3 The Challenge of (Self-)Consciousness: Kant, Artificial Intelligence and Sense-Making

Abstract: How do we make sense of the countless pieces of information flowing to us from the environment? This question, sometimes called the Problem of Representation, is one of the most significant problems in cognitive science. Some pioneering and important work in the attempt to address the problem of representation was produced with the help of Kant’s philosophy. In particular, the suggestion was that, by analogy with Kant’s distinction between sensibility and the understanding, we can distinguish between high- and low-level perception, and then focus on the step from high-level perception to abstract cognitive processes of sense-making. This was possible through a simplification of the input provided by low-level perception (to be reduced, for instance, to a string of letters), which the computer programme was supposed to ‘understand’. Most recently, a closer look at Kant’s model of the mind led to a breakthrough in the attempt to build programmes for such verbal reasoning tasks: these kinds of software or ‘Kantian machines’ seemed able to achieve human-level performance for verbal reasoning tasks. Yet, the claim has sometimes been stronger, namely, that some such programmes not only compete with human cognitive agents, but themselves represent cognitive agents. The focus of my paper is on this claim; I argue that it is unwarranted, but that its critical investigation may lead to further avenues for how to pursue the project of creating artificial intelligence.

Acknowledgements: The first draft of this paper was presented to the ‘Recent literature on Kant’ session of the 2017 UK Kant Society Annual Conference, which took place at the University of St Andrews. I am grateful to participants in the audience, particularly to Richard Evans, for discussion and feedback. Special thanks are owed to the editors of this volume, especially to Dieter Schönecker, for inviting me to contribute this chapter, for their patience and support during the writing of the piece, and for some excellent feedback. Needless to say, all remaining errors are my own.

Sorin Baiasu, Keele University, UK

© 2022 Sorin Baiasu, published by De Gruyter. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. https://doi.org/10.1515/9783110706611-003
1 Introduction

The problem of representation – of how we make sense of the countless pieces of information flowing to us from the environment – is one of the most significant problems in cognitive science. At least two issues have been identified as important for a solution to the problem of representation, both being applications of the problem to more specific parts of the cognitive process of representation. Using a terminology which, as we will see later in this chapter, is in some respects problematic, the two issues have been presented in some of the most influential texts in the literature, as follows. The first was described as an issue of processing information from various sensory modalities, this being the function of the so-called ‘low-level’ perception. Again, this is an issue of selecting from, and making sense of, information provided to us by the senses. The second was described as an issue of extracting meaning from the raw material by making sense of it at a conceptual level, a job assigned to ‘high-level’ perception. A significant step forward in the attempts to answer the problem of representation was made when it was acknowledged that low-level perception, high-level perception and more abstract cognitive activities are interacting in the process of representation and are difficult to be separated from each other.1

Even more progress was made when the problem of the integration of high-level perception and more abstract cognitive activities was separated from the question of how to integrate low- and high-level perception. This separation took place through a simplification of the input, of which high-level perception and more abstract cognitive activities were expected to make sense. Finally, more recently, there was some further progress noted in cognitive science, this time of particular relevance for Kantian studies. Thus, concerning the issue of integration of high-level perception and more abstract cognitive processes and with direct relevant for the general problem of representation, it has been suggested

---

1 As an approximation, the distinction between low-level perception and high-level perception can be understood along the lines of Kant’s distinction between subjective and objective perception. (KrV: A320/B377 – see n3 below for the convention used for references to Kant’s Critique of Pure Reason.) Both ‘low-level’ and ‘high-level perception’ are regarded as spectra of processes, some more concrete and some more abstract. For instance, an example of low-level perception takes place when the light impinges on the retina. A further example is the processing of brightness contrasts and of light boundaries in the visual field. High-level perception may include object recognition or, more abstractly, relation recognition. An example of a more abstract cognitive process would be the understanding of a complex situation, such as a love affair.
that a good solution could be found in Immanuel Kant’s theoretical work, particularly the *Critique of Pure Reason.*

The resulting ‘Kantian machine’, a computer equipped with the ‘Kantian programme’, yielded, for the same problems, results comparable with those obtained by human beings. The claim, however, was not simply that cognitive science has now the ability to match (and perhaps outdo) human performance also for the specific type of task under consideration – this would have been a significant claim in its own right to be sure, although the same obtained for other type of tasks (think of computation or even more complex tasks, such as playing chess). The claim was a much stronger one, namely, that cognitive science has now the ability to create cognitive agents. In this paper, I examine this claim and argue that it is, at this point, unwarranted, but that the argument presented here may suggest further avenues for how to pursue the general project of creating artificial intelligence.

In the next section, I introduce the representation problem and focus on the progress made in cognitive science by those attempting to answer it. Sections 3 and 4 continue the examination of this progress with particular attention to a recent attempt to create a machine that can solve certain tasks (for instance, verbal reasoning exercises) through a programme which follows the architecture of the mind presented by Kant in his theoretical philosophy. Particularly encouraging are the results the Kantian machine has in the attempt to solve specific tasks relevant for the problem of representation; the performance of the Kantian machine here is comparable to that of human beings. Sections 5 and 6 focus critically on a further claim that the Kantian machine is in fact a cognitive agent. The focus is on the distinction between sensory and cognitive agency, and on what is needed for an agent to be a cognitive being. The final section draws the conclusions of my argument.

### 2 The Representation Problem

In their seminal text, “High-Level Perception, Representation, and Analogy: A Critique of Artificial Intelligence Methodology”, David Chalmers, Robert French and Douglas Hofstadter identify our capacity of making sense of the vast amount of information constantly flowing to us from our environment as “one of the deepest problems in cognitive science”. (Chalmers *et al.* 1992, p. 185) The problem, more

---

2 See, for instance, Evans’s paper in this volume (Evans 2022), but also some of his earlier texts, including Evans (2017).
exactly, is to understand how this capacity functions and, even more precisely, to understand human perception’s ability to bring order to the multiplicity of “raw data”. (Chalmers et al. 1992, p. 185) It is from these ‘data’ that we select, organize and interpret specific inputs in order to obtain more or less abstract day-to-day perceptions, whether movement in the visual field, emotion in a tone of voice, what we should do in a game or why a particular political phenomenon is currently unfolding.

The result of the process of perception is representation. By analogy with Kant’s distinction between sensibility and the understanding, in the Critique of Pure Reason, a distinction is identified by Chalmers et al. as implicitly functioning in cognitive science. This is the distinction between high- and low-level perception. Thus, according to them, “[t]oday Kant’s model seems somewhat baroque”; yet, “its fundamental insight” is still “valid”. (Chalmers et al. 1992, p. 186) This fundamental insight is Kant’s distinction between a faculty (sensibility) “whose job it is to pick up raw sensory information” and a faculty (the understanding) “which is devoted to organizing these data into a coherent, meaningful experience of the world”. (Chalmers et al. 1992, p. 186) The focus for Chalmers et al. is on representations produced by high-level perception.

That this Kantian distinction is taken to be the rough model for the distinction between low- and high-level perception is not surprising, but it is not clear the language of ‘low’ and ‘high’ is very useful in connection with Kant. To be sure, sensibility and the understanding are distinct faculties in Kant, and there are specific senses in which what the understanding does, relies indeed essentially, for specific purposes, on what sensibility provides. For instance, the understanding is supposed to provide rules for the synthesis of the intuitions given by sensibility. Kant is famous for claiming that speculative metaphysical claims are the result of conceptual rules used without reference to intuitions. Yet, there also are specific senses in which what sensibility produces relies, for specific purposes, on the work of the understanding. For instance, intuitions are also the result of a synthesis of a manifold of sensations, a synthesis which is made possible by (the understanding’s) transcendental unity of apperception.

Be that as it may, Chalmers et al.’s reading of Kant is not the focus here and I doubt their brief comments on Kant were intended as hermeneutically illuminating for Kantian scholarship. Moreover, their discussion does not dwell on this distinction, although they do challenge some of its aspects, but, as already

---

3 In what follows, references to Kant’s Critique of Pure Reason follow the pagination of the first and second editions, abbreviated A/B accordingly. The English and German editions I have used are listed under Literature.
mentioned, the focus is on representation formed by ‘high-level’ perception. The focus on ‘high-level’ perception is not surprising; after all, the problem of making sense or the problem of representation is supposed to concern the process which yields coherent, meaningful experiences of the world. This, however, seems to be primarily the task Kant assigns to the understanding, although, as already mentioned, not to the exclusion of sensibility.

The approach Chalmers et al. take is, in the first instance, critical. They object to the dominant answers provided at that time to the “representation problem”. (Chalmers et al. 1992, p. 188) What is criticised is the dominant or traditional approach in artificial intelligence, which from the start identifies not only the structure, which ‘high-level’ perception-based representations are supposed to embody, but also the data considered to be relevant for a particular problem. Far from offering a solution to the problem of representation, the claim is that the traditional approach bypasses it, since it starts from data which have already been represented in a form close to the sought-for representation. (Chalmers et al. 1992, pp. 192–8)

One case considered is a model of scientific discovery, BACON, claimed to be able to discover, among other laws, Kepler’s third law of planetary motion. (Langley et al. 1987) This model is embodied in a computer programme, which allegedly starts from the same conditions as those of the human discoverers and is able to provide as output a formulation of Kepler’s third law of planetary motion. Another case is that of the structure-mapping engine (SME) for analogy-making. (Falkenhainer et al. 1989) The SME programme, for instance, is claimed to be able to ‘discover’ an analogy between an atom and the solar system. Yet, as in the case of BACON, the process of analogy-making bypasses representation, since data have been represented in such a way that the common structure is almost immediately apparent.

This objection formulated by Chalmers et al. to the traditional methodology in artificial intelligence has at least one plausible reply. It could be argued that the process of high-level perception can be separated from even more abstract cognitive processes, such as nomic formulation or mapping. Programmes, such as BACON or SME, focus on the latter and leave the task of solving the problem of ‘high-level’ perception for other researchers. The main difficulty for this reply, however, is that these two research tasks cannot be separated so easily. First, perception depends on analogical processes, since it depends on interpreting new situations in terms of old ones. Secondly, however, it is not possible to introduce a temporal separation between ‘high-level’ perception and more abstract cognitive processes, such as mapping. Briefly, this is because perception does not take place in a vacuum, but it is directed to further specific tasks, such as drawing an analogy or formulating
a law. Hence, the task of understanding ‘high-level’ perception will not have one response, which a separate research project could target. (Chalmers et al. 1992, pp. 198–200)

As mentioned, Chalmers et al. begin their seminal text with a critical discussion of the answers to the representation problem in contemporary cognitive science. Yet, their paper has also a constructive or reconstructive part. In this part, they try to integrate ‘high-level’ perception and the more abstract cognitive process of sense-making (in this case, analogy-making). There is, however, a second problem of integration, namely, integration of ‘low-’ and ‘high-level’ perception. The two problems of integration are related and both of them would need to be solved for a proper approach to the problem of representation.

The strategy adopted by Chalmers et al. is to deal with the second problem of integration (namely, between ‘low-’ and ‘high-level’ perception), in order to isolate the first problem of integration (‘high-level’ perception and abstract cognitive process of sense-making). They, then, also claim to have a solution for this first problem of integration. To isolate the first problem, a restriction is introduced, which is meant to answer the difficulty ‘low-level’ perception has to deal with, namely, making sense of the huge amount of information available in the real world in order to convey the relevant details to ‘high-level’ perception. The respective restriction is on the complexity of the input that will be considered for sense-making (in particular, analogy-making). Thus, the input is taken to be given by the domain of alphabetical letters. For instance, for the strings of letters abc and iijjkl, the programme will try to build representations which will make evident their common structure and find correspondences between the two representations. Through this restriction, the first problem of integration is being dealt with, and the focus can be then entirely on second problem of integration. The programme introduced as a solution to this second problem of integration is the Copycat programme.4 (Chalmers et al. 1992, pp. 201–10)

Imagine now a programme for which input is also considered from within the restricted domain of alphabetical letters. The computer is given a sequence of letters and has to predict the next letter in that set as part of a verbal reasoning task. It does this by making sense of the letters through the construction of a rule, on the basis of which it will predict the next element in the series. The programme is also influenced by Kant’s philosophy, but, unlike the case of Chalmers et al.’s Copycat, it regards Kant’s Critical philosophy much more

---

4 Copycat is described as having sprung “out of two predecessors, Seek-Whence and Jumbo”. (Hofstadter and Mitchell 1994, p. 34) The programme is intended as “universal” and, hence, as able to deal with the problems both Seek-Whence (Meredith 1986) and Jumbo (Hofstadter 1983) are supposed to tackle.
sympathetically and considers it as a much more useful tool. In addition, the outcomes yielded by this more Kantian programme are impressive – they are supposed to be comparable to human-level performance achieved for similar tasks.

Imagine, however, that the claim of such a Kantian programme’s supporters is not simply that, in this way, we obtain a machine with human-level performance. Instead, the claim is that, thus programmed, the computer is a cognitive agent. The Kantian programme, in this case, would not simply offer a potential breakthrough in the attempt to answer the problem of representation, but would also make a significant contribution to the attempts to create artificial intelligence more generally. In the next section, I introduce in more detail this Kantian programme.

3 The Kantian Programme

The standard position in the philosophy of mind for those working in the area of artificial intelligence is functionalism. In general, functionalism is the view that, in order for a creature or machine to count as a cognitive agent, what is important is not that they be made of some special kind of substance; instead, the requirement is that they function in a particular way (usually describable in the language of computer science). Every creature or machine that realises that function will count as a cognitive agent.

3 One instance of this Kantian programme, which I will also present in sections 3 and 4, is developed by Richard Evans et al. in a series of texts, including “Kant on Constituted Mental Activity” (2017); also see Evans (2022; in this volume). The claim that his Kantian programme will lead to the creation of a cognitive agent, a claim I will introduce shortly, is not to be found in all of Evans’s texts. For instance, it is absent in Evans et al. (2020) and also in Evans (2022). In the latter text, he does talk about ‘cognitive agent’, but in relation to Kant. In other words, there is an acknowledgement that Kant examines the conditions of cognitive agency, but there is no claim that his Kantian programme or machine (called in this text the “apperception engine”) is a cognitive agent. In email correspondence, Evans acknowledged my discussion in this paper (presented in 2017 – see note 1) as having had a real impact on his thinking.

6 Functionalism can be easily distinguished from traditional mind-body dualism (e.g., René Descartes’s), since, as already mentioned, it makes no claim that minds are made of some special kind of substance. It can also be distinguished from identity theory (e.g., J. J. C. Smart’s), since it does not assume mental states are states of brains and, hence, made of the matter brains are made of. Moreover, unlike behaviourism, (e.g., B. F. Skinner’s), it accepts the reality of internal mental states, rather than reducing it to the behaviour of the whole organism.
Now, the Kantian model of cognitive agency takes Kant’s *Critique of Pure Reason* to include a description of a rule-induction process, a process for the creation of rules on the basis of an input.\(^7\) In our exemplary case, as we have seen in the previous section, the input is a letter sequence given as part of a verbal reasoning task.\(^8\) On this interpretation of Kant, the claim is that, if the rule-induction process meets certain constraints, the internal activities of the process count as cognitive activities.

The idea of *counting* certain activities as activities of a specific kind is central for social practices. Social practices, such as taking part in a competition, will count as practices of a particular kind, if certain necessary and sufficient conditions are met. For instance, among other conditions for a sprint race, included is also the requirement that sprinters run in their respective designated lanes. If this condition is not met, if a person runs crossing lanes, then her activity of running cannot count as participation to a sprint race. In general, to count as social activity of type \(X\), one specific kind of action (or one set of specific kinds of actions) \(Y\) will have to be performed in the appropriate circumstances or in the apposite context \(Z\). By analogy, certain activities count as mental activities, when specific acts or processes are performed or take place under certain circumstances. For instance, having a representation of a red circle is a mental activity, which is counted as such when a plurality of sensory perturbations is experienced by an agent, who applies then a specific rule to these perturbations, in the appropriate context.

Now, counting-as does not happen automatically. Although the necessary and sufficient actions or processes may take place in the right context, these can only count as a specific type of activity, if somebody counts them as such. Hence, in order to count as an activity of type \(X\), a specific kind of action or set of actions \(Y\), performed in the appropriate context \(Z\), would have to be counted.

---

\(^7\) It is unclear, by the way, whether functionalism, as standardly presented in contemporary philosophy of mind, would also be the view Kant would subscribe to. I have discussed this question elsewhere (Baiasu 2018) and it would go beyond the scope of this paper to rehearse those arguments here. I mention only that there are many interpretations of Kant on offer: materialism (Brook 1994), mere immaterialism (Ameriks 2000), ‘dual-aspect’ theory (Nagel 1989), epistemological dualism (Schlicht 2016) and transcendental functionalism (Lau 2014). Not even the transcendental version of functionalism proposed by Lau is clearly compatible with standard functionalism.

\(^8\) Another set of data, where the Kantian programme is supposed to function with achieved human-level performance, is given by sensory readings in a two-dimensional grid world. The rules constructed must make sense of the data. Momentary apprehensions are combined into persisting objects, which change over time according to intelligible rules and interact with other objects according to intelligible rules. (Evans 2017, p.51).
as such by an agent \( W \). Hence, the agent \( W \) must not only be aware of the fact that \( Y \) is taking place and that \( Z \) obtains, but he or she must also be aware that \( Y \) counts as \( X \) in \( Z \). In other words, the agent will need to be aware of the rule for the activity which is counted as taking place (for instance, the social practice of sprint racing) and will need to apply this rule.

For instance, if a plurality of sensory perturbations is taking place in a certain context, then an agent may count those sensory perturbations as a certain kind of activity. Consider a thermometer which measures the temperature of the water in a recipient (say, an aquarium). When the temperature of the water changes, the mercury in the thermometer rises. The internal activity of the thermometer (the perturbation manifested by the rise of the level of mercury) counts (in the right context, say, when the thermometer is not faulty) as a change in the properties of the water, which is external to the thermometer. What is important for us to note for our purposes here is that the agent who counts certain sensory perturbations of the thermometer as a change in the water’s temperature is \( me \). Contrast this with the situation in which we consider the thermometer on its own. The water’s temperature will produce the same sensory perturbations of the mercury, whose level will rise. Yet, we usually do not consider the thermometer as able to count the plurality of its sensory perturbations leading to the rise of the mercury level as its representation of a change of the properties of the external world (the rise of the water’s temperature).

Consider the claim that a thermometer is representing the temperature of the water in the aquarium. On some accounts, we can say that \( A \) is a representation of \( B \) only if \( A \) expresses a thought about \( B \) or only if the relation between \( A \) and \( B \) is an intentional one.\(^9\) If either of these is correct, it follows that, when we talk about the thermometer’s mercury level as representing the water’s temperature, then we assert that the thermometer is expressing a thought about temperature or that thermometer and temperature are in a relation of intentionality. This, however, is at least awkward, if not completely implausible.

What may attenuate the implausible character of such claims (although they may still be awkward) is the idea of derivative intentionality, more exactly, the fact that this is intentionality we attribute to the thermometer. In general, as noted by John Haugeland, symbols have their intentionality derivatively when they inherit it from something else that has the same content already; for instance, a

\(^9\) This is a premise identified by some commentators in Roger Scruton’s argument against photography and film as art (1983); for instance, Robert Stecker formulates it in terms of thought, but also considers a possible reply from Scruton, who could formulate the premise in terms of intentionality. (Stecker 2009, pp. 122–3).
secret signal between you and me has a particular meaning derivatively from the stipulation we made about the signal, when we agreed to have it. (Haugeland 1990, p. 385) Haugeland emphasises the need for an original intentionality, as a way of stopping the infinite regress of a series of derivative intentionalities. Thus, in the case of the secret signalling, if our intentionality were also derived and the intentionality of the source of our intentionality were in its turn derived, and we simply moved from derivative intentionality to derivative intentionality, an infinite regress would threaten. (Haugeland 1990, p. 385)

Haugeland rightly notes that the issue of a potential regress can be answered by reference to an original, non-derivative intentionality; some intentionality must be non-derivative. The significant problem does not concern the transfer or derivation of intentionality from the original, non-derivative instance, but the very possibility of this original intentionality. One question concerns the nature of original intentionality. Haugeland rejects the assumption sometimes held that only mental intentionality is original. (Haugeland 1990, pp. 385–6) Thus, it seems clear that some intentionality must be non-derivative; it also seems plausible that at least some instances of original intentionality are mental in character; it does not follow, however, Haugeland concludes, that all original intentionality should be mental.

We have seen that functionalist accounts of the mind regard as cognitive any processes which realise specific functions. On the Kantian-programme-based account of cognitive agency, a computer equipped with the Kantian programme will count as a cognitive agent and, hence, as an agent with original intentionality. For such accounts, therefore, it is important to subscribe to a version of Haugeland’s scepticism about the character of original intentionality, namely, to subscribe to scepticism about the standard view that cognitive agents cannot be machines. Otherwise we could not take a machine equipped with the Kantian programme to be a cognitive agent with original intentionality, although the plausible assumption is that a cognitive agent (and perhaps any agent more generally) has original intentionality. Moreover, these accounts also rely on scepticism about another claim, sometimes made in these debates, namely, that only biological organisms can achieve original intentionality. A computer or robot, which is programmed to function in accordance with the Kantian programme is expected to achieve original intentionality and represent an instance of cognitive agency.

10 To be sure, what ‘having mental character’ means is an issue for ongoing debate. The argument in this paper does not rely on any particular response to this question.
In the following section, I would like to focus more closely on those aspects which are specified by the supporter of the Kantian-programme-based account of cognitive agency as characteristic for the cognitive agent.

4 Cognitive Agency

Consider a merely sensory agent, such as the thermometer in the example above. There are, to be sure, vigorous debates on the nature of agency,\textsuperscript{11} so my understanding of what an agent is will aim to be sufficiently general to avoid the main issues. I take agency to refer very generally to a capacity to do something. For instance, the thermometer can measure the temperature of the water. As the name indicates, a sensory agent has sensors. The behaviour of the merely sensory agent depends on the state of its sensors. To make this more evident, consider a slightly more complex instance of a merely sensory agent – a thermostat. We can assume that the context is the same – an aquarium – and we assume it is an aquarium housing tropical fish, who need a temperature between 25° and 27° Celsius. When the thermostat’s sensors indicate low temperature, the thermostat will react by switching on the aquarium heater. The thermostat does react to the perturbations of its temperature sensors, but it is an automatic response. Switching the aquarium heating on is not done because the thermostat counts the perturbations of its gauge as its representation of the water’s low temperature; we may do so, but the thermostat itself only responds blindly according to the way it was programmed.

Unlike a merely sensory agent, a cognitive one is a sensory agent who, in addition to behaving in response to his sensings, also counts his sensings as his representation or presentation of an external world – hence, an agent who also has original intentionality. The move from a merely sensory agent to a cognitive agent can be understood at least in part as the result of the way in which the cognitive agent combines the plurality of sensings obtained by the sensors (or provided as input). Consider now a programme that includes rules of composition, which combine parts into wholes. For instance, the programme may be permitted to combine a group of sensings representing a nose with a group of sensings representing an ear under the totality of a face.

These rules of composition are defeasible, but they enable the programme to place groups of sensings under one element. For instance, the rule for the totality of the face mentioned above is defeasible, since specific features of a

\textsuperscript{11} For an introductory presentation, see Markus Schlosser (2019).
context may undermine the composition of a face out of the representation of one ear and of one nose. We may discover that what the programmed machine senses is not a person’s face, but, say, a cubist painting. Sensings can be combined not only by composition, but also by connection. Rules of connection are normative: they specify which combination must be made, once a group of sensings is regarded as composing a whole, and which combinations should be avoided. For instance, if a group of sensings composes a nose, then it must also compose part of the face and, moreover, it should not be regarded as composing an ear.

On the Kantian model of cognitive agency, the activities of combination are similar to social practices, as discussed in the previous section. Thus, both composing and connecting are done by doing something else, in the same way in which I cannot simply run a sprint race, but I do so indirectly by performing some actions (including keeping within my designated lane). Similarly, activities of combination would be performed by constructing rules, which permit or obligate me to combine representations in a certain way and then apply these rules. Hence, combinations are performed indirectly, too, through the construction and application of rules.

To sum up, the Kantian-programme-based account of cognitive agency claims that a cognitive agent counts his sensings as representations of the external world; moreover, the claim is that, in order to do so, the cognitive agent needs to combine these in a certain way, and, in order to combine his sensings in the appropriate way, he needs to construct rules and apply them. We can take Kant to suggest the same; for instance, according to Kant,

when we think of a triangle as an object, we do so by being conscious of the assembly [Zusammensetzung] of three straight lines according to a rule whereby such an intuition can always be exhibited [dargestellt].

(KrV: A 105)

Kant can be read here to claim that, in order to count the three lines as a triangle, we need to apply a general rule for counting certain sets of three lines as triangles. We cannot combine sensings directly, without constructing and applying a rule, because a combination without rules does not satisfy the condition of unification, which is very important. According to this condition, unguided combination would not produce a unity of experience that I could call mine. To be able to combine sensings without rules would mean that there would be no need for a self that would have the sensings. Yet, without a self, the notion of cognitive agency would no longer apply.

This argument offered by the supporter of the Kantian-programme-based account of agency moves rather quickly from the assumption of a unification without rules to the absence of a self to whom experience and its unity are supposed to belong. This, as we will see, will be reflected by the objection I will
formulate in the next sections. Presumably, however, the argument could be reconstructed as follows: the unity of experience, which is needed in order for a cognitive agent to have experience, is a unity of presentations; for Kant, however, any presentation should be connected with a possible consciousness of the cognitive agent’s having it. My presentation of this window, for instance, is connected with a possible consciousness that I am seeing this window. If there is a presentation of a door, but I cannot connect it with a possible consciousness that I am seeing this door (for instance, it is your presentation of a door), then I cannot unify these presentations as part of my experience of a house.

The defeasible rule which connects the presentations of a window and of a door to yield the experience of a house is indeed necessary for the unity of experience required by the experience of the house. Without that rule, it is indeed the case that I cannot have the unity of the experience of a house, which I can call mine. If a unity of experience would still be possible without this rule, the self-related requirement imposed by the application of the rule (namely, that the presentations be the presentations had by the same conscious agent) may not be needed.

This interpretation seems also confirmed by the following passage in Kant:


Hence the original and necessary consciousness of one’s own identity is at the same time a consciousness of an equally necessary unity of the synthesis of all appearances according to concepts – these concepts being rules that not only make these appearances necessarily reproducible, but that thereby also determine an object of our intuition of these appearances, i.e., determine a concept of something wherein these appearances necessarily cohere [zusammenhängen]. For the mind could not possibly think its own identity in the manifoldness of its presentations, and moreover think this identity a priori, if it did not have present to it the identity of its act – the act that subjects all synthesis of apprehension (a synthesis that is empirical) to a transcendental unity, and thereby first makes possible the coherence [Zusammenhang] of those presentations according to a priori rules. (KrV: A 108)

Kant can be read here (as we will see, quite controversially) to claim that we cannot perceive the unity of the self in our sensings; we can perceive objects, which are determined by connecting the sensings under rules. The unity of the self is achieved by what persists through the sensings, and these are the constraints on the rules applied. A cognitive agent on this model does not have a primitive ability to combine representations, but can only combine them through a rule which says that she may or must do so. Hence, the cognitive agent sets down rules, which she will then follow in order to make sense of her sensings and perceive the world. Hence, in order for a sensory agent to count her sensings as representations of an external world (and, hence, in order for her to be a cognitive agent with original intentionality), she will have to construct and apply rules of combination satisfying various constraints.
A rule is understood as a general relation, which applies in many situations and specifies that a certain activity is permitted or obligatory. This is not an explicit, linguistically formulated conditional. Although we may employ language in order to describe a rule, the rule is an implicit procedure for generating a representation. Yet, although implicit, nor is a rule the same as a disposition. A disposition indicates a high probability for specific behaviour under certain conditions. A rule, by contrast, requires or allows the performance of an activity.

Rules of composition are perceptual rules – they are rules for apprehending particular configurations as parts of objects. For instance, one such rule may state that, if sensors meet certain conditions, the agent will count them as his representation of an ear. Rules of connection may form concepts or make judgements. Some rules, for instance, may form a concept by stipulating that it is a sub-concept of another concept or excludes another concept. Some rules may connect concepts conditionally, depending on external factors – for instance, that if a tree gets no water, then it dies. Finally, some rules may specify that if something is counted as a man, then it must also be counted as mortal, or if someone is Caesar, then he is a general.

Apart from constructing rules of combination, the cognitive agent needs also to apply them, in order to count the result as representation of the external world. One important aspect about rules in general is that their applicability is not conditional on the applicability of some further condition. Were a rule to be of this kind, an infinite regress would threaten, because, in order to determine whether or not the additional condition applies, another rule would be required, which, in its turn, would also require another rule with a potential for an infinite number of such iterations. Hence, rules are themselves responsible for determining whether or not they apply. If a rule applies in a particular situation, then it is will either be a rule of composition or a rule of connection. In the case of the former, the agent knows she may perform the combination activity, but whether or not she will perform is not decided on the basis of a further rule, since this will again lead to an infinite regress.

Kant certainly also makes this point about the applicability of rules; Kant explains that, in general logic, to show whether something stands under a rule or not, another rule would be needed. In its turn, this would also require for its application another rule. Hence, Kant concludes that the understanding can be instructed and equipped through rules, but the power of judgement cannot be taught, but only practiced. (KrV: A 133/B 172) So where, in the Kantian-programme-based account of agency, can we find the power of judgement? The answer seems to be that the power of judgement just is the application of rules, which contain procedures for determining whether or not they apply. The practice of rule application just is the faculty of judgement.
The rules of composition will be applied with the help of the imagination. The imagination is considered a faculty involving spontaneity, since it has the choice over whether to form a particular combination under rules of composition or not. However, the imagination is part of sensibility, since it applies rules of composition on sensings, which sensibility has provided. For rules of connection, there is no longer the same latitude in the application of rules, since rules of connection obligate the agent, rather than merely permitting her, to perform the respective mental activity.

So far, we have seen that a cognitive agent is a sensory agent, who counts her sensings as representing an external world. We have also seen that, to count these sensings as representing an external world, the agent must combine those sensings together in the right way. Finally, to combine the sensings together in the right way, the agent must construct and apply rules that satisfy a set of constraints. These constraints are severally necessary and jointly sufficient in order for the agent to count the plurality of sensory perturbations as representing an external world, since this is what it means for the agent to have original intentionality, and this is required for cognitive agency.

For the Kantian-programme-based account of cognitive agency, these constraints are provided by Kant in the *Critique of Pure Reason*. If one aim of the account of cognitive agency is to explain original intentionality, then one aspect which needs to be explained is the unity of cognitions. Cognitions can be unified in time, and this process involves four aspects: constructing moments in time, generating intermediate moments of time, providing a total ordering of moments of time and generating the totality of time (by excluding moments that are impossible).

These aspects, it is claimed, determine constraints on the construction and application of rules. The starting point is the highest principle of all synthetic judgements, which is a top-level constraint, according to which knowledge claims must be about items of possible experience. The four aspects mentioned above follow. They generate constraints on the types of rules that can be constructed and further constraints on the results of applying the rules.

We now have a relatively comprehensive picture of the Kantian computational architecture, the realisation of which (for instance, in a computer programmed on the basis of this architecture) is supposed to be a cognitive agent with original intentionality. The input provided is the result of a simplification, which, in the case of Chalmers et al., was meant to avoid the problem of the integration of ‘low’ and ‘high-level’ perception. In the remaining part of this paper, my question will be whether we can indeed regard the resulting programmed machine as a cognitive agent, that is, as a sensory agent endowed with original intentionality.
5 Original Intentionality and (Self-)Consciousness

As we have seen, A is said to have original intentionality in representing p, when A himself counts his activities as A’s representing p. By contrast, a merely sensory being, a being without original intentionality (think of the example of the barometer or of the thermostat) cannot count a plurality of its sensory perturbations as its representation of a change of properties in the external world. Now, in order for a being to count its activities as its representing something, it needs to be able to represent those activities as (its) representing activities. After all, this process of counting activities as representing the external world is an interpretive process. Yet, on a Kantian account of representation or presentation [Vorstellung], we can only present something and, hence, have presentations, when these presentations are connected with a possible consciousness of having them. As Kant famously puts it,

The I think must be capable of accompanying all my presentations. For otherwise something would be presented to me that could not be thought at all – which is equivalent to saying that the presentation either would be impossible, or at least would be nothing to me.

(KrV: B 131–2)

As is often pointed out, Kant does not require that every presentation be actually accompanied by this consciousness of having it. All that is needed is the possibility of this consciousness. It follows that, in order for a being to count its own activities as its representing something, the representing process of counting-as should be accompanied by the possibility of a consciousness of this process. This is a point Kant also makes explicit in the first edition of the Critique:

All presentations have a necessary reference to a possible empirical consciousness. For if they did not have this reference, and becoming conscious of them were entirely impossible, then this would be tantamount to saying that they do not exist at all. (KrV: A 117 n)

Here Kant presents a similar reductio, starting from the contradictory of what is to be demonstrated (that is, starting by assuming that there is a presentation without a necessary reference to a possible empirical consciousness) and concluding that, in order to exist, presentations must have a reference to the possibility of consciousness. The argument is that, if a presentation might not have reference to a possible empirical consciousness, then it might be impossible for a person to become conscious of that presentation. Yet, if a person cannot become conscious of a presentation, then that presentation cannot be that person’s presentation and, hence, cannot exist as that person’s presentation.12

12 Consider the objection (which I owe to Dieter Schönecker; see his paper in this volume, Schönecker 2022) that the unifier need not be an I; a computer’s control unit might suffice.
Perhaps the clearest way to see how the argument works is by an analogy with a fictional idea. If, say, I create a character for a short story, then that character’s existence depends on the representing activity I perform through imagination. If it is impossible for me to be conscious of this character, then I cannot imagine it and, hence, it cannot exist as my creation. The same goes for any of my presentations or representations – if I cannot be conscious of them, then I cannot represent them and, hence, they cannot exist.

Now, I have been using ‘consciousness’ without additional qualification; the claim that presentations have a necessary reference to a possible empirical consciousness may suggest that we may have presentations of which we are not actually conscious. The point here, however, is not to debate the possibility of unconscious or subconscious presentations in Kant. When Kant makes a claim to a possible empirical consciousness as necessary for every presentation, the implication is not that we have presentations of which we are not conscious. The point is rather that, in some cases, the link between me and my presentation is not the focus of my attention. I may well be conscious of a particular presentation without being explicitly aware of it as my presentation. This is then a presentation of which I am conscious, but which is not accompanied by the I think or by a consciousness of the presentation as mine.

Presumably the implication will be that, if a presentation cannot exist as a person’s presentation, it does not mean that it cannot be part of a unity the computer will offer through its processing. The average calculator, which synthesises ‘2’, ‘2’ and ‘+’ offering as an answer ‘4’, seems to perform precisely this synthesis without any I. But consider the simplified similar process of dropping two balls next to two others; in order to see the four balls as the result of a synthesis, I would need to be aware of the initial two balls, of the additional two and of the process of addition. If I am simply presented with four balls, it will not be self-evident that this is the result of a synthesis, and the calculator or the mechanism which leads to the result of four (say, four balls) will not be able to take this result as representing the result of an addition; it will just go through a mechanical process.

This issue is sometimes discussed when the history of psychoanalysis is researched and presented, but the topic is not directly relevant for this paper.

In what follows, I rely on aspects of my interpretation of Kant’s account of the transcendental unity of apperception in Kant and Sartre: Re-discovering Critical Ethics. (2011: Ch. 1, esp. §§12–3) In that text, I draw a parallel between Kant’s transcendental apperception and Sartre’s (self-)consciousness (as non-reflective self-consciousness) – hence, the title of this paper. See also José Luis Bermúdez (1994, p. 234–7). For the relevant aspects on which I focus here, Stephen Engstrom’s account offers a similar interpretation, although Engstrom does not link his account of the transcendental unity of apperception to the issue of identity and the Paralogisms. (2013, pp. 52–3 for what Engstrom takes to be the link to the problem of identity) The closest to my focus in this paper is the discussion by Melissa McBay Merritt. (2009) Merritt thinks that the philosophy of mind Kant advances in the Deduction is part of an enlightenment epistemology, which requires that a subject be able to recognise herself as the source of her
This can be easily noticed in cases where we are absorbed by the object of some activity or other to the point of ‘forgetting about ourselves’. We are certainly aware of the object of activity and of the activity itself, but we are not aware of them as our activities or objects. Asked what we are doing, we then change the focus of attention, reflect on ourselves and regard the previous object of consciousness as our presentation. Again, applied to the presenting process of counting-as, which is crucial for the notion of original intentionality, what this implies is that a cognitive agent can only take its sensory perturbations as representing activities of an external world, if this presentation of the sensory perturbations is linked to a possible consciousness of the presentation as the agent’s presentation.

This, however, is a condition for a second condition, which is a fundamental requirement for cognition in general. According to Kant, cognition relies on the “principle of the original synthetic unity of apperception”. (KrV: B 137) Kant regards this principle as “the primary pure cognition of understanding [. . .] entirely independent of all conditions of sensible intuition”. (KrV: B 137) The role of this synthesis is to make possible the unification of the manifold both under concepts and under intuitions. As Kant puts it, “not only do I myself need this condition in order to cognise an object, but every intuition must be subject to it in order to become an object for me.” (KrV: B 138) Kant takes ‘cognition’ to refer to intuitions, concepts or (in the case of “the proper meaning of the term” – KrV: A 78/B 103) a synthesis of intuitions and concepts. In all these cases, a synthesis is required – whether of sensations, of marks or of intuitions and marks – and this synthesis is made possible by the a priori synthesis of the transcendental unity of apperception. Thus, Kant sees this principle of the synthetic unity of apperception to be a condition which makes possible “all thought”.

As we have seen, each of our presentations is potentially accompanied by a consciousness of the presentation as our presentation. Cognitions (whether intuitions, concepts or a mixture of the two) are presentations, which refer to an object. For this objective reference, however, the understanding needs to synthesise under a rule the manifold of the elements, which constitute the cognitions. She reconstructs “Kant’s argument for the apperception principle [the principle of the original synthetic unity of apperception]” as connected “with the notion of cognitive agency”. (2009, p. 63).

15 “Cognition is either intuition or concept (intuitus vel conceptus). An intuition refers directly to the object and is singular; a concept refers to the object indirectly, by means of a characteristic that may be common to several things.” (KrV: A 320/B 376–7) For further discussion of Kant’s account of cognition and the distinction between cognition and knowledge, see Marcus Willaschek and Eric Watkins (2020).
respective cognition. This rule is given by the original synthetic unity of apperception. This is an a priori unity, which is created by apperception. Every aspect of my experience – whether a sensation, a conceptual mark, an intuition, a concept or a combination of these – is, as we have seen, an aspect of which I am potentially aware as an aspect of my experience. Hence, every aspect of my experience presupposes a potential link to a form of self-consciousness. The original synthetic unity of apperception is the mode in which aspects of our experience are formed, so that they are potentially linked to the same consciousness, namely, my identical consciousness.

I have called this original apperception ‘(self-)consciousness’, since it is a form of self-consciousness, but one in which I am not reflecting on a self or ego, but on an aspect of experience as an aspect of my experience. The self, therefore, is still not explicitly posited and the identity of this self is only presupposed by the a priori synthetic unity of apperception or (self-)consciousness. It is unclear, however, that a Kantian machine would be able to meet this condition of (self-)consciousness or at least this is what I will claim in the next section.

6 (Self-)Consciousness and the Kantian Machine

As we have seen, (self-)consciousness is what makes possible for me to reflect on a presentation I have and to make it explicit that it is my presentation. The presentation which I have and of which I am conscious as a presentation is presented as my presentation (since I can realise that it is I who thinks the thought represented by this presentation). Moreover, given the principle of the synthetic unity of apperception, this presentation of which I am now conscious as my presentation is already part of a formal unity, given by the identity of the I in the ‘I think the thought represented by this presentation’. This is how I understand Kant’s very condensed explanation of the principle of the synthetic unity of apperception:

For it says no more than that all my presentations in some given intuition must be subject to the condition under which alone I can ascribe them – as my presentations – to the identical self, and hence under which alone I can collate them, as combined synthetically in one apperception, through the universal expression I think. (KrV: B138)

Kant explains here that the principle of the synthetic unity of apperception, which he claims to be also the primary pure cognition of the understanding is itself an analytic judgement. It claims that in order for me to have a cognition, the presentations constituting that cognition must be unifiable as my
presentations and, hence, must meet the condition that they are the presentations of an identical self. What follows is that cognition in general relies on the possibility of reflection, as given by (self-)consciousness or transcendental apperception. This is needed for the synthetic unity of apperception, for cognition and, hence, also for cognitive agency with its capacity to make sense.

Now to see why the Kantian machine seems unable to meet this fundamental condition for cognitions and, hence, for cognitive agency, consider a different thermostat. Apart from the temperature sensor, this thermostat has also a sensor for atmospheric pressure. The thermostat no longer regulates temperature in an aquarium, but in a recipient where the liquid should not go over the boiling point. It is well known that the boiling point occurs at a lower temperature when the pressure is reduced. So the thermostat measures continuously air pressure and the temperature of the liquid in order to avoid the boiling of the liquid. This example is useful, since we have a system which is regulated by two types of rule. There is first the temperature-related rule, which determines the thermostat to switch on a cooling element, when the temperature of the liquid is close to the boiling point. There is, then, the pressure-related rule, which tells the thermostat what the boiling temperature is, depending on the pressure of the air around the recipient. The second rule is in fact a higher-order rule relative to the first, since it modifies the first rule. In fact, it can be seen as a constraint on the first rule.

Level of complexity aside, the second thermostat has a structure similar to the Kantian machine. Recall that the Kantian machine receives a relatively simple input, has certain rules (of combination, more exactly, composition and connection) and some constraints on the rules. Nevertheless, there seems to be no doubt that the second thermostat is a merely sensory agent, not a cognitive agent yet; by contrast, the claim of the supporter of the Kantian-programme-based machine is that the Kantian machine is a cognitive agent. As a cognitive agent, this machine should have original intentionality, that is, should be able to count its own perturbations as its representation of the external world.

Yet, in discussing Kant’s account of cognition, we have seen that, apart from rules and constraints, Kant’s account also presupposes as a significant condition (self-)consciousness. This is a capacity of the agent to represent her representations as her own. This condition is important for the synthetic unity of apperception, which Kant takes to be the fundamental condition for cognition in general and, hence, presumably, a fundamental condition for a

---

16 Some of the things a Kantian machine would need to do in order to be a cognitive agent seem to be doable by an extended model of what Evans calls in this volume the “Apperception Engine”. (See esp. sub-subsection 4.4.6.) The claims there are programmatic and go in the same direction as this paper’s argument.
cognitive agent too. Yet, there is no further aspect presented as part of the Kantian machine, which would play the role or function in the way in which (self-)consciousness is supposed to.

Nevertheless, as we have seen, Kant takes (self-)consciousness to be very significant for cognition. He thinks a cognition, such as an intuition, can be a synthetic unity of presentations only if I can ascribe these presentations – as my presentations – to the same I, and, hence, only if I can combine them synthetically as one cognition. This is also a presentation, which I can count as my presentation of the world, since the synthesis of its components is not mechanically performed according to a rule, but it is the result of a collation that I perform given the potential for each component to be presented by me as my thought.

By contrast, the account of the Kantian programme presented in Sections 3 and 4 explains unity solely by reference to rules, in particular, first, the construction and application of rules of combination, and, secondly, the constraints on the generation of the rules and on the results of the applications of the rules. Thus, to consider the first source of unity, there is a claim that presentations cannot be combined directly, but we need rules to combine them. A combination without rules, the argument goes, would not satisfy the condition of unification, would not produce the unity of experience that I can call mine, and a self that would have the presentations would no longer be needed. The second source of unity, as already mentioned, is given by the constraints on the construction and application of rules. The Kantian highest principle of all synthetic judgements is considered a top-level constraint, which stipulates that knowledge claims must be about items of possible experience.

While it is correct that rules and constraints are needed for synthesis, it is also the case, as we have seen, that a synthetic unity is not possible without the condition which makes it possible for me to collate the elements to be synthesised as mine. (Self-)consciousness is, therefore, also a necessary condition for cognitive agency, a condition which is distinct from those related to rules and constraints. At the same time, it is a condition for which the Kantian programme provides no account.

As an illustration, consider again the second thermostat: As we have seen, we also have rules and constraints in that case, yet, it is quite plausible to see this thermostat as a merely sensory agent. The gap between the merely sensory agent and the cognitive agent is not bridged by any additional specific aspect of the Kantian programme.
7 Conclusion

Consider a Kantian computer\textsuperscript{17} – a computer programmed with all the rule-generating procedures provided by Kant in the Schematism, Transcendental Deduction and Analytic of Principles. Assume that it provides good answers to verbal reasoning tasks – in fact, as good as those provided by human beings. Does this not mean the computer has to unify the various inputs as part of its experience? Does this not mean it can count its ‘sensings’ as its representing the external world?

In this paper, I have argued that these claims implicit in these rhetorical questions are unwarranted. To be sure, the significance of the successes of the Kantian machine in solving verbal reasoning tasks and other tasks relevant for the problem of representation is not questioned here, but it is taken for granted. What I have argued is unwarranted is the further claim that a Kantian machine is more or less the same as a cognitive agent, that is, an agent with original intentionality. The problematic assumption seems to be the assumption that some of the necessary conditions of experience presented by Kant are sufficient conditions.

I have argued that combining through rules is necessary for my experience, but it is not sufficient. Hence, it might be sufficient to equip a computer with rules which enable it to respond better to various tasks than other computers usually do, but from here to getting the computer to count its ‘sensings’ as its representing the external world is a gap that at least for Kant requires (self-) consciousness. It is, however, unclear how the Kantian machine could be seen as being equipped with this type of apperception. This, however, might be a line of inquiry that is worth pursuing in the attempt, both in cognitive science and philosophy, to solve the problem of representation.

References

KrV Kritik der reinen Vernunft/Critique of Pure Reason


\textsuperscript{17} For instance, as presented in the previous footnote, an extended Apperception Engine.


