider a hypothetical machine learning algorithm (or collection of algorithms) that can answer questions about the semantic values of words in a natural language based on some kind of data. This technology can be used in the following way: you wonder whether a certain wall is mauve. You open the "Ultimate Dictionary" app and point your phone at the wall. You ask whether the wall is mauve and it gives you an answer. This is, of course, also an answer to whether the wall is in the extension of 'mauve'.

Imagine that the "Ultimate Dictionary" provides you with an answer: No, the wall is lavender. Because so much language is context dependent, the "Ultimate Dictionary" app will have to consider information about your situation. This information about your situation might include information structure in your conversation, lighting, etc.; it could get this information from your phone which it is monitoring all the time thanks to surveillance capitalism.

We are going to consider a hypothetical class of machine learning algorithms that could be used to decrease vagueness in an app like "Ultimate Dictionary". There are probably many subclasses of algorithms that could accomplish this task in unimagined ways, but we consider three:

A. Natural Language Processing (NLP) Algorithms
B. Classification Algorithms
C. Artificial Theorist Algorithms

[14] See Jullien et al. (2022) an example of work in this direction, however.

[15] See Russell and Norvig (2020) for AI applications that can be used in each of these processes and see Zuboff (2019) for examples of AI applications used in surveillance capitalism. See Marcus and Davis (2019) for criticism of exaggerated claims about what current machine learning algorithms are capable of. Note that even Marcus and Davis are optimistic that abilities like the ones we describe in these examples *are* technologically possible, however they argue that it will require a kind of knowledge representation to be integrated with traditional machine learning algorithms.

The first subclass, NLP algorithms, covers the area of machine learning dedicated to algorithms that process written and spoken language. A famous recent example is GPT-3, which has, among other amazing feats, authored a paper about why people should not be afraid of it and replied to a team of philosophers debating its philosophical significance (GPT-3, 2020; Zimmerman, 2020). One important subclass of NLP algorithms are question and answer (Q&A) algorithms, which provide information in the form of answers to queries from users. If everyday people are going to utilize AI applications to decrease vagueness in the ways imagined below, then they will probably need to use some kind of Q&A algorithm. Beyond that, there is little in the NLP literature that would inspire an "Ultimate Dictionary" app, other than the idea that the advances we have already made in generating algorithms that can display mastery of language make it plausible that something like an "Ultimate Dictionary" is technologically possible.[16]

The second subclass—classification algorithms—are ubiquitous. For example, a neural network might be trained to identify photos with cats. When given a photo as input, it provides a prediction about whether that photo has a cat. Algorithms like this are trained using input/output pairs known to be true (e.g., where the theorist knows whether there is a cat in the photo or not), and the algorithm is trained, in part, by using this information. There are thousands of different machine learning classification algorithms, but one thing it is easy to miss is their semantic significance. Consider again the example algorithm that classifies photos by whether they contain an image of a cat. The important point is that this algorithm, by virtue of classifying photos in this way, also classifies whether photos are in the extension of the phrase 'photo with a cat'. If the average human is, say, 90% correct at identifying photos with cats and our algorithm is, say, 95% correct, then it could be used to provide us with new information. From the algorithm's answer we might infer some information about the world—whether a photo has a cat. Or we might infer some information about language—whether a photo is in the extension of 'photo with a cat'. Perhaps one of these is basic, and the other is inferred. We will not take a stand on this issue. But this sort of connection between claims about the world and claims about meanings or extensions is already well documented in the philosophical literature.[17]

The third subclass we have called *artificial theorists*. These are machine learning algorithms that construct scientific theories given certain kinds of data. For example, one genetic algorithm—which generates small changes over and over and selects the best variant based on how it performs—inferred general equations of motion from data about positions and velocities of a collection of objects (Schmidt & Lipson, 2009). Others try to accomplish similar goals with other sorts of algorithms (Wu & Tegmark, 2019).[18] Another example algorithm generates novel and accurate theoretical hypotheses about which physical materials have scientifically

[16] See Pater (2019), Pearl (2019), and Potts (2019) for other examples of how machine learning can aid natural language semantics.

[17] See Thommasson (2018) for an example.

[18] See also de Silva et al. (2020).

significant properties (e.g., that $CsAgGa_2Se_4$ is a thermoelectric) simply by looking at word embeddings in papers published in the materials science literature (Tshitoyan et al., 2019). In each of these cases, a machine learning algorithm generates general theoretical hypotheses that explain limited observed behaviour. Although we know of no examples of algorithms that generate natural language semantic theories from linguistic data, we are also unaware of any reason to think that they could not be a possibility. That is, one could find a machine learning algorithm that functions as an artificial semantic theorist, just as some computer scientists have already found algorithms that function as artificial physicists ("AI Physicist" is actually the name Wu and Tegmark (2019) use to describe their project). We are not going to speculate further about how the "Ultimate Dictionary" app works, but what we have so far ought to make it clear that it is technologically possible.

Consider an objection: just because some AI algorithm presupposes that our words work a certain way does not mean we should think that this assumption is true. In other words, our argument seems to be: surveillance capitalism relies on AI that uses sharp boundaries when gathering and processing information. On that assumption (and assuming humans accept the meaning constraints imposed by AI) the triumph of AI will be the end of vagueness. But there is no reason to think that the assumptions made by AI about what our words mean are true.

Reply: The objection misconstrues our arguments and assumptions. AI plays a role in collecting, formatting, and processing information about us. Some of this information pertains to the semantic features of our words (e.g., sharp but unknown boundaries of vague predicates). It is this information and the technological possibility of using it to make accurate predictions about the determinate boundaries of the words, not the assumptions made by AI about it, that is incompatible with what epistemicism says about vagueness. This conclusion does not depend on any assumptions made by AI. We do rely on conclusions defended by linguists about which aspects of language use in fact determine what our words mean, and we have argued for certain claims about what is technologically possible to achieve using AI given what we already know to be achievable. However, assumptions made by AI algorithms play no role.[19]

Surely there are readers who want to object: look around—there are no apps like the one just described. If we understand semantics so well and have so much new data on usage, then why don't we know anything new? Where are these trumpeted advances?[20]

Reply: This objection mistakes the epistemicist's position. We obviously do not have an "Ultimate Dictionary" app available right now. However, such a thing is not just technologically possible but likely in the near future. It is up to the epistemicist to argue that this is technologically impossible. The epistemicist must defend the

[19] See Marcus and Davis (2019) for discussion of common misunderstandings about how AI interacts with language.

[20] Acknowledgement redacted.

claim that our science and technology simply cannot advance to that point. And that seems utterly implausible.[21]

Consider a real-world example of vagueness resolution by AI. Botanists employed by the Sun Yat-Sen arboretum in Nanking, China collected and labelled the plants therein and made their dataset—now called the Flavia dataset—available to other researchers (Wu et al., 2007). Datasets like this one have been used to train algorithms that can classify plants better than many experts.[22]

4.1 Vagueness Resolution by AI

- At time 1, Plant 1 is unclassified. No one knows what it is, but it is, in fact, an orchid. The relevant *unknown* usage patterns together with *known* empirical facts determine that it is an orchid. It is an epistemic borderline because if all the relevant language usage patterns were known, then Plant 1 could be properly classified as an orchid. Experts know Plant 1's biochemistry and genetics but this is not enough to know that it is an orchid.
- Later, at time 2, an AI algorithm is trained using experts' classifications and previous reliable predictions. The AI algorithm predicts that Plant 1 is an orchid. Now Plant 1 is no longer an epistemic borderline—it is known to be an orchid. The word 'orchid' has become less vague.

Contrast the process of vagueness resolution by AI to normal scientific inquiry.

4.2 Regular Science

- At time 1, Plant 2 is unclassified. No one knows what it is, but it is, in fact, an orchid. The relevant *known* usage patterns together with *unknown* empirical facts determine that it is an orchid. Plant 2 is *not* an epistemic borderline because everyone involved knows all the relevant usage patterns, but no one involved knows the empirical fact about its genetics and biochemistry.
- Later, at time 2, an expert does genetic and biochemical tests on Plant 2 and determines that Plant 2 is an orchid. Plant 2 is now known to be an orchid. However, Plant 2 was never an epistemic borderline. Thus, no change in the vagueness of 'orchid' has occurred through this process of regular science.

Discovering a new fact about some non-linguistic object one is studying (i.e., regular science) does not decrease vagueness, but some technology, when applied in the right way, can decrease vagueness, as in the example above.

We anticipate an objection to this example: although the plant identification app does not treat the meaning of 'orchid' as if it depended on a huge number of

[21] Even critics of contemporary AI like Marcus and Davis (2019) offer no reason to think that this is technologically impossible.

[22] See Bonnet et al. (2018) and Sinha et al. (2021).

linguistic actions, the example agrees with the epistemicist that the extension of ‘orchid’ is modally fragile. For example, if the botanist working at the arboretum in question had made a different decision about the identity of the *Platanthera bifolia* the app would have classified it differently and the extension of ‘orchid’ would have been slightly different.

Reply: If the botanist in question had decided that a certain orchid plant is not an orchid, then that would have been a mistake. The AI algorithm powering the app would have also made a mistake. The plant at the arboretum is still an orchid, even if an expert denies it. In order to have modal fragility the epistemicist desires, one would need it not to be a mistake. That is, the field of botany would have to be different in order to make it the case that the plant in question is really not an orchid. And a world where the entire field of botany is that different is not a nearby possible world. Therefore, there is no modal fragility in this example, and we think that the idea has no place in a proper understanding of vagueness.

5 Technological Epistemicism

So far, we have argued that the information about word usage being recorded by surveillance capitalists is incompatible with epistemicism about vagueness. One interpretation is that epistemicism is right and surveillance capitalism is bringing about the end of vagueness by making it the case that knowledge of sharp semantic borders is now technologically possible. However, we think a better interpretation is that vagueness does not disappear simply because some knowledge is now in principle available. On this alternative interpretation, epistemicism is false and there is still vagueness. On this alternative interpretation, vagueness is not *in-principle* ignorance about determinate boundaries—that is one place the epistemicist went wrong.

There is a version of epistemicism that deals with these issues better than the traditional version. The idea is that if we knew more of the relevant data, and we know how to process it, then we could know the semantic values of our words and sentences with higher confidence and precision. There might potentially be many kinds of relevant data, depending on the kind of algorithm used to inform us about the fine details of the meanings of our words. In the previous section, we proposed three different ways that AI might be used to decrease the vagueness of our words (NLP, classifiers, and artificial theorists).

Maybe we will eventually eliminate all the uncertainty over usage. We can use the term *technological epistemicism* for the empirical claim that we or some kind of intelligent entity like us will eventually eliminate all the borderlines and thereby eliminate all uncertainty over determinate extensions.[23] We do not know whether technological epistemicism is true. However, we do think that insofar as there is a fact of the matter as to what we mean, then a suitable technology plus sufficient scientific understanding of semantic properties will eventually eliminate all the

[23] See Greenough (2003, pp. 252–253) for a suggestion with a similar spirit.

borderlines *that can be eliminated*. Whether that is all the borders that exist remains to be seen.

The technological epistemicist about vagueness says that as our knowledge of the relevant usage data increases and our ability to model semantic values on the basis of those data increases, our knowledge of the semantic values of our words increases in quality as well. Hence, the technological epistemicist can hold on to the idea that vague words have sharp boundaries, and that vagueness is ignorance of those boundaries. However, the idea that this ignorance is “in principle” or permanent must be jettisoned.

For what it is worth, we think that *vagueness pluralism* is the right view: there are different kinds of vagueness, some are *epistemic*, some are *semantic*, and some are *metaphysical*.[24] Perhaps these types of vagueness are associated with particular vocabularies, but it might be that a single word/topic displays all three. For example, it seems plausible to think that the meaning-constituting patterns of use for ‘red’ make it the case that the following claims are true. Thing 1 is known to be determinately red. It is not a borderline case in any sense. Thing 2 is determinately red, but no one knows this yet because we do not yet have available the information about our relevant patterns of usage. Hence, Thing 2 is an *epistemic* borderline case—there is a fact of the matter, we do not yet know it, but we can come to know it. Assume that Thing 3 is not determinately red because our patterns of usage do not determine every case. Nevertheless, there are facts about light, reflection, perception, etc. such that we could use ‘red’ in a way that would determine that Thing 3 is determinately red. Hence, Thing 3 is a *semantic* borderline case—there are relevant facts of the matter but our linguistic practice does not confer a meaning on ‘red’ that is precise enough to determine either that Thing 3 is red or that Thing 3 is not red. Finally, Thing 4 is not determinately red because the relevant bits of the world are indeterminate. There is no fact of the matter, so there is no way for us to have given ‘red’ a meaning that would decide Thing 4. Hence, Thing 4 is a *metaphysical* borderline case. We can imagine a word that displays epistemic vagueness, semantic vagueness, and metaphysical vagueness all at the same time.[25]

Of course, this is just an example to illustrate how vagueness pluralism might work as a background framework for technological epistemicism. Moreover, nothing else in this section or the paper depends on vagueness pluralism. Nevertheless, a vagueness pluralist might want to reject technological epistemicism as an account of *all* vagueness, but accept that technological epistemicism is true of all *epistemic* vagueness. If so, then this theorist would be committed to the view that our words have sharper boundaries than we know now, but not completely sharp boundaries. If that is right, then surveillance capitalism is a portent of the end of *epistemic* vagueness, but not other kinds of vagueness. That is what we think is the most plausible overall account, but we do not argue for it here.

In sum, the epistemicist claims that the technological possibility of knowledge of linguistic usage that is enough to know sharp boundaries for our words is

[24] As far as we know, this view of *vagueness pluralism* has not been defended in print.

[25] See Kölbel (2010) on semantic vagueness and Barnes (2010) on metaphysical vagueness.

incompatible with vagueness. Hence, if this were to become technologically possible, then it would be the end of vagueness, according to the epistemicist. Instead, we have offered technological epistemicism, which implies that there is currently vagueness. However, as our technological capacities grow, vagueness can decrease, and we might reach a point at which our epistemic vagueness is eliminated for those with access to the technology.

6 Explainable Artificial Intelligence (XAI)

We have argued in detail that AI can provide us with knowledge of the determinate, but previously unknown, boundaries to our words. That is, AI can bring about an end to (one kind of) vagueness. But, humans will not come to *know* about these boundaries if we refuse to *believe* what the AI in question tells us—if we do not trust the AI in question. Thus, one important line to pursue from here is: what would we do if we were given access to this information about the meanings of our words? One day, surveillance capitalism might be no more, but humanity could still be able to collect or utilize similar information about relevant aspects of language usage. For example, this information might be provided to everyone for free.

Imagine a speaker calls a wall "mauve". Someone else in the audience says, "actually, it is not mauve—you can see the Ultimate Dictionary app says that it is lavender." What is the next step for the speaker?

One move is to concede. Another move is to reject the appeal to technology entirely. Still another move would be to say something like, "well, it should be mauve, and that's what I'm going to call it." That is an interesting avenue down which one finds things like metalinguistic negotiation[26] and conceptual engineering.[27] But, that is not our focus here. The speaker could instead ask for a *justification* or a *reason* for the claims made about what 'mauve' means.[28] That is where our story involves how we might come to trust AI.[29]

To get a feel for the issues that currently dominate discussion of XAI, imagine that a machine learning algorithm employed by a hospital says that limited resources should be spent on treating a particular patient and not treating another. The untreated patient is expected to die. This patient and family will demand to know why their loved one should go untreated. One answer is: the machine learning algorithm said so. That kind of answer is obviously unpersuasive to most humans.

Another answer one could give to the patient and family is that the machine learning algorithm has been optimized to make medical decisions that are in the overall best interest of the community over time, given their limited resources. If the untreated patient were to be treated, then something even worse than the untreated

[26] See Plunkett (2015) for an overview.

[27] See Cappelen (2018) for a survey.

[28] This sort of move might be motivated by the knowledge norm of assertion—see Williamson (2000) for discussion.

[29] For a recent philosophical discussion of trust, see Hawley (2019).

patient's fate would have to occur instead make up for it. This sort of reason might be convincing to some people, but many will notice that it could be given—just as it is—to *everyone* who demands a reason for a decision from the algorithm. Instead, people will demand to know the reason for *this particular decision*, not for decisions in general.

It is at this point that the hospital might appeal to XAI (Carter et al., 2019; Doshi-Velez & Kim, 2017; Gilpin et al., 2019; Gunning, 2017; Molnar, 2021; Olah et al., 2017, 2018). In the machine learning literature, these are often called *interpretable* machine learning algorithms. This new movement focuses on explaining the behaviour of machine learning algorithms and other artificial systems and designing new ones that provide humans with explanations for the behaviour of the algorithms. The explanations in question need to be understandable in human terms—they cannot be just listing lines of code or activation parameters.

Christoph Molnar's recent monograph includes a detailed discussion of the kinds of projects in XAI and interpretable machine learning. In his terminology, a machine learning *algorithm* is fed *data* and thereby trains a *model*. The model is a program that benefits from what was learned about the data. The model is then used to make predictions (e.g., whether some email is spam, whether an object in a photograph is a face). Molnar offers the following kinds of questions addressed by research on XAI and interpretable machine learning:

- How does the algorithm create the model?
- How does the trained model make predictions?
- How do parts of the model affect predictions?
- Why did the model make a certain prediction for a particular input?
- Why did the model make specific predictions for a group of inputs? (Molnar, 2021, pp. 24–26).

The answers to these questions are what is missing in the above examples involving mauve and the hospital. These are the answers that would stand a chance of satisfying people's questions about the behaviours of AI systems. And these sorts of reasons are the ones that will decide whether artificially intelligent systems earn our *trust*.

If we go back to the case of mauve, then we see the same array of options. What matters is the explanation given for why the machine learning algorithms employed by surveillance capitalism decided that the cut off for 'mauve' was such and such. These sorts of reasons are the ones that will decide whether AI systems earn our trust to serve as standards in human conversations about what we mean.

In the mauve/lavender example, a person uses the Ultimate Dictionary app to determine that a certain wall surface is in the extension of 'lavender' and not in the extension of 'mauve'. If those involved want to understand why the app gave this answer, they might be able to rely on some additional resources. Imagine that the app has an "Explain Answer" option, which activates an XAI algorithm. In what follows, we can suppose that the Ultimate Dictionary works because it is really a massive compilation of classifiers (this would be option B from Sect. 4). If so, then we might use an XAI method that works for classifiers.

For example, SHapely Additive exPlanations (SHAP) is an XAI algorithm that works for classifiers and focuses on the features that are involved in a particular prediction in order to explain that prediction.[30] A *feature* is an input to a machine learning model (in Molnar's terms explained above), where a prediction is the output. In the example of the plant identification model, the model is a probabilistic neural network and the features include the diameter, length, width, area, and perimeter of the leaf, which are used to compute additional features like narrowness (i.e., diameter / length) and rectangularity (i.e., (length x width) / area). These features of each item are inputs into the model, which then uses the feature values (the numbers) together with its own internal calculations to predict whether the item in question has the target property (or which target property it has). In this case, the plant type is the target property. The SHAP algorithm is used to interpret each prediction from a classifier by looking at how much weight each feature had in making that prediction. So, in our example, when the app says that the plant in question is *Platanthera bifolia*, it made that prediction based on estimating the plant's features. These estimated feature values were then fed into the model, which in turn calculated the target property of *Platanthera bifolia*.

The SHAP algorithm helps explain the predictions of another model (e.g., the plant classifier or the Ultimate Dictionary). It explains these predictions by figuring out which features were involved the most in making the prediction in question. In the plant classifier identifying *Platanthera bifolia*, perhaps it was the narrowness of the leaves combined with the colour of the flowers. In this way, SHAP and other XAI algorithms can explain the behaviour of "regular" machine learning algorithms. It is important to remember that SHAP is a completely different algorithm that uses ideas from game theory that allow the features of the underlying model to compete with one another to see which one has the most influence in a given prediction. It takes as input certain properties of the underlying model (e.g., plant classifier) and provides as output the features most involved in a certain prediction made by that underlying model.

Returning to our example conversation with the Ultimate Dictionary app, once the "Explain Answer" button is pressed, another algorithm—an XAI algorithm that we can suppose is SHAP—looks at the "Ultimate Dictionary" model. We have supposed that the Ultimate Dictionary is composed of a bunch of classifiers, so SHAP might takes as input the properties of the "Colour" classifier. These might be derived from an image taken of the wall in question together with information about the conversation in question (to disambiguate and to settle contextual features of colour terms). The output of the "Explain Answer" algorithm is which features of the "Colour" classifier were most involved in the prediction that the wall is lavender and not mauve. In this way, the "Explain Answer" algorithm offers a reason or an explanation for the answer given by the "Ultimate Dictionary" app. And this answer is specific to the unique prediction in question. SHAP explains why the app provided the output it did by citing the most important inputs used in the calculation of the answer. This is just what one would expect from a human expert in the

[30] See Lundberg and Lee (2016).

same situation, and that is why XAI is so important for humans to be able to have knowledge of the determinate boundaries of vague words. Without XAI to explain the answers about the boundaries of our words, we might not readily believe what the algorithms tell us about those boundaries. And without believing, there is no knowing, at least according to many popular accounts of knowledge.

In conclusion, the importance of AI and surveillance capitalism in the study of vagueness should be apparent now. Humans using AI will be able to collect enough of the right kinds of data and process them so that we will be able to find more determinate semantic values for many words. If ordinary people can access these precise semantic values, what will happen? Whether they matter for ordinary people, whether they are taken as authoritative, whether people trust them: these depend on how well people can understand the decisions made by the artificial systems involved in collecting and processing the data. If the XAI movement succeeds, then we should expect people to trust AI systems to make important decisions, like what we mean and whether we live or die. In other words, people might come to trust *artificial* decisions and opinions in their lives, but only if they can understand the reasons for them in *human* terms.

References

Alpaydin, E. (2020). *Introduction to machine learning* (4th ed.). MIT Press.

Altschuler, D., Parsons, T., & Schwarzschild, R. (2019). *A Course in semantics*. MIT.

Bacon, A. (2018). *Vagueness and thought*. Oxford University Press.

Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33). https://escholarship.org/uc/item/5rk7z59q

Ball, D., & Rabern, B. (Eds.). (2018). *The science of meaning*. Oxford University Press.

Barnes, E. (2010). Ontic vagueness: A Guide for the perplexed. *Nous, 44*, 607–627.

Bonnet, P., Goëau, H., Hang, S. T., Lasseck, M., Šulc, M., Malécot, V., Jauzein, P., Melet, V., You, C., & Joly, A. (2018). Plant identification: Experts vs. machines in the era of deep learning. In A. Joly, S. Vrochidis, K. Karatzas, A. Karppinen, & P. Bonnet (Eds.), *Multimedia tools and applications for environmental & biodiversity informatics* (pp. 131–150). Cham: Springer. https://doi.org/10.1007/978-3-319-76445-0_8

Caie, M. (2012). Vagueness and semantic indiscriminability. *Philosophical Studies, 161*, 365–377.

Cappelen, H. (2018). *Fixing language: An Essay on conceptual engineering*. Oxford University Press.

Carter, S., Armstrong, Z., Schubert, L., Johnson, I., & Olah, C. (2019). Activation Atlas. *Distill*. https://distill.pub/2019/activation-atlas.

Castroviejo, E., McNally, L., & Sassoon, G. W. (Eds.). (2018). *The Semantics of gradability, vagueness, and scale structure*. Springer.

Chierchia, G., & McConnell-Ginet, S. (2000). *Meaning and grammar: Introduction to semantics* (2nd ed.). MIT.

Davidson, D. (1967). The logical form of action sentences. In N. Rescher (Ed.), *The logic of decision and action.* University of Pittsburgh Press.

de Silva, B., Higdon, D., Brunton, S., & Kutz, J. N. (2020). Discovery of physics from data: Universal laws and discrepancies. *Frontiers in Artificial Intelligence*. https://doi.org/10.3389/frai.2020.00025

Dietz, R., & Murzi, S. (Eds.). (2010). *Cuts and clouds: Vagueness, its nature and logic*. Oxford University Press.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. Preprint retrieved from arXiv:1702.08608v2.

Dowty, D., Wall, R., & Peters, S. (1980). *Introduction to Montague semantics*. Kluwer.

Égré, P. (2015). Explanation in linguistics. *Philosophy Compass, 10*, 451–462. https://doi.org/10.1111/phc3.12225

Fine, K. (2020). *Vagueness: A global approach*. Oxford University Press.

Gillon, B. (2019). *Natural language semantics: Formation and valuation*. MIT.

Gilpin, L., Bau, D., Yuan, B., Bajwa, A., Specter, M., & Kagal, L. (2019). Explaining explanations: An overview of interpretability of machine learning. Preprint retrieved from arXiv:1806.00069. https://doi.org/10.48550/arXiv.1806.00069

Gomez-Torrente, M. (1997). Two problems for an epistemicist view of vagueness. *Philosophical Issues, 8*, 237–245. https://doi.org/10.2307/1523008

GPT-3. (2020, September 8). A robot wrote this entire article. Are you scared yet, human? *The Guardian*. https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

Greenough, P. (2003). Vagueness: a minimal theory. *Mind, 112*(446), 235–281.

Gunning, D. (2017). Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA), Program Update. US Department of Defense, November 2017. https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

Hawley, K. (2019). *How to be trustworthy*. Oxford University Press.

Hawthorne, J. (2006). Epistemicism and semantic plasticity. *Metaphysical Essays* (pp. 185–210). Oxford University Press.

Ienca, M., & Vayena, E. (2018, March 30). Cambridge Analytica and online manipulation. *Scientific American*. https://blogs.scientificamerican.com/observations/cambridge-analytica-and-online-manipulation/

Jullien, M., Valentino, M., & Freitas, A. (2022). Do transformers encode a foundational ontology? Probing abstract classes in natural language. Preprint retrieved from arXiv:2201.10262. https://doi.org/10.48550/arXiv.2201.10262

Kennedy, C. (2011). Vagueness and comparison. In P. Égré & N. Klinedinst (Eds.), *Vagueness and language use* (pp. 73–97). Palgrave Macmillian.

Ko, B. C. (2018). A Brief review of facial emotion recognition based on visual information. *Sensors, 18*, 401. https://doi.org/10.3390/s18020401

Kölbel, M. (2010). Vagueness as semantic. In R. Dietz & S. Murzi (Eds.), *Cuts and clouds: Vagueness, its nature, and its logic.* Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199570386.003.0018

Kosinski, M., & Wang, Y. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology, 114*, 246–257.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *PNAS, 110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. arXiv:1705.07874 [cs.AI]

MacFarlane, J. (2020a). Lecture 1: Vagueness and communication. *Journal of Philosophy, 117*(11/12), 593–616. https://doi.org/10.5840/jphil2020117111240

MacFarlane, J. (2020b). Lecture II: Seeing through the clouds. *Journal of Philosophy, 117*(11/12), 617–642. https://doi.org/10.5840/jphil2020117111241

MacFarlane, J. (2020c). Lecture III: Indeterminacy as indecision. *Journal of Philosophy, 117*(11/12), 643–667. https://doi.org/10.5840/jphil2020117111242

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Pantheon Books.

Molnar, C. (2021). *Interpretable machine learning: A Guide for making black box models explainable*. Lean Publishing.

Mosdell, M. (2015). When to think like an epistemicist. *Canadian Journal of Philosophy, 45*(4), 538–559. https://doi.org/10.1080/08912963.2015.1112114

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature Visualization. *Distill*. https://distill.pub/2017/feature-visualization

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The Building blocks of interpretability. *Distill*. https://distill.pub/2018/building-blocks

Oms, S., & Zardini, I. (Eds.). (2019). *The Sorites paradox*. Oxford University Press.

Pater, J. (2019). Generative linguistics and neural networks at 60: foundation, friction, and fusion. *Language, 95*(1), 41–74.

Pearl, L. S. (2019). Fusion is great, and interpretable fusion could be exciting for theory generation. *Language, 95*(1), 109–114.

Plunkett, D. (2015). Which concepts should we use?: Metalinguistic negotiations and the methodology of philosophy. *Inquiry, 58*(7), 1–47. https://doi.org/10.1080/0020174X.2015.1080184

Potts, C. (2019). A case for deep learning in semantics. *Language*. https://doi.org/10.1353/lan.2019.0019

Raffman, D. (2014). *Unruly words: A Study of vague language*. Oxford University Press.

Ripley, D. (2021). Précis of uncut. *Análisis Filosófico, 41*(2), 235–260. https://doi.org/10.36446/af.2021.462

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A Modern approach* (4th ed.). Pearson.

Salles, S. (2021). *Vagueness as arbitrariness: Outline of a theory of vagueness*. Springer.

Sassoon, G. (2013). *Vagueness, gradability and typicality: The Interpretation of adjectives and nouns*. Brill Publishing.

Schmidt, M., & Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science, 324*(5923), 81–85. https://doi.org/10.1126/science.116589

Sennet, A. (2012). Semantic plasticity and epistemicism. *Philosophical Studies, 161*(2), 273–285.

Sinha, S.K., Kumar, S., Kumar, S., Katiyar, G., & Chandola, R. (2021). Plant identification using machine learning. In *2021 Asian conference on innovation in technology (ASIANCON)* (pp. 1–4). https://doi.org/10.1109/ASIANCON51346.2021.9544670

Sorensen, R. (2001). *Vagueness and contradiction*. Oxford University Press.

Sorensen, R. (2018). Vagueness. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy* (Summer 2018 ed.). https://plato.stanford.edu/archives/sum2018/entries/vagueness/

Tonhauser, J., & Matthewson, L. (2015). Empirical evidence in research on meaning. http://ling.auf.net/lingbuzz/002595. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.737.4344&rep=rep1&type=pdf

Thommasson, A. (2018). *Ontology made easy*. Oxford University Press.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., & Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature, 571*, 95–98. https://doi.org/10.1038/s41586-019-1335-8

Williamson, T. (1994). *Vagueness*. Routledge.

Williamson, T. (2000). *Knowledge and its limits*. Oxford University Press.

Wills, J. (2017). *Tug of war: Surveillance capitalism, military contracting, and the rise of the security state*. McGill-Queen's University Press.

Wright, C. (2021). *The Riddle of vagueness: Essays 1975–2020*. Oxford University Press.

Wu, S. G., Bao, F. S., Xu, E. Y., Wang, Y., Chang, Y., & Xiang, Q. (2007). A Leaf recognition algorithm for plant classification using probabilistic neural network. In *IEEE 7th International symposium on signal processing and information technology* (pp. 11–16). https://doi.org/10.1109/ISSPIT.2007.4458016

Wu, T., & Tegmark, M. (2019). Toward an artificial intelligence physicist for unsupervised learning. *Physical Review E, 100*, 033311.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *PNAS, 112*(4), 1036–1040. https://doi.org/10.1073/pnas.1418680112

Zimmerman, A. (2020, July 30). Philosophers On GPT-3 (updated with replies by GPT-3). *Daily Nous*. https://dailynous.com/2020/07/30/philosophers-gpt-3/

Zuboff, S. (2019). *The age of surveillance capitalism: The Fight for a human future at the new frontier of power*. Profile Books.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Alison Duncan Kerr[1] · Kevin Scharp[2]

Alison Duncan Kerr
adk10@st-andrews.ac.uk

[1] Arché Philosophical Research Centre, St Andrews Institute for Gender Studies, University of St Andrews, St Andrews, Fife, Scotland, UK

[2] Arché Philosophical Research Centre, Centre for Exoplanet Science, University of St Andrews, St Andrews, Fife, Scotland, UK