Richard Evans[1]

Self-Legislating Machines: What can Kant Teach Us about Original Intentionality?

Abstract: In this paper, I attempt to address a fundamental challenge for machine intelligence: to understand whether and how a machine's internal states and external outputs can exhibit original non-derivative intentionality. This question has three aspects. First, what does it take for a machine to exhibit original *de dicto* intentionality? Second, what does it take to exhibit original *de re* intentionality? Third, what is required for the machine to *defer* to the external objective world by respecting the word-to-world direction of fit? I attempt to answer the first challenge by providing a constitutive counts-as understanding of *de dicto* intentionality. This analysis involves repurposing Kant's vision of a self-legislating agent as a specification of a machine that reprograms itself. I attempt to answer the second and third challenges by extending Kant's synchronic model of *de dicto* intentionality with Brandom's interpretation of Hegel's diachronic model of *de re* intentionality, using Hegel's notion of recollection to provide an understanding of what is involved in achieving deference to the external world.

Keywords: original intentionality, Kant, *de dicto* and *de re* intentionality, Hegel, Brandom.

Introduction

In recent years, modern machine learning techniques have produced remarkable machines capable of excelling in a wide variety of domains, including playing go at a super-human level (Silver et al 2017), accurately predicting protein structure (Jumper et al 2021), and uttering novel sentences that seem to exhibit few-shot understanding of novel concepts (Brown et al 2020). But a fundamental question still remains: when we ascribe meaning to the outputs of one of these machines, is that significance *merely imputed by us*, or do these outputs mean something *for the machine itself*?

A parrot repeatedly squawks a phrase it has overheard from human conversation. This phrase means something to the humans, but it does not mean anything for the parrot itself. Here, the intentionality of the parrot's squawk is *derivative*. When GPT-3 (Brown et al 2020) is described as a "stochastic parrot" (Bender et al 2021), the suggestion is that the natural language output of the machine is similarly derivative: the output might mean something *to us*, but it does not mean anything *for the machine itself*. Here, the accusation is that the output of GPT-3 only has significance because we external observers project that meaning onto it; if we external observers were to cease doing so, the machine's output would cease to mean anything.

[1] Richard Evans. DeepMind. Email: richardevans@deepmind.com

The distinction between original and derivative intentionality was introduced by John Haugeland. Intentionality is derivative if it is attributed by someone else, by *another* agent who is doing the counting-as:

> At least some outward symbols (for instance, a secret signal that you and I explicitly agree on) have their intentionality only derivatively - that is, by inheriting it from something else that has the same content already (e.g. the stipulation in our agreement). And, indeed, the latter might also have its content only derivatively, from something else again; but obviously, that can't go on forever. Derivative intentionality, like an image in a photocopy, must derive eventually from something that is not similarly derivative; that is, at least some intentionality must be original (non derivative). (Haugeland 1990: 385)

The distinction between original and merely derivative intentionality applies to internal states as well as to external outputs. Consider the humble barometer, a simple sensory device that can detect changes in atmospheric pressure. If the mercury rises, this means the atmospheric pressure is increasing; if the mercury goes down, the pressure is decreasing. Now *we* count the mercury's rising as the machine responding to the atmospheric pressure. We count, in other words, a process that is internal to the instrument (the mercury rising) as representing changing properties of an external world (atmospheric pressure increasing). But although we count the internal process as representing an external process, the barometer itself does not. The barometer is incapable of counting the internal process as a representation because - of course - it is incapable of counting anything as anything.

A fundamental challenge, then, for us who seek to understand machine intelligence, is to understand whether and how a machine's internal states and external outputs can exhibit original non-derivative intentionality.

Before we can answer this question, however, we need to recall a distinction between two aspects of intentionality: *de dicto* intentionality (where we express a proposition) and *de re* intentionality (where we represent an object). Consider two ways of talking about Oedipus' mental state after an argument with an old man on the road to Thebes. Compare the *de dicto* ascription "Oedipus believed that the old man had provoked him and deserved to die" with the *de re* ascription "Oedipus believed, of his father, that he had provoked him and deserved to die". In the *de dicto* ascription, the embedded sentence inside the "that" clause expresses a *proposition* that the thinker is responsible *for*; in a *de re* ascription, the embedded noun-phrase inside the "of" operator represents the *object* that the thinker is responsible *to* — but not necessarily under a description that the thinker would himself endorse.

Once we distinguish between these two dimensions of intentionality, our question about original intentionality divides into two parts: not only must we ask whether and how a machine can exhibit original *de dicto* intentionality, but we must also ask whether and how a machine can exhibit original *de re* intentionality.

When we ask whether a machine is capable of exhibiting original *de re* intentionality, if it is capable of having thoughts that are about an external mind-independent world, a further question arises: how does the machine respond when the world and its beliefs about the world go out of sync?

Brian Cantwell Smith emphasizes the importance of this:

> If we are going to build a system that is itself genuinely intelligent, that knows what it is talking about, we have to build one that is itself deferential - that itself submits to the world it inhabits, and does not merely behave in ways that accord with our human deference. To do that, it will have to know (i) that there is a world, (ii) that its representations are about that world, and (iii) that it and its representations must *defer* to the world that they represent. (Smith 2019: 79)

Now the original notion of deference, the "home language game", is the situation in which one agent defers to *another agent*. So, for example, one agent defers to another agent because the latter is more knowledgeable in a certain area, or because she has higher status. But here, in this crucial passage, Smith is using deference to describe a stance our machines should have to *the world*.

What Smith means by "deference to the world" is to adopt the *word-to-world* direction of fit: a representation defers to the world if, when the representation and the piece of reality diverge, then the agent endeavors to change the representation to match the piece of reality, rather than changing the world itself. This is in contrast to a world-to-word direction of fit: when a desire, for example, is out of sync with the world, we attempt to change the world to fit the desire, and not the other way around. The final part of our challenge, then, is to understand how a machine can defer to the world by respecting the word-to-world direction of fit.

This paper attempts to answer the following question: whether (or how) a machine can exhibit original non-derivative intentionality. We have seen that this question involves three sub-questions: (1) What does it take for a machine to exhibit original *de dicto* intentionality? (2) What does it take to exhibit original *de re* intentionality? (3) What is required for the machine to defer to the external objective world by respecting the word-to-world direction of fit? In the sequel, I will attempt to answer the first challenge by providing a constitutive counts-as understanding of *de dicto* intentionality: *if the right norms are in play*, then this particular sub-agential activity *counts as* perceiving that this object has this quality, for example; or, this rule-formation activity *counts as* forming a belief. This counts-as analysis involves repurposing Kant's vision of a self-legislating agent as a specification of a machine that reprograms itself. Next, I attempt to answer the second and third challenges by extending Kant's synchronic model of *de dicto* intentionality with Hegel's diachronic model of *de re* intentionality, using Hegel's notion of recollection to provide an understanding of what is involved in achieving deference to the external world.

Counting-as

The account of intentionality to be presented here explains *de dicto* thought in terms of the counts-as relation: if the right norms are in play, then certain sub-agential activities count as constituting certain *de dicto* mental attitudes. Before I begin this analysis, I need to assemble various reminders about the counts-as relation itself. These features of the counts-as relation will turn out to be essential to the account of *de dicto* intentionality presented below.

In a certain context, an object under one description can also count as falling under another description. To take a well-worn example: in the right circumstances, this wooden horse-shaped piece counts as a knight. Obviously, the counts-as relation applies to people as well as to inanimate objects. For example, in this ceremony, this person counts as the officiator.

Note that the counts-as relation applies to *actions* as well as objects. For example, running away in this particular context (the field of battle, when the enemy are charging) counts as desertion. Note also that the counts-as relation applies to *events* (mere happenings) as well as to intentional action. For example, this deluge counts as a world record-beating flood.

Note that multiple objects are often counted-as jointly and simultaneously; the counts-as claims come as a package. It isn't just that a *single* object counts as satisfying some further description, but rather that a *collection* of objects jointly counts as satisfying a collection of descriptions, as long as multiple norms apply to that whole collection of objects. For example, this wooden horse-shaped piece only counts as a knight if this wooden castle-shaped piece also counts as a rook. Here, there is a multitude of objects (pieces and players), a multitude of counts-as ascriptions (the various roles of the pieces e.g. the Black queen), and a multitude of norms. And, as well as the multitude of objects that count as fulfilling roles, there are also a multitude of events (the various pushings of various wooden pieces) that count as various moves (e.g. moving the knight to queen's bishop three).

Counts-as constitution is more than mere classification. If all sheep are mammals, then in some very weak sense every sheep "counts as" a mammal. But that is not how I am using the term here. The extra thing that is needed to distinguish a full-blooded counts-as conditional from a trivial classification is that in a counts-as conditional there is an extra condition that needs to be satisfied, and this extra condition is *normative*, describing what an object or agent *should do*. For example, the wooden horse-shaped piece only counts as a knight in a certain context - the game of chess - but that context is just short-hand for a large collection of interlocking norms that interrelate a large collection of objects (the pieces and the players). Those norms include the requirement that one player should move after the other has moved, that neither player may castle after they have already moved their king, and so on and so forth.

Note that I am focusing on a notion of counts-as where the condition is normative, but where the condition is that the norm is *in play* - not that the norm is *satisfied*. For example, for this wooden piece to count as Black's king involves, among other things, that Black is permitted to move it one square in any direction, not permitted to jump over other pieces, and obligated to move it out of the way when it is threatened. But suppose an omniscient being assiduously watched every chess game ever played, and observed that only 7.4% of those actual games were completely free of violations of the rules of chess. Do we conclude that *the remaining 92.6% of the games did not actually count as chess games?* That in those cases, the wooden horse shaped piece did not actually count as a knight? - Of course not. The piece counts as a knight because of the norms that are in play - not the norms that are satisfied. Another example: a broken dishwasher is still a dishwasher. Even in its current woeful dilapidated state, it still counts as a dishwasher because it *should* clean the dishes — even if, right now, my dishes come out dirtier than when they went in.

Now some philosophers (for example, Hegel, Wittgenstein, and Brandom) believe the realm of the normative is essentially social, claiming that a norm can only be in play if there is a multi-agent social practice that institutes that norm. In this paper, I assume Kant's alternative

individualist approach in which a single agent is fully capable of adopting and satisfying a norm entirely on her own, without the need to be part of a wider multi-agent community. I discuss this issue in more depth in the final section.

De dicto intentionality

In this section, I argue that *de dicto* intentionality can be explained using the counts-as relation when a certain set of norms are in play. These are the norms of cognition described by Kant in the first half of the first *Critique*.

The result of this analysis will be a multitude of counts-as claims connecting sub-agential processes with *de dicto* cognitions. When Kant's cognitive norms apply, ascribing this attribute to this intuition counts as *perceiving that a particular object has a particular quality* (to take an example at the non-discursive level). Or (to take an example at the discursive level): adding this rule connecting these concepts counts as *forming a belief.*

The analysis will rely on three key features of the counts-as relation that were described above: (i) the counts-as relation can apply to a multitude of objects/actions/events at once; (ii) it can apply to activities as well as to objects; (iii) as well as applying to intentional action, the counts-as relation can also apply to mere happenings, including sub-agential processes.

The key idea is to explain *de dicto* intentionality using the following structure: *If certain norms are in play, then each of these sub-agential activities count as a cognition with de dicto intentionality*. Filling in this sketch involves answering three questions: What are the constitut*ed* activities? What are the constitut*ing* activities? What are the norms? I shall address each in turn.

The constituted activities

The constitut*ed* activities are the various cognitions with *de dicto* intentionality that we want to explain. These include perceptions of a particular moment: perceiving that a particular object has a particular quality, or perceiving that a particular object falls under a general property. They also include perceptions of relations between moments in time: perceiving that an object's having one particular quality is simultaneous with that object's having another (compatible) quality, or perceiving that an object's having one particular quality is succeeded by that object's having a different (incompatible) quality. As well as perceptions, the explananda also include beliefs: believing that an individual object falls under some general concept, believing that a pair of objects fall under some relation, or believing that every object satisfying a certain condition also satisfies a further property. Note that the explananda only contain the elements of cognitive reasoning, not the elements of *practical* reasoning; I am not considering feelings of pleasure and pain, desires, intention, or intentional action.

The constituting activities

The constitut*ing* activities are the various sub-agential processes that will be used to explain the *de dicto* cognitions. I use Kant's terminology from the first *Critique*. The activities include: receiving an impression; intuiting; attributing a particular attribute to an intuition; placing an object in space; comparing two attributes of intuition; connecting two determinations using the

relations of simultaneity, succession, and incompatibility; subsuming a particular object (or tuple of objects) under a predicate; forming a rule connecting concepts. I shall go through these in turn, but the treatment must be condensed for reasons of space.[2]

The sub-agential processes listed above are enabled by various distinct faculties. The faculty of sensibility is responsible for constructing intuitions, i.e. representations of a particular object (e.g. this particular jumper), or a particular attribute (also known as a trope, or mode) of a particular object at a particular time (e.g. the particular dirtiness of this particular jumper at this particular time). Sensibility has a passive aspect allowing it to receive impressions, and an active aspect allowing it to construct intuitions. The intuitions constructed can either be pure (for constructing space and time), or empirical (for constructing objects). In Kant's cognitive architecture, the mind is not given objects by sensibility, but has to *construct* objects in order to unify the information arriving from various sensory modalities. When hearing a buzzing sound, and seeing a black-and-yellow striped object, for example, it is a highly non-trivial achievement to bind these different pieces of information together into a single object.

The faculty of imagination is responsible for connecting intuitions together. There are three operations for connecting intuitions together to form a *determination*[3]: ascribing a particular attribute (trope) to a particular object of intuition, placing an object in a particular region in space, and comparing two attributes for relative intensity. The imagination also provides three relations for connecting determinations together: simultaneity, succession, and incompatibility.

The power of judgment is responsible for subsuming a particular attribute under a general predicate. It is responsible for subsuming this particular shade of red under the general predicate "red", for example. This relation is of course many-many: many intuitions are subsumed under one predicate, and one intuition is subsumed under many different predicates.[4]

Finally, the capacity to judge is responsible for connecting concepts together into a rule. A rule might be of the form: for all intuitions *X*, if I subsume *X* under predicate *p*, then I must also subsume *X* under predicate *q*.

It is at this point, where the capacity to judge is constructing rules, that Kant's picture gets rather complicated, because there are *two levels* at which norms are operating. On the one hand, there are the lower-level norms that are instituted by the agent when the capacity to judge constructs a rule. On the other hand, there are the meta-level norms of cognition that are always in play if the various sub-agential processes are to count as cognitions with *de dicto* intentionality (Pollok 2017). Understanding Kant's vision of the mind requires understanding these two different levels of normativity. To get clear on these two levels, let me digress for a moment to focus on self-legislating machines that are subject to higher-level norms.

---

[2] For a more thorough treatment of this part of Kant's architecture, see (Evans 2022).

[3] In this paper, I use "determination" to mean a combination of intuitions, in contrast to a judgment (which is a combination of concepts). If the meta-norms are satisfied, a determination counts as a perception, while a judgment counts as a belief.

[4] Throughout, I focus on unary predicates for ease of exposition. But of course there are also binary predicates which subsume pairs of objects, and ternary predicates which subsume triples, and so on and so forth.

Self-legislating machines

Imagine the simplest example of a self-legislating machine: initially, the machine constructs an if-then condition-action rule, *freely* choosing one from an infinite set of candidate rules; subsequently, the machine is *duty-bound* to follow the rule it has adopted. For example: the machine operates in a simple two-dimensional grid of cells. The rule it adopts is: if there is a red cell above me, then move right. Once the rule is adopted, it is obeyed unthinkingly: the machine continues to move to the right until there is no longer a red cell above it.

Now we humans, of course, describe rules in natural language. But we are not assuming the machine understands English. Instead, these condition-action rules are expressed in a *simple machine language* that can be interpreted by a simple computer program.[5]

Now we humans, of course, are consciously aware of some of the rules we follow. But this machine is not. It follows the rule, but it is not conscious of the rule. After all, the machine is not conscious at all. Thus the rules operate entirely at a *sub-agential level*.

Now we humans, of course, are always free to ignore the rules we adopt. But here, in this imagined example, we are deliberately suppressing this possibility. Our self-legislating machine is *entirely unquestioning* at the rule-following level.

Next, let us consider a somewhat more complicated extended example. Imagine now a self-legislating machine that can add and remove rules. We also give it a *meta-constraint* that it must always have exactly one rule in play at any time. Suppose, the machine starts off with the rule: if there is a red cell above me, then move right. But after five time-steps, it removes this rule, and replaces it with a new one: if the cell to my right is yellow, then move up.

At the lowest level, the machine is entirely constrained: once it has adopted a rule, it is duty bound to follow it. At the intermediate level, the machine is entirely free: it can adopt *any rule it likes* from an infinite set of possible rules. At the highest level, the machine is again entirely constrained: it has no choice but to add and remove rules so that it has exactly one rule in play at every moment.

Next, let us modify the example one more time. Imagine now that the action to be performed when a rule fires is not a physical action modifying the external world, but a cognitive sub-agential activity. Specifically, imagine the case where the activity to be performed is to *subsume an individual under a predicate*. Now an example rule might be: if you are subsuming this object of intuition under predicate *p*, then you must also subsume this object under predicate *q*.

Imagine further that the machine follows a *set* of such rules at any moment, rather than just an individual rule. Suppose also that the machine has a priority ordering to deal with cases of conflict (cases where two rules fire that propose subsuming the same individual under incompatible predicates).

So far, this is all entirely unconstrained. We need something that is going to constrain the initial set of subsumptions that we start with. The faculty that provides this initial constraint is *receptivity*: the machine has sensors that passively accept input from the external environment, and these impressions ground the initial subsumptions.

But, still, nevertheless, even with the constraint that the original subsumptions are grounded in receptivity, this picture is still woefully under-determined. To prevent severe

[5] The condition-action rules could, for example, be defined in the Teleo-Reactive language (Nilsson 1993).

under-determination, Kant imposes a meta-constraint on this free-for-all process of adding and removing rules under the daunting title of the “synthetic unity of apperception”: the machine’s cognitions must always achieve a certain sort of unity.

This is Kant’s vision of the self-legislating machine. We are free to add and remove rules at any moment. But we must always ensure, following the meta-level norm of unity, that the various activities we perform make sense of the whole sequence of sensory inputs. In this picture, the machine has enormous freedom - it may construct any sort of first-level rules it pleases. But that doesn’t mean that “whatever seems right is right”: the machine must always respect the meta-level norms collected under the umbrella of the synthetic unity of apperception.

The meta-norms of cognition

At the heart of Kant’s cognitive architecture, then, is the idea that when various sub-agential activities jointly satisfy various meta-norms, those activities also count as cognitions exhibiting original *de dicto* intentionality. What are the meta-norms, and how will we know if they are satisfied?

The meta-norms of cognition are of three types: first, the cognitions need to *explain* the succession of impressions that is given by sensibility; second, the various cognitions need to be *connected* together; third, the connected cognitions need to achieve a distinctive form of *unity*:

> But in addition to the concept of the manifold and of its synthesis, the concept of combination also carries with it the concept of the *unity* of the manifold. [B130][6]

Connecting the intuitions together is relatively straightforward. Intuitions are connected into determinations using three operations of the imagination: attributing a trope to an individual, placing an object in space, and comparing two attributes for intensity. Determinations are connected together using the relations of simultaneity, succession, and incompatibility. (One of the striking things about Kant’s vision is that modal relations of incompatibility apply to *determinations* as well as to judgments: perceiving that this particular object has this particular quality is incompatible with perceiving that the object has that particular quality. Here, modal and temporal operators connect lived experience just as much as they connect discursive propositions).

In order for the intuitions to count as connected together, the objects of intuition must be connected in space, the attributes of intuition (tropes) must be related in intensity, and the determinations must be connected in empirical time (via the temporal relations of simultaneity and succession) and pure time (via the modal relation of incompatibility) [A145/B184ff].

The constraints of unity are somewhat more complex, and for reasons of space my treatment is condensed.[7]

Although intuitions are connected in imagination, these connections are - so far - merely arbitrary: there are many possible ways of connecting intuitions together, and no reason to

---

[6] References to the *Critique of Pure Reason* use the standard A/B format.

[7] For a fuller treatment of the constraints of unity, see (Evans 2022).

prefer one over the other. In order for our sub-agential activities to count as cognitions that exhibit intentionality, that represent an external world that is independent of our thought processes, the *arbitrary* connections generated by the imagination must be supplemented by *necessary* connections generated by the understanding.

Imagine someone trying to connect his intuitions together. Suppose he has "intuition dyslexia" – he is not sure if this intuition is the object and this other intuition is the attribute, or the other way round. Or he has two determinations in a relation of succession, but he is not sure which is earlier and which is later. The intuitions are swimming before his eyes. He needs something that can pin down which intuitions are assigned which roles, but what could perform this function? Kant's fundamental claim is that it is only the *judgment* that can fix the positioning of the intuitions. Moreover, this is not just one role of the judgment amongst many – this is the primary role of the judgment:

> a judgment is *nothing other* than the way to bring given cognitions to the objective unity of apperception [B141] (my emphasis)

More specifically, the relative positions of intuitions in a determination can only be fixed by forming a judgment that necessitates this particular positioning. This judgment contains concepts that the intuitions fall under, and the position of the intuitions in the determination are indirectly determined by the positions of the corresponding intuitions in the judgment. Thus:

> The same function that gives unity to the different representations in a judgment also gives unity to the mere synthesis of different representations in an intuition. [A79/B104-5]

There is a parallel claim one level up, at the level of complex judgments: the relative positions of determinations in a relation of simultaneity/succession/incompatibility can only be fixed by forming a complex judgment that necessitates this particular positioning. This complex judgment contains two constituent judgments that the two determinations fall under, and the position of the determinations in the connection are indirectly determined by the positions of the corresponding judgments in the complex judgment. So, for example, if two determinations attributing two tropes to the same object are considered to be incompatible, then there must be an exclusive disjunctive judgment featuring two incompatible predicates which the two tropes are subsumed under.

In order for our intuitions to be about a mind-independent world, the connections between them must be necessary connections. Unfortunately, the faculty of imagination is unable on its own to provide the requisite necessity. In Kant's architecture, the faculties of the power of judgment (subsuming intuitions under concepts) and the capacity to judge (connecting concepts together into rules) are *only needed* to *confer* the necessity on the synthesis of intuitions. At one level, of course, there is a symmetric dependence between the intuitive and the discursive: "Thoughts without content are empty, intuitions without concepts are blind" [A50-51/B74-76].[8] But at a deeper level, there is a striking asymmetry: concepts and judgments

[8] But note the striking asymmetry even here between the two types of deficiency: blindness is a deficiency of a *living being*, while emptiness is a deficiency of a mere *container*.

are merely means to an end, while the unity of intuitions is the end that cognition is striving towards:

> In whatever way and through whatever means a cognition may relate to objects, that through which it relates immediately to them, and at which *all thought as a means is directed as an end*, is intuition. [A19/B33, my emphasis.]

The spontaneity sandwich

This, then, is Kant's account of original *de dicto* intentionality. When the various meta-norms of cognition are in play, the various sub-agential activities count as cognitions exhibiting original (*de dicto*) intentionality.

This picture is an intriguing mixture of utter rigidity combined with total freedom. At the lowest level, the machine has no choice about the succession of impressions it receives from sensibility. Further, the machine has no choice, once it has adopted a rule, in whether or not to follow it. But at the intermediate levels, the machine has an enormous, dizzying array of choices: in terms of pure intuition, the machine is free to divide up space and time any way it likes. In terms of empirical intuition, the machine is free to construct any objects it likes. In terms of the imagination, the machine is free to connect up intuitions together into determinations in any way it pleases, and is also free to connect up those determinations together in any way it pleases. In terms of the power of judgment, it is free to construct any general procedure for mapping attributes (tropes) to predicates. Finally, in terms of the capacity to judge, the machine is free to form any rules it likes. But at the very highest level, by contrast, the machine is entirely constrained: it must satisfy the various unity conditions. This is non-negotiable.

So Kant's vision of the mind sees the subject as a *spontaneity sandwich*: at the very lowest level, there is no freedom at all (it has no choice in receptivity, or in whether or not to follow a rule that it has adopted); similarly, at the very highest level, there is no freedom at all: if its sub-agential activities are to count as achieving original intentionality, then it must respect the norms of cognition. But in the middle — when it comes to the imagination, the power of judgment, and the capacity to judge — the agent is entirely free, wildly gloriously free, to do whatever it likes. The filling of the sandwich is unconstrained spontaneity.

Perhaps the following analogy may be helpful. You have joined a large company as the head of a whole division, and your new boss has entrusted you with considerable freedom: *you can do whatever you like* as long as it increases profits. You have freedom to institute new practices, new expectations, even a new culture, within your division. You have the power to institute new norms (that your underlings will unhesitatingly follow) as long the overall activity of the division satisfies the high-level directive that was handed down to you.

A computer implementation of Kant's cognitive architecture

In order to get clear about the details of Kant's cognitive architecture, I found it helpful to descend from the abstract philosophical level to the concrete computational level, and attempted to turn Kant's blueprint into a working computer program: the Apperception Engine.[9]

[9] The architecture is described in (Evans 2021a), (Evans 2021b), and (Evans 2022). The source code is available at https://github.com/RichardEvans/apperception.

Although crude in places, the implementation has some rather appealing features. In particular, the strong inductive bias provided by Kant's unity constraints enables the machine to learn from a much smaller number of datapoints, significantly outperforming state of the art neural networks in data-hungry situations.

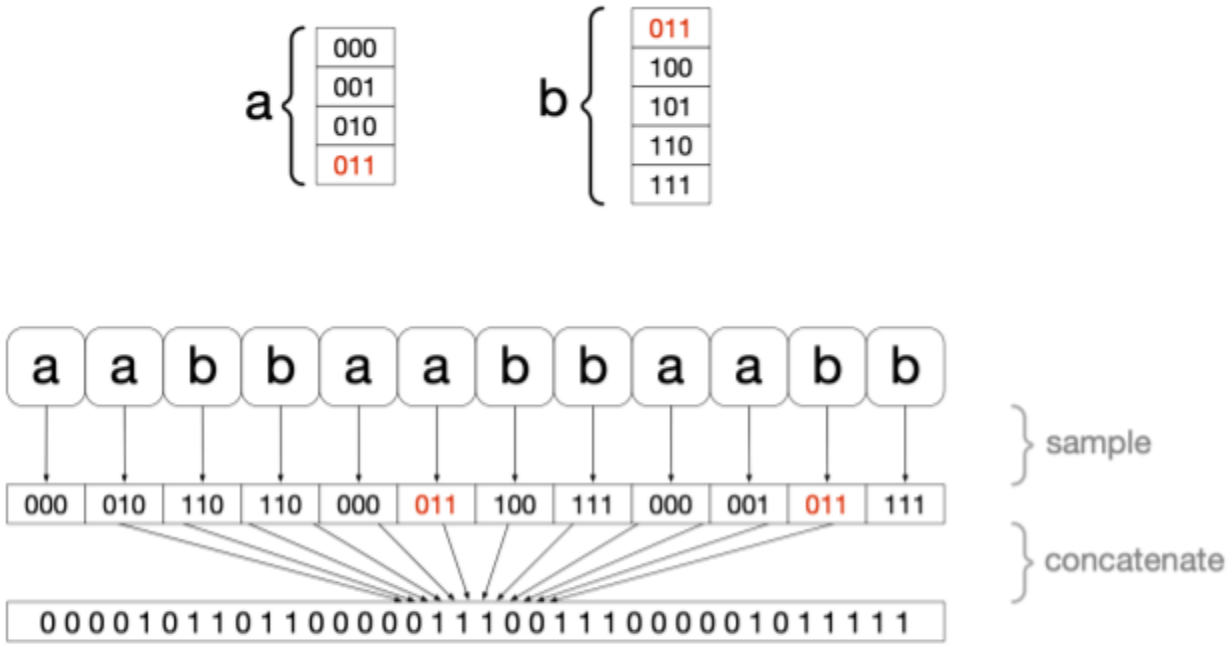

Figure 1. Generating noisy sequences.

Consider, for example, the problem in Figure 1. Here, we start with a simple repeating symbolic sequence *aabbaabbaabb…* Then we transform the sequence of discrete symbols into a sequence of noisy vectors. Each occurrence of symbol "a" can be mapped to 000, 001, 010, or 011. Each occurrence of symbol "b" can be mapped onto 011, 100, 101, 110, or 111. Note that the vector 011 is ambiguous and can either represent an "a" or a "b". Once we have replaced each symbol with one of the vectors, we have a sequence of 3-bit vectors. Finally, we concatenate the vectors together, producing a sequence of bits. Note that after concatenation, the fact that the input was originally composed of a 3-bit chunks has been lost.

The task for the machine is to make sense of the sequence of bits. It must (somehow) parse the sequence into chunks, discern the underlying regularity, and use that regularity to correctly predict future bits in the sequence.

When the Apperception Engine is presented with this sequence of bits, it finds an interpretation that makes sense of the sequence of raw impressions. See Figure 2. The machine invents two predicates, *p* and *q*, representing whether the symbol is an "a" or "b". To map raw intuitions to predicates, it uses a binary neural network to implement the power of judgment. To construct rules expressing judgments, it uses a program synthesis system to implement the capacity to judge. Modern neural networks, by contrast, were entirely unable to make sense of these sequences. For further details of this experiment, and others, see (Evans 2021b).

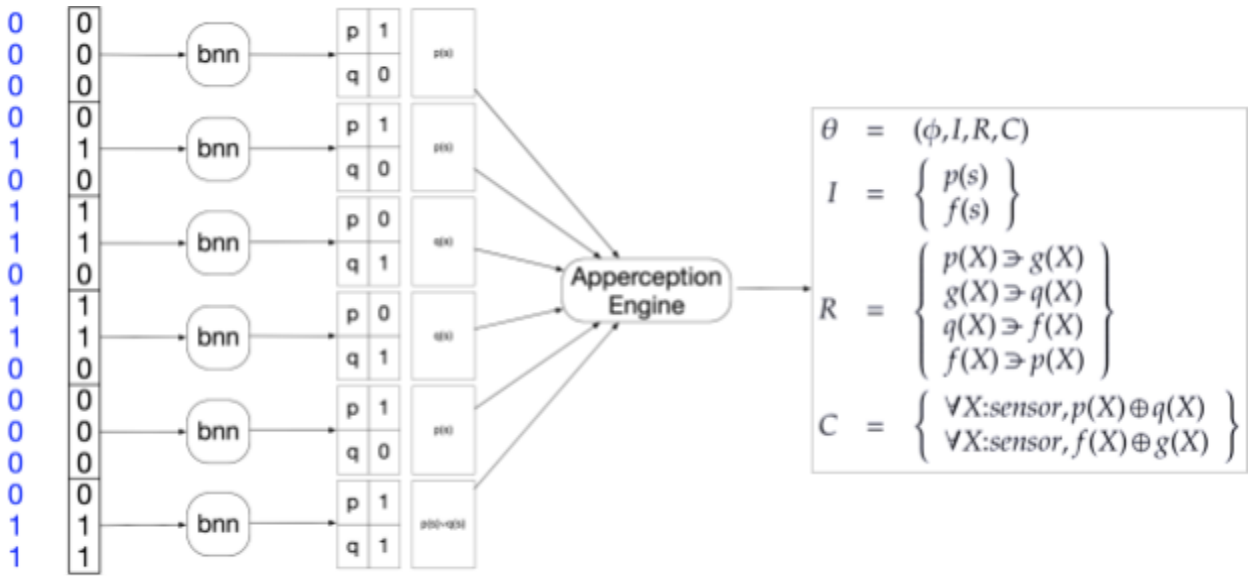

Figure 2. Making sense of the raw sequence involves mapping raw impressions into concepts, and constructing rules that explain and predict.

When looking at the output of this machine, it not only seems permissible to say that this machine believes that the object has switched from *p* to *q* – it is *unavoidable*: once you understand what the machine is doing, and how it is doing it, there is really no choice but to see its rules as representing beliefs about the world in which it finds itself. (There are, of course, lower-level descriptions of what is going on in the machine. But that is true of us humans, too.)

De re intentionality and deference

The aim of this paper is to show that a certain sort of self-legislating machine is capable of exhibiting original intentionality, achieving *de dicto* intentionality, *de re* intentionality, and deference to the world. So far I have argued that a self-legislating machine that is subject to the norms collected under the umbrella of the "synthetic unity of apperception" performs activities that count as various forms of *de dicto* intentionality: perceiving that a particular object has a particular quality, for example, or forming a belief.

So far, we have focused on the static case, where a machine at a particular fixed moment in time tries to make sense of a sequence of past events. Next, I shall move from this synchronic perspective to a diachronic perspective, in order to expand the constitutive account of *de dicto* intentionality into an account that also explains *de re* intentionality and deference. The following explanation of *de re* intentionality is based on aspects of Brandom's reading of Hegel in his remarkable work *A Spirit of Trust* (Brandom 2019).

The machine is given a sequence of impressions $x_1, \ldots, x_n$. On the *n*th time-step, it constructs a theory to make sense of that sequence. (Here, a "theory" is a set of cognitions that includes intuitions, determinations, subsumptions, and judgments, and "making sense" means that the theory satisfies the various unity conditions collected together under the umbrella of the synthetic unity of apperception.) Later, it receives *further* impressions $x_{n+1}, \ldots, x_m$.

Now it must make sense of the whole sequence $x_1, \ldots, x_n, x_{n+1}, \ldots, x_m$.

One fortunate possibility is that the original theory also makes sense of the new section $x_{n+1}, \ldots, x_m$ of the sequence. In this case, the machine does not have any additional work to do. It just applies the original theory to the longer sequence and is done.

Another less fortunate possibility, however, is that the new section $x_{n+1}, \ldots, x_m$ of the sequence reveals inadequacies in the original theory: the original theory cannot make sense of the latest part of the sequence. Now the machine has no choice but to construct a new theory to explain both the new section and also the old. It would not be ok for it to introduce a discontinuity here, and use one theory to explain $x_1, \ldots, x_n$, and a second distinct theory to explain $x_{n+1}, \ldots, x_m$. That would introduce a discontinuity (a rift, a schism) between the two time periods, and that would violate the unity of time. Rather, the machine must construct a new theory that explains the *entire* sequence $x_1, \ldots, x_n, x_{n+1}, \ldots, x_m$ that includes both the old sequence $x_1, \ldots, x_n$ and the new sequence $x_{n+1}, \ldots, x_m$ as subsequences.

Now when constructing a new theory to make sense of the entire sequence, the machine is free to re-use any of the materials from the old theory. It can re-use as much of the ontology (the objects, the types of the predicates) and the judgments (the rules) that it can. In some unusual cases of Kuhnian paradigm shift, the new subsequence forces the machine to construct a radically different ontology, and much of the previous ontology is discarded. But in more quotidian cases, much of the old theory can survive intact, and the machine merely needs to revise a small number of judgments. According to Brandom's reading of Hegel (Brandom 2019), such cases of belief revision provide the raw materials needed to construct an account of *de re* intentionality. Consider a simple example:

> A naïve subject looks at a stick half-submerged in the water of a pond and perceptually acquires a belief that the stick is bent. Upon pulling it out, she acquires the belief that it is straight. Throughout she has believed that it is rigid, and that removing it from the water won't change its shape. These judgments are jointly incompatible. (Brandom 2019: 76)

Brandom's reconstruction of Hegel sees the process of handling such cases as involving three separable stages. The first stage is acknowledgement: the agent acknowledges the incompatibility between the judgments; in Brandom's example, the subject acknowledges that "the stick is bent" is incompatible with "the stick is straight" (Brandom 2019: 77). The second stage is rectification: the subject constructs a new theory that explains the whole sequence of sensory impressions; in this particular example, she posits a theory involving the refraction of light to conclude that the stick was actually straight throughout the episode (Brandom 2019: 77). The third stage is recollection: the subject uses her current understanding of the world, the new theory she has just constructed, to retrospectively evaluate the truths of the judgments in the previous theories she had constructed; in this particular example, she uses her endorsement of "the stick is straight" to justify her rejection of "the stick is bent" (Brandom 2019: 681).

But what, if anything, does this three-stage process have to do with *de re* intentionality? Observe, following Wittgenstein (Wittgenstein 1953) that performing a certain activity can implicitly also count as performing a mental activity:

- By recoiling from the hot stove, I am implicitly counting it as hot; there doesn't have to be some antecedent mental event causing my arm to move: I just recoil my arm, and that movement *just is* my counting it as hot.
- By not answering the question, I am implicitly counting the question as unworthy of my consideration; there doesn't have to be some antecedent mental event, some prior feeling of contempt, causing me not to answer the question: I just refuse to answer the question, and my contempt for the question *just is* my not answering it.

In both cases, I *implicitly* count something as something by *doing* something. Brandom's version of Hegel uses this implicit counting-as phenomenon to re-characterize the three stages of belief revision: In the first stage, acknowledging the incompatibility between "the stick is bent" and "the stick is straight" is implicitly *treating the two beliefs as being about the same object*. In the second stage, adopting a new theory is implicitly *seeing the world as it really is*. In the third recollective stage, the subject looks back from the viewpoint of her new judgment (that the stick was straight all along) to re-evaluate her earlier judgment (that the stick was bent) as false; this looking back, this re-evaluation, is implicitly *deferring one's belief to the world that the belief is about*.

Our present task is to understand how an account of *de dicto* intentionality can be expanded to include an account of *de re* intentionality and deference to the world. In the first stage, acknowledging an incompatibility counts as implicitly treating two judgments as being co-referential (Brandom 2019: 76), and if two judgments refer to the *same* thing, then they refer to *some* thing. In the third stage, rejecting "the stick is bent" in favor of "the stick is straight" counts as implicitly deferring to the world: treating the reality of the straight stick as authoritative over my beliefs (Brandom 2019: 685).

Now the subject may not have the sophisticated meta-conceptual vocabulary to talk about co-referentiality, the distinction between appearance and reality, and deferring to the world. But nevertheless her various cognitive activities count as implicitly seeing the world in these terms: by re-evaluating her previous beliefs, she *is* deferring to the world, even if she would not put it that way herself (Brandom 2019: 79). She may not yet have the meta-conceptual vocabulary to make this explicit, but she already has all the behavioral capacities needed to understand that vocabulary if and when it is introduced (Brandom 2010).

Brandom's Hegel on determinate content in Kant

At a very high level, Hegel's response to Kant's cognitive architecture involves making two moves.The first move is historical: to expand Kant's account from the synchronic to the diachronic case. The second move is social: to replace Kant's individualistic self-legislating agent with multiple agents synthesizing a community through reciprocal recognition. In this paper, I have tried to reappropriate Hegel's diachronic model of recollection within a Kantian framework, while studiously avoiding Hegel's second move from the single- to the multi-agent case. This pick and mix approach is justified, I believe, because Hegel's first move is compatible with everything that Kant said and thought, while the second move is, I shall argue, a step in the wrong direction.

The first move, expanding our area of concern from the synchronic to the diachronic, is not merely *compatible* with Kant; it is something he was well aware of, even if he did not always focus on it. In *What is Enlightenment*? (Kant 1784), he is emphatic that the cognitive agent must never be satisfied with a statically defined set of rules - but must always be modifying existing rules and constructing new rules. He stresses that adhering to any statically-defined set of rules is a form of self-enslavement:

> Precepts and formulas, those mechanical instruments of a rational use, or rather misuse, of his natural endowments, are the ball and chain of an everlasting minority.

Later, he uses the term "machine" to describe a cognitive agent who is no longer open to modifications of his rule-set. He defines enlightenment as *the continual willingness to be open to new and improved sets of rules*. He imagines what would happen if we decided to fix on a particular set of rules, and forbid any future modifications or additions to that rule-set. He argues that this would be disastrous for society and also for the self. In *The Metaphysics of Morals*, he stresses that the business of constructing moral rules is an ongoing never-ending task: "virtue can never settle down in peace and quiet with its maxims adopted once and for all" (Kant 1797: 409).

Just as for moral rules, just so for cognitive rules; Kant's cognitive agent is always constructing new rules to make sense of a pattern which is new in every moment:

> There is no unity of self-consciousness or "transcendental unity of apperception" apart from this effort, or conatus towards judgment, *ceaselessly affirmed and ceaselessly threatened with dissolution* in the "welter of appearances" (Longuenesse 98: 394).

However, while Hegel's first move may be compatible with everything Kant said or thought, Hegel's second move (from the single- to the multi-agent case) is incompatible with Kant's fundamental premise that a single agent can institute norms on its own. Kant thinks that a single agent is capable of instituting norms on its own, while Brandom's Hegel believes that it is only when *multiple* agents recognise each other that determinate content can be achieved. Consider the following representative passage:

> [Kant's] model says that it is up to me whether I am committed—for instance, to the coin's being copper. But if the relations of material incompatibility and consequence that articulate the concept copper I have applied in undertaking the commitment are *also* up to me, then I have undertaken no determinate commitment at all. As Wittgenstein says: "If whatever is going to seem right to me is right, that only means that here we can't talk about 'right.'" [PI §258] Concepts with determinate contents serve as normative standards for assessing whether the subject who applies them has fulfilled the rational responsibilities undertaken thereby—has acknowledged incompatibilities and drawn appropriate conclusions. Hegel wants to know how it is that the subject has access to such determinately contentful normative standards. If they cannot be the products of the attitudes of the one who applies them in judgment, where do they come from? He does not find an adequate answer in Kant. (Brandom 2019: 701).

The idea here is that the "home language game" of commitment is the case where one agent commits to *another*. When I commit to another, that other may (or may not) later release me from that commitment. But what, according to Brandom's Hegel, does it even mean to release *myself* from a commitment? How can I distinguish, in my own case, between *failing* to honor a commitment I have adopted, and *releasing myself* from that commitment? If I adopt a rule at one moment, but discard it at the next moment, in what sense was I ever truly *committed* to it?

> But if the relations of material incompatibility and consequence that articulate the concept I have applied in undertaking the commitment are also up to me, then *I have undertaken no determinate commitment at all*. (Brandom 2019: 701).

Certainly, once Kant's cognitive architecture is extended to the diachronic case, things are more complex: the subject is no longer stuck with the rules she initially constructed, but can add and remove rules freely at any moment. But just because the situation is more complex does not mean that determinacy has been lost. The conditional quoted here is false: *even if* the relations of material incompatibility and consequence are up to me (and they are), there is *still* determinate commitment, for two reasons.

First, the rule, once it is adopted, *compels the subject to apply it to every moment of time that she constructs*. Note that there are *two* notions of time in play here: there are the moments of time *at which* the subject *constructs rules* (call these moments of external time), and there are the moments of time that the subject *constructs*, the moments at which she *applies* rules (call these moments of internal time). In the diachronic case, she can revise any rule, at any moment of external time. But once she has adopted a rule, she must apply it at every moment of internal constructed time.

The second reason why Kant's subject institutes determinate content is that, as well as the adopted rule that rigidly constrains behavior at the level *below*, there is also the constraint from *above*: she must adopt rules that, when applied, enable her intuitions to achieve unity. The synthetic unity of apperception is the "supreme principle" of the understanding [B136], ruling forever over the subject's spontaneity. It is true that no judgment is immune from the possibility of revision – the subject is free to throw away any belief at any moment (of external time) – but the supreme constraint of unity is always there, standing over her, insisting that her new rule set be sufficient to enable her to unify her intuitions. The Kantian subject does not suffer from "whatever seems right is right" because her spontaneity (her ability to construct, adopt and reject any rule she pleases) is sandwiched between the rigid bottom layer (where rules are rigidly applied[10]), and the rigid top layer (where the supreme principle of the synthetic unity of apperception insists that at every moment of external time, her ruleset achieves unity over every moment of internal constructed time).

If the gentle reader is unconvinced by these general claims, I urge her to read (Evans, 2021b) and (Evans 2022), to look and see, in practical cases of worked examples, how

[10] I am not saying that the subject *always does* obey the rules she has constructed, but simply that she *must*. We should not confuse the hardness of the logical must with some sort of super-strong exceptionless causal law (Wittgenstein 1953).

determinate conceptual content is instituted by a single agent constructing and applying rules according to Kant's architectural constraints.

It is a commonplace of pop psychology that the weakness you accuse your opponent of is the very same weakness that you are dimly aware you suffer from yourself. Indeed, when Brandom's Hegel accuses Kant's theory of having insufficient resources to institute determinate content, I wonder if that criticism can be applied, with more force, to Hegel himself. If an agent isn't the right sort of thing to institute determinate content on its own, it is not clear how adding more things of the same type could help. If "whatever seems right to *me* is right", then "whatever seems right to *us* is right". Brandom interprets Kant as operating merely at the conceptual level:

> For Kant, to be aware in the narrower sense is to synthesize a constellation of commitments that exhibits a distinctive kind of unity: apperceptive unity. This is a rational unity—and hence, he thinks, a discursive unity, in the sense of one that is conceptually articulated. (Brandom 2019: 678).

But this is to seriously mis-read what Kant is aiming at: the synthetic unity of apperception is first and foremost a constraint on *intuitions*, not concepts. The fundamental requirement is that the subject's intuitions are connected in a unity. Now it turns out that, for subjects like us (those that have distinct faculties of sensibility and understanding), the only way to achieve this unity is to form concepts and judgements. But these concepts and judgments are merely *a means to an end*[11]: concepts and judgments are merely the glue we use to bind our intuitions together.

Brandom's Hegel sees the subject as operating almost entirely at the discursive level. She constructs judgments out of concepts, and then, because concepts and judgments are not enough on their own to institute determinate content, she connects with other subjects via reciprocal recognition, and connects with other moments in time via recollection. Kant's subject, by contrast, achieves determinate content because her concepts are always connected, via the power of judgment, to the raw intuitions provided by sensibility that relate directly (but non-discursively) to the world.[12]

[11] See the earlier discussion, arguing that there is a striking asymmetry between intuitions and concepts.

[12] Thanks to Tom Smith for many illuminating discussions. Thanks also to Michiel van Lambalgen, Nicholas Shea, Christopher Peacocke, Robert Long, Marek Sergot, Rob Craven, Murray Shanahan, Jose Hernandez-Orallo, Sorin Baiasu, Dieter Schonecker, Konstantin Pollok, Andrew Stephenson, David Hyder, Barnaby Evans, Tara Pesman, and Arie Soteman for insightful comments and thoughtful feedback on this project.

References

Bender, E.M. et al. 2021. “On the Dangers of Stochastic Parrots”. *Proceedings of the ACM*, pp. 610-623.
Brandom, R.B., 2010. *Between Saying and Doing*. OUP Oxford.
Brandom, R.B., 2019. *A Spirit of Trust*. Harvard University Press.
Brown, T. et al. 2020. “Language Models are Few-shot Learners”. *Neurips Proceedings*, pp.1877-1901.
Evans, R, et al. 2021(a). “Making Sense of Sensory Input”. *Artificial Intelligence*, *293*, p.103438.
Evans, R, et al. 2021(b). Making Sense of Raw Input. *Artificial Intelligence*, *299*, p.103521.
Evans, R, 2022. “The Apperception Engine”. *Kant and Artificial Intelligence*. de Gruyter.
Haugeland, J., 1990. “The Intentionality All-stars”. *Philosophical Perspectives*, *4*, pp.383-427.
Jumper, J. et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold”. *Nature*, pp.583-589.
Kant, I. 1784. “What is Enlightenment?” *Practical Philosophy*, pp. 11–22. Cambridge University Press.
Kant, I. 1788. *Critique of Pure Reason*. Cambridge University Press.
Kant, I. 1790. *Critique of the Power of Judgment*. Cambridge University Press.
Kant, I. 1797. “The Metaphysics of Morals”. *Practical Philosophy*, Cambridge University Press.
Longuenesse, B. 1998. *Kant and the Capacity to Judge*. Princeton University Press.
Nilsson, N., 1993. “Teleo-reactive Programs for Agent Control”. *Journal of Artificial Intelligence Research*.
Pollok, K. 2017. *Kant's Theory of Normativity*. Cambridge University Press.
Shanahan, M. 2005. Perception as Abduction. *Cognitive Science*. pp.103-134.
Silver, D. et al. 2017. “Mastering the Game of Go Without Human Knowledge”. *Nature*, pp.354-359.
Smith, B.C. 2019. *The Promise of Artificial Intelligence: Reckoning and Judgment*. MIT Press.
Wittgenstein, L. 1953. *Philosophical Investigations*. John Wiley & Sons.