Richard Evans

2 The Apperception Engine

Abstract: This paper describes an attempt to repurpose Kant's a priori psychology as the architectural blueprint for a machine learning system. First, it describes the conditions that must be satisfied for the agent to achieve unity of experience: the intuitions must be connected, via binary relations, so as to satisfy various unity conditions. Second, it shows how the categories are derived within this model: the categories are pure unary predicates that are derived from the pure binary relations. Third, I describe how Kant's cognitive architecture has been implemented in a computer system (the Apperception Engine) and show in detail what it is like for the system to construct a unified experience from a sequence of raw sensory input.

1 Introduction

This paper describes an attempt to repurpose Kant's *a priori* psychology as the architectural blueprint for a machine learning system.

Imagine a machine, equipped with sensors, receiving a stream of sensory information. It must, somehow, *make sense* of this stream of sensory data. But what, exactly, does this involve? We have an intuitive understanding of what is involved in "making sense" of sensory data – but can we specify precisely what is involved? Can this intuitive notion be formalized?

In machine learning, this is called the *unsupervised learning problem*. It is both fundamentally important and frustratingly ill-defined. This problem contrasts with the supervised learning problem where the sensory data come attached with labels. In a supervised learning problem, there is a clear learning objective, and there are a number of powerful techniques that perform very successfully. However, *the real world does not come with labels attached to sensory data*. We just receive the data. As Geoffrey Hinton said:[1]

> When we're learning to see, nobody's telling us what the right answers are – we just look. Every so often, your mother says "that's a dog", but that's very little information. You'd be lucky if you got a few bits of information – even one bit per second – that way. The brain's visual system has 10^{14} neural connections. And you only live for 10^9 seconds. So

1 Quoted in Kevin Murphy's *Machine Learning: a Probabilistic Perspective* (Murphy, 2012).

Richard Evans, Imperial College London

https://doi.org/10.1515/9783110706611-002

> it's no use learning one bit per second. You need more like 10^5 bits per second. And there's only one place you can get that much information: from the input itself.

In unsupervised learning, we are given a sequence of sensor readings, and want to make sense of that sequence. The trouble is we don't have a clear formalisable understanding of what it means to "make sense". Our problem, here, is *inarticulacy*. It isn't that we have a well-defined quantifiable objective and do not know the best way to optimize for that objective. Rather, we do not know what it is we really want.

One approach, the *self-supervised* approach, is to treat the sensory sequence as the input to a prediction problem: given a sequence of sensory data from time steps 1 to t, maximize the probability of the next datum at time $t+1$. But I believe there is more to "making sense" than merely predicting future sensory readings. Predicting the future state of one's photoreceptors may be *part* of what is involved in making sense – but it is not on its own sufficient.

What, then, does it mean to make sense of a sensory sequence? In this paper, I argue that the solution to this problem has been hiding in plain sight for over two hundred years. In the *Critique of Pure Reason* (Kant, 1781), Kant defines exactly what it means to make sense of a sequence: to reinterpret that sequence as a *representation of an external world composed of objects, persisting over time, with attributes that change over time, according to general laws*.

In this paper, I reinterpret part of Kant's first *Critique* as a specification of a cognitive architecture, as a precise computationally-implementable description of what is involved in making sense of the sensory stream. This is an interdisciplinary project and as such is in ever-present danger of falling between two stools: neither philosophically faithful to Kant's intentions nor contributing meaningfully to AI research. Kant himself provides[2]

> the warning not to carry on at the same time two jobs which are very distinct in the way they are to be handled, for each of which a special talent is perhaps required, and the combination of which in one person produces only **bunglers** [AK 4:388]

The danger with an interdisciplinary project, part AI and part philosophy, is that both potential audiences are unsatisfied. The computer science might reasonably ask: why should a two hundred year old book have anything to teach us now? Surely if Kant had anything important to teach us, it would already have been absorbed? The Kant scholar might reasonably complain: is it really necessary to

2 Translations are from the Cambridge Edition of the Works of Immanuel Kant (details at the end), with occasional modifications. With the exception of those to the *Critique of Pure Reason*, which take the standard A/B format, references to Kant are by volume and page number in the Academy Edition [*Immanuel Kants gesammelte Schriften*, 29 volumes, Berlin: de Gruyter, 1902-].

re-express Kant's theory using a computational formalism? We do not need these technicalities to talk about Kant. At best, it is an unnecessary re-articulation. At worst, misunderstandings are piled on misunderstandings, as Kant's ideas are inevitably distorted when shoe-horned into a simple computational formalism.

Nevertheless, I will argue, first, that contemporary AI has something to learn from Kant, and second, that Kant scholarship has something to gain when rearticulated in the language of computer science.

1.1 AI has Something to Learn from Kant

It is increasingly acknowledged that the strengths and weaknesses of neural networks and logic-based learning are *complementary*. While neural networks[3] are robust to noisy or ambiguous data, and are able to absorb and compress the information from vast datasets, they are also data hungry, uninterpretable, and do not generalize well outside the training distribution (Fodor and Pylyshyn, 1988; Marcus, 2018a; Lake et al., 2017; Evans and Grefenstette, 2018). Logic based learning, by contrast, is very data efficient, produces interpretable models, and can generalise well outside the training distribution, but struggles with noisy or ambiguous data, and finds it hard to scale to large datasets (Rocktäschel and Riedel, 2016; Evans and Grefenstette, 2018).

What we would really like, if only we can get it, is a system that combines the advantages of both. But this is, of course, much easier said than done. What, exactly, is involved in combining low-level perception with high-level conceptual thinking?

In the first *Critique* Kant describes, in remarkable detail, exactly what this hybrid architecture should look like. The reason why he was interested in hybrid cognitive architectures is because he was attempting to synthesise the two conflicting philosophical schools of the day, empiricism and rationalism. The neural network is the intellectual ancestor of empiricism, just as logic-based learning is the intellectual ancestor of rationalism. Kant's unification of empiricism and rationalism is a cognitive architecture that attempts to combine the best of both worlds, and points the way to a hybrid architecture that combines the best of neural networks and logic-based approaches.[4]

3 An introduction to neural networks is beyond the scope of this paper and we refer to Murphy (2012).

4 So far, so programmatic. The hybrid neuro-symbolic architecture is outlined in Section 3 and described in detail in (Evans et al., 2021a), and the ascription of this architecture to Kant in particular is justified in Section 2.

1.2 Kant Interpretation has Something to Learn from AI

Some of the most exciting and ambitious work in recent philosophy (Brandom, 1994, 2009, 2008; Sellars, 1967, 1968, 1978) attempts to re-articulate Kantian (and post-Kantian) philosophy in the language of analytic philosophy. Now this re-articulation is not merely window-dressing: it is not merely dressing up old ideas in the latest fashionable terminology, but rather an attempt to achieve a new level of perspicuity in a semi-formal language that was designed for clarity and precisiom.

My aim in this paper is to re-articulate Kant's theory at a further level of precision, by reinterpreting it as a specification of a *computational architecture*. Why descend to this particular level of description? What could possibly be gained? The computational level of description is the ultimate level of precise description. There is no more precise you can be: even a mere *computer* can understand a computer program. Computers force us to clarify our thoughts. They admit no waffling or vagueness. Hand-waving is greeted with a compilation error, and a promissory note is returned, unread.

The advantage of re-articulating Kant's vision in computational terms is that it gives us a new level of specificity. The danger is that, in an effort to shoe-horn Kant's theory into a particular implementable system, we distort his original ideas to the point where they are no longer recognisable. Whether this is indeed the unfortunate consequence, the gentle reader must decide.

1.3 Kant's Cognitive Architecture

The first half of the *Critique of Pure Reason* is a sustained exercise in *a priori* psychology: the study of the processes that must be performed if an agent is to achieve experience. For Kant, this *a priori* psychology was largely a means to an end – or, to be precise, two ends. One of his high-level goals was metaphysical: to enumerate once and for all the pure aspects of cognition – those features of cognition that must be in place no matter what sensory input has been received. The pure aspects of cognition include the pure forms of intuition (space and time, as described in the *Aesthetic*), the pure concepts (the categories, as described in the *Analytic of Concepts*), and the pure judgements (the synthetic *a priori* propositions, as described in the *Principles*). His other high-level goal was metaphilosophical: to delimit the bounds of sense, and finally put to rest

various interminable disputes,[5] by showing that the pure concepts can only be applied to objects of possible experience.

But I believe that, apart from its role as a means to his metaphysical and metaphilosophical ends, Kant's peculiar brand of psychology has independent interest in its own right, *as a specification of a cognitive architecture*. According to Kant's specification, making sense of a sensory sequence involves constructing a symbolic causal theory that explains the sensory sequence and satisfies a set of unity conditions. According to our interpretation, making sense of sensory input is a type of program synthesis, but it is *unsupervised* program synthesis, constrained in such a way as to achieve the *synthetic unity of apperception*.

To test this hypothesis, we need to implement this architecture in a computer program, and test it on a wide array of examples. Kant's theory is intended to be a general theory of what is involved in achieving experience, so – if it actually works – it should apply to *any* sensory input. To test the viability of this architecture, then, we need to actually implement it, and evaluate it in a large and diverse set of experiments.

Our computer implementation of Kant's cognitive architecture is called the APPERCEPTION ENGINE.[6] Our system is able to produce interpretable human-readable causal theories from very small amounts of data because of the strong inductive bias provided by Kant's unity constraints. We have tested this system in a variety of experiments, and found it shows promise as a machine for making sense of unlabelled sensory input.

In this paper, I shall first (Section 2) extract some core theses from the first half of the *Critique*, and assemble them into a specification of a cognitive architecture. Next (Section 3), I describe some examples of the APPERCEPTION ENGINE in action. I show one worked example in detail.[7] Finally (Section 4), I discuss the various interpretive decisions that were made, and defend them against alternatives. One of the things that makes a computational implementation challenging is that it forces one to pick a specific interpretation of Kant, since the computer has zero tolerance for vagueness or equivocation.

5 He wanted to "put an end to all dispute" [A768/B796].

6 The APPERCEPTION ENGINE is described in detail in (Evans et al., 2021b,a; Evans, 2020). The source code is available at https://github.com/RichardEvans/apperception.

7 For the various other experiments, see (Evans et al., 2020) and (Evans et al., 2021a).

2 Achieving Experience

In the first half of the *Critique of Pure Reason*, Kant focuses on the following fundamental question:

> What activities must be performed if the agent is to achieve **experience**?[8]

Note that this is a question about intentionality – not about knowledge. Kant's question is very different from the standard epistemological question:

> Given a belief, what else has to be true of the agent for us to count that belief as knowledge?

Kant's question is *pre-epistemological*: he does not assume the agent is given a belief. Instead, we see his belief as an *achievement* that cannot be taken for granted, but has to be *explained*:

> Understanding belongs to all experience and its possibility, and the first thing that it does for this is not to make the representation of the objects distinct, *but rather to make the representation of an object possible at all* [A199, B244-5]

Kant asks for the conditions that must be satisfied for the agent to have any possible cognition (true *or* false) [A158, B197]. Note that this is not an empirical psychological question about the processes that *human beings* happen to use, but rather a question of *a priori* psychology:[9] what must a system – any physically realised system at all[10] – do in order to achieve experience?[11]

In this paper, I will try to distill Kant's answer to this fundamental question, and reinterpret his answer as the specification of a cognitive architecture.

8 The subtitle of the *Transcendental Deduction* in the First Edition is: "On the *a priori* grounds for the possibility of experience." [A95].

9 In this project, I side with Longuenesse (Longuenesse, 1998), Waxman (Waxman, 2014), and others in interpreting the first half of the *Critique* as *a priori* psychology. *Contra* Strawson (Strawson, 2018), I believe that *a priori* psychology is a legitimate and important form of inquiry, and that if we try to expunge it from Kant's text, there is not much left that is intelligible.

10 There are a number of places in the *Critique* where Kant seems to restrict his inquiry to just humans e.g., [B138-9]. But Kant uses the term "human" to refer to any agent who perceives the world in terms of space and time and has two distinct faculties of sensibility and understanding. This is a much broader characterisation than just *homo sapiens*.

11 Because the second question is broader, it is more relevant to the project of artificial intelligence (Dennett, 1978).

2.1 Achieving Experience by Unifying Intuitions

A central claim of the *Transcendental Deduction* is that:

(1) In order to achieve experience, I must unify my intuitions. [A110]

Before we can assess the truth of such a claim, we first need to understand what it means. (i) What does Kant mean by an experience? (ii) What are intuitions? (iii) What does it mean to unify them? I shall consider each in turn.

2.1.1 What does Kant mean by 'Experience'?

Kant's notion of **experience** ('Erfahrung') is close to our usual use of the term. I shall list some features of this term as Kant uses it. First, experience is *everyday*. It is not an unusual peak state that people only achieve occasionally, like enlightenment or ecstasy. Rather, it is a state that most of us have most of the time when we are awake. Second, experience is *unified*. At any one time, I am having *one* experience [A110]. I cannot have multiple simultaneous experiences. I may be conscious of multiple stimuli, but they are all part of one experience. Third, experience is *articulated* (Stephenson, 2013). It is not a mere 'blooming, buzzing confusion' (James et al., 1890). Rather, experience is composed of distinct objects with distinct properties. Fourth, experience is *not (merely) conceptual*. It is not just a collection of beliefs. It is, to anticipate, a unified combination of intuitions and concepts. Fifth, experience is *not necessarily veridical*. It purports to represent the world accurately, but may fail to do so (Longuenesse, 1998; Stephenson, 2013; Waxman, 2014).

Experience is not something we should take for granted. Rather, experience is *an achievement*. When I open my eyes, I see various objects, with various properties that change over time. But this experience is a complex achievement that only occurs if a myriad of underlying processes work exactly as they should do. The central contribution of Kant's *a priori* psychology is to describe in detail the underlying processes needed in order for experience to be achieved.

2.1.2 What does Kant mean by 'Intuition'?

An **intuition** ('Anschauung') is a representation of a particular object[12] (e.g., this particular jumper) or a representation of a particular attribute[13] of a particular object at a particular time (e.g., the particular dirtiness of this particular jumper at this particular time).

Intuitions are produced by the faculty of **sensibility** [A19/B33]: the receptive faculty that detects sensory input. Sensibility provides the agent with a *plurality* of intuitions [B68], which the mind needs to make sense of.

Intuitions are private to the individual. My intuitions are different from yours. It is not just that we do not share intuitions we *cannot* share intuitions, as they are essentially private. To see this, consider four possible relations between an action and its object:

1. The object existed before and after the action (e.g., kicking the football).
2. The object existed before but not after the action (e.g., destroying the evidence).
3. The object existed after but not before the action (e.g., making a cake).
4. The object existed neither before nor after, but only during the action, because the object is only an *aspect* of the action

Let us focus on the fourth. When I draw a circle in the air, this thing – the circle – only exists for the duration of the activity because it is an *aspect* of the activity. Or consider "the contempt in his voice ": this thing, this contempt, only exists for the duration of his vocal utterance because it is an aspect of the utterance.

The way I read Kant, the object of intuition is a type (4) object: it only exists as part of the act because it is an aspect of the act.[14]

---

12 [B76].

13 A186/B229: "The determinations of a substance that are nothing other than *particular ways for it to exist* are called accidents." Note that whenever Kant talks about "existence" in the Analogies, he is really talking about a particular *way of existing*. See e.g., A160/B199: "synthesis is either mathematical or dynamical: for it pertains partly merely to the intuition, partly to the existence of an appearance in general". Here, "the existence of an appearance" means the particular way of existing of an appearance (e.g., the particular dirtiness of this particular jumper).

14 Kant interpreters differ on whether intuitions are relations between conscious minds and actual existing material objects (Allais, 2009;Gomes, 2013; McLear, 2016), or whether the object of an intuition is just a mental representation that in no way implies the existence of a corresponding external physical object (Longuenesse, 1998; Stephenson, 2015, 2017).The interpretation in this project fits squarely within the latter, representational interpretation. My reason for preferring the representational interpretation is based on a general interpretive

> But in order to ognize something in space, e.g., a line, I must **draw** it. [B137]

Now because intuiting is a private mental act (no other agent can perform the same token-identical act), and because the object of intuition is a type (4) object that only exists as an aspect of the act, it follows that the object of intuition inherits the privacy of the intuiting act of which it is an aspect. Nobody else can have my particular object of intuition because this object is an aspect of my activity of intuition, and nobody else can perform this particular activity.

Intuitions are distinct from concepts. While an intuition is a representation of a particular object, a concept is a general representation that many intuitions fall under [B377]. For Kant, intuitions and concepts are distinct types of representation. While empiricists saw concepts as a special type of intuition that is used in a general way, and while rationalists saw intuitions as a special type of concept that is maximally specific, Kant understood intuitions and concepts to be entirely distinct *sui-generis* types of representation. His reasons for thinking intuitions and concepts are entirely distinct are: (i) they come from distinct faculties (sensibility and understanding respectively); (ii) while intuitions are private to an individual, concepts can be shared between individuals; (iii) while intuitions are immediately directed to an object (the particular object only exists as an aspect of the activity of intuiting, just as the circle only exists as an aspect of the activity of drawing a circle in the air), concepts are only mediately related to objects via intuitions [A68/B93, B377].

The intuition occupies a unique place in Kant's *a priori* psychology: it is the ultimate goal of all thought,[15] the final end that all cognition is aiming at. All the other aspects of thought (e.g. concepts and judgements) are only needed in so far as they help to unify the intuitions:

---

prejudice: whenever there are two ways of reading Kant, and one of those interpretations relies on fewer prior capacities, thus requiring the mind to do more work to achieve the coherent representation of an external world that we take for granted in our everyday life, then prefer that interpretation. The relational view takes for granted a certain type of cognitive achievement: the ability of the mind to be about an external object. The representational view, by contrast, sees this intentionality, this mind-directedness, as something that requires work to be achieved. Thus, simply because it is more demanding and asks harder questions, it should be preferred. Further, and not coincidentally, the representational view can be implemented in a computer program, while it is entirely unclear how we could begin to implement any relational view that takes for granted the ability for the mind's thoughts to be directed to particular external physical objects.

15 In this paper I focus on Kant's theoretical philosophy rather than his practical philosophy, and thus "thought" here means cognitive thought aimed at making sense of the world – rather than feelings, volitions, intentions, etc.

> 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.]

2.1.3 What does Kant mean by 'Unifying' Intuition?

Recall Kant's key claim that:

(1) In order to achieve experience, I must unify my intuitions.

Here, the *explanandum* is a mental state (experience), while the *explanans* is a process (the process of unifying the intuitions). But what, exactly, does this process involve, and how will we know when it is finished?

The process of unifying intuitions can be unpacked as a particular type of synthesising process that satisfies a particular constraint, the constraint 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]

I shall first consider the synthesising process in general, and then turn to the unity constraint. The activity of synthesis may seem frustratingly metaphorical or ill-defined:

> The inadequacies of such locutions as "holding together" and "connecting" are obvious, and need little comment. Perceptions do not move past the mind like parts on a conveyor belt, waiting to be picked off and fitted into a finished product. There is no workshop where a busy ego can put together the bits and snatches of sensory experience, hooking a color to a hardness, and balancing the two atop a shape. (Wolff, 1963, p. 126)

What exactly does it mean to unify intuitions? What is the glue that binds the intuitions together? As I read Kant, the only thing that can bind intuitions together is the *binary relation*.[16] Synthesising intuitions means connecting the intuitions together using binary relations so that the resulting undirected graph is fully connected.[17] The synthesising process is the job of the faculty of **productive imagination**[18] [A78/B103; A188/B230], described in Section 2.2 and formalized in Section 2.4.

---

16 The precise binary relations involved are listed in the *Schematism* and described in detail in Section 2.2.

17 A graph is fully connected if there is a path of (undirected) edges between any two nodes. See West et al. (2001) for an introduction to graph theory.

18 Kant distinguished between the productive and reproductive imagination [A100-2]. Here, I focus exclusively on the productive imagination. The reproductive imagination's job is to recall

But there is much – much more – to unifying intuitions than just connecting them together with binary relations. The extra requirement that must be satisfied for a connected binary graph to count as a unification of intuitions is that the graph satisfies Kant's *unity conditions*. While there are many ways to connect intuitions together via binary relations to form a connected graph, only a small subset of these satisfy the various conditions of unity that Kant imposes. These unity conditions are satisfied by the faculty of **understanding** [A79/B104], and are described in detail in Sections 2.3.1, 2.3.2, 2.3.3, and 2.3.4.

The second claim, then, unpacks what it means to unify intuitions:

(2) Unifying intuitions means combining them using binary relations to form a connected graph, in such a way as to satisfy various unity conditions (described in detail in Sections 2.3.1, 2.3.2, and 2.3.4).

2.1.4 The Status of Claim 1

Claim (1), then, is the claim that an agent can only achieve experience – everyday conscious experience of a single articulated world – if it can unify its intuitions by connecting them together in a relational graph that satisfies various (as yet unspecified) unity conditions.

Let us break this down into two claims:

(1a) In order to achieve experience, my intuitions must be unified.
(1b) In order for my intuitions to be unified, I[19] must unify them.

Claim (1a) can be interpreted with at least two levels of strength. A strong interpretation treats the claim as *definitional*: experience *just is* unified intuition. A weaker interpretation sees the claim as merely a necessary condition: experience *requires* unified intuition, but it also needs more besides. In this project, I

earlier determinations and reproduce them. This capacity is taken for granted in the current implementation: I assume the whole sequence of sensory input has been given as a whole, so the agent does not need to recall earlier elements.

19 I do not, of course, mean that the agent deliberately and consciously performs various activities that result in the intuitions being unified. Rather, I mean that various *sub-personal* processes within the agent must occur in order for there to be a unified person at all.

adopt the stronger interpretation, and there is reason to think that Kant endorsed this stronger interpretation too.[20]

The second claim (1b) is not entirely trivial. An alternative possibility is that my intuitions arrive, via the faculty of sensibility, *already unified*. But Kant clearly rules out this alternative.[21] So, then, if my intuitions do not arrive already unified, and if I cannot pay or persuade somebody else to unify them for me,[22] then I must unify them myself. This is a task that only I can do.

2.2 Synthesis

In this section, I describe the relations that are used by the imagination to connect the intuitions together [A78/B103].

When Kant talks about pure synthesis [A78/B104], he means connecting intuitions by **pure** relations[23] that apply to all intuitions in all situations.[24] Why does Kant insist that synthesis can only use pure relations to connect intuitions? Because the unity conditions (that will be described in Sections 2.3.1, 2.3.2, and 2.3.4) are conditions that must apply to *every possible synthesis* of intuitions. If the unity conditions are to apply to every possible synthesis, they can only reference relations that feature in every possible synthesis, and these are the pure relations.

There are three[25] operations that bind intuitions together:

- **containment**: in(X, Y) means that object X is (currently) in object Y (e.g., the package is in the kitchen)
- **comparison**: X<Y means that attribute X is (currently) less than attribute Y (e.g., the weight of the package is less than the weight of the spoon)
- **inherence**: *det*(X, Y) means that attribute Y (currently) inheres in object X (e.g., this particular heaviness (of 2.3 kg) is an attribute of this particular parcel)

---

20 "[Experience] is therefore a synthesis of perceptions." [A176/B218] "There is only one experience, in which all perceptions are represented as in thoroughgoing and lawlike connection."[A110].

21 "Yet the **combination** (*conjunctio*) of a manifold in general can never come to us through the senses, and therefore cannot already be contained in the pure form of sensible intuition." [B129].

22 Nobody else can get anywhere near my intuitions because they are aspects of my private mental acts. See Section 2.1.2.

23 Pure relations are opposed to *impure* relations, such as *father-of*, that only apply contingently.

24 Kant enumerates the pure relations in the *Schematism*.

25 The containment operation is described in the *Axioms of Intuition*, the comparison operation in the *Anticipations of Perception*, and the inherence operation in the *First Analogy*.

When two intuitions are bound together by one of the three operations, the result is a **determination**. Thus, in (a, b), a < b, and *det*(a, b) are all determinations. Determinations hold at a particular moment or moments in time; they do not persist indefinitely [A183-4,B227].

The constituents of determinations are intuitions, representations of individuals; these are either particular objects, or particular attributes of those objects. To hold *det*(a,b) is to ascribe particular attribute b to particular object a (for example, to ascribe this particular dirtiness to this particular jumper).

It is absolutely essential, I believe, for understanding Kant's architecture that we distinguish clearly between attributes and concepts. Attributes are a type of intuition representing the particular way in which a particular object exists at a particular moment. Concepts, by contrast, are general representations. A number of different attributes typically fall under the same concept. Consider, for example, the particular dirtiness of this particular jumper, and the particular dirtiness of this particular laptop. Both attributes fall under the concept "dirty", but they are nevertheless distinct attributes: this jumper's particular dirtiness is different in myriad subtle ways from the dirtiness of my laptop.

Just as an attribute is a different kind of representation from a concept, just so a determination is a different kind of thought from a judgement. Seeing the particular dirtiness of the particular jumper at this particular moment (a determination) is very different from believing that the particular jumper is dirty (a judgement). In the former, I notice an individual property of an individual object. In the latter, I subsume a concept representing an individual object (the particular jumper) under a general concept ("dirty").

A determination is not a judgement, but a *way of perceiving*: I *see* the baby in the cot (containment); I *feel* the cup being heavier than the spoon (comparison); I *hear* the contemptuousness of the utterance (inherence). In each case, the argument of the perceptual verb is a *noun-phrase*, not a that-clause (Sellars, 1978).

Since a determination is a way of perceiving, it does not have a truth-value:

> For truth and illusion are not in the object insofar as it is intuited, but in the judgment about it insofar as it is thought. Thus it is correctly said that the senses do not err; yet not because they always judge correctly, but because they do not judge at all. Hence truth, as much as error, and thus also illusion as leading to the latter, are to be found only in judgments, i.e., only in the relation of the object to our understanding . . . In the senses there is no judgment at all, neither a true nor a false one. [A293-4/B350] See also
>
> [*Jäsche Logic* 9:53]

As well as the three pure operations that bind intuitions together, there are three[26] pure relations that bind determinations together:

- **succession**: $succ(P_1, P_2)$ means that P_1 is succeeded (at the next time-step) by P_2
- **simultaneity**: $sim(P_1, P_2)$ means that P_1 occurs at the same moment as P_2
- **incompatibility**: $inc(P_1, P_2)$ means that P_1 and P_2 are incompatible

When two determinations are bound together by one of the three relations, the result is a **connection.**[27] Thus, $succ(in(a, b), in(a, c))$ means that a's being in b is succeeded by a's being in c, and $inc(det(a, b), det(a, c))$ means that attributing b to a is incompatible with attributing c to a.

2.2.1 The Justification for this Particular Set of Operations and Relations

Why these particular pure relations? What makes this particular list special? The justification for this list is that the three pure operations and the three pure relations together constitute a *minimal set of binary operators that together are sufficient to construct the forms of space and time* [A145/B184ff].[28]

According to Kant, intuitions and determinations do not arrive with space and time coordinates attached [B129]. The job of sensibility is just to provide us with intuitions, but not to arrange them in objective space/time. It is the function of *synthesis*, the job of the imagination, to connect the intuitions together, using the pure operations and relations described above, so as to construct the objective spatio-temporal form:

> since time itself cannot be perceived, the determination of the existence of objects in time can only come about through their combination in time in general, hence only through *a priori* connecting concepts. [A176/B219]

To see that sensibility does not provide us with objects of intuition that are already positioned in space and time, consider a robot with a camera that provides a two-dimensional array of pixels for each visual snapshot. The robot receives information about the location of each pixel in egocentric two-dimensional space, and it must determine the positions of objects in *three-*

26 The succession and simultaneity relations are described in the second and third *Analogies*, and incompatibility is discussed in the *Postulates of Empirical Thought*.

27 "Experience is possible only through the representation of a necessary connection of perceptions." [B218].

28 This claim holds for a suitably qualified minimal notion of space. See Section 2.3.1.

dimensional space. Suppose a yellow pixel is left of a red pixel. Does the yellow pixel represent an object that is in front of the object represented by the red pixel, or behind? The visual input does not provide this information – the robot must decide itself. Next, consider time. Suppose the robot receives a sequence of visual impressions as its camera surveys the various parts of a large house [B162]. Do these subjectively successive impressions count as various representations of one moment in objective time, or do they represent different moments of objective time? The sensory input arrives ordered in subjective space/time but not in objective space/time.[29] In order to place our intuitions in objective space/time, the imagination needs to connect them together using the pure relations described above.[30]

The three pure operations together with the three pure relations constitute a minimal set that is sufficient for generating the form of objective space/time. The containment operation *in* allows us to combine intuitions into a spatial field (a minimal representation of space that abstracts from the number of dimensions (Waxman, 2014)) [A162/B203ff]. The comparison operation $<$ allows us to compare two different attributes; if we generate an intermediate attribute between two comparable attributes, we can generate an intermediate moment in time between two observed moments [A165/B208ff], thus filling time [A145/B184]. The inherence operation allows us to ascribe different attributions to an object at different times. The simultaneity and succession relations allow us to order determinations in time. Finally, the incompatibility relation allows us to test when sets of determinations are compossible.

> Now one sees from all this that the schema of each category contains and makes representable: in the case of magnitude, the generation (synthesis) of time itself, in the successive apprehension of an object; in the case of the schema of quality, the synthesis of sensation (perception) with the representation of time, or the filling of time; in the case of the schema of relation, the relation of the perceptions among themselves to all time (i.e., in accordance with a rule of time-determination); finally, in the schema of modality and its categories, time itself, as the correlate of the determination of whether and how an object belongs to time. The schemata are therefore nothing but *a priori* **time-determinations** in accordance with rules, and these concern, according to the order of the categories, the **time-series**, the **content of time**, the **order of time**, and finally the **sum total of time** in regard to all possible objects. [A145/B184ff]

29 Kant makes this claim many times in the *Principles*. See [A181/B225], [A183/B226], etc.

30 In (Waxman, 2014) Chapter 3, Wayne Waxman makes a powerful case that intuitions do not arrive from sensibility already unified. They arrive as a mere multitude, and it is the job of the imagination to unify them in space/time. In other words, what the empiricist takes as "given" (the unified field of sensory input) is not actually "given" but rather has to be *achieved* by a mental process.

The third key claim, then, is:

(3) Synthesis involves (i) connecting intuitions together via containment, comparison, and inherence operations to form determinations; and (ii) connecting determinations together via succession, simultaneity, and incompatibility relations.

2.3 The Unity Conditions

So far, I have described how intuitions are connected together using the various pure binary relations. But there is more – much more – to synthetic unity than mere connectedness of intuitions. In this section, I describe the four types of unity condition that Kant imposes.[31]

(4) There are, in total, four types of unity condition in Kant's system: (i) the unity conditions for the synthesis of mathematical relations, (ii) the unity conditions for the synthesis of dynamical relations, (iii) the requirement that the judgements are underwritten by determinations, and (iv) the conceptual unity condition.

I shall go through each in turn.

2.3.1 The Unity Conditions for the Synthesis of Mathematical Relations

Kant divides the pure relations into two groups: the **mathematical relations** (containment and comparison) and the **dynamical relations** (inherence, succession, simultaneity, and incompatibility). The mathematical relations control the arbitrary synthesis of homogeneous elements,[32] while the dynamical relations control the necessary synthesis of heterogeneous elements[33] [B201n].

Kant says that the mathematical relations combine "what does not necessarily belong to each other" while the dynamical relations combine what "necessarily

31 There are many ways to connect intuitions together via binary relations to form a connected graph. If there are n nodes, then there are $2^{\binom{n}{2}}$ simple undirected graphs. The number of simple connected graphs for n nodes is the integer sequence $A001187$ which starts $1, 1, 1, 4, 38, 728, 26704, 1866256, \ldots$. See http://oeis.org/A001187. But only a small fraction of these satisfy the various unity conditions that Kant imposes.

32 Observe that *in* relates two objects of intuition, while $<$ relates two intuition attributes.

33 Observe that *det* relates two different types of intuition, an attribute and an object.

belongs to one another" [B201n]. This means that the agent has freedom to synthesise using containment and comparison in a way that is unconstrained by the conceptual realm of the understanding, but the synthesis using the dynamical categories is constrained by judgements produced by the understanding.[34]

I shall start with the unity conditions for the mathematical relations, before moving to the unity conditions on the dynamical relations. The fundamental unity condition for the mathematical relations is that the intuitions are combined in a fully connected graph. There are two further specific conditions, one for containment and one for comparison.

The unity condition for containment requires that there is some object, the maximal container, which contains all objects at all times [A25/B39]. Slightly more formally, the first unity condition for the synthesis of mathematical relations is:

---

(5)(a) There exists some intuition x such that for each object of intuition y, for each moment in time, there is a chain of *in* determinations between y and x.

Of course, objects can move about, from one container to another, but at every moment, the objects must always be contained in the maximal container.

Satisfying this unity condition means positing both pure objects (spatial regions with a mereological structure) and also impure objects (appearances) which are in the spatial regions.

Once objects have been placed in the containment hierarchy, and once we know which intuitions fall under which concepts, then we have all the information we need for *counting*. In order to count how many pens are in the box, I need to be able to tell whether each object falls under the concept "pen", and I also need to be able to tell which objects are actually in the box and which are outside. Thus, as Kant says, the pure schema of magnitude is "number, which is a representation that summarizes the successive addition of one (homogeneous) unit to another" [A142/B182]. The appearances are homogenous since they fall under the same concept, and we know which appearances to count and which to ignore by choosing a particular container in the containment hierarchy.

Now this containment hierarchy is a *necessary aspect* of any spatial representation: if we fix the positions and extensions of objects in 3D space, then the

---

34 See also [B110]: "the first class (mathematical categories) has no correlates which are to be met with only in the second class". Here, the correlates are the judgements that are required to underwrite the dynamical connections, but that are not required to underwrite the mathematical compositions.

containment hierarchy is also fixed. But, of course, the converse does not hold: specifying the containment hierarchy does not determine all the spatial information. Suppose, for example, that *x* and *y* are both in container *z*. We know that *x* and *y* are in the same container, but we do not know if *x* is above *y*, or below it. We do not know how near *x* is to *y*, etc.

The containment hierarchy is a distinguished sub-structure of the spatial world. If we abstract from our spatial representation all the aspects that are peculiar to our human form of intuition, all that is left is the containment hierarchy. As Kant says:

> Thus if, e.g., I make the empirical intuition of a house into perception through apprehension of its manifold, my ground is the necessary unity of space and of outer sensible intuition in general, and I as it were draw its shape in agreement with this synthetic unity of the manifold in space. This very same synthetic unity, however, *if I abstract from the form of space, has its seat in the understanding, and is the category of the synthesis of the homogeneous in an intuition* in general, i.e., the category of quantity, with which that synthesis of apprehension, i.e., the perception, must therefore be in thoroughgoing agreement. [B162]

And again:

> The pure *image* of all magnitudes (quantorum) for outer sense is space . . . The pure *schema* of magnitude (quantitatis), however, as a concept of the understanding, is number. [A142/B182]

Of course, a spatial representation performs many functions. It allows us, for example, to position and orient the parts of our bodies to manipulate other objects. But the function of space that is highlighted in the First Critique is space as the *medium in which appearances are unified*. Now space-qua-unifier-of-intuitions has fewer essential properties than space-qua-form-of-human-outer-sense. Qua unifier of intuitions, the key property of space is that it supports a containment hierarchy, in which we can tell which objects are in which containers. Kant makes it clear, when he first introduces space in the *Aesthetic*, that the function of space that he is focusing on is its ability to support the containment hierarchy:

> For in order for certain sensations to be related to something outside me (i.e., to something in another place in space from that in which I find myself), thus in order for me to represent them as outside one another, thus not merely as different but as in different places, the representation of space must already be their ground) [A23/B38]

Space, qua unifier, is just the medium in which appearance can be placed together, the medium that allows me to infer from "I am intuiting *x*" and "I am intuiting y" to "I am intuiting x and y". This abstract unifying space just is the

containment hierarchy: "space is the representation of coexistence (juxtaposition)" [A374].

To summarize, although Kant's notion of space was the standard (at the time) three-dimensional space of Euclidean geometry (B41), when he was thinking of space as the medium in which appearances can be unified, he focused on a substructure in which many of the features of space have been abstracted away: the containment hierarchy.[35]

The unity condition for comparison[36] simply requires that:

(5)(b) The comparison operator < forms a strict partial order.

Of course, we do not insist that < is a *total* order: although the dirtiness of this jumper can be compared with the dirtiness of this mug, the *weight* of this jumper need not be comparable with the dirtiness of this mug.

We do not, also, insist that < is *dense*.[37] This is because we follow Kant in wanting to allow *finite* models.[38]

2.3.2 The Unity Conditions for the Synthesis of Dynamical Relations

I have described above the unity conditions for the synthesis of mathematical relations (containment and comparison). Next we turn to the conditions Kant imposes on the synthesis of *dynamical* relations (inherence, succession, simultaneity, and incompatibility). This is perhaps the most important, the most original, and the most difficult part of the *Transcendental Analytic*. In fact, one of the major reasons that Kant rewrote the *Transcendental Deduction* in the B edition is precisely to re-express this condition as clearly as possible. In this

35 For a related position, see Waxman (Waxman, 2014) Section 4B: "It as if the mere use of the word 'space' is enough for many to reflexively read into Kant's doctrine virtually every meaning commonly attached to the term, or at least everything one supposes to remain after factoring in the adjective 'pure'. It becomes a space with all the features attributed to it by Euclid or Newton and so a space a priori incompatible with the features that have been or will be ascribed to space by later mathematicians and physicists. But . . . the unity of sensibility clearly does not require that pure space be determinately flat hyperbolic or elliptical, three-dimensional or ten-dimensional or any other number of dimensions, Ricci-flat or Ricci-curved, etc".

36 See [A143/B182-3] and [A168/B210].

37 A relation R is dense if Rxy implies there exists a z such that Rxz and Rzy.

38 (Pinosio, 2017) page 119.

section, I shall first explain Kant's general strategy before going into the specific details of how he handles each of the pure dynamical relations.

Kant was dissatisfied with the presentation of the Transcendental Deduction in the A edition. In the B edition, he changed the exposition significantly by splitting the proof into two parts (concluding in § 20 and § 26).[39] The first part of the Transcendental Deduction, culminating in § 20, relies heavily on a new explanation of the categories that was added to § 13 in the B edition:

> I will merely precede this with the **explanation of the categories**. They are concepts of an object in general, by means of which its intuition is regarded as **determined** with regard to one of the **logical functions** for judgments. Thus, the function of the **categorical** judgment was that of the relationship of the subject to the predicate, e.g., "All bodies are divisible." Yet in regard to the merely logical use of the understanding it would remain undetermined which of these two concepts will be given the function of the subject and which will be given that of the predicate. For one can also say: "Something divisible is a body." Through the category of substance, however, if I bring the concept of a body under it, it is determined that its empirical intuition in experience must always be considered as subject, never as mere predicate; and likewise with all the other categories. [B128-9]

There are many other places where Kant makes similar claims.[40] What exactly is the claim here, and how exactly does Kant justify it?

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 *judgement* that can fix the positioning of the intuitions. Moreover, this is not just one role of the judgement amongst many – this is the *primary* role of the judgement:

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

39 The first half aims to show that we are always permitted to apply the pure concepts to intuitions, while the second half aims to show that the pure judgements (the synthetic *a priori* claims of the *Principles*) always hold.

40 For example, in a note added to Kant's copy of the first edition: "Categories are concepts, through which certain intuitions are determined in regard to the synthetic unity of their consciousness as contained under these functions; e.g., what must be thought as subject and not as predicate." He also makes similar claims in the *Metaphysik von Schon*, quoted in *Kant and the Capacity to Judge*, p.251, and *Prolegomena* § 20.

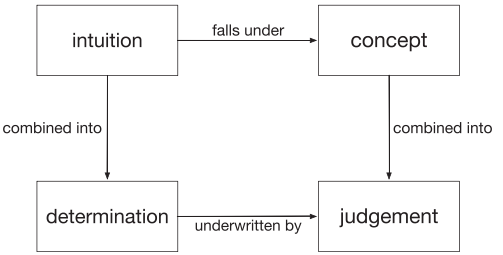


Figure 2.1: Intuitions are combined into determinations, just as concepts are combined into judgements. An intuition falls under a concept, just as a determination is underwritten by a judgement.

More specifically, the relative positions of intuitions in a determination can only be fixed by *forming a judgement that necessitates this particular positioning*. This judgement 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 judgement. See Figure 2.1. 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*. The same understanding, therefore, and indeed by means of the very same actions through which it brings the logical form of a judgment into concepts by means of the analytical unity, also brings a transcendental content into its representations by means of the synthetic unity of the manifold in intuition in general. [A79/B104-5]

There is a parallel claim one level up, at the level of complex judgements: the relative positions of determinations in a connection can only be fixed by forming a complex judgement that itself contains a pair of judgements as constituents[41] that necessitates this particular positioning. This complex judgement contains two constituents – judgements – that the two determinations fall under, and the position of the determinations in the connection are indirectly determined by the positions of the corresponding judgements in the complex judgement.

What justification does Kant provide for this claim? His argument goes something like this: the aim of the dynamical relations is to order the intuitions and determinations in *objective* space-time. Now we can only achieve objectivity by imposing *necessity* on the combination.[42] But the faculty of imagination

41 Kant is emphatic on this point: "hypothetical and disjunctive judgments do not contain a relation of concepts but of judgments themselves." [B141].

42 "Our thought of the relation of all cognition to its object carries something of necessity with it." [A104] The concept of an object is "the concept of something in which [the appearances] are necessarily connected" [A108].

is entirely incapable of imposing necessity. All the imagination can do is connect the intuitions using the pure relations – it cannot impose necessity on those connections.[43] In fact, the only element that can provide the desired necessity is the judgement.[44] Thus, the only way dynamical relations can be ordered in objective space-time is by indirectly positioning them, using judgements that impose the necessity that the connections require.[45]

In terms of the cognitive faculties responsible for the various processes, the **capacity to judge**[46] is responsible for constructing the judgements, and the faculty of **the power of judgement**[47] is responsible for constructing the subsumptions that decide which intuitions fall under which concepts.

This, then, is the general claim, as it applies to all the dynamical relations. Next, I shall describe the various forms of judgement that are needed to underwrite the various dynamical relations: inherence, succession, simultaneity, and incompatibility.

Inherence must be backed up by a categorical judgement. The first of the four conditions of dynamical unity is that the positions of intuitions in an inherence determination must be backed up by a corresponding judgement:[48]

---

(6)(a) If I form an inherence determination, ascribing a particular attribute *a* to a particular object *O*, then I must be committed to a judgement "this/some/all *X* are *P*", where *O* falls under *X*, and *a* falls under *P*.

Suppose, for example, I am seeing the particular dirtiness of this particular jumper. This inherence determination is a combination of two bare particulars:

---

43 "Apprehension is only a juxtaposition of the manifold of empirical intuition, but no representation of the necessity of the combined existence of the appearances that it juxtaposes in space and time is to be encountered in it." [A176/B219].

44 "This word [the copula "is"] designates the relation of the representations to the original apperception and its *necessary unity*, even if the judgement itself is empirical, hence contingent." [B142].

45 Here, the agent "binds" itself in two distinct but related senses. First, it binds its intuitions together via the pure relations. But this binding at the intuitive sensible level must be underwritten by a second binding at the conceptual discursive level: it is only because the agent binds itself to a rule relating concepts that the binding of intuitions achieves the necessity required for objectivity. See (Evans et al., 2019).

46 The capacity to judge (*Vermögen zu urteilen*) generates judgements from concepts. See [A81/B106] and (Longuenesse, 1998).

47 The power of judgement (*Urtheilskraft*) is responsible for deciding whether an intuition falls under a concept. See [A132/B171] and (Kant,1790).

48 In each of the unity conditions that follow, I restrict to the case of unary predicates. The extension to binary, ternary, and so on is straightforward but complicates the presentation.

this particular jumper and this particular instantiation of dirtiness. Now it is essential, in seeing the inherence correctly, that this particular dirtiness is the attribute and this particular jumper is the object in which the attribute inheres. Things would be very different indeed if the intuition of the dirtiness is the object, and the intuition of the jumper is the attribute.[49]

Kant's fundamental claim is that it is only because I form some corresponding categorical judgement that I am able to fix the positions of the two arguments of the inherence operator *det* [B128-9]. In this case, suppose I have formed the judgement "Some jumper is dirty." Now my intuition of this particular jumper falls under the concept "jumper", and my intuition of this particular dirtiness (of this particular jumper at this particular moment) falls under the concept "dirty". Thus, I am able to fix the positions of the two arguments to the inherence operator indirectly, via the judgement and the falls-under relation. I see the positions of the intuitions in the inherence *through* the corresponding judgement.

Now of course I do not need to use that precise judgement "Some jumper is dirty" to fix the positions of the intuitions in the inherence determination. I could have used "Some jumper is revolting", or "This jumper is dirty", and so on and so forth. All that is needed is *some* categorical judgement where the two intuitions fall under the two concepts.

Succession must be backed up by a causal judgement. The second condition of dynamical unity is that every succession of determinations must be backed up by a causal judgement:

---

(6)(b) If I form a succession, in which one determination (say, particular object O having particular attribute a) is followed by another determination (say, O having incompatible attribute b), then I must have formed a conditional judgement relating judgements describing the two determinations (say, "If $\phi(X)$ holds then X changes from P to Q", where object O falls under concept X, attribute a falls under concept P, attribute b falls under concept Q, and $\phi(X)$is a sentence featuring free variable X.)

Suppose, for example, I see the jumper's cleanliness followed by the jumper's dirtiness. It is essential, when seeing this succession, that I see the order

49 It is perhaps tempting to argue that it is just obvious which is the attribute and which is the object of the inherence: we can tell from the *types* of the two intuitions which one is which. Above, I said that there are two types of intuitions: intuitions of objects and intuitions of particular attributes. But this distinction only applies *after* a judgement has been constructed which allows the intuitions to be positioned; before that, these intuitions are not yet dignified with these roles as intuitions of objects or intuitions of particular attributes; they are just indeterminate intuitions. In other words, this response just begs the question, assuming that we have already access to the very positioning assignments that we are struggling to achieve.

correctly. Seeing the cleanliness followed by the dirtiness is very different from seeing the dirtiness followed by the cleanliness.

Kant claims[50] that it is only because I form some corresponding causal judgement that I am able to fix the positions of the two determinations in the succession relation [A189/B232]. Suppose, for example, I have formed the causal rule that if I wallow about in the mud, then my clothing will transform from clean to dirty. Now my intuition of this jumper falls under the concept "clothing", my intuition of this particular cleanliness falls under the concept "clean", and my intuition of this particular dirtiness falls under the concept "dirty". Thus, I am able to fix the positions of the two determinations in the succession relation indirectly, via the causal judgement and the falls-under relation.

Simultaneity must be backed up by a pair of causal judgements. The third condition of dynamical unity is that every simultaneity of determinations must be backed up by a pair of causal judgements:

---

(6)(c) If I form a simultaneity, in which one determination (say, particular object O_1 having particular attribute *a*) is simultaneous with another determination (say, object O_2 having attribute *b*), then there must be a pair of causal judgements describing determinations of the two objects (say, one of which states that an attribute of O_2 (simultaneous with *a*) causally depends on an attribute of O_2, and another of which states that an attribute of O_2 (simultaneous with *b*) causally depends on an attribute of O_1.)

Suppose, for example, I have two determinations simultaneously, one involving the sun, and one involving the moon. Now since simultaneity is a symmetric relation, it does not matter which of the two determinations is placed where in the *sim* relation. But it does matter whether we ascribe simultaneity or succession to the pair of determinations. When we are presented with a subjective succession of determinations, should we ascribe them to the same moment (of objective time) or to two successive moments (of objective time)?[51]

Kant's claim here is that in order to choose simultaneity over succession, we need to form a pair of judgements describing, for both objects, how some attribute of that object causally depends on some attribute of the other [A212/

50 Not all commentators agree with this way of reading Kant. Beatrice Longuenesse, for example, believes that we do not have to have already formed a causal judgement – we just need to acknowledge that we should form a causal judgement. For Longuenesse, perceiving a succession means being committed to look for a causal rule – it does not mean that I need to have already found one (Longuenesse, 1998).

51 "The apprehension of the manifold of appearance is always successive. The representations of the parts succeed one another. Whether they also succeed in the object is a second point for reflection, which is not contained in the first." [A189/B234].

B259]. I do not dwell on this principle, because it is the most controversial,[52] hard to understand, and does not feature in our computer implementation.

Incompatibility must be backed up by a disjunctive judgement. Kant talks throughout the *Postulates* about the possibility of an *object* – not of the possibility of a sentence being true. It is easy to see this as a category error, or as elliptical: perhaps "the object is possible" is short-hand for "it is possible that the object exists"? This temptation must be resisted. Kant predicates possibility/actuality/necessity of *determinations* as well as of judgements. When we connect two determinations with the *inc* connective, we are making a modal connection between two elements, two ways of seeing, elements that *do not have a truth value*.

Kant claims[53] that every incompatibility between determinations must always be backed up by a disjunctive[54] judgement:

---

(6)(d) If I form an incompatibility in which one determination (say, particular object *O* having attribute *a*) is incompatible with another (say, particular object *O* having attribute *b*), then I must have formed an exclusive disjunctive judgement stating that two judgements describing the two determinations are incompatible (say, "All *X* are either (exclusive disjunction) *P* or *Q* or . . .", in which *O* falls under *X*, *a* falls under *P*, and *b* falls under *Q*.)

Suppose, for example, I see this jumper's cleanliness as incompatible with the jumper's dirtiness. Now this is, to repeat, an incompatibility between determinations, ways of seeing, not an incompatibility between *judgements*. But Kant claims that this incompatibility between determinations must be underwritten by an exclusive-or disjunctive judgement. Suppose, for example, I have formed the judgement that every article of clothing is either clean or dirty. Now my intuition of this particular cleanliness falls under the concept "clean", my intuition of this particular dirtiness falls under the concept "dirty", and my intuition of this particular jumper falls under the concept "article of clothing." Thus, the exclusive disjunctive judgement (expressing an incompatibility between concepts) justifies the incompatibility relation between determinations.

---

52 See e.g., (Longuenesse, 1998) p.388.

53 "The schema of possibility is the agreement of the synthesis of various representations with the conditions of time in general (e.g., since opposites cannot exist in one thing at the same time, they can only exist one after another)." [A144/B184].

54 Recall that for Kant, disjunctions are *exclusive*: "*p* or *q*" means either *p* or *q* but not both.

2.3.3 Making Concepts Sensible

As well as the unity condition requiring that determinations are underwritten by judgements, there are also unity conditions in the other direction, requiring that judgements are supported by corresponding determinations.

> It is thus just as necessary to make the mind's concepts sensible (i.e., to add an object to them in intuition) as it is to makes its intuitions understandable (i.e., to bring them under concepts). [A51/B75]

The requirement here is that judgements cannot "float free" of the underlying intuitions. Instead, each judgement must be backed up by a corresponding determination.

More specifically (and restricting ourselves to unary predicates):

---

(7) If I form a judgement, ascribing a concept *P* to a particular object *X*, then there must be a corresponding inherence determination ascribing particular attribute *a* to particular object *O*, where *O* falls under *X* and *a* falls under *P*.

It might seem that this condition is trivially satisfied given that the agent starts with intuitions and determinations, and forms judgements to make them intelligible. But this is not always so: sometimes the agent constructs new *invented objects* to make sense of the sensible given and ascribes properties to these invented objects. In such cases, condition (7) requires that as well as subsuming object *o* under concept *P*, there is also a corresponding particular individual attribute *a* that inheres in *o*.[55]

2.3.4 Conceptual Unity

In addition to the synthetic unity described above, Kant also requires that one's concepts be unified by being connected together via judgements. I shall first consider a weak form of this constraint, before describing a stronger version.

A judgement connects various concepts together. For example, the judgement "some bodies are divisible" connects the concepts of "body" and "divisible". Let us say two concepts are *together* if there is some judgement in

---

55 The experiment of Section 3.1 shows just such an example where an invented object is postulated, and particular individual attributes of that object are posited in imagination to make the concepts sensible.

which they both feature. Define *together** as the transitive closure of *together*. Now the weak constraint of conceptual unity is that every pair of concepts are *together**.

Kant uses a significantly stronger constraint. His requirement is that the concepts are not just connected, but that they are connected into a *hierarchy* of genera and species.[56] In order that one's concepts form a *system* in this sense, we focus exclusively on the judgement form of exclusive disjunction [A70/B95]. Consider a judgement of the form "every X is either (exclusive) P or Q". This does not merely state that P and Q are exclusive; it also states that P and Q form a *totality*: the totality of concepts that together capture X. By bringing concepts under the *xor* judgement form, we bring them into a hierarchical community with a genera-species structure.[57]

The condition of conceptual unity is the requirement that:

(8) Every concept features in some disjunctive judgement.

2.4 Taking Stock

It is time to take stock. For Kant, the fundamental mental representation is the intuition, a representation of an individual element (e.g. a particular object or a particular attribute of a particular object). All the other types of representation serve only to unify the intuitions into a coherent whole.

Intuitions can be combined into determinations using the three pure operations of containment, comparison, and inherence. Further, determinations can be combined into connections using the pure relations of succession, simultaneity, and incompatibility. (See Section 2.2).

In order for the connections of determinations to achieve unity,[58] multiple conditions must be satisfied. The mathematical operations (of containment and

56 See (Longuenesse, 1998, p.105).

57 "What the form of disjunctive judgment may do is contribute to the acts of forming categorical and hypothetical judgments the perspective of their possible systematic unity", (Longuenesse, 1998), p.105.

58 In Section § 16 of the B deduction, Kant distinguishes four types of unity using two cross-cutting distinctions: analytic versus synthetic unity, on the one hand, and original versus empirical unity, on the other. Analytic unity is achieved when the mind has the ability to subsume each of its intuitions and determinations under the unary predicate "I think". Synthetic unity is achieved when the intuitions and determinations are connected together via the pure

comparison) must form a structure of the appropriate sort (Section 2.3.1), the dynamical functions (of inherence, succession, simultaneity, and incompatibility) must be underwritten by judgements of the appropriate sort (Section 2.3.2), the judgements must be underwritten by determinations of the appropriate sort (Section 2.3.3), and the concepts used in judgements must form their own unity (Section 2.3.4).

Why these unity conditions in particular? One of the remarkable things about Kant's philosophy is its systematicity. Instead of being content with merely enumerating the pure concepts of the understanding, Kant insists on showing how the pure concepts form a *system*, by showing that these are all and only the *a priori* concepts needed to make sense of experience.[59] The same systematicity requirement applies to the unity conditions: he must show that these are *all and only* the unity conditions needed for the synthesis of intuitions to achieve objectivity. To see that the unity conditions described above form a system, observe that there are two realms of cognition: the sensible intuitions and the discursive concepts. There are exactly four possible conditions involving these two realms: (i) a requirement that the intuitions achieve their own individual unity, (ii) a requirement that the intuitive realm respects the conceptual, (iii) a requirement that the conceptual realm respects the intuitive, and (iv) a requirement that the conceptual realm achieves its own individual unity. Here, (i) is the requirement that the synthesis of apprehension forms a fully connected graph satisfying 5(a) and 5(b) (Section 2.3.1). Condition (ii) is the requirement that the connections between intuitions are underwritten by corresponding judgements (Section 2.3.2). Condition (iii) is the requirement that the judgements respect the intuitions (Section 2.3.3). The final condition (iv) is the requirement that the discursive realm of judgement achieves conceptual unity (Section 2.3.4).

If our agent does all these things, and satisfies all these conditions, then it has achieved *experience*: it has combined the plurality of sensory inputs into a coherent representation of a single world. Achieving experience requires four faculties: sensibility (to receive intuitions), the imagination (to connect intuitions together

relations of Section 2.2 in such a way as to satisfy the unity conditions of Sections 2.3.1, 2.3.2, 2.3.3, and 2.3.4. Synthetic unity is the more fundamental concept, as it is presupposed by analytic unity [B133]. The distinction between empirical and original unity is the difference between a particular unity achieved by a particular mind when confronted with a particular sensory sequence, and what is in common between all unities achieved by all minds no matter which sensory sequence they are provided with. In this paper, I focus on the general conditions common to all minds when achieving synthetic unity.

59 See (Longuenesse, 1998, p.105).

using the pure relations as glue), the capacity to judge (to generate judgements), and the power of judgement (to decide whether an intuition falls under a concept).

According to our interpretation, intuitions are formed by sensibility, entirely independently of the understanding.[60] Further, intuitions can be connected (via the pure relations of Section 2.2) by the imagination, without the need for the understanding.[61] But intuitions can only constitute *experience* if the intuitions are brought under concepts (via the power of judgement) and the concepts are combined into judgements (via the capacity to judge): experience requires understanding working in concert with sensibility and the imagination to bring the connected intuitions into a unity. Thus, both sensibility and understanding need each other if they are to jointly achieve experience.[62]

Here are the core claims, brought together in one place for ease of reference:

1. In order to achieve experience, I must unify my intuitions.
2. Unifying intuitions means combining them using binary relations to form a connected graph, in such a way as to satisfy the various unity conditions.
3. Synthesis involves (i) connecting intuitions together via containment, comparison, and inherence operations to form determinations; and (ii) connecting determinations together via succession, simultaneity, and incompatibility relations.
4. There are, in total, four types of unity condition that Kant imposes: (i) the unity conditions for the synthesis of mathematical relations, (ii) the unity conditions for the synthesis of dynamical relations, (iii) the requirement that the judgements are underwritten by determinations, and (iv) the conceptual unity condition.
5. The unity conditions for the synthesis of mathematical relations are:
   (a) There exists some intuition x such that for each object of intuition y, for each moment in time, there is a chain of *in* determinations between y and x.
   (b) The comparison operator $<$ forms a strict partial order.

60 "Appearances can certainly be given in intuition without functions of the understanding." [A90/B122]. "The manifold for intuition must already be given prior to the synthesis of the understanding and independently from it." [B145].

61 "Synthesis in general is, as we shall subsequently see, the mere effect of the imagination, of a blind though indispensable function of the soul, without which we would have no cognition at all, but of which we are seldom even conscious" [A78/B103].

62 "Thoughts without content are empty, intuitions without concepts are blind." [A50-51/B74-76]. But note the striking asymmetry between the types of deficiency when one activity is performed without the other: *blindness* is a deficiency of a living conscious being, while *emptiness* is a deficiency of a mere *container*. This asymmetry confirms the interpretation in Section 2.1.2 that unity of intuition is the final end of all thought, and conceptual thought is merely a means to that end.

6. The unity conditions for the synthesis of dynamical relations are:
   (a) If I form an inherence determination, ascribing a particular attribute a to a particular object o, then I must be committed to a judgement "this/some/all X are P", where o falls under X, and a falls under P.
   (b) If I form a succession, in which one determination (say, particular object o having particular attribute a) is followed by another determination (say, o having incompatible attribute b), then I must have formed a conditional judgement "If $\phi(X)$ holds and X is P then X becomes Q at the next time-step", where object o falls under concept X, attribute a falls under concept P, attribute b falls under concept Q, and $\phi(X)$ is a sentence featuring free variable X.
   (c) If I form a simultaneity, in which one determination (say, particular object o_1 having particular attribute a) is simultaneous with another determination (say, object o_2 having attribute b), then there must be a pair of causal judgements, one of which states that an attribute of o_1 causally depends on an attribute of o_2, and another of which states that an attribute of o_2 causally depends on an attribute of o_1.
   (d) If I form an incompatibility in which one determination (say, particular object o having attribute a) is incompatible with another (say, particular object o having attribute b), then I must have formed a judgement "All X are either (exclusive disjunction) P or Q or . . .", in which o falls under X, a falls under P, and b falls under Q.
7. The requirement that the conceptual realm respects the intuitive is the condition that if I form a judgement, ascribing a concept P to a particular object X, then there must be a corresponding inherence determination ascribing particular attribute a to particular object o, where o falls under X and a falls under P.
8. The unity condition for conceptual unity is the requirement that every concept must feature in some disjunctive judgement.

In this section, I shall formalise the task of achieving synthetic unity of apperception. The formalism introduced is necessary for the derivation of the categories below.

2.5 Achieving Synthetic Unity

Let $\mathcal{I}$ be the set of intuitions, $\mathcal{D}$ the set of determinations, and C the set of connections. The signature of the three pure operations of containment, comparison, and inherence are:

$$in : \mathcal{I} \times \mathcal{I} \to \mathcal{D}$$

$$< : \mathcal{I} \times \mathcal{I} \to \mathcal{D}$$

$$det : \mathcal{I} \times \mathcal{I} \to \mathcal{D}$$

The signature of the three pure relations of succession, simultaneity, and incompatibility are:

$$succ : \mathcal{D} \times \mathcal{D} \to C$$

$$sim : \mathcal{D} \times \mathcal{D} \to C$$

$$inc : \mathcal{D} \times \mathcal{D} \to C$$

For example, if a, b, c are intuitions of type $\mathcal{I}$, then $det(a, b)$, $in(a, b)$, and $b < c$ are determinations of type $\mathcal{D}$; and $succ(det(a, b), det(a, c))$ and $sim(in(a, b), b < c))$ are connections of type C.

The input that the mind receives from sensibility is a sequence of individual determinations from $\mathcal{D}$. Note that the input is not a sequence of *sets* of determinations that are already assumed to be simultaneous, but a sequence of *individual* determinations. Kant insists on this:

> The apprehension of the manifold of appearance is always successive. The representations of the parts succeed one another. Whether they also succeed in the object is a second point for reflection, which is not contained in the first . . . Thus, e.g., the apprehension of the manifold in the appearance of a house that stands before me is successive. Now the question is whether the manifold of this house itself is also successive, which certainly no one will concede. [A189/B234ff]

Here, Kant asks us to imagine an agent surveying a large house from close range. Its visual field cannot take in the whole house in one glance, so its focus moves from one part of the house to another. Its sequence of visual impressions is successive, but there is a further question whether a pair of (subjectively) successive visual impressions represents the house at a single moment of objective time, or at two successive moments of objective time.[63]

Given a sequence $(d_1, \ldots, d_t)$ of individual determinations, constructed from a set I of intuitions using the three pure operations (containment, comparison, and inherence), the task of making sense of sensory input is is to construct a synthetic unity – a tuple $(J, D, \kappa, \upsilon, \theta)$ – satisfying various conditions, where:

- J is a set of intuitions that must include I but also includes new intuitions that were constructed by the productive imagination

63 See also (Longuenesse, 1998, p.359).

- D is a set of determinations that must include $d_1, \ldots, d_t$ but also includes new determinations that were constructed by the productive imagination
- $\kappa \subseteq C$ is a set of connections between determinations
- $\upsilon \subseteq \mathcal{I} \times P_1$ is the falls-under relation (also known as subsumption) between intuitions and unary predicates P_1, between pairs of intuitions and binary predicates P_2, etc.
- θ is a collection of judgements

The connections κ are generated by the faculty of *imagination*. Note that not all the determinations in κ need come from the original sequence $(d_1, \ldots d_t)$. Some of the determinations may involve new invented objects constructed by *pure intuition* (for spaces and times) or by the *imagination* (for hypothesised unperceived empirical objects). The connections must satisfy the following conditions:

- For every pair of intuitions in J, there is a chain of determinations in D connecting one to the other
- If d_i, d_{i+1} are successive determinations in $(d_1 \ldots, d_t)$, then either $sim(d_i, d_{i+1})$ or $succ(d_i, d_{i+1})$ must be in κ
- The determinations are fully connected: every determination in D is κ-connected to every other determination via some path of undirected edges.

While the falls-under relation υ is generated by *the power of judgement*, the theory θ is a collection of judgements that is generated by the *capacity to judge*. The formal language for defining judgements, Datalog$^{\ni}$, is described in (Evans et al., 2021b), but in brief: judgements are either rules or constraints. Rules are either arrow rules $\alpha_1 \wedge \ldots \alpha_n \rightarrow \alpha_0$ (stating that if $\alpha_1 \ldots, \alpha_n$ all hold, then α_0 also holds at the same time-step), or causal rules $\alpha_1 \wedge \ldots \alpha_n \ni \alpha_0$ (stating that if $\alpha_1 \ldots, \alpha_n$ all hold, then α_0 also holds at the *next* time-step). Constraints are either *xor* judgements $\alpha_1 \oplus \ldots \oplus \alpha_n$ (stating that exactly one of the α_i hold) or a uniqueness constraint $\forall X, \exists! Y, r(X, Y)$ (stating that for each X there is exactly one Y such that $r(X, Y)$).

Figure 2.2 shows two different ways of grouping the four faculties, according to two cross-cutting distinctions. According to one distinction, sensation and imagination both fall under *sensibility* because both faculties process intuitions.[64] The power of judgement and the capacity to judge both fall under the *understanding* because both faculties process concepts. According to the

64 "Now since all of our intuition is sensible, the imagination, on account of the subjective condition under which alone it can give a corresponding intuition to the concepts of understanding, belongs to **sensibility**." [B151].

other distinction, sensation falls under *receptivity* because it is a purely passive capacity that merely receives what it is given. The other three faculties fall under *spontaneity*[65] because the agent is free to construct *whatsoever it pleases*, as long as the resulting construction satisfies the various unity conditions.

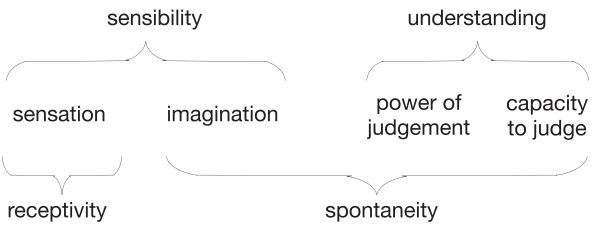
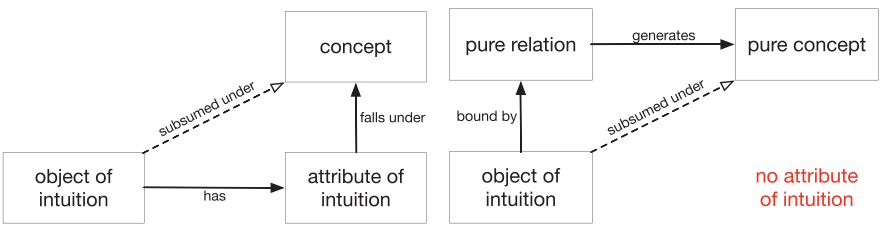


Figure 2.2: The relationship between the four faculties.



(a) The explanation of subsumption for an empirical concept

(b) The explanation of subsumption for a pure concept

Figure 2.3: Both diagrams provide an explanation for an object being subsumed under a concept. In (a), the concept is empirical, and the explanation goes via the intermediary of an attribute of intuition. In (b), the concept is pure, there is no corresponding attribute, and the explanation goes via the another intermediary: a pure relation.

We have now assembled the materials needed to define the task of synthetic unity.

Given a sequence $(d_1, \ldots, d_t)$ of determinations, the **task of achieving synthetic unity of apperception** is to construct a tuple $(J, D, \kappa, \upsilon, \theta)$ as described above that satisfies the unity conditions of Sections 2.3.1, 2.3.2, 2.3.3, and 2.3.4.

---

65 See [A51/B75], [B133], [B151].

2.6 The Derivation of the Categories

The problem of the pure categories is explained in the opening paragraphs of the *Schematism*:

> In all subsumptions of an object under a concept the representations of the former must be **homogeneous** with the latter, i.e., the concept must contain that which is represented in the object that is to be subsumed under it, for that is just what is meant by the expression "an object is contained under a concept." . . . Now pure concepts of the understanding, however, in comparison with empirical (indeed in general sensible) intuitions, are entirely unhomogeneous, and can never be encountered in any intuition. Now how is the **subsumption** of the latter under the former, thus the **application** of the category to appearances possible, since no one would say that the category, e.g., causality, could also be intuited through the senses and is contained in the appearance? [A137/B176 ff]

For empirical concepts, an object's being subsumed under a concept can be explained in terms of a particular attribute that the object has which falls under the concept. See Figure 2.3(a). Suppose, for example, my intuition of this particular jumper is subsumed under the concept "dirty". This subsumption is explained by (i) the object of intuition having, as one of its determinations, a particular attribute of intuition (my representation of the particular dirtiness of this particular jumper at this particular moment), and (ii) the attribute of intuition falling under the concept "dirty". The problem, for the pure concepts such as *Unity*, *Reality*, *Substance*, and so on, is that there is no corresponding attribute of intuition, so the explanation of the subsumption in Figure 2.3(a) is not applicable. What, then, justifies or permits us to subsume the objects of intuition under the pure concepts?

According to Kant, what justifies my subsuming an object under a pure concept is the existence of a *pure relation*[66] that the object is bound to. See Figure 2.3(b). Here, the subsumption of the object under the pure concept is explained by (i) the object of intuition being bound to the pure relation, and (ii) the pure concept being derivable from the pure relation. Note that in both Figures 2.3(a) and (b) there is an intermediary that explains the object being subsumed under a concept, but it is a different sort of intermediary in the two cases:

> Now it is clear that there must be a third thing, which must stand in homogeneity with the category on the one hand and the appearance on the other, and makes possible the application of the former to the latter. This mediating representation must be pure (without anything empirical) and yet intellectual on the one hand and sensible on the other. Such a representation is the transcendental schema. [A138/B177]

66 I.e. one of the six pure relations introduced in Section 2.2.

The "transcendental schema" is just another term for what I have been calling a pure relation: *in*, $<$, *det*, *succ*, *sim*, and *inc*.

This, then, is the outline of Kant's argument explaining how the pure concepts (categories) apply to objects of intuition. The next stage is to show, in detail, for each pure concept, exactly how it is derived from the corresponding pure relation. The derivation is straightforward and Kant did not see the need to spell it out.[67] But for the sake of maximal explicitness, we shall go through each in turn.

Starting with the title of *Relation*, intuition X falls under the pure concept **substance** if there exists an intuition Y such that $det(X, Y)$ is a determination in κ [B128-9]. Likewise, X falls under the pure concept **accident** if there exists an intuition Y such that $det(Y, X)$ is a determination in κ. Determination d falls under the pure concept **cause** if there exists a determination d' such that $succ(d, d')$ is in κ [A144/B183]. Likewise, determination d falls under the pure concept **dependent** if there exists a determination d' such that $succ(d', d)$is in κ. A set D of determinations falls under the pure concept **community** if for each d, d' in D, $sim(d, d')$ is in κ [A144/B183-4].

Moving to the title of *Modality*, a set D of determinations falls under the pure concept **possible** if there is some sequence of sensor readings, and some theory θ that makes sense of those readings, such that D is contained in one of the states of the trace of θ [A144/B184]. A set D of determinations is **actual** if it is contained in one of the states of the trace of the best theory that explains the sensor readings that have been received.[68] A set D of determinations is **necessary** if it is contained in every state of the trace of the best theory that explains every possible sensory sequence.

Moving next to the title of *Quality*, intuition X falls under the pure concept of **reality** if there exists an intuition Y such that $Y < X$ [A168/B209]. Likewise, intuition X falls under the pure concept of **negation** if there does not exist an intuition Y such that $Y < X$.

---

67 In (Brandom, 2009), Brandom describes how new unary concepts can be derived from given relations. So, for example, if we have the binary relation $P(x, y)$ representing that x admires y, then we can form the new unary predicate $Q(x)$ defined as $Q(x) = R(x, x)$. Here, $Q(x)$ is true if x is a self-admirer. In a similar manner, the unary categories are derived from the pure relations of Section 2.2.

68 "The postulate for cognizing the **actuality** of things requires **perception**, thus sensation of which one is conscious – not immediate perception of the object itself the existence of which is to be cognized, but still its connection with some actual perception." [A225/B272].

Moving, finally, to the title of *Quantity*, the categories of *Unity*, *Plurality*, and *Totality* are slightly more involved because they are implicitly indexed by a predicate p. A container is a **unity** of p's if it contains all the objects that fall under p. In other words, X falls under the pure concept of unity if for all Y, $(Y,p) \in \upsilon$ implies $in(Y,X)$. A container is a **plurality** of p's if all the objects within it fall under p. In other words, X falls under the pure concept of plurality if for all Y, $in(Y,X)$ implies $(Y,p) \in \upsilon$. A container is a **totality** of p's if it contains all and only the objects that fall under p.[69]

Returning to the overall argument for the derivation of the categories, Kant's deontic[70] argument can be summarized as:

- Achieving experience requires that I connect the intuitions using the pure relations.
- If I connect the intuitions using the pure relations, then I may apply the pure concepts (the categories) to the objects of intuition.
- Therefore, achieving experience permits me to apply the pure concepts to the objects of intuition.

Thus the *quid juris* question [A84/B116] has been answered. Note, however, that my permission to apply the pure concepts to objects of intuition is conditioned on my *activity*, the activity of trying to achieve experience. Hence Kant's conclusion that the categories are only permitted to apply to objects of experience.[71]

Kant insisted that the categories are not innate. The pure unary concepts are not "baked in" as primitive unary predicates in the language of thought. The only things that are baked in are the fundamental capacities (sensibility, imagination, power of judgement, and the capacity to judge) together with the pure relations of Section 2.2. The categories themselves are *acquired* – derived from the pure relations *in concreto* when making sense of a particular sensory sequence. But they are *originally* acquired [*Entdeckung*, Ak. VIII, 222–23; 136.][72] because they are *always* derivable from *any* sensory sequence. The pure concepts, then, are not innate but originally acquired (Longuenesse, 1998).[73]

---

69 Kant says that a totality is a plurality considered as a unity [B111].

70 The argument is deontic in that it relies on the concepts of obligation and permission. Kant tries to show that we are permitted to apply the pure concepts to objects of experience, and his justification is that we are obligated to perform the activity of achieving synthetic unity.

71 "The category has no other use for the cognition of things than its application to objects of experience." [B145].

72 This is quoted in (Longuenesse, 1998).

73 Some cognitive scientists (e.g. Gary Marcus (Marcus, 2018b)) place Kant on the nativist side of the nativist versus empiricist debate. But the key question for Kant is not what humans are born with, but what agents *must do* in order to make sense of the sensory input. It is a normative

3 Experiments

The cognitive architecture described above has been implemented in the APPERCEPTION ENGINE. The computer system is described in (Evans et al., 2021b) and (Evans et al., 2021a). In this section, I describe one experiment in detail.

3.1 The Sensory Input

In this experiment, there are two light sensors that can register various levels of intensity. If we take readings of both sensors at regular intervals, we get Figure 2.4. Here, the top row shows a human-readable discretised version of the sensor readings, revealing a simple regular pattern. The bottom row shows a fuzzier version of the same pattern where each sensor reading was perturbed with random noise. It is this second fuzzier version that is used in this experiment. But the sensory input, as presented in Figure 2.4(b), shows the sensory readings after they have already been assigned to particular moments in time. In Kant's theory, this time-assignment is not something that is given to the system, but rather is a hard-won *achievement*. In Kant's theory, the sensory input is presented as a sequence of *individual* sensory readings, and the agent has to decide how the various readings should be combined together into moments of objective time. So the actual input to the Kantian agent is shown in Figure 2.5. Here, the agent is given a sequence of individual sensory readings, and must choose how to combine them together into a succession of simultaneous readings. While Figure 2.5 shows the sequence of individual readings in subjective time, Figure 2.6 shows a variety of different ways of parsing the raw sequence into moments. The bottom row of Figure 2.6 shows the correct way of parsing the sequence in Figure 2.5; this correct parse corresponds to Figure 2.4(b).

The input, then, is the sequence shown in Figure 2.5. In our implementation, the continuous sensor readings are first discretised into binary vectors. The total sequence $(d_1, \ldots, d_{50})$ is a list of 50 inherence determinations. Note that the readings do not simply alternate between a and b. Sometimes there are multiple a's or

question of *a priori* psychology, not an empirical question about ontogenetic development. From Kant's perspective, the list of innate concepts proposed by cognitive scientists (spelke and Kinzler, 2007) is a "mere rhapsody" [A81/B106] unless they can be unified under a *common principle*. Nativists compile their list of innate concepts by looking at what human babies can do. But the capacities that evolution has hard-wired to help us in our particular situation are not *maximally general*. For example, babies can distinguish faces from other shapes before they are born, but the concept of a face is not a pure concept in Kant's sense.

b's in a row. The subjective sequence records the sequence of items the agent is attending to (he can only attend to one sensation at a time), and the agent might attend to either sensor at any moment of subjective time. Given this sequence in subjective time, we must reconstruct the moments of objective time by connecting the determinations using the relations of simultaneity and succession.

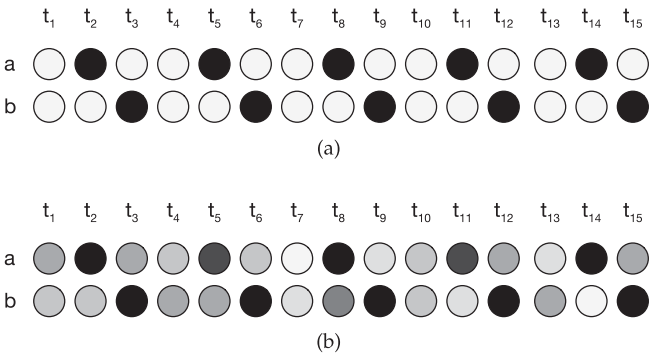


Figure 2.4: A simple sequence involving two sensors. (a) shows a noise-free version, where the pattern is clearly apparent. (b) shows the fuzzy version with random noise that is used in this experiment.

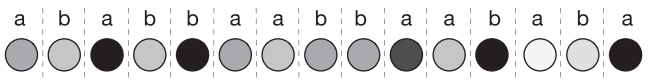


Figure 2.5: The input to the APPERCEPTION ENGINE is a sequence of individual readings. The engine must choose how to group the individual readings into groups of simultaneous readings.

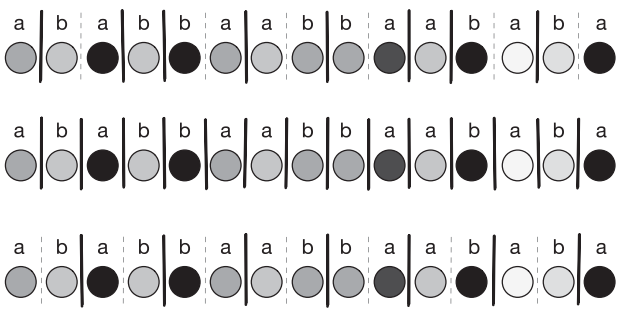


Figure 2.6: We show three ways of parsing the individual readings (in subjective time) into a succession of simultaneous readings (in objective time). The thin dashed lines divide the readings in subjective time, while the thicker lines group the individual readings into sets of simultaneous readings in objective time. The bottom row of the three represents the correct ground-truth way of grouping the readings.

3.2 The Model

Given the sensory sequence, the agent must construct an interpretation that makes sense of the sequence. The interpretation consists of:

1. A **synthesis of intuitions**. This contains a set of determinations (that must include the original sensory sequence, but can also include determinations involving other invented intuitions) connected together via the pure relations of *sim*, *succ*, and *inc*.
2. A **collection of subsumptions**. This is a set of mappings from intuitions of individual objects to general concepts. The mapping is implemented as a binary neural network.
3. A **set of judgements** that connect the concepts together.

I shall go through each in turn.

3.2.1 The Synthesis of Intuitions

The given sequence $(d_1, \ldots, d_{50})$ is a sequence of individual determinations in subjective time. We need to produce a sequence of sets of determinations in objective time. For each consecutive pair d_t, d_{t+1}, they can either be simultaneous or successive.

In our example, this choice rule gives us 2^{49} possibilities.[74] Once the *sim* and *succ* relations are provided, this determines the positions of the determinations in objective time.

3.2.2 The Set of Subsumptions

A subsumption maps an intuition (a bit vector) to a concept (symbol). We implement the power of judgement using a binary neural network parameterised by Boolean weights.

The neural network's input is a binary vector and the output is a binary vector of length $|P|$ (where $|P|$ is the number of unary predicates). The neural network implements a multilabel classifier mapping binary vectors to $2^{|P|}$.

74 The current implementation assumes that any pair of consecutive sensor readings are either simultaneous or successive. This precludes the possibility that there are intermediate time-steps between the two consecutive readings. In future work, I plan to expand the choice rule to allow this further possibility, so that it is possible to abduce intermediate time-steps.

3.2.3 The Set of Judgements

Kant's faculty of understanding is implemented as a program synthesis system that takes as input a stream of sensory information, and produces a theory (a set of judgements) that both explains the sensory stream and also satisfies various unity conditions. For details, see (Evans et al., 2021b).

3.2.4 Filling in the Unperceived Details

Kant's requirement that judgements should be underwritten by determinations is implemented by adding a choice rule for each predicate p, stating that if an object X satisfies predicate p at T, then there is some particular attribute *Attr* ascribed to X at T (where *Attr* falls under p).

3.2.5 Finding the Best Model

When the three sub-systems (the imagination, power of judgement, and understanding) described above are implemented in one system, many different interpretations are found. In order to decide between the various interpretations, we use the following preferences:

1. We prefer shorter theories over longer theories, all other things being equal.
2. We prefer more discriminatory neural networks which assign fewer intuitions to the same concept.

See (Evans et al., 2021a) for the mathematical details of how these two desiderata are weighted and compared.

3.3 Results

The interpretation found by the APPERCEPTION ENGINE consists of a tuple $(J, D, \kappa, \upsilon, \theta)$ consisting of a synthesis of intuitions, a collection of subsumptions, and a set of judgements. We shall consider each in turn.

The synthesis of intuitions κ. When confronted with the sensory sequence of Figure 2.5, the engine produces a set κ of connections using the pure relations of *sim*, *succ*, and *inc*. Here is an excerpt:

sim(([1, 0, 0], a, 1), ([1, 0, 1], b, 2))	succ(([1, 0, 1], b, 2), ([0, 0, 1], a, 3))	inc(([1, 0, 0], a, 1), [0, 0, 1], a, 3))
sim(([0, 0, 1], a, 3), ([1, 0, 1], b, 4))	succ(([1, 0, 1], b, 4), ([0, 0, 0], b, 5))	inc(([1, 0, 1], b, 2), ([0, 0, 0], b, 5))
sim(([0, 0, 0], b, 5), ([1, 0, 0], a, 6))	succ(([1, 0, 0], a, 6), ([1, 0, 1], a, 7))	inc(([1, 0, 0], a, 6), ([0, 0, 1], a, 10))
sim(([1, 0, 1], a, 7), ([1, 0, 0], b, 8))	succ(([1, 0, 0], b, 8), ([1, 0, 0], b, 9))	inc(([1, 0, 1], a, 7), ([0, 0, 0], a, 15))

Here, the determinations are triples containing an attribute (a binary vector of length 3, representing a particular shade of gray), an object (here *a* or *b*), and an index (from 1 to 15) in *subjective* time. This index is needed so that two determinations sharing the same object and attribute at different moments of time are nevertheless treated as distinct.

Figure 2.7 shows how the *succ* and *sim* relations produce objective time from subjective time.

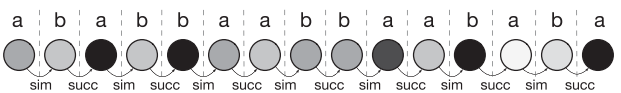


Figure 2.7: How the objective temporal sequence is constructed from the subjective temporal sequence via the pure relations of *sim* and *succ*.

The falls-under relation υ. The APPERCEPTION ENGINE constructs two unary predicates, p and q, and subsumes the binary vectors under them. The binary neural network implements a multilabel classifier, mapping binary vectors to subsets of $\{p, q\}$. The subsumptions υ produced by the engine are:

$$\begin{array}{cc} [0,0,0] \mapsto \{q\} & [0,0,1] \mapsto \{q\} \\ [0,1,0] \mapsto \{q\} & [0,1,1] \mapsto \{p,q\} \\ [1,0,0] \mapsto \{p\} & [1,0,1] \mapsto \{p\} \\ [1,1,0] \mapsto \{p\} & [1,1,1] \mapsto \{p\} \end{array}$$

Note that $[0,1,1]$ is considered ambiguous.

Figure 2.8 shows the subsumptions generated by the engine. Note the introduction of an invented object, *c*, that was not part of the sensory input.

The set of judgements θ**.** Along with the synthesis of intuitions and the collection of subsumptions, the APPERCEPTION ENGINE also generates a theory θ, containing a set of judgements that explain the dynamics of the system. The theory constructed for the problem of Figure 2.5 is $\theta = (\phi, I, R, C)$, where ϕ is a type signature, I, is a set of initial conditions, R is a set of conditionals, and C

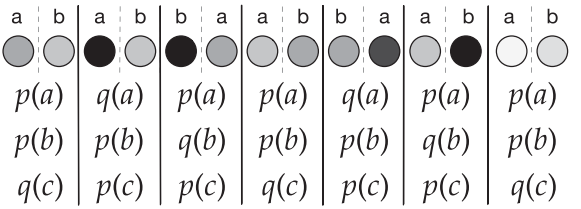


Figure 2.8: The subsumptions generated by the engine. The dashed lines divide subjective time, while the solid lines divide moments of objective time. The atoms generated at each moment are displayed below.

is a set of constraints. The type signature ϕ consists of types T, objects O, and predicates P where:

$$T = \{sensor, space\}$$
$$O = \{a{:}sensor, b{:}sensor, c{:}sensor, s_1{:}space, s_2{:}space, s_3{:}space, s_w{:}space\}$$
$$P = \{p(sensor), q(sensor), in(sensor, space), in_2(space, space), r(space, space)\}$$

The initial conditions I, rules R and constraints C are:

$$I = \left\{ \begin{array}{llll} p(a) & p(b) & q(c) & \\ in(a, s_1) & in(b, s_2) & in(c, s_3) & \\ in_2(s_1, s_w) & in_2(s_2, s_w) & in_2(s_3, s_w) & in_2(s_w, s_w) \\ r(s_1, s_2) & r(s_2, s_3) & r(s_3, s_1) & \end{array} \right\}$$

$$R = \left\{ \begin{array}{l} q(X) \Rightarrow p(X) \\ in(X, S_1) \wedge in(Y, S_2) \wedge r(S_1, S_2) \wedge q(X) \Rightarrow q(Y) \end{array} \right\}$$

$$C = \left\{ \begin{array}{l} \forall X{:}sensor,\ p(X) \oplus q(X) \\ \forall X{:}sensor, \exists! Y{:}space,\ in(X, Y) \\ \forall X{:}space, \exists! Y{:}space,\ in_2(X, Y) \\ \forall X{:}sensor, \exists! Y{:}sensor,\ r(X, Y) \end{array} \right\}$$

Here, the sensors a and b are given as part of the sensory input, but c is an invented object, constructed by the imagination. The invented objects s_1, s_2, and s_3 are three parts of space, constructed by pure intuition. The three spaces are all parts of the spatial whole s_w.

The unary predicates p and q are used to distinguish between a sensor's being on and off. The *in* relation places sensors in space, and the in_2 relation places spaces inside the spatial whole. The r relation is used to define a one-dimensional

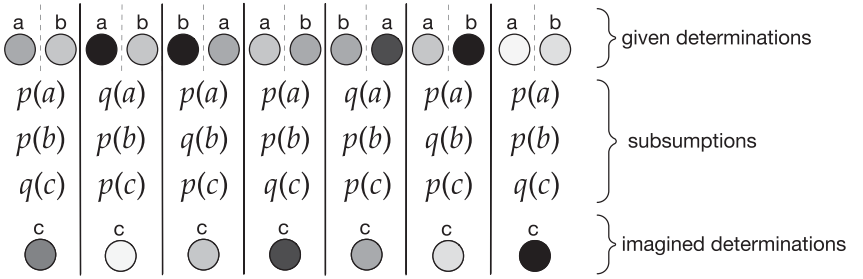


Figure 2.9: The determinations imagined by the engine. Here we show the given determinations (top row), the subsumptions (middle row), and the imagined determinations (bottom row) that are generated to satisfy condition (7): the requirement that every judgement needs to be underwritten by a determination. Thus, for example, the atom $q(c)$ in time step 1 needs to be underwritten by an inherence determination attributing a particular shade of q-ness to object c.

space with wraparound.[75] Note that our "spatial unity" requirement is rather minimal: we just insist that there is *some* containment structure connecting the intuitions together. It is not essential that the space constructed has the particular three-dimensional structure that we are accustomed to. Any spatial structure will do as long as the intuitions are unified , Chapter 3. In terms of Kant's distinction between the *form of intuition* and the *formal intuition* [B160n], the relation r describes the form of intuition (relations between objects) while the particular spaces (s_1, s_2, s_3, and s_w) represent the formal intuitions.

Note that the given objects of sensation (the sensors a and b) are not directly related to each other. Rather, they are *indirectly related* via the spatial objects and the *in* and r relations.

The rules describe how the unary properties p and q change over time. The first rule states that objects that satisfy q at one time-step will satisfy p at the next time-step. The second rule describes how the q property moves from one sensor to its right neighbour.

The constraints are constructed to satisfy conceptual unity (Section 2.3.4). The first insists that every sensor is either p or q but not both. The second requires that every sensor is contained within exactly one spatial region.

Filling in the unperceived details. In order to make concepts sensible (Section 2.3.3), the engine must ensure there is a determination corresponding to every judgement. In particular, the judgements involving invented unperceived

75 Note that, in this example, the spatial structure is static. But see Evans et al. (2020) for examples where objects move around.

object c must be underwritten by corresponding determinations. This means that for each time step at which $p(c)$ (respectively $q(c)$) is true, there must be an inherence determination $det(c, \alpha)$ ascribing particular attribute α to c, where c falls under p (respectively q).

Satisfying this condition means *imagining* particular attributes assigned to c for each moment of objective time. One set of determinations satisfying this condition is shown in Figure 2.9.

Thus, the unperceived object c is not merely subsumed under a predicate, but is also involved in a determination. *Even though c is an external object with which the agent has no sensory contact, it is cognised as satisfying particular perceptual determinations*. This is, I believe, the truth behind the Kant-inspired claim that "perception is a kind of controlled hallucination" (Clark, 2013).

Note that requirement (7) of Section 2.3.3 insists that object c must be involved in *some* determination, but does not – of course – insist on any *particular* determination. The productive imagination is free to construct any determination it pleases.

Discussion. Figure 2.10 shows the whole experiment, from the original input to the complete output consisting of a synthesis of intuitions, a collection of subsumptions, and a set of judgements. It is gratifying to see the APPERCEPTION ENGINE discerning a discrete intelligible structure behind the continuous noisy input. It started with a fuzzy sensory input, and perceived, amongst all the noise, an underlying system involving two discrete unary predicates, p and q, and devised a simple theory explaining how p and q change over time.

Let us pause to check that the interpretation of Figure 2.10 satisfies the various conditions (Section 2.4) required to achieve synthetic unity:

- The determinations are connected together via the relations of *succ*, *sim*, and *inc* to form a fully connected graph, as required in Section 2.2.
- The containment condition 5(a) of Section 2.3.1 is satisfied by the initial conditions I of Figure 2.10. Here, s_w is the spatial whole in which all other objects are contained, directly or indirectly.
- The $<$ relation is not needed in this particular example. The empty relation trivially satisfies the condition 5(b) that $<$ is a strict partial order.
- The requirement 6(a) of Section, that every inherence determination is underwritten by a judgement, is satisfied by the theory θ together with the subsumptions υ. Consider, for example, the first determination in the given sequence: $det(a, [1, 0, 0])$, ascribing the binary vector $[1, 0, 0]$ (representing a particular shade of gray) to object a. Note that $[1, 0, 0] \mapsto p$ according to υ, and since a is an object of type *sensor*, the determination is underwritten by the judgement $\exists X{:}sensor, p(X)$.

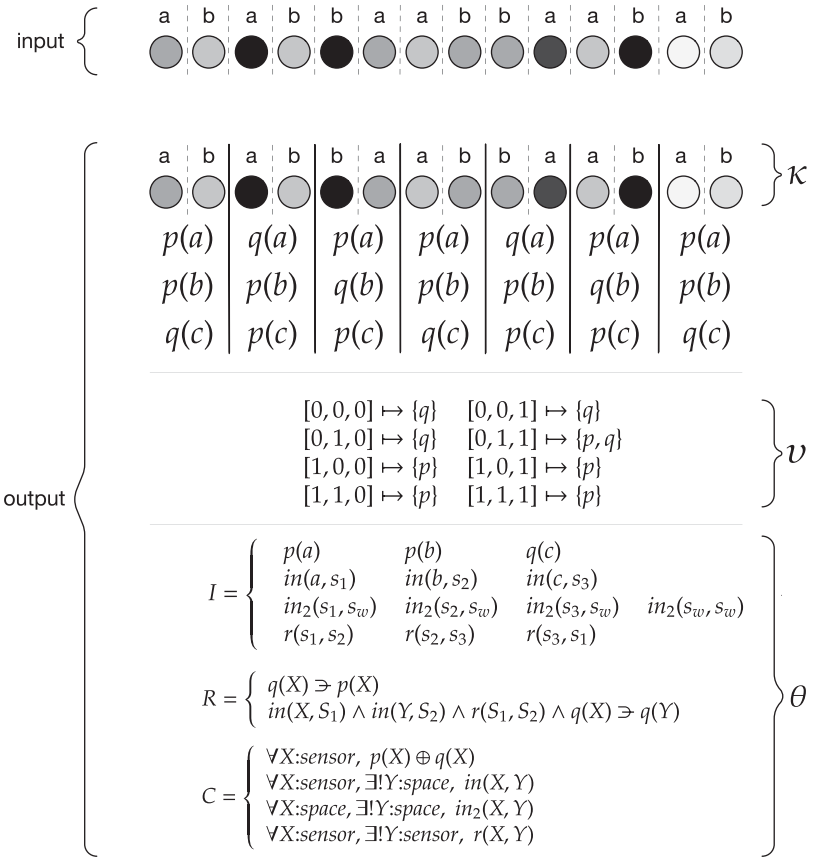


Figure 2.10: The result of applying the APPERCEPTION ENGINE to the input of Figure 2.5. The dashed lines divide moments of subjective time, while the solid lines divide moments of objective time. We show the synthesis of intuitions κ, the subsumptions υ, and the theory θ. We also show the ground atoms at each step of objective time, generated by applying the subsumptions υ to the raw input.

– The requirement 6(b) of Section, that every succession is underwritten by a causal judgement, is satisfied by the theory θ together with the subsumptions υ. Consider, for example, the succession:

$$succ(([0,0,1],b,4),([1,1,0],b,5))$$

This represents the succession of $det(b,[0,0,1])$ by $det(b,[1,1,0])$ (i.e., b changing from one particular shade of gray to another). Note that $[0,0,1]\mapsto q$

and $[1,1,0]\mapsto p$ according to the subsumptions υ, and rules R contain the causal judgement $q(X) \supset p(X)$.

- The requirement 6(c) of Section is not used in our initial implementation of the Apperception Engine. See Section 4.4 for a discussion.
- The requirement 6(d) of Section, that every incompatibiity is underwritten by a constraint, is satisfied by the constraints C in θ together with the subsumptions υ. Consider, for example, the incompatibility:

$$inc(([1,0,0],a,1),[0,0,1],a,3))$$

This incompatibility between determinations is underwritten by the constraint $\forall X{:}sensor,\ p(X) \oplus q(X)$, together with the mappings $[1,0,0]\mapsto p$ and $[0,0,1]\mapsto q$.

- The requirement 7 of Section 2.3.3 is satisfied by the inherence determinations featuring invented object c as shown in Figure 2.9.
- The requirement 8 of Section 2.3.4, that every predicate features in some *xor* or uniqueness constraint, is satisfied by the theory θ of Figure 2.9. Here, predicates p and q feature in the constraint $\forall X{:}sensor,\ p(X) \oplus q(X)$, *in* features in the constraint $\forall X{:}sensor, \exists! Y{:}space,\ in(X,Y)$, and so on for the other binary relations.

3.4 Perceptual Discernment and Conceptual Discrimination

Compare the interpretation of Figure 2.10 with the alternative degenerate interpretation of Figure 2.11. Both interpretations satisfy the unity conditions, but they do so in very different ways. While Figure 2.10 discerns a difference between the inputs – dividing them into two classes, p and q – and constructs a theory that explains how p and q properties interact over time, Figure 2.11, by contrast, fails to discern any difference between the input vectors. Because Figure 2.11 is coarser and less discriminating, mapping all input vectors to p and none to q, it can make do with a much simpler theory: if everything is always p and never q, we do not need a complex theory to explain how objects transition between p and q.[76]

76 The Apperception Engine considers and evaluates many different theories when presented with the sensory input of Figure 2.5. It prefers the interpretation of Figure 2.10 over the degenerate interpretation of Figure 2.11 precisely because the former discriminates finer. In (Evans et al., 2021a), I explain how one interpretation is preferred to another if, other things being equal, the first makes more fine-grained perceptual discriminations. I justify the preference using simple Bayesian considerations.

In Kant's theory of synthetic unity, as we interpret it, this phenomenon holds across the board. In order to discern a fine-grained discrimination between sensory input, we must provide a theory that underwrites that distinction, a theory that explains how the various properties that we have discriminated actually interact. Fine-grained perceptual discrimination requires an articulated theory (a collection of concepts and judgements) that underpins the distinctions made at the sensible level. Intuitions without concepts are blind.

There is a recurrent myth that humans have fallen from a state of pre-conceptual grace (Jaynes, 2000). At some mythic earlier time, humans were not saddled with the conceptual apparatus we now take for granted, and – precisely because they were unburdened by concepts and judgements – were able to perceive the world in all its glory, with a fine-grained vividness we moderns can only dream of. It is as if there is only a finite amount of consciousness to go round; because we modern concept users waste some of that consciousness on the conceptual side of our experience, there is less consciousness remaining to spend on the sensible side. The mythic earlier man, by contrast, is able to spend all his consciousness on the sensible level. Thus for him, in his state of pre-conceptual grace, the colours are brighter.

If Kant is right, this myth gets things exactly the wrong way round. Consciousness is not a zero-sum game between sensibility and understanding, in which one side's gains must be the other side's losses. Rather, perceptual discrimination at the sensible level requires conceptual discrimination from the understanding. *The more intricate the theories we are able to construct, the more vividly we are able to see.*

4 Discussion

4.1 Rigidity and Spontaneity

There is a popular image of Kant as a rigid rule-bound automaton whose daily routine was so tightly scheduled you could use it to calibrate your clock. According to this popular image, Kant's philosophy (both practical and theoretical) is as rigid and rule-bound as his unusually unremarkable personal life. What is most unfair about this gross mischaracterisation is that it omits the critical fact that, for Kant, the rules I am bound to are rules that *I myself create*.

Spontaneity and self-legislation are at the heart of Kant's philosophy, both practical and theoretical. In his practical philosophy, I am free to construct *any maxims whatsoever* – as long as they satisfy the universalisability conditions of

the categorical imperative. In his theoretical philosophy, I am free to construct *any rules whatsoever* – as long as they satisfy the unity conditions. When confronted with a stream of raw sensory input, the Kantian agent constructs a set of connections between intuitions, a set of subsumptions mapping intuitions to concepts, and a set of judgements connecting concepts together. The agent is completely free to construct *any* set of connections between intuitions, *any* set of subsumptions, and *any* set of judgements – so long as the package jointly satisfies the unity conditions (Sections 2.3.1, 2.3.2, and 2.3.4). These conditions of unity are not unnecessary extraneous requirements that Kant insists on for some personal Puritan preference – they are the absolutely minimal conditions necessary for it to be *you* who is doing the constructing. According to Kant, the conditions that need to be satisfied to interpret the sensory input as a coherent

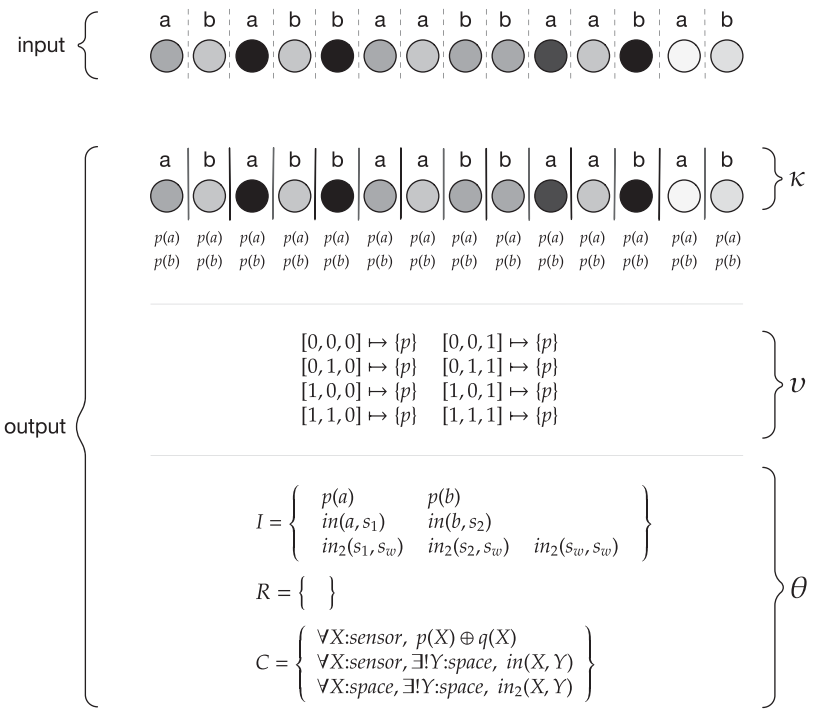


Figure 2.11: An alternative degenerate interpretation of the input of Figure 2.5. Here, all sensory input is mapped, indiscriminately, to *p*. Because no discriminations are made, and nothing changes, the induced theory is particularly simple. Note in particular that the set of dynamic rules *R* is empty, hence nothing changes.

representation of a single world are *exactly the same conditions* that need to be satisfied for there to be a *self* who is perceiving that world.[77]

Unlike the popular image, Kant's vision of the mind is one of remarkable freedom. I am continually constructing the program that I then execute. The only constraint on this spontaneous construction is the requirement that there is a single person looking out. In our computer implementation, this spontaneity is manifest in a particular way: when given a sensory sequence, the APPERCEPTION ENGINE constructs an unending sequence of increasingly complex interpretations, each of which satisfies Kant's unity conditions. The engine must decide, somehow, which of these interpretations to choose.[78]

4.2 Rigidity and Diachrony

Wittgenstein is sometimes interpreted as denying the possibility of any rule-based account of cognition. Throughout the *Investigations* (Wittgenstein, 2009), Wittgenstein draws our attention, again and again, to cases where our rules *give out*:

I say "There is a chair" What if I go up to it, meaning to fetch it, and it suddenly disappears from sight? – "So it wasn't a chair, but some kind of illusion". – But in a few moments we see it again and are able to touch it and so on. – "So the chair was there after all and its disappearance was some kind of illusion". – But suppose that after a time it disappears again – or seems to disappear. What are we to say now? *Have you rules ready for such cases* – rules saying whether one may use the word "chair" to include this kind of thing? But do we miss them when we use the word "chair"; and are we to say that we do not really attach any meaning to this word, because *we are not equipped with rules for every possible application of it*? (*Investigations*, § 80)

Our rules for the identification of chairs cannot anticipate every eventuality, including their continual appearance and disappearance – but this does not mean we cannot recognise chairs. Or, to take another famous example, we have rules for determining the time in different places on Earth. But now suppose someone says:

> It was just 5 o'clock in the afternoon on the sun (*Investigations*, § 351)

77 "The *a priori* conditions of a possible experience in general are at the same time conditions of the possibility of the objects of experience." [A111].

78 Our way of deciding between the various interpretations is based on the theory size and the fine-grainedness of the perceptual classifier. See (Evans et al., 2021a). This is one place where we attempt to go beyond Kant's explicit pronouncements, since he does not give us guidance here.

Again, our rules for determining the time do not cover all applications, and sometimes just *give out*. They do not cover cases where we apply the time of day on the sun. Since any set of rules is inevitably limited and partial, we must continually improvise and update.

This point is important and true, but is fully compatible with Kant's vision of the cognitive agent. Such an agent is *continually constructing* a new set of rules that makes best sense of its sensory perturbations. It is not that it constructs a set of rules, once and for all, and then applies them rigidly and unthinkingly forever after. Rather the process of rule construction is a continual effort.

Kant describes an *ongoing process* of constructing and applying rules to make sense of the barrage of sensory stimuli:

> There is no unity of self-consciousness or "transcendental unity of apperception" apart from this effort, or conatus towards judgement, *ceaselessly affirmed and ceaselessly threatened with dissolution* in the "welter of appearances" (Longuenesse, 1998, p.394)

Kant's apperceptual agent is continually constructing rules so as to best make sense of the barrage of sensory stimuli. If he were to cease constructing these rules, he would cease to be a cognitive agent, and would be merely a *machine*.

In *What is Enlightenment?* (Kant, 1784), Kant 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.

Some of Wittgenstein's remarks are often interpreted as denying the possibility of *any* sort of rule-based account of cognition:

> We can easily imagine people amusing themselves in a field by playing with a ball so as to start various existing games, but playing many without finishing them and in between throwing the ball aimlessly into the air, chasing one another with the ball and bombarding one another for a joke and so on. And now someone says: The whole time they are playing a ball-game and following definite rules at every throw. (*Investigations* §83)

Now there is a crucial scope ambiguity here. Is Wittgenstein merely denying that there is a set of rules that captures the ball-play at every moment? Or is he making a stronger claim, claiming that there is *some* moment during the ball-play that cannot be captured by *any set of rules at all*? I believe the weaker claim is more plausible: we make sense of the world by applying rules, but we need to continually modify our rules as we progress through time. Wittgenstein's passage in fact continues:

> And is there not also the case where we play and make up the rules as we go along? And there is even one where we alter them, as we go along.

Here, he does not consider the possibility of there being activity that cannot be explained by rules – rather, he is keen to stress the *diachronic* nature of the rule-construction process: one set of rules at one moment in time, a modified set of rules at a subsequent moment. Thus Wittgenstein's remarks on rules should not be seen as precluding any type of rule-based account of cognition, but rather as emphasising the importance of always being open to revising one's rules in the light of new information. As T. S. Eliot once observed:[79]

> For the pattern is new in every moment
> And every moment is a new and shocking
> Valuation of all we have been

4.3 Basic Assumptions

The APPERCEPTION ENGINE in its current form, and its limitations as described below, are a result of some fundamental decisions that were made early on in the project, answers to some basic questions about how to interpret and implement Kant:

1. When Kant says that every succession of determinations must be underwritten by a causal rule, does he mean that (i) there must be a causal rule that the agent believes? Or, much weaker, (ii) the agent must merely believe there *is* a causal rule?
2. When Kant says that judgements are rules, does he mean (i) explicit rules formed from discrete symbols? Or could he mean that some judgements are just (ii) implicit rules?
3. How expressive are Kant's judgements in the *Table of Judgements*? Does he just allow (i) simple definite clauses? Or does he also allow (ii) geometric rules (with disjunctions or existentials in the head)?

79 Four Quartets, *East Coker*.

4. Given that the understanding involves two separable capacities – the capacity to subsume intuitions under concepts and the capacity to combine concepts into rules – how should these two capacities be implemented? Should there be (i) one system that performs both, or (ii) two separate systems, with one passing its output to the other?

The design of the APPERCEPTION ENGINE was based on choosing option (i) at each of the four decision points. I shall attempt to justify each decision in turn.

4.3.1 Succession and Causal Rules

In the *Second Analogy*, Kant writes:

> If, therefore, we experience that something happens, then we always presuppose that something else precedes it, which it follows in accordance with a rule. [A195/B240]

Now this claim has a crucial scope ambiguity: does it mean that (i) whenever there is a succession there is a rule which the agent believes that underwrites the succession? Or does it mean that (ii) whenever there is a succession the agent believes that there is some rule that underwrites the successsion, even if the agent does not know what the particular rule is?

Some commentators have assumed the second, weaker interpretation. For example, Longuenesse believes that I do not have to have already formed a causal judgement to perceive a succession – I just need to acknowledge that I should form a causal judgement. For Longuenesse, perceiving a succession means *being committed to look for a causal rule* – it does not mean that I need to have *already found one*:

> The statement that "everything that happens presupposes something else upon which it follows according to a rule" does not mean that we cognize this rule, but that we are so constituted as to search for it, for its presupposition alone allows us to recognize a permanent to which we attribute changing properties. (Longuenesse, 1998, p.366)

Others, including Michael Friedman (Friedman, 1992) take the first, stronger interpretation.

I do not have the space or time to enter into the exegetical fray, but would like to make one observation. If we take the first, stronger interpretation, then any implementation of Kant's theory will be a system that can be used to predict future states, retrodict past states, and impute missing data. This ability to fill in the blanks in the sensory stream is only available because the agent *actually constructs rules* to explain the succession of appearances. If we had implemented

the second, weaker interpretation, then the agent would merely believe that there was *some* rule – it would not have been forced to find the rule, it would have been content to know that the rule existed somewhere. Such an agent would not be able to anticipate the future or reconstruct the past.

4.3.2 Explicit or Implicit Rules

When Kant says that judgements are rules, does he mean that judgements are (i) explicit rules formed from discrete symbols? Or could he mean that some judgements are just (ii) implicit rules (e.g., a procedure that is implicit in the weights of a neural network)?

The first interpretation, assuming judgements are explicit rules using discrete symbols in the language of thought,[80] is a form of what Brandom calls *regulism* (Brandom, 1994, p.18). The second interpretation allows for rules that are universal (they apply to all objects of a certain type), necessary (they apply in all situations), but *implicit*: the rule may not be expressible in a concise sentence in a natural or formal language. For a concrete example of the second interpretation, consider the *Neural Logic Machine* (Dong et al., 2019). This is a neural network that simulates forward chaining of definite clauses but without representing the clauses explicitly. The "rules" of the Neural Logic Machine are implicit in the weights (a large tensor of floating point values) of the neural network and cannot be transformed into concise human-readable rules. Nevertheless, the rules are universal and necessary, applying to all objects in all situations.

Most commentators believe that Kant's rules are explicit rules composed of discrete symbols.[81] I do not want to contribute to the exegetical debate, but rather want to provide a practical reason for preferring the first interpretation in terms of explicit rules. Part of the attraction of the APPERCEPTION ENGINE as described above is that the theories found by the engine can be read, understood, and verified. For example, the theory learned from the *Sokoban* trace is not just correct, but *provably* correct. If we need to understand what the machine is

80 In this project, I follow Jerry Fodor in assuming that our beliefs are expressed in a language of thought (Fodor, 1975) which is symbolic and compositional. Moreover, I assume that the language of thought is something like Datalog$^{\ni}$, but somewhat more expressive (Piantadosi, 2011).

81 But there is a note, inserted in Kant's copy of the first edition of the first Critique [A74/B99], which suggests that judgements need not be explicit: "Judgments and propositions are different. That the latter are *verbis expressa* [explicit words], since they are assertoric".

thinking, or need to verify that what it is thinking is correct, then we must prefer explicit rules.

Another, perhaps more fundamental, reason for preferring explicit rules is that they enable us to test whether Kant's unity conditions (see Section 2.4) have been satisfied. In order to test whether every succession is underwritten by a causal judgement (Section), for example, we need to be able to inspect the rules produced. It is unclear how a system that operates with merely implicit rules can detect whether or not Kant's unity conditions have actually been satisfied.

4.3.3 The Expressive Power of Kant's logic

Commentators disagree about the expressive power of Kant's judgements. Some think Kant's logic is restricted to Aristotelian syllogisms over judgements containing only unary predicates. If this were so, Kant's logic would indeed be "terrifyingly narrowminded and mathematically trivial"[82] Similarly, many commentators (for example, MacFarlane [42], p.26; also [55]) assume or claim that Kant's logic is highly restrictive in that it does not support nested quantifiers. Others[83] argue that Kant must have a more expressive logic in mind, a logic that includes at least nested quantifiers of the form $\forall\exists$.

There is, of course, a tradeoff between the expressiveness of the logic and the tractability of learning theories in that logic: the more complex the judgement forms allowed, the harder it is to learn. Geometric logic, for example, is highly expressive[84] but it is also undecidable (Bezem, 2005). Datalog, by contrast, is decidable, and has polynomial time data complexity dantsin 2001 complexity.

Because of this tradeoff, in this work we opted for a simpler logic (i.e. Datalog$^{\ni}$ rather than geometric logic) in order to make it tractable to synthesise theories in that logic. One of the central pillars of our interpretation is that Kant's fundamental notion of spontaneity is best understood as *unsupervised program synthesis*. To test out this claim, it was necessary to build a system that is capable of generating theories to explain a diverse range of examples. Thus, in this

82 (Hazen, 1999), quoted in (Achourioti and Van Lambalgen, 2011).

83 See in particular (Achourioti and Van Lambalgen, 2011; Achourioti et al., 2017), and also (Evans et al., 2019).

84 More generally, (Dyckhoff and Negri, 2015) shows that, for each set Σ of first-order sentences, there is a set of sentences of geometric logic that is a conservative extension of Σ.

Figure 2.12: Top-down influence from the symbolic to the sub-symbolic. Here the ambiguous image (the image used to represent both the 'H' of 'THE' and the 'A' of 'CAT') is disambiguated at the sub-symbolic level using knowledge (of typical English spellings) at the symbolic level.

project, we used an extension of Datalog to define a simple range of judgements. We do not claim that logic adequately represents the range of judgements expressible in Kant's *Table of Judgements*: after all Datalog$^{\ni}$ contains no negation symbol, no existential quantifier, and no modal operators. In future work we plan to extend this language with stratified negation as failure, disjunction in the head, and existential quantifiers, to increase its expressive power.

4.3.4 One System or Two?

The understanding involves two distinguishable capacities: the capacity to subsume intuitions under concepts (the power of judgement), and the capacity to combine concepts into rules (the capacity to judge). These two capacities take different sorts of input: the power of judgement takes raw intuitions and maps them to discrete concepts, while the capacity to judge operates on discrete concepts. This difference could suggest that we need a hybrid approach involving two distinct systems for the two capacities: one system (perhaps a neural network) for mapping intuitions to concepts and another (perhaps a symbolic program synthesis system) for combining concepts into rules. According to this suggestion, the output of the first system is fed as input to the second system.

A concern with this hybrid approach is that it is very unclear how to support top-down information flow from the conceptual to the pre-conceptual. There is much evidence that expectations from the conceptual symbolic realm can inform decisions at the pre-conceptual sub-symbolic realm. See, for example, Figure 2.11.[85] Here, part of the image is highly ambiguous: the 'H' of "THE" and the 'A' of "CAT" use the same ambiguous image, but we are able to effortlessly disambiguate (at the sub-symbolic level) by using our knowledge of typical English spelling at the symbolic level.

85 This example is adapted from (Chalmers et al., 1992).

Thus, it is essential that the high-level constraints – the conditions of unity – are allowed to inform the low-level sub-symbolic processing. This consideration precludes a two-tier architecture where a neural network transforms intuitions into concepts, and a symbolic system searches for unified interpretations. In such an architecture, it is not possible for the low-level neural network to receive the information it needs from the high-level system. The only information that the neural network will receive in a two-tier approach is a *single bit*: whether or not the high-level symbolic system was able to find a unified interpretation. It will not know *why* it was unable, or *which constraints* it was unable to satisfy. This is insufficient information.

Because of this concern, we opted for a different architecture, in which a *single system* jointly performed both tasks: both mapping intuitions to concepts and combining concepts into rules.[86]

4.3.5 Alternative Options

The particular design decisions taken in the APPERCEPTION ENGINE represent one way of answering the four questions above. But there are many other possible architectures. One option, for example, would be to represent the rules implicitly (Dong et al., 2019), and to use a single neural network to jointly learn to map intuitions to concepts and to learn the weights of the implicit rules. Another option would be to use a hybrid architecture in which a neural network, trained on gradient descent, maps intuitions to concepts, while another symbolic system combines concepts into rules. These alternative options have issues of their own, as I hope the discussion above makes clear, but the point remains that the APPERCEPTION ENGINE is certainly not the only way to implement Kant's cognitive architecture.

4.4 Moving Closer to a Faithful Implementation of Kant's *a priori* Psychology

This project is an attempt to repurpose Kant's *a priori* psychology as the architectural blueprint for a machine learning system, and as such has the real potential to irritate two distinct groups of people. AI practitioners may be irritated

86 Of course, our single system itself contains both a neural network mapping intuitions to concepts and a program synthesis component that constructs sets of rules. But this counts as a single architecture rather than a hybrid architecture because our binary neural network is implemented in ASP and the weights are found using SAT, rather than gradient descent.

by the appeal to a notoriously difficult eighteenth-century text, while Kant scholars may be irritated by the indelicate attempt to shoe-horn Kant's ambitious system into a simple computational formalism. The concern is that Kant's ideas have been distorted to the point where they are no longer recognisable.

In what ways, then, does the APPERCEPTION ENGINE represent a faithful implementation of Kant's vision, and in what ways does it fall short?

I shall focus, first, on the respects in which the computer architecture is a faithful implementation of Kant's psychological theory. Kant proposed various faculties that interoperate to turn raw data into experience: the imagination (to connect intuitions together using the pure relations as glue), the power of judgement (to decide whether an intuition falls under a concept), and the capacity to judge (to generate judgements from concepts). Throughout, Kant emphasized the spontaneity of the mind: the faculties are free to perform *whatever activity they like*, as long as the resulting system satisfies the various unity conditions described in the *Principles*.

The APPERCEPTION ENGINE provides a unified implementation of the various faculties Kant describes: the imagination is implemented as a set of non-deterministic choice rules, the power of judgement is implemented as a neural network, and the capacity to judge is implemented as an unsupervised program synthesis system. These sub-systems are highly non-deterministic: the imagination is free to synthesise the intuitions *in any way whatsoever*, the power of judgement is free to map intuitions to concepts *in any way it pleases*, and the capacity to judge is free to construct *any rules at all* – so long as the combined product of the three faculties satisfies the various unity conditions (implemented as *constraints*).

Thus, while contingent information flows bottom-up (from sensibility to the understanding), necessary information flows top-down, as the unity conditions of the understanding are the only constraints on the operations of the system. As Kant says: "through it [the constraint of unity] the understanding determines the sensibility [B160-1n]". This is, I believe, a faithful implementation of Kant's cognitive architecture at a high level.

Next I shall turn to the various respects in which the computer architecture described above falls short of Kant's ambitious vision of how the mind must work. I shall focus on six aspects of Kant's cognitive architecture that are not adequately represented in the current implementation.

4.4.1 The Representation of the Input

The way in which raw data is given to the APPERCEPTION ENGINE is different from how Kant describes it. Kant describes a cognitive agent receiving a *continuous* stream of information, making sense of each segment before receiving the next. The APPERCEPTION ENGINE, by contrast, is given the entire stream as a single unit. If the APPERCEPTION ENGINE is to operate with a continuous stream, it will have to synthesise a new theory *from scratch* each time it receives a new piece of information.

In the *A Deduction*, Kant describes three aspects of synthesis: the synthesis of apprehension in the intuition, the synthesis of reproduction in the imagination, and the synthesis of recognition in a concept. The synthesis of reproduction in the imagination involves the ability to recall past experiences that are no longer present in sensation. The APPERCEPTION ENGINE does not attempt to model the synthesis of reproduction. Rather, it assumes that the entire sequence is given.

The form of the raw data is also different from how Kant describes it. In Section 3.1, the raw data is provided as a sequence of *determinations*: assignments of raw attributes to persistent objects (sensors). Here, we assume that the agent is provided with the sensor, as a persistent object. But in Kant's architecture, the construction of determinations featuring persistent objects is a hard-won achievement – not something that is given. What *is* given, in Kant's picture, is the activity of sensing and the ability to tell when a particular sensing performed at one moment is the same sensing activity performed at another (the "unity of the action"). Thus, in Kant's picture, the agent is provided with a more minimal initial input than that given to our system, and so his agent has more work to do to achieve experience.

4.4.2 The Representation of Space and Time

The way space is represented in the APPERCEPTION ENGINE is different from how Kant describes it. For Kant, space is a single *a priori* intuition. He starts with space as a totality, and creates sub-spaces by division ("limitation" [A25/B39]). In the APPERCEPTION ENGINE, by contrast, we start with objects representing spatial regions, and compose them together using the containment structure (Section 2.3.1).

Similarly, with time, Kant starts with the original representation of the whole of time, and constructs sub-times by division [A32/B48]. In the APPERCEPTION ENGINE, by contrast, the sequence of time-steps are determined by the given input, and it is not possible for the system in its current form to construct

new moments of time that are intermediate between the given moments. Relatedly, it is not possible to represent continuous causality (e.g. water slowly filling a container) in our formalism. In future work, we plan to enrich Datalog$^{\exists}$ so that it can represent continuous change{\mdottt}

4.4.3 The Minimal Conception of Space

The Apperception Engine unifies objects by placing them in a containment structure: each object is in some spatial region which is itself part of some larger spatial region, until we reach the whole of space. In Section 2.3.1, I argued that this containment structure is a central component of any notion of space. But there is much more to spatial relations than the containment structure: just knowing that x and y are in does not tell us anything about the relative positions of x and y.

Kant had a much more full-blooded conception of space than just a containment structure: he assumed three-dimensional Euclidean space [B41]. In future work, I plan to provide the Apperception Engine with three-dimensional space,[87] thus providing a stronger inductive bias, which should help the system to learn more data-efficiently.

4.4.4 The Expressive Power of the Logic

In the *Transcendental Deduction*, Kant argued that the relative positions of intuitions in a determination can only be fixed by forming a judgement that necessitates this particular positioning [B128]. The Apperception Engine attempts to respect this fundamental requirement by insisting that the various connections between intuitions are backed up by judgements of various forms (Section 2.3.2). However, the forms of judgement supported in Datalog$^{\ni}$ are a mere subset of the forms enumerated in the *Table of Judgements* [A70/B95]. Datalog$^{\ni}$ supports universally quantified conditionals, causal conditionals, and xor constraints (corresponding to Kant's disjunctive judgement). But it does not support negative judgements, infinite judgements, particular judgements, singular judgements, or

87 Perhaps by providing an axiomatisation of Euclidean space using Tarski's formalisation, or somesuch (but note that axiomatising Euclidean geometry requires ternary predicates, which are not currently handled in the Apperception Engine). But Tarski assumes points as primitive, where a point is defined as a vector of real numbers. It would be closer to Kant's program, I believe, to axiomatise space starting from the notion of *limitation*, without assuming real numbers as given.

modal judgements. In future work, we plan to extend the expressive power of Datalog$^{\ni}$ to capture the full range of propositions expressible in the *Table of Judgements*.[88]

4.4.5 The Role of the Third Analogy

The *Third Analogy* states that whenever two objects' determinations are perceived as simultaneous, there must be a two way interaction between the two objects. This does not mean, of course, that there must be a direct causal influence between them, but just that there must be a chain of indirect causal influences between them.

This requirement has not been implemented in the APPERCEPTION ENGINE. This is because it would make it very hard for the system to find any unified interpretation at all if every time it posited a simultaneity between determinations it also had to construct some rules whereby one determination of one object indirectly caused some determination of the other object. Longuenesse (Longuenesse, 1998) has a different understanding of the second and third *Analogies*, and does not believe that we need to *have actually formed a causal rule* in order to perceive succession or simultaneity. In her interpretation, we merely need to *believe that there is a causal rule to find* (see Section for a discussion). However, in our interpretation, in which the rule must actually be found before a temporal relation can be assigned, the *Third Analogy* does seem restrictively strong. In future work, we hope to address this issue and find a way to respect the simultaneity constraint.

4.4.6 Consciousness and Analytic Unity

The first *Critique* contains various discussions of various aspects of self-consciousness. But no aspect of self-consciousness is implemented in the APPERCEPTION ENGINE. In the *B Deduction*, Kant distinguishes the synthetic unity of apperception (the connecting together of one's intuitions via the pure relations in such a way as to achieve unity) from the analytic unity of apperception (the ability to subsume any of my cognitions under the predicate "I think"). He claims that synthetic unity of apperception is a necessary condition for achieving analytic

88 By contrast, the geometric logic used in (Achourioti and Van Lambalgen, 2011; Achourioti et al., 2017) is much more expressive.

unity [B133-4]. Although the APPERCEPTION ENGINE aims to implement the synthetic unity of apperception, no attempt has been made to implement the analytic unity of apperception.

Kant is clear to distinguish between inner sense and explicit self-consciousness [B154]. Inner sense is the aspect of sensibility in which the mind perceives its own mental activity: it notices the formation of a belief, for example, or the application of a rule. Inner sense provides us with intuitions that must be ordered in time. Explicit self-consciousness, by contrast, is the construction of a theory that makes sense of the sequence of perturbations produced by inner sense. In inner sense I become aware of some of the cognitions I am having, and in explicit self-consciousness, I posit a theory that explains the dynamics of my own mental activity – although this hypothesized theory may or may not reflect accurately the actual mental processes I am undergoing [B156]. In future work, I plan to extend APPERCEPTION ENGINE so that (some of) its own activity is perceptible via inner sense, so that the system is forced to construct a theory to make sense of its perceptions of its own mental activity.

There are, then, various aspects of Kant's theory of mental activity that are not captured in the current incarnation of the APPERCEPTION ENGINE. There is, I think it is fair to say, more work still to do.

5 Conclusion

In the *Critique of Pure Reason*, Kant asks: what activities must be performed by an agent – *any* finite resource-bounded agent – if it is to make sense of its sensory input. This is not an empirical question about the particular activities that are performed by *homo sapiens*, but an *a priori* question about the activities that any agent must perform. Kant's answer, if correct, is important because it provides a blueprint for the space of *all possible minds* – not just our particular human minds with their particular human foibles.

If Kant's cognitive architecture is along the right lines, this will have significant impact on how we should design intelligent machines. Consider, to take one important recent example, the data efficiency of contemporary reinforcement learning systems. Recently, deep reinforcement learning agents have achieved super-human ability in a variety of games, including Atari (Mnih et al., 2013) and Go (Silver et al., 2017). These systems are very impressive, but also very data-inefficient, requiring an enormous quantity of training data. DQN (Mnih et al., 2013) requires 200 million frames of experience before it can reach human performance on Atari games. This is equivalent to playing non-stop for 40 days.

AlphaZero (Silver et al., 2017) played 44 million games to reach its performance level.

Pointing out the sample complexity of these programs is not intended to criticise these accomplishments in any way. They are very impressive achievements. But it does point to a fundamental difference between the way these machines learn to play the game, and the way that humans do. A human can look at a new Atari game for a few minutes, and then start playing well. He or she does not need to play non-stop for 40 days. A human's data efficiency at an Atari game is a consequence of our inductive bias: we start with prior knowledge that informs and guides our search.

It is a commonplace that the stronger the inductive bias, the more data-efficiently a system can learn. But the danger, of course, with injecting inductive bias into a machine, is that it biases the system, enabling it to learn some tasks quicker, but preventing it from learning other tasks effectively. What we really want, if only we can get it, is inductive bias that is maximally general. But what are these maximally general concepts that we should inject into the machine, and how do we do so?

Neural net practitioniers, for all their official espousal of pure empiricist anti-innatism, do (in practice) acknowledge the need for certain minimal forms of inductive bias. A convolutional net (LeCun et al., 1995) is a particular neural architecture that is designed to enforce the constraint that the same invariants hold no matter where the objects appear in the retinal field. A long short-term memory (LeCun et al., 1995) is a particular neural architecture that is designed to enforce the constraint that invariants that are valid at one point in time are also valid at other points in time. But these are isolated examples. *What, then, are the maximally general concepts that we should inject into the machine, to enable data efficient learning?*

The answer to this question has been lurking in plain sight for over two hundred years. In the first *Critique*, Kant identified the maximally general concepts, showed how these concepts structure perception itself, and identified the conditions specifying how the pure concepts interoperate. Kant's principles provide the maximally general inductive bias we need to make our machines data-efficient.[89]

> In the history of human inquiry, philosophy has the place of the central sun, seminal and tumultuous: from time to time it throws off some portion of itself to take station as a science, a planet, cool and well regulated, progressing steadily towards a distant final state. – Austin, *Ifs and Cans* (Austin, 1956)

89 Thanks to Dieter Schönecker and Sorin Baiasu for thoughtful feedback.

References

Achourioti, T. and Van Lambalgen, M. (2011). A formalization of Kant's transcendental logic. *The Review of Symbolic Logic*, 4(2):254–289.

Achourioti, T., van Lambalgen, M., et al. (2017). Kant's logic revisited. *IfCoLog Journal of Logics and Their Applications*, 4:845–865.

Allais, L. (2009). Kant, non-conceptual content and the representation of space. *Journal of the History of Philosophy*, 47(3):383–413.

Austin, J. L. (1956). Ifs and cans. *Proceedings of the British Academy*.

Bezem, M. (2005). On the undecidability of coherent logic. In *Processes, Terms and Cycles: Steps on the Road to Infinity*, pages 6–13. Springer.

Brandom, R. (1994). *Making It Explicit*. Cambridge, MA: Harvard University Press.

Brandom, R. (2009). How analytic philosophy has failed cognitive science. *Towards an Analytic Pragmatism (TAP)*, pages 121–133. Proceedings of the Workshop on Bob Brandom's Recent Philosophy of Language: Towards an Analytic Pragmatism (TAP-2009). Genoa, Italy, April 19–23, 2009. Edited by Cristina Amoretti, Carlo Penco, Federico Pitto.

Brandom, R. B. (2008). *Between Saying and Doing*. Oxford: Oxford University Press.

Chalmers, D. J., French, R. M., and Hofstadter, D. R. (1992). High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental & Theoretical Artificial Intelligence*, 4(3):185–211.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204.

Dantsin, E., Eiter, T., Gottlob, G., and Voronkov, A. (2001). Complexity and expressive power of logic programming. *ACM Computing Surveys (CSUR)*, 33(3):374–425.

Dennett, D. C. (1978). Artificial intelligence as philosophy and as psychology. *Brainstorms*, pages 109–26. MIT Press.

Dong, H., Mao, J., Lin, T., Wang, C., Li, L., and Zhou, D. (2019). Neural logic machines. *arXiv preprint arXiv:1904.11694*.

Dyckhoff, R. and Negri, S. (2015). Geometrisation of first-order logic. *Bulletin of Symbolic Logic*, 21(2):123–163.

Evans, R. (2017). Kant on constituted mental activity. *APA on Philosophy and Computers*, 16(2): 41–53.

Evans, R. (2019). A Kantian cognitive architecture. In *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, pages 233–262. Springer.

Evans, R. (2020). *Kant's cognitive architecture*. PhD thesis, Imperial College London.

Evans, R., Bošnjak, M., Buesing, L., Ellis, K., Pfau, D., Kohli, P., and Sergot, M. (2021a). Making sense of raw input. *Artificial Intelligence*, 299:103521.

Evans, R. and Grefenstette, E. (2018). Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research (JAIR)*, 61:1–64.

Evans, R., Hernandez-Orallo, J., Welbl, J., Kohli, P., and Sergot, M. (2020). Evaluating the Apperception Engine. *arXiv preprint arXiv:2007.05367*.

Evans, R., Hernández-Orallo, J., Welbl, J., Kohli, P., and Sergot, M. (2021b). Making sense of sensory input. *Artificial Intelligence*, 293:103438.

Evans, R., Sergot, M., and Stephenson, A. (2020). Formalizing Kant's rules. *Journal of Philosophical Logic*, 49(4):613–680.

Fodor, J. A. (1975). *The Language of Thought*. Cambridge, MA: Harvard University Press.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1–2):3–71.

Friedman, M. (1992). *Kant and the Exact Sciences*. Cambridge, MA: Harvard University Press.

Gomes, A. (2013). Kant on perception: Naive realism, non-conceptualism, and the B-deduction. *The Philosophical Quarterly*, 64(254):1–19.

Hazen, A. P. (1999). Logic and analyticity. In *The Nature of Logic*, pages 79–110. CSLI.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8): 1735–1780.

Hofstadter, D. R. (1995). *Fluid Concepts and Creative Analogies*. New York, NY: Basic Books.

Hofstadter, D. R., Mitchell, M., et al. (1994). The copycat project: A model of mental fluidity and analogy-making. *Advances in connectionist and neural computation theory*, 2(31–112):29–30.

James, W., Burkhardt, F., Bowers, F., and Skrupskelis, I. K. (1890). *The Principles of Psychology*, volume 1. Macmillan London, Transactions of the Charles S. Peirce Society, 19(2).

Jaynes, J. (2000). *The Origin of Consciousness in the Breakdown of the Bicameral Mind*. Boston, MA: Houghton Mifflin Harcourt.

Kant, I. (1781). *Critique of Pure Reason*. Cambridge University Press.

Kant, I. (1784). What is enlightenment? In *Practical Philosophy*, pages 11–22. Cambridge University Press. Translated and edited by Mary J Gregor.

Kant, I. (1790). *Critique of the Power of Judgment*. Cambridge: Cambridge University Press.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40. Cambridge: Cambridge University Press.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, 3361(10):1995.

Longuenesse, B. (1998). *Kant and the Capacity to Judge*. Princeton, NJ: Princeton University Press.

Marcus, G. (2018a). *The Algebraic Mind*. Cambridge, MA: MIT press.

Marcus, G. (2018b). Innateness, AlphaZero, and artificial intelligence. *arXiv preprint arXiv:1801.05667*.

McLear, C. (2016). Kant on perceptual content. *Mind*, 125(497):95–144.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing Atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Murphy, K. P. (2012). *Machine Learning: a Probabilistic Perspective*. MIT press.

Piantadosi, S. (2011). *Learning and the Language of Thought*. PhD thesis, Massachusetts Institute of Technology.

Pinosio, R. (2017). *The Logic of Kant's Temporal Continuum*. PhD thesis, University of Amsterdam.

Rocktäschel, T. and Riedel, S. (2016). Learning knowledge base inference with neural theorem provers. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*, pages 45–50.

Sellars, W. (1967). Some remarks on Kant's theory of experience. In *In the Space of Reasons*, pages 437–453. Cambridge, MA: Harvard University Press.

Sellars, W. (1968). *Science and Metaphysics*. Oxfordshire: Routledge.

Sellars, W. (1978). The role of imagination in Kant's theory of experience. In *In the Space of Reasons*, pages 454–466. Harvard University Press.

Shanahan, M. (2005). Perception as abduction. *Cognitive science*, 29(1):103–134.
Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359.
Spelke, E. S. and Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, 10(1):89–96.
Stephenson, A. (2013). *Kant's Theory of Experience*. PhD thesis, University of Oxford.
Stephenson, A. (2015). Kant on the object-dependence of intuition and hallucination. *The Philosophical Quarterly*, 65(260):486–508.
Stephenson, A. (2017). Imagination and inner intuition. *Kant and the Philosophy of Mind*.
Strawson, P. (2018). *The Bounds of Sense*. Routledge.
Waxman, W. (2014). *Kant's Anatomy of the Intelligent Mind*. Oxford: Oxford University Press.
West, D. B. et al. (2001). *Introduction to graph theory*, volume 2. Hoboken, NJ: Prentice hall Upper Saddle River.
Wittgenstein, L. (2009). *Philosophical Investigations*. Oxford: John Wiley & Sons.
Wolff, R. (1963). *Kant's Theory of Mental Activity*. Cambridge, MA: Harvard University Press.