Is AI case that is explainable, intelligible or hopeless?

Łukasz Mścisławski
Wrocław University of Science and Technology, Poland

H. Cappelen, J. Dever, *Making AI Intelligible: Philosophical Foundations*,
Oxford University Press, Oxford 2021, 175+viii pp.

The bored reader may sigh: another book in which philosophers create an artificial problem without completely understanding artificial intelligence (AI). It is entirely legitimate here to ask whether philosophers' throwing themselves (and—to be honest—not only them) at various AI problems is fruitful and the subject matter is really important? Does AI itself, in order to be somehow tamed (to function in a more or less communicative way with us)—need such a far-reaching intellectual effort, by not only technical in its nature? Some justification can be found in the surprising effectiveness of AI systems in relation to the tasks set before them, which arouses understandable interest and is sometimes heavily exploited in the media. Nevertheless, some revolutions of extreme importance, involving digital systems, as Paweł Polak (2015, p.151) rightly pointed out, proceed without special publicity. Perhaps, at least to a large extent, this would also be the fate of AI systems if it were not for the fact that decisions can be made based on them about important issues in the lives of ordinary people.

Herman Cappelen and Josh Dever, in their book *Making AI Intelligible: Philosophical Foundations*, provide a positive answer to both questions. The scope of the subject matter addressed in the book is

Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)
No 73 (2022), pp. 357–369 • CC-BY-NC-ND 4.0

quite narrow and is mainly concerned with the issue of the possibility of linking the content on which humans can operate with the way AI systems function and deliver results. The authors raise a number of important issues that become more pressing as AI systems penetrate more and more new areas of human functioning. It also turns out that attempting to theoretically justify the answers to the questions that arise is far from easy, despite the existence of an extremely rich set of different philosophical traditions, equipped with powerful tools developed to solve various problems.

Chapter 1 (*Introduction*) presents the primary task of the book: an attempt to answer the question of whether philosophical theories of meaning, content and language can be helpful in understanding, explaining and—perhaps—improving AI systems?

According to the authors, the answer to the question posed in this way is positive. They begin their argument by presenting a situation in which a decision concerning a fictitious person is made by an AI system. The question is about the possibility of granting credit. The answer is negative. This raises a simple question: why? The authors point out that knowing how AI systems work does not directly translate into understanding the results provided by such systems. Much worse, however, is the attempt to reconstruct what is the rationale for such and not such a result (pp.4–10).[1] For all the effectiveness of such tools, the fundamental difficulty, on the unraveling of which the authors devote practically the entire book, lies in the fact that it is not very clear whether and how the information processing processes taking place in AI systems are related to the content on which humans can operate. Attempting to answer this question can be seen as a waste of time and the answer itself as adding little—at least from the point

[1] All page numbers without Author and year refer by default to (Cappelen and Dever, 2021).

of view of those designing and implementing such systems. This kind of working conclusion seems to conclude the exemplary dialogue with a hypothetical AI specialist, to which the whole of Chapter 2 is devoted (*Alfred (The Dismissive Sceptic). Philosophers, Go Away*!). Nevertheless, already in Chapter 1, the authors make important points: the need for a good content theory for AI systems and the pressing issue of relating the output of such systems, expressed through a group of sentences in some natural language, to the content determined in that language by that very group of sentences (pp.20–21). However, the answer to the question of why to explore such a seemingly insignificant issue turns out to be very important, given the increasingly widespread use of AI-based decision-making systems. Although no such statement is made explicitly, an attempt to reconstruct such an answer from an already preliminarily sketched, fictional dialogue between a philosopher and a specialist about AI could be as follows: the link between the results provided by AI, their justification and the content on which humans operate is important, as AI systems are increasingly being incorporated into decisions concerning existential human affairs. These include the possibility (or not) of e.g. taking out a loan, health matters, but also making a diagnosis or the adjudication of being a criminal suspect (pp.36–38). Hence, the suggestion that reliance on these systems simply because it is well-written software, based on sophisticated mathematical apparatus, seems insufficient at best (p.37). It should be emphasised that the authors are not concerned with some kind of embedded content in AI systems. Rather, they are concerned with understanding what the content is in a given complex system and how that content was obtained by that system (pp.22–23). This is particularly true for AI systems, the results they provide and their interpretation (p.18). Here is right place to highlight is one of the minor shortcomings of the work. The aforementioned thesis, that the link between the results provided by AI, their justification and the content on which humans operate is important, as AI systems are increasingly being incorporated into decisions concerning existential human affairs and the suggestion that reliance on these systems only because it is well-written software, based on sophisticated mathematical apparatus, is not formulated explicitly. Moreover, with regard to AI-based systems, the issue of trusting the software is one thing, but there are also other problems: the problem of AI bias (which luckily is addressed by Author, however rather in technical context and with reference to content issues), the issue of quality and ethics and the value system used in AI training (e.g. Spence, 2021) and the fundamental question of the correctness of the mathematical model (and its adequacy to the simulated area of reality). Although the last issue is not the direct focus of the authors' research, it seems that some mention of such difficulties would be most welcome.

Chapter 3 (*Terminology*) is devoted to introducing basic concepts, which is important for the clarity of the overall discussion and introduces the reader to aboutness, representation, and attempts to outline the connections between these concepts and AI, metasemantics and philosophy of mind. Building on a previous dialogue with a sceptic, the authors note that, in essence, software or devices in themselves say nothing. The analogy is with AI systems. Moreover, an attempt to build an understanding of the results provided by AI systems, based on knowledge of their internal structure and operating principles, does not necessarily shed much light here. This, in turn, leads to the conclusion that there is a need for a stronger interplay between the metaphysics of content and theories of AI, and a suggestion to look more closely at the possibility of using the tools provided by the externalist branch in the philosophy of language. The authors see the lack of a wider discussion of this possibility in the literature as a gap that needs to be filled (pp.53–58; cf. Krzanowski and Polak, 2022).

In Chapter 4 (*Our Theory. De-Anthropocentrized Externalism*), the authors make a presentation of their own conception, which they describe as de-anthopocentrised externalism. They base their proposal on two basic claims: 1) the content of AI systems should be explained externalistically; and 2) existing externalist approaches are anthropocentric. The first thesis is based on the observation that content, related to action, is not a problem at the level of software or computation. It is a problem at the environmental and sociological level. The second thesis is the observation that however philosophers have developed impressive models of human language and human mental states. However, this is not the case with AI systems—the operation of software on specific hardware, both in the computational layer and in terms of the functioning of the hardware, is fundamentally different from what can be described by such means. A de-anthropocentised metasemantics is therefore needed here (pp.59–71). However, some additional rule of thumb is also needed for the future selection of appropriate measures and the development of an effective content theory of AI systems. Here, the authors propose a meta-metasemantic principle: interpreter-centric knowledge-maximization. Two important issues also arise here, which can also serve as a kind of guideline in the search for appropriate tools for further research: a) it is the human knowledge and not the AI system that is important, so the idea is to maximise human knowledge; and b) the perspective of interests is important here, bearing in mind that human interests may differ from those of the an AI system[2] (pp.75–79).

[2] In simple terms: a human may want to know why he or she was classified negatively in given aspect, while an AI system may seek to optimise the data in some way (e.g. finding the minimum of a function).

The background thus outlined serves the authors to attempt to apply already existing philosophical tools when it comes to relating content to the results provided by AI systems. The test task is to classify a particular person into a particular category. The basic problematic therefore involves AK1) referring to that particular person; AK2) attributing to that person a test characteristic (adjudicating that person as having that characteristic); AK3) criteria for classifying that person into a particular category (adjudicating attribution to a category). Chapter 5 (*Application. The Predicate 'High Risk*') attempts to unravel these issues using an externalist approach based on proposals of Kripke. In doing so, they draw attention to the fundamental difficulties of such approaches: the problem of the anchoring event, the problem of defining chain of transmission and issuses connected with the problem of being part of communicative chain, when AI systems are involved (p.82-88). Additional difficulties are posed in relation to AK3 by the possible variability of classifications and models and the fact of context dependence. Unsurprisingly, it is very important to note that in the case of systems based on machine learning, the final correctness of the answers given depends on those that the human training such systems deems to be true, i.e. on human decisions. Another problem is that AI systems do not seem to have capacity to representing that could be analogous to human's ability of representing using proper names. Such a situation generates serious problems of communicative and epistemic nature (pp.99–105). Cappalen and Dever make an analogous analysis when it comes to the potential application of the Mental Files Framework[3] and attempt to extend the

[3] Murez and Recanti shall be characterized as *devices of direct reference whose deployment makes it possible to entertain singular thoughts, i.e. thoughts that are about particular objects rather than about whatever possesses certain features or satisfies such and such a description* (cf. Murez and Recanati, 2016, p.267).

findings of Evans and Recanati to Epistemically Rewarding Relations. Chapter 6 (*Application. Names and the Mental Files Framework*) is devoted to this. The addition of the knowledge-maximization rule to the framework in question raises the question of whether it is about maximising general or specific knowledge (p.114). Ultimately, however, it appears that with the philosophical tool in question the case is similar to that of the Kripke-style framework, a simple application to AI systems may not be feasible. In particular, there is need to abstract the notion of an epistemically rewarding relationship. The main difficulties in the context of the philosophical framework in question, however, are the need to focus on particular epistemic goals and activities and the fact that what the results provided by an AI system are about depends on the aims of the interpreter. Hence the results of the considerations in Chapters 4 and 5 are reinforced and shows the organic nature of the internalist view of AI: you cannot bite off all the facts about the classification content of a machine learning system by looking only at the internal implementation of that system (pp.115–116).

It should be emphasized that the analyses carried out of the positions presented in the chapters under discussion are very detailed and thorough. The bigger surprise is Chapter 7 (*Application. Predication and Commitment*), which turns out to be a fundamental twist. The authors conclude that the attempts presented earlier to link human-understandable content to the way AI systems function are far from sufficient. They therefore pose the thesis that this kind of linkage is not only a denotation of something, but also an act of adjudication. They thus introduce the reader to the foundations of the Act Theoretic View, which seems a legitimate step insofar as, following Soames and Hanks, they assume that propositions do not have intrinsic representational properties. This in turn—at least to some extent—gets rid of the problems of semantic externalism. After a brief introduction, the whole chapter is essentially devoted to an attempt to investigate whether it is possible to use such a tool to solve the problem of interpreting the performance of AI systems and the results they provide. (pp.117–121). The metasemantic tool that Cappalen and Dever choose to use in relation to predication is the Teleofunctionalist Hypothesis (TFH). They formulate TFH as the statement that a mental act is the act of predication because of its teleofunctional role in giving rise to judgements that guide action. Using proposed tools gives them also possibility to not committ to any particular architecture of analyzed system (pp.123–125).

In doing so, they point out that the TFH approach also presents peculiar difficulties. An example of this is the changing objectives that ultimately help to provide an answer with a given content.[4] However, the aforementioned independence from specific AI system architectures should be considered a very strong advantage. They also propose to consider the relationship between TFH and commitment (or assertion) and try to infer some some norms that could be results of theory based on such approaches. At the end of the chapter, they propose an outline for such research project that could attempt to explore theories of assertion and commitment for humans and AI. Although one of the authors (Cappalen) is sceptical about the category of assertion itself, for the inquisitive reader this outline will undoubtedly provide some inspiration for their own research.

The last chapter (*Four Concluding Thoughts*) contains a kind of explication of the threads that, however scattered, appeared in the previous chapters. The first point is that AI systems are dynamic realities in the senescence that they have a kind of dynamic purpose. Such a situation requires a little more knowledge of technical details

[4] E.g. positive or negative classification for a mortgage.

on the part of philosophers, which, for example, a is essential when dealing with issues related to the scoring mechanism and the problem of the number of layers used for characterization of content (p.141, pp.140–148).

The second point is for the authors to consider the applicability of the philosophical description associated with 'active externalism', as proposed by Clark and Chalmers, and the concept of extended mind presented by them (pp.148–157). The problem is important because, as the authors point out, the effort to understand extraneous content, and the content contained in AI systems can be considered as such, turns into the issue of understanding the determination of content within our extended mind (p.156). A point to be made here, however, is that it seems to be one thing to determine content and another to understand what these extensions of the mind operate on and what is the relation of the extended mind as a whole to the content on which humans operate.

The third point is an attempt to completely change the position, which here Cappalen and Dever refer to as a content-driven approach. The authors sketch an attempt to justify an application to the issues considered in the book from the point of view of the No-Content-Just-Evidence approach. In doing so, however, they draw attention to the problems of Adversial Perturbations, ML system bias and the important fact that coincidental convergence is not justified enough to treat AI systems as reliable for new cases (pp.157–162). This raises the question of the justification of trust in AI systems. This brings the authors, at least in a sense, to the fourth point and some kind of connection with *Explainable AI* stream. It is appropriate to cite their objections to this stream: a) without content there are no reasons nor explanations, and AI systems 'says' something that is contentful. Also reasons are contentful themselves; b) very explainability is also a process of determining specific content. Hence there is great need to say something about content and its connection with explanation in context of AI systems (pp.162–165).

What can be seen as very strong point of book under review is the observation that talking about issues referring to functioning of AI systems there are many anthropomorphism used. Meanwhile, in the case of AI, the matter is quite complicated. As Authors put it:

> In philosophy, consideration of alien languages either starts with the assumptions that the aliens share with us a basic cognitive architecture of beliefs, desires, reasons, and actions, or (as Davidson does) concludes that if the aliens aren't that much like us, then whatever they do simply can't count as a language. Our point is that the aliens are already among us, and they're much more alien than our idle contemplation of aliens would have led us to suspect. Not only that, but they are weirdly alien—we have built our own aliens, so they are simultaneously alien and familiar. (p.17)

This shows how difficult it is to connect more or less obvious for a man content of sentences used by his language with, as it seems, complete alien world of AI systems, of which technical structure and algorithms, paradoxically, we know almost all.

It should be emphasized that presentations of ideas and argumentations that lead to using externalistic metasemantics in interaction with AI, are very clear and is one of the strongest points of the book. The book has, however, also some shortcommings. They do not decrease the value of the work of Cappelen and Dever, nevertheless they are confusing and hinder reading of this fascinating book.

The issue of the relationship between human understanding of and operation on content and how AI systems function is, on the one hand, an extremely important task, but also a very difficult one. It could be said that one of the weaknesses of the book is the lack of attempt to outline what the authors actually mean by the content. This could have helped the reader grasp more of what the difficulty of the whole endeavour is, above all in comparison with operating on, for example, colours within graphic systems.

Perhaps also devoting some space to other seemingly obvious issues would have helped not only the readers but also the authors in their endeavour. It would seem that already with regard to the aforementioned notion of content, it would at least be appropriate to outline overtly some ontological background within which the authors conduct their analyses. There would then be a chance for various hidden assumptions to see the light of day and this would give an opportunity to assess their impact on the overall argumentation carried out. One such assumption, by no means obvious, is to treat AI systems as designed and made intentionally, as opposed to human (p.70). However, that humans are not created intentionally seems to be a very strong ontological assumption.

The starting point for the strategy proposed by the authors, as could be seen in the brief presentation of the individual chapters, is an attempt to use attribution mechanisms that appear to be human-specific (which is the case in Chapter 5). However, the authors are aware that this kind of tactic cannot be applied across the board, as witnessed in Chapter 4, and the de-anthropocentric perspective proposed therein. However, when, as they rightly point out, this path does not yield satisfactory results, they change the tools with which they try to get their way (Chapter 6 and later Chapter 7). While this is understandable, it seems that the link between abandoning the previous path of trying to deal with the problem under investigation and the choice of subsequent tools is not sufficiently justified. A side effect of such a situation may be a feeling that the reading of the book is piecewise smooth. The authors also do not make any remarks, whether the failure of the given tool in question is a permanent or whether it is only temporary situation and we need more research in given area or wait for more advanced technologies.[5]

The book may leave you feeling unsatisfied a little bit. At the moment when the narrative gains momentum, the book ends with few important and accurate remarks on explainable AI, but also is leaving the reader with only an outline of further possibilities for continuing the plot. However, this is quite understandable due to the fact that the issues raised by the authors are extremely extensive. Attempting to cover all possible approaches to the issues raised would fundamentally break the frame of any book of reasonable length. Nevertheless, in their work Cappelen and Dever is very inspiring, it poses many very important questions, tries to find solutions and provokes independent study.

Abstract

This article is a review of the book *Making AI Intelligible. Philosophical Foundations*, written by Herman Cappelen and Josh Dever, and published in 2021 by Oxford University Press. The authors of the reviewed book address the difficult issue of interpreting the results provided by AI systems and the links between human-specific content handling and the internal mechanisms of these systems. Considering the potential usefulness of various frameworks developed in philosophy to solve the problem, they conduct a thorough analysis of a wide spectrum of them, from the use of Saul Kripke's work to a critical analysis of the explainable AI current.

[5] The answer to that question seems to be particularly in case of explainable AI.

Keywords

AI, externalism, metasemantics, content.

Bibliography

Cappelen, H. and Dever, J., 2021. *Making AI Intelligible: Philosophical Foundations* [Online]. 1st ed. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780192894724.001.0001.

Krzanowski, R. and Polak, P., 2022. The Meta-Ontology of AI systems with Human-Level Intelligence. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, (73), pp.197–230.

Murez, M. and Recanati, F., 2016. Mental Files: an Introduction. *Review of Philosophy and Psychology* [Online], 7(2), pp.265–281. https://doi.org/10.1007/s13164-016-0314-3.

Polak, P., 2015. Bezgłośna komputerowa rewolucja w naukach eksperymentalnych [recenzja]. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, (58), pp.151–157.

Spence, E., 2021. *Stoic Philosophy and the Control Problem of AI Technology: Caught in the Web*, *Values and identities*. Lanham: Rowman & Littlefield Publishers.