Kantian Review (2024), **00**, 1–4
doi:10.1017/S136941542400027X

BOOK REVIEW

Hyeongjoo Kim and Dieter Schönecker (eds), *Kant and Artificial Intelligence*. Berlin/Boston: De Gruyter, 2022. pp. vii + 290. ISBN 9783111355696 (pbk) $21.99

If Kant's analysis in the *Critique of Pure Reason* is valid, it should be possible to use his metaphysics as the basis for the architecture of a form of artificial intelligence (AI) which is in at least some of its essential aspects a person just like us soft-tissue humans. Such is the claim of Richard Evans in his contribution to the edited collection, *Kant and Artificial Intelligence*. This volume also includes perceptive contributions on issues such as AI and self-consciousness, robot ethics, AI autonomy, and AI and beauty, but it is Evans' chapter that is the most striking, especially as he is a researcher for Deep Mind, one of the main developers of AI today.

The very useful opening chapter by Tobias Schlicht introduces the main theories supporting the possibility of AI: functionalism, which highlights causal relations; enactivism, which stresses 'entangled and intertwined embodied activities'; autopoiesis, according to which organisms are 'subjects having purposes according to values encountered in the making of their living' (pp. 12–13); and predictive processing, which focuses on hypothesis testing. Schlicht then argues that cognition in Kant's sense presupposes consciousness, before going on to provide a commendably clear description of the operation and limitations of deep neural networks, which learn via feedback mechanisms and are central to recent developments in AI. His chapter concludes by setting out some of the more prominent views in the AI research community about the possibility of AI.

In his chapter on building a Kantian AI, Evans' starting point is the following:

> In the *Critique of Pure Reason*, Kant asks: what activities must be performed by an agent – any finite resource-bounded agent – if it is to make sense of its sensory input . . . ? Kant's answer, if correct, is important because it provides a blueprint for the space of all possible minds. (p. 99)

Accordingly, he re-expresses Kant's *Critique of Pure Reason* approach to synthesis in a form amenable to adoption by computer, translates this into symbolic logic (the language of computers), and then reports the results so far of implementing this on computer. But is his architecture based on a defensible interpretation of Kant's metaphysics? And how far has he got in building a Kantian AI?

One would expect a faithful modern version of the *Critique of Pure Reason* to be expressed in a contemporary idiom, but it still needs to constitute a structure that is consistent with Kant's metaphysical insights and arguments. Evans starts by presenting Kant as arguing that to make sense of sensory experience is 'to reinterpret that sequence as a representation of an external world composed of objects persisting over time, with attributes that change over time, according to general laws' (p. 40). He sees Kant as enumerating 'all the pure aspects of cognition – those features of cognition that must be in place no matter what sensory input has been received'

(p. 42), which include the pure intuitions of space and time, the categories, and the principles of pure understanding. It follows that a Kantian AI would have to achieve experience via the operation of these pure intuitions, categories and principles. But can this be done?

Evans thinks not, or at least not using Kant's eighteenth-century formulation. Instead, he focuses on what a computer actually has to do to implement Kantian synthesis and in this light reformulates Kant's list of pure aspects of cognition into a set of six pure relations, by which he means relations that feature in every possible synthesis and are defined in terms of the operations that instantiate them: containment, comparison, inherence, succession, simultaneity, and incompatibility.

1. The containment operation (X is in Y) combines intuitions into spatial fields ('containers') which at this stage are not characterised by dimensions. 'The unity condition for containment requires that there is some object, the maximal container, which contains all objects at all times' (p. 55). This maximal container is seen as equivalent to Kant's pure intuition of space, space being understood by Evans as 'the medium in which appearances can be placed together, the medium that allows me to infer from "I am intuiting x" and "I am intuiting y" to "I am intuiting x and y"'.
2. The comparison operation compares attributes (X is brighter than Y).
3. The inherence operation (object X has attribute Y) ascribes attributes to objects, including different attributes to the same object at different times.
4. The succession operation establishes succession via cause and effect.
5. The simultaneity operation establishes what is happening at the same moment by applying mutual causation.
6. The incompatibility operation establishes what cannot be happening at the same place and time.

Evans reports that the processing structure that results, the Apperception Engine, engages in Kantian synthesis by means of nondeterministic choice rules, a neural network, and an unsupervised programme synthesis system.

But does it work? Evans describes an experiment that exemplifies his results, but unfortunately his account is not clear enough to enable evaluation of his claim that these demonstrate the power of his approach. More and clearer accounts of the activity of the Apperception Engine are needed. At the same time, its scope is limited. It is given its sensory stream as a single unit rather than as a continuous stream over time. It does not model the synthesis of reproduction. It does not cover the full range of the table of judgements. No aspect of self-consciousness is implemented.

The following chapters are more sceptical than Evans about the possibility of a Kantian AI. Sorin Baiasu argues that cognition as Kant understands it requires self-consciousness, which AIs do not have. Lisa Benossi and Sven Bernecker maintain that robots cannot be moral agents because they lack an autonomous will. Dieter Schönecker puts the case that computers cannot act morally not only because they lack free will, their actions being totally determined by the laws of physics, but also because they lack the moral feelings necessary for acting from duty. Larissa Berger argues that AIs cannot appreciate beauty because 'the beautiful is that which,

https://doi.org/10.1017/S136941542400027X Published online by Cambridge University Press

without concepts, is represented as the object of a universal satisfaction' (AA 5:211, cited p. 269), which AIs cannot have because they cannot have mental states with phenomenal character. These are perceptive accounts, although it is not clear that AIs will never acquire self-consciousness, feelings, and free will.

The remaining chapters are quite diverse. Hyeongjoo Kim analyses the work of John McCarthy, a major figure in the development of AI, and argues that his view that 'AI as a technical entity is an imitation of the computational ability of human intelligence for problem solving in the empirical physical world' (p. 129) is a form of transcendental realism. Elke Elisabeth Schmidt argues that a Kantian analysis of the trolley problem leads to the conclusion that it is 'not permissible for the Kantian programmer of an autonomous vehicle to make his algorithm steer the car around a group of people if doing so results in hitting a smaller group or a single person' (p. 216). Ava Thomas Wright argues that AIs as moral agents should focus on duties of right, defined as actions which 'can coexist with the freedom of every other under universal law' (AA 6:230, cited p. 225) because these are easier to determine than duties of virtue, being restricted to the public outward aspects of one's actions, and because this focus would sidestep concerns relating to AI capacity for freedom and forestall the possibility of paternalistic meddling by moralistic AIs. Claus Dierksmeier, in quite an empirical chapter, inspects and evaluates from a Kantian point of view two online professional collaboration platforms that use AI algorithms designed to enhance member autonomy. The analyses in these chapters are important given that AIs, whether Kantian or not, will increasingly take decisions in areas in which humans use moral reasoning.

AIs are our future. If we want AIs that are in their essentials like us, basing their architecture on Kant's analysis of cognition is one promising option. Evans' Apperception Engine, which no doubt has been developed further since his chapter was written, makes a start along the road to unity of apperception. But questions remain. Is the shift from pure intuitions, categories, and principles to his six pure relations justified? For example, Evans acknowledges that containment – synthesising objects representing spatial regions – is not the same as Kant's pure intuition of space, which starts with space as a totality and creates sub-spaces by division. Would the operation of his six pure relations, if implemented fully, add up to synthetic unity of apperception, by which he means 'the connecting together of one's intuitions via the pure relations in such a way as to achieve unity' (p. 98)? If so, might this imply a transcendental I? Such a being would start to look quite like a person.

What would a well-built Kantian AI look like? I imagine a small robot sitting in the middle of a parking lot looking at what is in front of it and putting what it senses into the form of 'an external world composed of objects persisting over time, with attributes that change over time, according to general laws' (p. 40). But it would not move. Why would it? It would have no motivations beyond synthesising intuitions in spontaneous acts of synthesis, on and on, with no particular purpose. But AIs are given purposes, or more precisely performance standards, to govern their actions. If these are connected with the processes of synthesis, these processes will start to act in concert with the AI's purposes. The robot would look around, focus, and act. Again, this begins to look a bit like a person.

https://doi.org/10.1017/S136941542400027X Published online by Cambridge University Press

It seems that Kantian AIs may be on their way. This book gives philosophers an excellent overview of recent developments in this area. It is highly recommended.

Hugh Compston
Emeritus Professor, Cardiff University, Cardiff, UK
Email: compstonh@icloud.com

https://doi.org/10.1017/S136941542400027X Published online by Cambridge University Press