[PENULTIMATE VERSION. FINAL VERSION: https://doi.org/10.1093/pq/pqad088]

Making AI Intelligible: Philosophical Foundations. By HERMAN CAPPELEN and JOSH DEVER. (Oxford University Press, 2021. pp. viii + 175. Price £34.99.)

Linguistic outputs generated by modern machine-learning neural net AI systems seem to have the same contents — i.e., meaning, semantic value, etc. — as the corresponding human generated utterances and text. Building upon this essential premise, Herman Cappelen and Josh Dever's *Making AI Intelligible* sets for itself the task of addressing the question of *how* AI generated outputs have the contents that they seem to have (henceforth, 'the question of AI Content'). In pursuing this ambitious task, the book makes several high-level, framework observations about how a meta-semantic account of the content of AI generated outputs should proceed.

Two such observations form the philosophical core of the book and are particularly noteworthy. The first concerns *where* to look for content. The book focuses on machine-learning neural net AI systems (e.g., Open AI's ChatGPT or Google's LaMDA), which are interestingly different from symbolic AI systems (e.g., calculators). Given knowledge of their programming, one can find out how the computations performed by symbolic AI systems on an input (that represents, say, a number) should result in the output represent, say, the square root of the input.

However, it is not obvious how a corresponding content level story may be constructed by examining the computational architecture of machine-learning neural net AI systems (henceforth, simply 'AI systems'). The computations that an AI system performs on an input are determined by the system itself during its training phase; and even to the programmers of AI

systems, it remains a mystery why performing one set of computations on the input rather than some other yields the desired output (pp. 42-42, 60-61).

Grounding AI Content in the computational architecture of such systems is therefore not a simple, straightforward project. Cappelen and Dever (C&D) contend that besides being difficult, this is also a project headed in the wrong direction. Humans are the 'original neural network systems' (p. 22). So, the question of how the content of *human* speech is determined — a question with a rich history in the philosophical debate between internalism and externalism — should naturally and reasonably be expected to bear upon the question of AI Content.

The question of AI Content, C&D argue, is 'not a problem at the level of computational detail' (p. 62). Their argument for this claim draws on the argument for the dominant externalist tradition, according to which the content of human speech is not entirely determined by 'internal' factors (e.g., the biological makeup of the human brain) but in part also by social and environmental factors. Analogously, the question of AI Content must be answered in part by appealing to factors that go beyond the computational architecture of the system e.g., the details of the system's training process, the purpose for which the system is employed, etc.

The second observation concerns what C&D call the 'anthropocentric' nature of existing meta-semantic theories. Although existing theories can be helpful in addressing the question of AI Content, their focus has almost exclusively been the sounds, symbols, or gestures produced by *humans* (to the exclusion of animals and other entities that may possibly use a language). 'Content' has implicitly been assumed to stand for 'content of human generated speech', leading theorists to appeal to features and capabilities restricted to humans e.g., the appeal to intentions in causal-historical accounts, human evolutionary history in teleosemantics etc.

If existing meta-semantic theories are to be relevant for addressing the question of AI Content, C&D argue, then they must be '*de-anthropocentrized'* (p. 69). This involves abstracting away from human-specific elements — e.g., intentions or human biological evolution — while retaining the useful aspects of the theory.

The decision concerning which aspects count as useful (and with what to replace the human-specific elements) depends on the purposes that the theory is expected to serve. To make such decisions, C&D adopt a version of Williamson's principle of knowledge maximization: the purpose of a meta-semantic account is to maximize our knowledge *qua* interpreters of AI generated outputs (p. 76). A large part of the book (chapters. 5-7) is devoted to providing hands-on examples of how existing externalist accounts — e.g., causal-historical chain account for predicates, mental files framework for names — can be relevant in addressing the question of AI Content, and how one may attempt to 'de-anthropocentrize' existing meta-semantic accounts.

The book makes a strong case for expanding the scope of semantic investigation to non-human language using entities, and also provides a framework in which such investigations can proceed. This is the strongest philosophical merit of the book and where it breaks new ground. Another strong feature of the book is the many connections it makes between the question of AI Content and some broader human concerns e.g., the fact that important decisions concerning our health and finances increasingly rely on AI generated outputs (Chapter 1), and the challenge posed by AI generated outputs to our need for a reasoned engagement with the world, which is at the centerstage of the 'Explainable AI' movement (pp. 25-26, pp. 162-166).

The limited argumentative scope of the book, however, also represents a missed opportunity. Readers looking for an extended discussion or refutation of some interesting and plausible alternatives to the central assumption of the book may be disappointed. One such alternative

discussed elsewhere in the literature, including by the same set of authors, is that AI outputs have *alien* contents: the features of reality that AI systems track and represent via their outputs are incomprehensible to humans. Another alternative is that AI outputs have *no* contents: they are just statistically probable arrangements of words that do not represent nor are about anything.

The skeptical possibility that the outputs of AI systems may not have contents (of the sort that we take human speech to have) remains a live possibility until the end of the book, and one may worry that this takes away some motivation behind the book's main project. Particularly because providing a meta-semantic account of AI outputs is a massive, resource intensive project (which is sensitive to the technological features of specific AI systems and the purposes for which they were deployed) and it is not entirely clear that a meta-semantic project of the sort that the book undertakes is indispensable for addressing some of the broader questions of human interest that are used to motivate the project (e.g., a meta-semantic account does not *ipso facto* justify, but only makes available an explanandum in the form of a contentful AI output)

Making AI Intelligible begins an original, creative, and ambitious project which contributes both to the scientist's search for alternative methods to make sense of a phenomenon, as well as the philosopher's search for her own blind spots.

NIKHIL MAHANT
Central European University, Vienna