Chapter 13
A Kantian Cognitive Architecture



Richard Evans

Abstract In this paper, I reinterpret Kant's *Transcendental Analytic* as a description of a cognitive architecture. I describe a computer implementation of this architecture, and show how it has been applied to two unsupervised learning tasks. The resulting program is very data efficient, able to learn from a tiny handful of examples. I show how the program achieves data-efficiency: the constraints described in the *Analytic of Principles* are reinterpreted as strong prior knowledge, constraining the set of possible solutions.

Keywords Kant · Critique of pure reason · Rule induction · Unsupervised learning · Data efficiency · Cognitive architecture · Computational modeling · Original intentionality · Cognitive agency

13.1 Introduction

In this paper, I shall reinterpret part of Kant's *Critique of Pure Reason* as a specification of a cognitive architecture. I will describe a computer implementation of this architecture, and show how this program has been applied to an open problem in AI.

Now this project may seem, on the face of it, absurd: why should a book written in the eighteenth century have anything to teach us now? I will argue that this is not as unpromising as it might, at first, appear. Kant still has something to teach us. His insights have not yet been fully absorbed into cognitive science or AI.

I shall describe two Kant-inspired computer programs that are able to perform unsupervised learning from a tiny handful of examples. Now the ability to learn

R. Evans (✉)
Imperial College London, London, UK

DeepMind, London, UK
e-mail: RichardPrideauxEvans@imperial.ac.uk

© Springer Nature Switzerland AG 2019

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_13

from a handful of data, often called *data-efficiency*, requires strong prior knowledge. But if strong priors encode *domain-specific* information, then our machine-learning system is parochial, tied to a particular subject matter. What we really want, if only we could get it, is a set of strong priors that are also *domain-independent*.

In this paper, I shall show how the constraints described in Kant's *Analytic of Principles* can be reinterpreted as a set of strong, domain-independent priors: a set of constraints on any agent that is trying to turn its sensory input into a coherent unified experience of an external world. First, in Sect. 13.2, I shall pick out a particular argumentative strand that runs through the first half of the First Critique. I shall describe a set of conditions that, Kant holds, must be true of any agent that is able to make sense of its sensory data. Next, in Sect. 13.3, I shall reinterpret these Kantian claims as a specification of a cognitive architecture. The Kantian constraints are translated into constraints on the types of rules generated by a rule induction system. Finally, in Sect. 13.4, I shall describe two applications of this architecture. In one, the agent is placed in a simple two-dimensional grid-world, and must make sense of the sensory data he receives. In the second application, the agent's sensory input is a one-dimensional string of symbols. Making sense of these strings of symbols amounts to solving a standard verbal reasoning task. Surprisingly, the Kantian agent is able to achieve human-level performance in this verbal reasoning task, with no prior training data and no hand-engineered feature recognition.

13.2 Original Intentionality Via Synthetic Unity

I define a **sensory agent** as some sort of animal or device, equipped with sensors, whose actions depend on the state of its sensors. It might have a temperature gauge, a camera with limited resolution, or a sonar that can detect distance. The sensory agent is continually performing what roboticists call the *sense-act cycle*: it detects changes to its sensors, and responds with bodily movements.

A thermostat, for example, is a simple sensory agent. When it notices that the temperature has got too low, it responds by increasing the temperature. Now although the thermostat has a sense-act cycle, it does not experience[1] the world it is responding to. *We* count the perturbations of its gauge as representations of the temperature in the room it is in, but *it* does not. The gauge movements count as temperature representations *for us*, but not *for the thermostat*. Nothing counts as anything for the thermostat. It just responds *blindly*.

The thermostat does not, in other words, have *original intentionality* (Haugeland 1990). *We* might interpret some of its activities as representations, but *it* does not.

[1]In making this claim, I am assuming a suitably red-blooded notion of "experience". Of course, for some sufficiently thin notion of "experience", the thermostat must "experience" the world in order to act at all. But there is a difference between merely responding to a stimulus and *making sense* of that stimulus: reinterpreting the stimulus as a representation of a coherent external world. The latter is "experience" in the strong sense I am using it.

We can distinguish between derivative and original intentionality using the activity of counting-as[2]:

- x has derivative intentionality in representing p if an agent y (distinct from x) counts x's activity as x's representing p
- x has original intentionality in representing p if x himself counts x's activity as x's representing p

What distinguishes an agent with original intentionality, a **cognitive agent**, from a mere sensory agent is that the former *counts its own sensings as* representations of a determinate external world. It interprets its own sensory perturbations as a representation of a coherent unified world of external objects, interacting with each other.

One of Kant's fundamental questions is:

> *What does a sensory agent have to do, in order for it to count its own sensory perturbations as experience, as a representation of an external world?*

What, in other words, must a sensory agent do to be a cognitive agent?

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

> Given a set of beliefs, what else has to be true of him for us to count his beliefs as knowledge?

Kant's question is *pre-epistemological*: he does not assume the agent is "given" a set of beliefs. Instead, we see his beliefs 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* (Kant 1781)(A199, B244-5)[3]

Kant asks for the conditions that must be satisfied for the agent to have any possible cognition (true *or* false) (Kant 1781)(A158, B197). Kant's question, in the first person, is:

> *What do I have to do, in order to count these sensory perturbations as my experience?*

His answer, roughly, is:

> I count this plurality of sensings as my experience if I combine them together in the right way

[2]Note that I am not *defining* intentionality in terms of the activity of counting-as (which would be uninformative). Rather, I am using counting-as to *distinguish* between original and derivative intentionality. Later, counting-as will itself be explicated in terms of the construction and application of rules.

[3]All such references [A, B] are to the A and B editions of the *Critique of Pure Reason*, (Kant 1781).

What, then, does Kant mean by "combine", and what does he mean by "the right way"?

First, in Sect. 13.2.1, I will describe what Kant means by "combine". To anticipate, there are two types of combination, achieved by applying two types of rules (rules of composition and rules of connection). Second, in Sect. 13.2.2, I will describe what Kant means by the "right way". To anticipate, combining in the right way means connecting the cognitions together via certain relations, so that the plurality of cognitions becomes a totality connected in time. The constraints described in the *Analytic of Principles* are constraints on the construction and application of rules that Kant claims are severally necessary and jointly sufficient for an agent to construct a coherent representation. In Sect. 13.2.3, I will outline the general argument structure that underlines each of the four *Principles*.

13.2.1 The Basic Activity of Combination

The activity at the heart of Kant's theory is the mental act of **combination**, of bringing cognitions together, "running through and holding together this manifoldness" (Kant 1781)(A99). Kant explains what he means by "combination" in Kant (1781)(B201n): I combine a plurality of cognitions together when I subsume then under a "mark". Kant says little about what a "mark" is, given its load-bearing role in his theory. *"Merkmal"* is typically translated as "mark", but it can also be translated as "feature". Kant's mark is not a shared linguistic symbol. It is rather what computer scientists call a "gen-sym": a generated symbol, an atomic identifier. When it is first created, a mark is just an uninterpreted symbol. But by constructing inferential rules that relate this mark to others, the agent can elevate it into a concept.

Combining, then, is subsuming cognitions under a mark. For example: if this configuration of sensors is turned on, then I count their being-on as representing a nose. Or: if this pattern of sensors counts as representing a nose, and this other pattern counts as representing an eye, then the aggregate pattern of sensors counts as representing a face.

13.2.1.1 Combination Can Only Be Performed Indirectly Via the Construction and Application of Rules

Although this combining activity is fundamental, it cannot, according to Kant, be performed directly by the agent. The agent cannot just bring representations together *willy-nilly*.[4] Combining is not something he can *just do*. On the contrary, the *only way*, according to Kant, that the agent can perform the activity of combination

[4]By "willy-nilly", I mean without justification from the application of a rule. Kant's view is that the only mental actions that are justified are actions that result from applying a rule. What leaves room in this stern vision for spontaneity and autonomy is that the rules are not imposed from outside; rather, they are self-legislated.

is indirectly, by *applying general rules*[5] *that it has constructed*. This is Kant's surprising claim.

In a revealing footnote [B201n], Kant distinguishes between two types of rule of combination. **Rules of composition** are rules for combining parts into wholes, producing a part-whole graph united under one element: the totality. A rule of composition produces, if it applies, a defeasible *permission*[6] for the agent to group intuitions together under a mark. For example, if you count this group of sensings as representing an ear, and this group of sensings as representing a nose, then you may count this aggregate group of sensings as representing a face. Whether or not the rule-following agent makes use of this permission will depend on his concomitant commitments.

Rules of composition are described by *defeasible* conditionals. Wittgenstein stresses the defeasibility of such conditionals when discussing what counts as a friendly face:

> When we notice the friendly expression of a face, our attention, our gaze, is drawn to a particular feature in the face, the 'friendly eyes', or the 'friendly mouth etc. . . . *It is true that other traits in this face could take away the friendly character of this eye*, and yet in this face it is the eye which is the outstanding friendly feature (Wittgenstein 1958) (p.145–146)

This is defeasibility in action: in this situation, the features of this eye counts as his having a friendly face; but in another situation, the very same features plus some other additional facial features might count as something entirely different – mocking cruelty, for instance.

The second type of rule is a necessary rule that must be applied when it can be applied. **Rules of connection** produce *obligations* to group representations under a mark.[7] So, for example, if we count this structure as a nose, then we *must* also count it as a facial part – and if we count it as a nose, then we *must not* count it as an ear.

Kant's striking claim is that the mental act of combination is not a self-sufficient action, something you can *just do* – rather, you can only do it by applying these two types of rules. We are used to thinking of *social* activity as constituted: moving your knight to king's bishop three is something you can only do indirectly *by doing something else* – by pushing a wooden object in a certain direction. Similarly, requesting Bob to shut the door is not something you can *just do*: you can only do it by doing something else – perhaps by uttering a sequence of sounds, or by pointing at the door; there are an infinite number of different actions that could constitute

[5]Please note that these Kantian rules do not have to be linguistically articulated or consciously accessible. Rather, the rules that determine the activities of mental combination are implicit and consciously inaccessible, in the same way that the rules of a compiled Prolog program are inaccessible to the executing process.

[6]See Kant (1781)(B201n): "the synthesis of a manifold of what does *not necessarily* belong to each other".

[7]See Kant (1781)(B201n): "the second combination is the synthesis of that which is manifold insofar as they *necessarily* belong to one another".

such a request, but you have to do *one* of them – requesting is not something primitive you can do on your own. But we are not so used to thinking of fundamental *mental* activity as similarly constituted.

All the agent can do is *construct* general rules of the above form, permitting or obligating him to combine representations in a certain way, and then *apply* these rules, thus *indirectly* performing combinations via the construction and application of rules. This claim appears throughout the First Critique[8]:

> everything (that can even come before us as an object) necessarily stands under rules, since, without such rules, appearances could never amount to cognition of an object [B198, A159]

Why can't a cognitive agent perform the activity of combination *directly*, without needing to construct and then apply a rule? We will see why this is so in Sect. 13.2.2. The basic reason, to anticipate, is that combining without rules would not satisfy the condition of *unification* at the heart of K's theory. The unification condition is a set of constraints on the construction and application of *rules*, and so can *only* be applied to a rule-following and rule-constructing agent. Arbitrary combination of cognitions that was unguided by rules would not produce a unity of experience that I could call *mine*; instead, the combined representations would be a "mere play", "less even than a dream" (Kant 1781)(A112). If I could combine representations into intuitions without rules, then there would be no *self* to have the intuitions.

The Kantian rule-following agent is continually constructing the very software that it will then execute.[9] It is always constructing rules, and then interpreting those rules. In fact, the *only* way that it can perceive *anything* is by applying rules it has already constructed in order to make sense of the incoming barrage of sensations.[10]

The Kantian rule-follower, then, is a norm-giving agent who solemnly sets down rules that he will then obediently follow. He only allows himself to perform acts of mental combination if these acts are shown to be permitted by rules he has previously constructed.

[8] See also [A105], [A177, B220].

[9] In computational terms, think of a meta-interpreter that is able to construct pieces of code as *data*, and then execute these new pieces of code.

[10] Kant makes the same point in the *Metaphysical Deduction*: "The same function that gives unity to the different representations *in a judgement* also gives unity to the mere synthesis of different representations *in an intuition*, which, expressed generally, is called the pure concept of the understanding. The same understanding, therefore, *and indeed by means of the very same actions* through which it brings the logical form of a judgement 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" (Kant 1781)(A79, B104-5). In other words, there is only *one* process (a process of constructing and applying rules) which explains *both* how we form judgements *and* how we form intuitions.

13.2.1.2 Constructing and Applying Rules

The rule-following agent can perform two types of activity: he can construct a rule, and he can apply a rule he has already constructed. Kant says it is the job of the faculty of understanding to construct rules:

> Sensibility gives us forms (of intuition), but the understanding gives us rules. It is always busy poring through the appearances with the aim of finding some sort of rule in them. (Kant 1781)(A126)

Recall that there are two types of rule (rules of composition, and rules of connection), so there are two types of rule construction. Constructing rules of composition is forming perceptual rules, rules of apprehension for counting particular configurations as parts of objects. For example, the agent adds a new rule that, if some of its sensors satisfy such and such a condition, it may count them as representing an ear.

Constructing rules of connection is forming concepts or making judgements. Forming a concept is constructing a set of rules that describe the inferential connections between this concept and others. So, for example, to form the concept of "tree", we need rules of composition for saying under what conditions a set of sensury perturbations count as representing a tree. But we also need rules of connection for linking this concept with others. For example, if we count it as a tree, we must also count it as a plant.[11]

Making a judgement is also constructing a rule of connection. If we form the judgement that "All men are mortal", this is just to adopt the rule of connection: if I count a cognition as a man, then I must also count it as mortal. But this inferential understanding of judgement applies to categorical statements just as much as to hypothetical statements. To form the judgement that "Caesar is a general" just is to adopt the rule: if I count a cognition as Caesar, then I must also count it as a general (Longuenesse 1998).

This is why Kant says (Kant 1781(A126)) that the faculty of constructing rules is also the faculty of concept-formation and judging: both concept-formation and judging are just special cases of the more general ability to construct rules.

Next, I shall turn to the process of *applying* the rules that the understanding has constructed. If the rule applies in a particular situation, a norm is operative: either the agent must combine the representations under a certain mark (if the rule is a rule of connection), or it may do so (if it is a rule of composition). If it is a rule of composition, then all the agent knows is that he *may* perform the combination activity – he does not *have* to do so. Consider, for example, Jastrow's famous duck-rabbit (Fig. 13.1). Focus on the lines on the left of the image. Now we have two rules

[11] Some of the connection rules involved in characterising a concept do more than simply state that one concept is a sub-concept of another, or that one concept excludes another. Some of them relate the concept to another concept only conditionally – dependent on the existence of external factors. For example: "If the weather gets cold, trees lose their leaves", "If a tree gets no water, it perishes" (Longuenesse 1998). Some of the conceptual inference rules, in Kantian terms, are hypothetical rather than categorical.

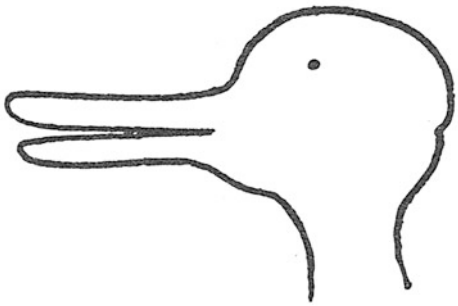


Fig. 13.1 Jastrow's duck-rabbit

of composition that apply to these lines: we can count these lines as a mouth, or as a pair of ears. Now there is a rule of connection that prevents us from applying both: if something is a mouth, then it is not a pair of ears. We may apply either rule of composition – but we must not apply both. What makes us decide which to apply?

Kant argues convincingly that it cannot be a further *rule* that tells us which to apply. If we needed rules to determine which rules to apply, then those determining rules would themselves need further rules to determine their application, and so on, generating a vicious regress (Kant 1781)(A133, B172).

Kant defines the *imagination* as the faculty responsible for **applying** the rules that the understanding has constructed. As the duck-rabbit picture shows, the imagination has some *choice* about how to apply the rules of composition (Kant 1781)(B151). This is why Kant says that both understanding and imagination involve *spontaneity* – the understanding has a choice about which rules to construct; and then, once it has constructed them, the imagination has a further choice about which rules of composition to apply:

> It is one and the same spontaneity that, there under the name of imagination and here under the name of understanding, brings combination into the manifold of intuition (Kant 1781)(B162n)

Note that it is only when applying rules of *composition* that the imagination has a choice about which to apply. When it comes to applying rules of *connection*, the rule-following agent is *obligated* to perform the required mental activity.

To summarise, a rule-following agent is a type of sensory agent who can combine representations by constructing and applying rules. Given that there are two types of activity (constructing and applying rules), and two types of rule (rules of composition and rules of connection), we have a square of operations.

	Rules of composition	Rules of connection
Constructing	Forming perceptual rules	Forming concepts and judgements
Applying	Forming intuitions	Inferring properties of objects

Recall our original question:

What must I do, in order to count these sensory perturbations as my experience?

The rule-following agent is a central part of Kant's answer:

- A sensory agent is a cognitive agent if he counts his sensings as representing an external world
- He counts these sensings as representing an external world if he combines those sensings together in the right way
- He combines his sensings together in the right way if he constructs and applies a set of rules that satisfy a set of (as yet unspecified) *constraints*

The next question, then, is: *what set of constraints on the construction and application of rules are severally necessary and jointly sufficient for counting this plurality of sensory perturbations as representing an external world?* The next section will describe the constraints involved.

13.2.2 Combining in the "Right Way"

The constraints on the activity of combination are specified in the *Analytic of Principles*. The argument justifying these particular constraints is spread through the *Transcendental Deduction*, the *Schematism*, and the *Principles*:

1. Counting this plurality of sensory perturbations as *my experience* requires *connecting* the representations together
2. Connecting the representations together requires bringing the representations under a *relation*
3. The only medium that can connect all my representations is *time*
4. Connectedness in time involves four activities[12]:
   (a) constructing moments in time
   (b) filling time
   (c) ordering time
   (d) constructing the totality of time

I shall go through these points in turn.

13.2.2.1 If I am to Make them Mine, I Must Connect them Together

For the representations to be mine, I must *make* them united:

> Combination does not lie in the objects, however, and cannot as it were be borrowed from them through perception and by that means first taken up into the understanding, but is rather only an *operation* of the understanding (Kant 1781)(B134-5)

[12]These activities are described in Kant (1781)(B185, A146) .

Nobody is going to unify my sensings for me – it is an *activity I must perform myself* if I am to experience anything at all. I must *work* to bring my representations together into a unity. Kant calls this requirement the "supreme principle of human cognition".

13.2.2.2 Connecting Cognitions Together Requires Uniting Them Under Relations

The only way we can connect cognitions together is by placing them under relations. Connecting the cognitions via relations is the only way "to make them fit for a thoroughgoing *connection* in one experience" (Kant 1781)(B185, A146).

13.2.2.3 The Only Medium That Can Connect All My Representations Is Time

Our representations are connected by bringing them together under certain relations. Some of my representations are intuitions about the external world, and others are thoughts about my own inner states (beliefs, pain, etc). Now although I can place my outer intuitions in space, my inner intuitions are not spatially-located. My inner intuitions are ordered in time, but not in space. The only medium in which I can place *all* my intuitions is time:

> There is only one totality in which all of our representations are contained, namely inner sense and its a priori form, time. (Kant 1781)(B194, B155)

So the *only* marks that can relate *all* my representations are relations involving time:

> Time is the formal condition of the manifold of inner sense, thus of the connection of all representations (Kant 1781)(A138, B177)

13.2.2.4 Connecting my Representations in Time Requires Four Activities of Time-Determination

Kant lists four activities of **time-determination** which he claims are necessary and sufficient for unifying all representations in time (Kant 1781)(B184, A145). The first activity is *constructing moments in time*: constructing successive apprehensions[13] of an object, and representing a moment in time as a collection of simultaneous apprehensions of objects, apprehensions that are organised in a part-whole totality. The second aspect of time-determination is *filling time*: determining cognitions with sufficient fine-grainedness that any intermediate moment in time can be constructed.

[13] I use the Kantian term **apprehension** to denote a time-slice of an enduring object at a particular moment in time. Throughout, I use "apprehension" and "object-slice" interchangeably.

Between any two moments, the agent can construct an intermediate moment. The third aspect is *ordering time*: placing moments of time in a determinate order, a total ordering on moments. The fourth aspect of time-determination is *constructing the totality of time*: specifying which of the various candidate combinations of representations are coherent moments in time, and which are impossible.

13.2.3 The General Structure of the Four Principles

The constraints described in the *Analytic of Principles* are constraints on the construction and application of rules needed to satisfy the four activities of time-determination above.

First, I shall lay out the common argument structure behind each principle before addressing them individually.

Let us distinguish, following Sellars (1968), between:

- (A) **representeds**: the things that are represented
- (B) **representings**: activities that are *of*, or *about* representeds
- (C) **rules**: general procedures that apply in many situations; when a rule is applied, it results in a representing activity

There are various connections between these three elements. First, the representeds (A) are the content of the representing activities (B): representings (B) are of, or about representeds (A). We are only able to have representeds (A) by doing representings (B). Second, we can only perform representings (B) by applying rules (C). It is only by applying rules (C) that we are able to perform (B).

Corresponding to these three types of construct are three types of constraint:

- **A-Constraint**: constraints on representeds: constraints on the content of the representing activity
- **B-Constraint**: constraints on representings: constraints on the activity of representing itself
- **C-Constraint**: constraints on the rules that are constructed and applied

The basic structure of the argument in each principle is:

1. start with A-constraints: constraints on the representeds
2. derive from the original A-constraints further A-constraints, using the additional premise that time is not directly perceived
3. move to B-constraints: constraints on the representings, on the activity of combination
4. move to C-constraints: constraints on the rules used to produce the activity of combination
5. finally, derive from the C-constraints additional necessary A-consequences: things that must be true of any representation that has achieved unity (Fig. 13.2)

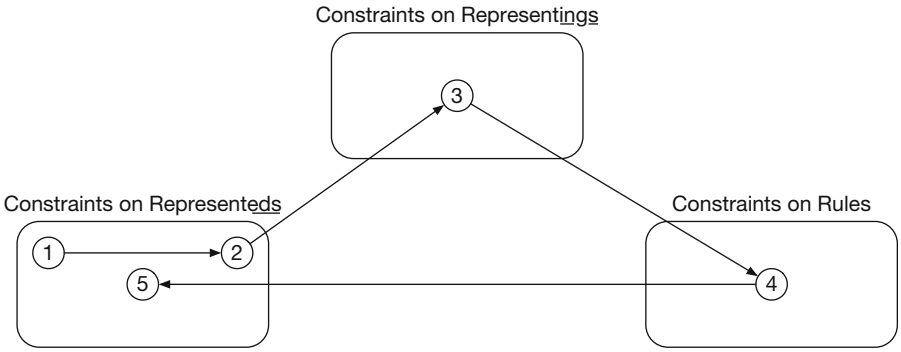


Fig. 13.2 The basic argument structure

13.2.3.1 Constraint 1: The Representeds Must Achieve a Certain Aspect of Time Determination

We start with a requirement on the representeds (A-Constraint): that the cognitions must satisfy an aspect of time-determination in order to achieve unity. Recall from Sect. 13.2.2.4 that there are four aspects of time-unity described in the *Schematism*: construction of moments, filling time, ordering time, the sum total of time.

We ask: what must be true for this group of representeds to achieve this aspect of time determination?

13.2.3.2 Constraint 2: There Must Be a Certain Relation That, If Satisfied by the Representeds, Allows the Representeds to Achieve a Certain Time Determination

In Constraint 2, we remain focused on A-constraints. We just move from a general constraint to a more specific constraint. In this phase, we use the additional premise that *time cannot be directly perceived*. Given that time cannot be directly perceived, the only way we can achieve the general high-level constraint is by the existence of a *relation* satisfying various properties.

We ask: what must be true of this plurality of representeds for there to be a relation satisfying various properties?

13.2.3.3 Constraint 3: The Activity of Combination (i.e. Rule-Application) Must Satisfy a Certain Property

In Constraint 3, we move from an A-constraint (on representeds) to a B-constraint (on the activity of representing). Recall that the representings are just acts of combination, and these acts of combination are just applications of rules. We ask:

what must be true of this *rule-application activity* for the produced representeds to satisfy the relation in Constraint 2?

13.2.3.4 Constraint 4: The Rules Must Satisfy a Certain Property

Now, given that the activity of rule application depends on the types of rules that were constructed, we move in Constraint 4 from a B-constraint on the activity of rule-application to a C-constraint on the types of rules that were constructed in the first place.

We ask: what properties must the rules satisfy if, when they are applied, the rule-application activity satisfies Constraint 3?

13.2.3.5 Constraint 5: Therefore, the Representeds Will Satisfy Various Additional Properties

Now, when we apply rules satisfying certain properties, the products of this rule application (i.e. the representeds) will necessarily have certain properties.

In this phase, we infer from constraints on (C) to additional necessary properties of (A). If we are interested in cognitive science, we are most interested in Kant's derivation of Constraints 1–4. If we are interested in metaphysics (specifically, Kant's argument for various synthetic a priori propositions) then we are interested in the final conclusion, Constraint 5.

13.2.3.6 An Analogy

Suppose you are leading a kindergarten dance class and you need, at some point during the performance, to get the group of five year olds to form a circle. This is the initial general Constraint 1.

Now suppose we are not permitted to draw a circle on the ground. The stage-manager is very strict, and does not permit any tampering with the stage. So the children must be encouraged to form a circle without any external markings to help them. What we want is a relation "right-of" between agents such that, for every agent, there is exactly one agent who is "right-of" him. Furthermore, we require that the transitive closure of "right-of" connects every two agents (Note: of course, this isn't strictly-speaking necessarily a circle. It might be a rather squishy shape. But it will be a closed shape). This is Constraint 2.

Now how do we get the five year olds to achieve this relation? The dance teacher will not be on stage with the children when the time comes for the children to form a circle. We can only get them to form a circle by giving them, in advance, a rule to apply. This is Constraint 3: we have moved from a constraint on the positions of their bodies to a constraint on their rule application.

Now what sort of rule would be suitable? One possible rule, for instance, is: when it is time to form a circle, find somebody whose left hand is free, and hold their left hand with your right hand. This is Constraint 4, a condition on the rules that generate the activity.

Now the conclusion (Constraint 5) is a derived fact about the physical orientation of the dancers, assuming they are following the rules correctly: nobody has a free hand, as each hand is holding some other hand.

This analogy may or may not be helpful, but there is one point of disanalogy that must be mentioned. In Constraint 4, the proposed rule is only one amongst many ways of achieving the right-of relation. In Kant's argument, by contrast, the fourth constraint on the construction of rules is *necessary*.

13.3 Reinterpreting Kant's Theory as a Specification of a Cognitive Architecture

To recap, Kant sees a cognitive agent as a particular type of rule-follower. The only way it can make sense of its sensory given is by applying rules it has already constructed. It is "always busy poring through the appearances with the aim of finding some sort of rule in them" (Kant 1781)(A126). Recall that there are two types of rule: rules of composition are defeasible rules that bring intuitions together; rules of connection are strict rules that operate by necessity, subsuming intuitions under marks. The Kantian cognitive agent must find a set of rules that, when applied to the sensory given, produce a set of representations that satisfy the constraints described in the *Analytic of Principles*. If the results of the rule application process do indeed satisfy the four *Principles*, then the resulting cognitions achieve a unity: they are unified in the medium of time. This is what it is for the agent to achieve original intentionality.

The central idea of this paper is to reinterpret Kant's cognitive agent as a rule-induction system, searching over the space of non-monotonic logic programs. Rules of composition are interpreted as defeasible rules in a non-monotonic logic, rules of connection are interpreted as strict rules. Given a set of rules (i.e. a non-monotonic logic program), a coherent interpretation is a stable model of that logic program. Each of the four *Principles* is interpreted by two constraints: a relational constraint on the stable models that are deemed acceptable, and a structural constraint on the forms of logical rules that are allowed. Making sense of one's sensor readings, according to this interpretation, is searching through the space of non-monotonic logic programs (restricted by the structural constraints) for a program that, when applied to the sensory given, produces at least one model that satisfies the relational constraints.

In the rest of this section, I shall try to explain this central idea in more detail.

13.3.1 Representing Kant's Rules in a Non-monotonic Logic

Consider a logic that contains both strict and defeasible rules. As well as strict rules of the form:

$$h \leftarrow b_1, \ldots, b_n$$

we also have defeasible rules of the form:

$$h \leftsquigarrow b_1, \ldots, b_n$$

The latter means that if $b_1, \ldots, b_n$ all hold, then you may conclude that h, as long as h is compatible with your other commitments. This defeasible implication can be translated away using negation-as-failure[14] and classical negation:

$$h \leftarrow b_1, \ldots, b_n, \mathsf{not}\ \neg h$$

Logics with defeasible rules, defaults,[15] or negation-as-failure, are **non-monotonic** in that $A \models p$ does not imply $A \cup B \models p$.

Defeasible rules are used to model Kant's rules of composition. Strict rules (rules containing no negation-as-failure) are used to model Kant's rules of connection.

13.3.2 Representing Apprehensions (Object-Slices) as Logical Terms

When representing objects that change over time, most work in knowledge representation assumes that the logical terms represent enduring objects, persisting over time.[16] To capture the fact that objects' properties change over time, the predicates are given an additional argument, to represent the time-index, or situation, at which the property holds.

In this model, by contrast, logical terms represent *apprehensions*: slices of objects *at a particular time*. Rules of composition (defeasible rules) produce new terms representing object-slices at a particular moment. The important thing about starting with object-slices and building up to objects (rather than starting with objects, and dividing them into slices) is that *it allows us to focus on the non-trivial task of reidentifying the same object over time*: is this object-slice which I construct from sensor-34 at time t_1 the same object as the intuition which I construct

[14] Here I assume the stable model semantics (Gelfond and Lifschitz 1988) for negation-as-failure.

[15] See Reiter (1980).

[16] For influential examples, see Kowalski and Sergot (1989) and McCarthy (1963).

from sensors 78 and 102 at time t_2? As Frege has emphasised, this question is and should be non-obvious. The reason for starting with object-slices and building up to enduring objects is that it makes explicit the achievement involved in answering this question of reidentification.

13.3.3 Translating Each Principle into a Relational and Structural Constraint

Recall that the general structure of the argument for each principle is:

1. start with a general A-constraint: a constraint on the representeds representing an aspect of time-determination needed for unity
2. derive from the original A-constraint a further A-constraint, using the additional premise that time is not directly perceived, requiring the existence of a certain relation with certain properties
3. move to a B-constraint: a constraint on the representings, on the activity of combination
4. move to a C-constraint: a constraint on the rules used to produce the activity of combination

When translating these constraints into our logic-programming formalism, we treat the different stages differently. Constraints 1 and 3 are not encoded explicitly. If we satisfy 2, then we automatically satisfy 1. If we satisfy 4, then we automatically satisfy 3. We shall focus, therefore, on constraints 2 and 4. Constraint 2 is translated into a *relational* constraint on the set of stable models. Constraint 4 is turned into a requirement on the *structure* of the rules that are generated.

13.3.4 Translating Constraint 2 into a Relational Constraint

Each principle requires the existence of a certain relation satisfying certain properties. But the requirements on the various relations are *interdependent*: it is not possible to formulate the requirement on one relation without invoking some of the others. So, to specify the constraints in the First and Third[17] Principles, we will quantify over four binary relations on object-slices:

– PartOf(x, y): object-slice x is part of object-slice y (this implies they are simultaneous)
– SameObject(x, y): slices x and y are slices of the same object at different times

[17] I omit, for reasons of space, discussion of the Second Principle, the *Anticipations of Perception*. The Fourth Principle does not need its own relation.

– Succeeds(x, y): object-slice x is the predecessor slice of y
– Simultaneous(x, y): object-slices x and y are slices from the same moment in time

The entire set of constraints is one existential second-order sentence stating the existence of four binary relations that together satisfy the constraints, a sentence of the form:

$$\exists\ \mathsf{PartOf}, \exists\ \mathsf{SameObject}, \exists\ \mathsf{Succeeds}, \exists\ \mathsf{Simultaneous}...$$

13.3.4.1 Constraint 2 in the Axioms of Intuition

Constraint 1 of the *Axioms of Intuition* is the requirement that all our representations must be grouped into *moments*. This is the first aspect of time-determination: the collection of moments together constitute the entire time-series (Kant 1781)(B184, A145). The *Axioms of Intuition* are responsible for:

> the generation (synthesis) of time itself in the successive apprehension of an object (Kant 1781)(B184, A145)

Now, in Constraint 2, we move to a requirement that a certain sort of relation must exist: there must be a PartOf relation that allows the apprehensions to be grouped into moments.

The justification for this derived constraint is that, since moments cannot be directly perceived,[18] the only way a moment can be represented is as a concept derived from a relation. Define PartOf* as the transitive closure of PartOf. Constraint 2 requires that there is one largest element (the totality) for each moment such that every apprehension (object-slice) at that moment is PartOf* the totality element:

$$\forall x\ \exists! y\ \mathsf{Simultaneous}(x, y) \wedge \forall z\ \mathsf{Simultaneous}(z, y) \rightarrow \mathsf{PartOf}^*(z, y)$$

Now we can derive an equivalence relation InSameTotalityAs as the reflexive transitive closure of PartOf, and we can define moments as equivalence classes of this InSameTotalityAs relation: a **moment** is a maximal set of apprehensions (object-slices) S such that, for every x, y in S, InSameTotalityAs(x, y).

13.3.4.2 Constraint 2 in the Analogies

Common to all three of the *Analogies* is one high-level constraint: there must be a total ordering on moments in time (Kant 1781)(A145, B184-5).

[18] This is claimed explicitly in a marginal note to the first edition.

Now time is not directly perceived. Our sensations do not arrive with a convenient time-stamp we can inspect. In order to *construct* an ordering on moments in time, we must construct an ordering on object-slices. In fact, we need three relations between object-slices:

- SameObject
- Simultaneous
- Succeeds

The *First Analogy* focuses on the SameObject relation, an equivalence relation between object-slices. A pair of object-slices satisfy the SameObject relation if they are different slices of the same persisting space-time worm. This relation is connected to the relations of the other two *Analogies*:

- if $\mathsf{Succeeds}^*(x, y)$, then $\mathsf{SameObject}(x, y)$
- if $\mathsf{Simultaneous}(x, y)$ but x and y are distinct apprehensions, then it cannot be that $\mathsf{SameObject}(x, y)$

Constraint 2 for the *First Analogy* is the requirement that persisting objects (substances) exist at every moment in time:

> For every object x at moment m, for every other moment $m' \neq m$, there must exist a unique y at m' such that $\mathsf{SameObject}(x, y)$

But, since time cannot be directly perceived, we need to re-express this requirement without making use of explicit moments of time. One way to re-express the requirement is:

$$\forall x\ \forall y\ \neg\mathsf{Simultaneous}(x, y) \rightarrow \exists!z\ \mathsf{SameObject}(z, x) \wedge \mathsf{Simultaneous}(z, y)$$

The Second and Third *Analogies* impose constraints on the Succeeds strict ordering and the Simultaneous equivalence relation:

$$\forall x\ \forall y\ \mathsf{SameObject}(x, y) \rightarrow \mathsf{Simultaneous}(x, y) \vee$$

$$\mathsf{Succeeds}^*(x, y) \vee \mathsf{Succeeds}^*(y, x)$$

Here, $\mathsf{Succeeds}^*$ is the transitive closure of Succeeds.

The Second and Third *Analogies* also require a more general constraint on *any* two object slices x and y, even if they are not part of the same object. Define a derived relation $<$ on object-slices as the smallest transitive relation satisfying:

- if $\mathsf{Succeeds}(x, y)$ then $x < y$
- if $x < y$ and $\mathsf{Simultaneous}(x, x')$ and $\mathsf{Simultaneous}(y, y')$, then $x' < y'$

In other words, the $<$ relation relates object-slices that are not necessarily part of the same persisting object. Intuitively, it means x occurs at an earlier moment than y. Now the more general requirement is:

$$\forall x\ \forall y\ \mathsf{Simultaneous}(x, y) \vee x < y \vee y < x$$

13.3.5 Translating Constraint 4 into a Restriction on the Structure of the Rules

In this section, we translate Constraint 4 into a condition on the *structure* of the rules that are generated. These constraints are translated into a set of templates[19] that the logic programs are built from.

13.3.5.1 The Structure of Rules in the Axioms of Intuition

The types of rule we need to generate to satisfy the *Axioms of Intuition* are rules of composition, rules that allow us to create object-slices from other object-slices. The format of this rule is:

$$\forall x_1, \ldots, x_n \; \phi(x_1, .., x_n) \rightsquigarrow \exists y \bigwedge_{i=1}^{m} \psi_i(y, x_1, \ldots, x_n)$$

This rule allows us to construct a composite object-slice, y, satisfying multiple properties $\psi_1, \ldots, \psi_m$, if the slices $x_1, \ldots, x_n$ satisfy ϕ. Here, PartOf may be one of the ψ_i. Note that this is defeasible implication: this rule permits us to infer the consequent, but does not necessitate it. Sometimes, there will be many possible defeasible rules of composition, not all of which are compatible. For example, this group of sensory readings could be interpreted as representing a nose; or it could be interpreted as representing an ear; but both defeasible rules could not fire at once with the same substitutions for variables $x_1, \ldots, x_n$.

13.3.5.2 The Structure of Rules in the First Analogy

The *First Analogy* requires that there are rules of composition that produce object-slices for objects at moments at which they cannot be perceived. We need a rule such that, for each slice x and each moment m, it produces a slice z at moment m such that SameObject(z, x). This rule of the "productive imagination" (Kant 1781)(B154) has the form:

$$\forall x, y \; \phi(x, y) \rightsquigarrow \exists z \; \mathsf{SameObject}(z, x) \wedge \mathsf{Simultaneous}(z, y)$$

This is an instance of the general structure of rules of composition in Sect. 13.3.5.1 above.

[19] This is called a "language bias" in the program induction literature.

13.3.5.3 The Structure of Rules in the Second Analogy

To count the subjective sequence of apprehensions as objectively successive, we must apply a causal rule that explains why this particular object's fluent properties changed in the way that they did, and also necessitates other object's fluent properties changing in similar ways in similar situations.

Define a three-place relation $\mathsf{Causes}(x, y, p)$ relating two object-slices x and y and a reified proposition p. This causal relation determines both which propositions become true and the $\mathsf{Succeeds}$ relation between slices:

$$\mathsf{Causes}(x, y, p) \rightarrow \mathsf{Succeeds}(x, y)$$
$$\mathsf{Causes}(x, y, p) \rightarrow \mathsf{Is}(p)$$

Here, $\mathsf{Is}()$ is a predicate holding of reified propositions exactly when those propositions are true.

Now the template for causal relations is:

$$\forall x, y\ \phi(x, y) \rightarrow \mathsf{Causes}(x, y, \psi(x, y))$$

where $\psi(x, y)$ is a reified proposition: a term describing the propositional result of the causal interaction in which object-slice x succeeds to y. For example, if a melting candle gets smaller by 1 cm each time-step, then the causal rule would relate two successive apprehensions of the candle, x and y, and $\psi(x, y)$ would state the height of the candle-slice y as a function of the height of the candle-slice x.

13.3.5.4 The Structure of Rules in the Third Analogy

The only way we can count our apprehensions as simultaneous is if we construct an interaction rule that reciprocally determines these particular slices' fluent properties.

Define a three-place relation $\mathsf{Interacts}(x, y, p)$ relating two object-slices x and y and a reified proposition p. The $\mathsf{Interacts}$ relation determines both which propositions become true and the $\mathsf{Simultaneous}$ relation between slices:

$$\mathsf{Interacts}(x, y, p) \rightarrow \mathsf{Simultaneous}(x, y)$$
$$\mathsf{Interacts}(x, y, p) \rightarrow \mathsf{Is}(p)$$

Again, $\mathsf{Is}()$ is a predicate holding of reified propositions exactly when those propositions are true.

Now the template for interactions is:

$$\forall x, y\ \phi(x, y) \rightarrow \mathsf{Interacts}(x, y, \psi(x, y))$$

where $\psi(x, y)$ is a reified proposition: a term describing the result of the interaction between simultaneous object-slices x and y. For example, if a touch-sensitive sensor is turned on when a moving object presses against it, the two slices x and y are the slices of the sensor and the moving object, and the proposition $\psi(x, y)$ is the fact that the sensor is turned on.

13.3.5.5 The Structure of Rules for the Postulates of Empirical Thought

In order to rule out certain propositions as impossible, we must construct necessary rules of incompatibility and rules of entailment that, when applied, rule out certain configurations:

> 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). (Kant 1781)(A144, A184)

These necessary rules of connection have the form:

$$\forall x_1, \ldots, x_m \neg(\phi_1(x_{1_1}, \ldots, x_{1_k}) \wedge \ldots \wedge \phi_n(x_{n_1}, \ldots, x_{n_k}))$$

For example:

$$\forall x, y, \neg(\mathsf{Nose}(x) \wedge \mathsf{Ear}(y) \wedge \mathsf{SameObject}(x, y) \wedge \mathsf{Simultaneous}(x, y))$$

13.3.6 Searching for Non-monotonic Logic Programs that Satisfy the Relational and Structural Constraints

Now, given a set of sensory perturbations (represented as logical atoms), the Kantian cognitive agent finds a set of rules, according to the templates specified by the structural constraints (detailed in Sect. 13.3.5), that generate some stable model satisfying the relational constraints (detailed in Sect. 13.3.4). If the agent finds such a set of rules, it has constructed a coherent interpretation of its sensory given, making a unity out of the plurality of sensings. Perception, then, is a form of program synthesis[20]: finding a logic program that, when applied, makes sense of the sensory perturbations we are given.

The Kantian agent's task is related to, but different from, the standard Inductive Logic Programming (ILP) task. In ILP, the learning agent is given a background theory, B, and two sets of examples: a set P of positive examples of the target

[20]Contrast with Shanahan (2005), who sees perception as a form of *abduction*.

predicate, and a set N of negative examples of the target predicate. The task of ILP[21] is to produce a logic program, R, a set of rules, such that

- $B \cup R \models p$, for all $p \in P$
- $B \cup R \not\models n$, for all $n \in N$

The Kantian agent's task is rather different from the ILP problem because the Kantian agent is not given a set of positive and negative examples of a target predicate. The Kantian agent performs *unsupervised learning*: he is given some input, of course. But the input he is given is a set of atoms representing the states of his sensors. He is not given a supervised signal of what he is supposed to learn. *He is free to construct any set of rules he chooses*[22] – as long as the rules satisfy the structural constraints (Sect. 13.3.5) and the resulting rule-set has a stable model satisfying the relational constraints (Sect. 13.3.4).

In the current implementation, the machine searches through the space of logic programs using a variant of an ILP algorithm described by Corapi et al (2012). Each structural constraint described in Sect. 13.3.5 above is converted into a template for generating ASP clauses, using negation-as-failure to implement the defeasible rules. The search-space is explored through a form of iterative deepening. I omit the technical details for reasons of space.

13.4 Experiments

I took a simplified[23] version of the Kantian cognitive agent, described above, and tested it in two domains: a two-dimensional grid world, and a one-dimensional world of string sequences. These are preliminary, proof-of-concept experiments. But they do show that a Kantian cognitive agent is capable of making sense of sensory data without supervision, from a tiny handful of examples.

[21] The problem description for finding *non-monotonic* logic programs from positive and negative examples is actually somewhat more complicated, as there may be multiple models, each with their own positive and negative instances. See Law et al (2014) for details.

[22] Hence Kant's emphasis on *spontaneity*: the Kantian agent is both less free (because he can *only* perform actions by applying rules) and more free (because he can construct *any set of rules he likes*) than the empiricist can possibly imagine.

[23] There are two major simplifications in the current implementation. The first is that the spatial framework needed to satisfy the *Axioms of Intuition* is given in advance, pre-specified, hand-coded. The agent is told that he is operating in a 2-dimensional grid world. The second major simplification is that the constraints involved in the *Anticipations of Perception* are ignored altogether: in the initial implementation, time is modelled as a series of discrete points, rather than being dense. In future work, I plan to overcome these limitations.

Table 13.1 Sensory data

	Sensor w	Sensor x	Sensor y	Sensor z
t_1	Off	Off	Off	Off
t_2	On	Off	Off	Off
t_3	Off	On	Off	Off
t_4	Off	Off	On	Off
t_5	Off	Off	Off	On
t_6	Off	Off	Off	Off

13.4.1 Making Sense of the Grid World

Imagine a robot in a simple grid world. The robot has an array of four sensors, w, x, y, z, arranged in a line. The robot is given sensory data for six time steps, as shown in Table 13.1. Here, the sensors are turned on, one at a time, from left to right. To get a feeling for the sensory world of the robot, clench one fist, close your eyes and ask someone to drag some object (perhaps, a pen) along your knuckles. The sensations you will get are similar[24] to the sensations the robot receives.

The robot must construct a set of rules that, when applied, makes sense of this sensory data. The robot has to construct *some* explanation of the sensations it is given. Since the on/off predicates are *fluent* properties of the sensors, there has to be, according to the Second and Third *Analogies*, some rule to explain the changing values of the fluent. The robot has some work to do to construct an explanation of the fluent values changing.

When the Kantian machine is given this sequence of sensory data, the first explanation it finds is this: there is an object m moving from left to right in the row directly above the sensors. (see Fig. 13.3). Whenever the object is directly above a sensor, it makes the sensor turn on (indicated in Fig. 13.3 by a shaded block). Otherwise, the sensor remains off.

This explanation requires the following rules to be constructed:

- a rule of composition, according to the template for the *Axioms of Intuition* in Sect. 13.3.5.1, constructing an object-slice for m at position $(x, y-1)$ for every sensor at (x, y) that is turned on: we count the sensor's being on as evidence for the existence of an object-slice at the position directly above the sensor
- a rule of composition, according to the template for the *First Analogy* in Sect. 13.3.5.2, generating an object-slice of the moving object for each time step when it is not perceived by the sensors (in this case, for time-steps 1 and 6)
- a causal rule, according to the template for the *Second Analogy* in Sect. 13.3.5.3, describing how the moving object moves one square from left to right each time step

[24] One important difference is that your tactile sensations are much more fine-grained: you receive a number of intermediate sensations as the object moves between your four knuckles. The robot just has four discrete boolean sensors (one for each knuckle).



Fig. 13.3 One interpretation of the sensory data. The sensors (w, x, y, z) are shown at each time step. E.g. $w1$ means sensor w at time step 1. The interpretation that the Kantian machine finds is that there is a moving object, m, moving from left to right, in the row above the sensors. The moving object turns on the sensor when the mover is directly above the sensor. Note that the moving object is not directly perceived; it is constructed (or invented) to make sense of the data

– an interaction rule, according to the template for the *Third Analogy* in Sect. 13.3.5.4, describing how the object presses on a sensor when it is directly above it, causing it to turn on

Note that, although these rules were constructed simply to make sense of the sensory data, they also allow us to infer that the moving object was at position (1, 1) at time t_1 (even though we cannot sense the object at that time) and to predict that the moving object will be at position (6, 1) at time t_6 (even though we cannot sense the object at that time). This is as it should be: a Kantian interpretation of the present will always also allow us to predict the future and retrodict the past.

To satisfy the *Axioms of Intuition*, the Kantian agent constructs the following defeasible rules of composition describing the moving object that is above the sensor that is turned on:

```
pred1(skf1(X)) :-
     is(on(X), T),
     is(at(X, P1), T),
     above(P2, P1),
     not -pred(skf1(X)).

is(at(skf1(X), P2), T) :-
     is(on(X), T),
     is(at(X, P1), T),
     above(P2, P1).
     not -is(at(skf1(X), P2), T).
```

These clauses generate an appearance from a sensor-reading: if we perceive that a sensor is on, then count its being on as evidence of the existence of another object, a moving object, that is above the sensor.

These rules are generated from the template in Sect. 13.3.5.1. The `skf1(.)` function is a skolem function replacing the existentially quantified variable y in that template.

Here, `pred1` is a gen-sym, a new predicate replacing ψ in the template of Sect. 13.3.5.1. The meaning of this predicate was not chosen in advance by an engineer – rather, it was constructed during the program-synthesis process. The meaning of this generated predicate is *entirely determined by its inferential role* in the clauses that were constructed by the machine. Examining all the rules in which it appears, it is clear that `pred1` is used to denote a new type of object: a type that is distinct from sensors. The various rules generated by the machine together define the behaviour of this new type of object: it exists in the row above the sensors, it moves from left to right, and when it is directly above a sensor, it turns it on.

To satisfy the *First Analogy*, the Kantian agent constructs the following clauses:

```
same_object(skf2(X,Y), X) :-
     appearance(X),
     appearance(Y),
     not exists_appearance_at_time(X, Y).
```

```
simultaneous(skf2(X,Y), Y) :-
    appearance(X),
    appearance(Y),
    not exists_appearance_at_time(X, Y).

pred1(skf2(X,Y)) :-
    appearance(X),
    pred1(X),
    not exists_appearance_at_time(X, Y).
```

These rules say: if we have already constructed an appearance X satisfying `pred1`, and we have no corresponding slice of X for the moment at which Y exists, then posit the existence of an unperceived slice, simultaneous with Y and conclude that it is also a `pred1`. These clauses are generated from the template in Sect. 13.3.5.2. Here, `skf2(.)` is a skolem function replacing the existentially quantified variable z in that template.

To satisfy the *Second Analogy*, the Kantian agent constructs the clause:

```
causes(X1, X2, at(X2, P2)) :-
    is(at(X1, P1)),
    left(P1, P2),
    pred1(X1),
    pred1(X2),
    not -succeeds(X1, X2).
```

This rule states that the moving object moves right by one grid square each time step. To satisfy the *Third Analogy*, it constructs:

```
interacts(X, Y, on(Y)) :-
    is(at(X, P1)),
    is(at(Y, P2)),
    above(P1, P2),
    pred1(X),
    sensor(Y),
    not -simultaneous(X, Y).
```

This rule states that, if a moving object is above a sensor, the moving object presses on the sensor, turning it on.

These rules, when applied to the sensory data, produce a stable model that satisfies the relational constraints in Sect. 13.3.4. The Kantian agent has found a coherent unified interpretation of its sensory data by generating instances of the a priori relations (PartOf, SameObject, Succeeds, Simultaneous) that connect the plurality of sensory perturbations into one cohesive unity. It is worth stressing that the Kantian agent has found a satisfying unified interpretation despite only being given a tiny amount of experiences: it had to construct a unified interpretation given a sequence of only six time-steps.

13.4.2 Making Sense of a One-Dimensional String of Letters: A Verbal-Reasoning Task

Hofstadter introduced the **Seek Whence** problem-set in Hofstadter (2008). In this task, the player is given a sequence of symbols, for example:

a, a, b, b, c, c, d, d, ...

The player needs to guess the next symbol. The *only* knowledge the player can use is the successor[25] relation: a $\prec$ b $\prec$ c $\prec$ ... Here are some example problems:

- a, b, c, d , *e*
- a, a, b, b, c, c, d, d , *e, e*
- a, k, b, k, k, c, k, k, k, d, k , *k, k, k*
- b, a, b, b, b, b, b, c, b, b, d, b, b, e, b , *b, f, b*

The Kantian agent, when confronted with a Seek Whence letter sequence, is forced (by the *Axioms of Intuition*) to interpret its sensory data in terms of moments of time, composed of apprehensions (object-slices) that are spatially related to each other. It is forced (by the *First Analogy*) to connect the apprehensions together into enduring objects, persisting through time. It is forced (by the *Second Analogy*) to interpret changes to these enduring objects in terms of causal rules that explain those changes.

The Kantian agent, in other words, is *doomed* to reinterpret its sensory data in terms of objects persisting through time, changing state according to intelligible causal laws. Surprisingly, the Kantian constraints are enough, on their own, to achieve human-level success on these verbal-reasoning tasks. Consider the following example:

b, a, b, b, b, b, b, c, b, b, d, b, b, e, b , *b, f, b*

Hofstadter called this example the "theme song" of the Seek Whence project, because of its ambiguity. The long string of *b*'s encourages us to mis-parse the sequence, while the true parsing Hofstadter intends is a sequence of triples, *bxb* for increasing x.

Figure 13.4 shows the results of the Kantian machine's deliberations on this particularly tricky Seek Whence problem. The Kantian agent parses this sequence in the way Hofstadter intends. Note how the sequence is interpreted as three objects, persisting over time, changing state in different ways. The left and right objects (Objects 1 and 3 in the diagram) remain the same, while the middle object (Object 2) increases its value every time-step. Note the causal rule (`update`) that is constructed to explain the state change of the middle object. Making sense of the letter sequence means constructing a program that, when applied, reinterprets the sequence as a set of objects, persisting through time, changing state according to intelligible laws.

[25] We assume, for simplicity, that the alphabet is cyclic, so that the successor of z is a.

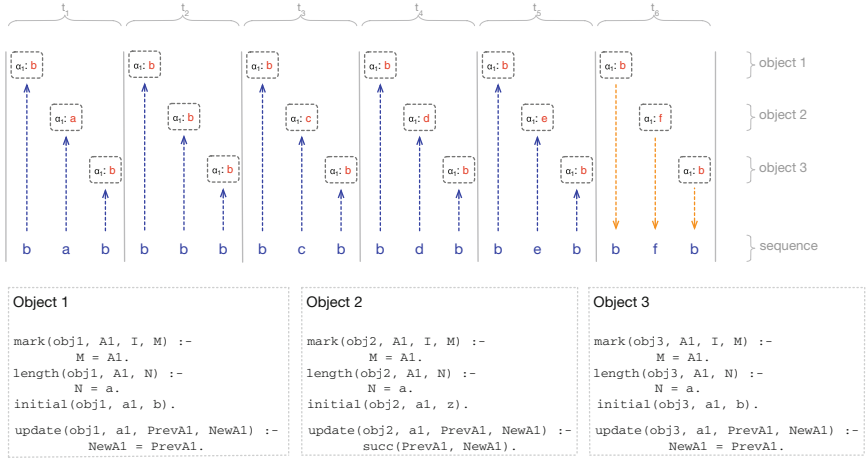


Fig. 13.4 Applying the Kantian machine to a tricky seek whence problem

The Kantian agent was tested on three data-sets: the "Blackburn Dozen" from Meredith (1986), the "Hofstadter Fifteen" from Hofstadter (2008), and the C-test from Hernandez-Orallo and Minaya-Collado (1998). Overall, the Kantian agent scored 86% correct[26] on the three data-sets, reaching human-level performance. This compares favourably with the only other known attempt to solve this problem, in Meredith (1986), that achieved 25% (Fig. 13.5).

13.5 Conclusion

The modern debate between deep-learning practitioners and advocates of logic-based approaches resembles the eighteenth century debate between empiricists and rationalists. One of Kant's driving forces was reconciliatory: to capture the insights of both empiricism and rationalism in one unified system. It seems at least possible that his work, his extraordinarily ambitious system, might have something useful to say about the modern-day reincarnation of this old debate.

This paper represents a first, tentative step towards a full computer implementation of Kant's vision: a self-legislating agent, bound by the rules he has

[26] We need to be careful with notions of "correctness" in sequence induction tasks. There are always infinitely many ways of continuing a finite series, even if some appear more "natural" to us than others. In the case of the "Blackburn Dozen" and the "Hofstadter Fifteen", the authors specified the intended continuation. I did not use these intended continuations when evaluating correctness. Instead, I gave the questions to 100 people, as an online form, and took the mode as the "correct" continuation. The Kantian constraints provide a way of formally specifying what is "natural" about the "natural" continuations.

Sequence	Human	Kantian Agent
b,b,b,c,c,b,b,b,c,c,b,b,b,c,c,...	b	b
b,a,a,b,b,b,a,a,a,a,b,b,b,b,b,...	a	a
b,a,b,e,b,a,a,a,e,b,b,b,b,e,a,...	e	-
b,c,a,c,a,c,b,d,b,d,b,c,a,c,a,...	e	a
a,b,b,c,c,d,d,e,e,f,f,g,g,...	h	h
a,a,b,a,b,c,a,b,c,d,a,b,c,d,e,...	a	a
b,a,c,a,b,d,a,b,c,e,a,b,c,d,f,...	a	a
a,b,a,c,b,a,d,c,b,a,e,d,c,b,...	g	g
c,b,a,b,c,b,a,b,c,b,a,b,c,...	b	b
a,a,a,b,b,c,e,f,f,g,g,g,h,h,i,...	s	-
a,a,b,a,a,b,c,b,a,a,b,c,d,c,...	a	a
a,a,b,c,a,b,b,c,a,b,c,c,a,a,...	a	a
a,a,b,c,a,b,b,c,a,b,c,c,a,b,...	a	a
a,b,b,c,c,a,a,b,c,c,a,a,b,b,...	a	a
a,b,b,c,c,a,b,b,c,a,b,c,c,a,...	a	b

Fig. 13.5 The Kantian agent's performance on Hofstadter's dataset (Hofstadter 2008)

himself constructed, reinterpreting his sensory perturbations as a coherent unified experience of an external world.

The implemented system is able to perform *unsupervised learning*: making sense of its sensory input without labeled data or rewards. This system takes the raw sensory stimuli and creates a program that, when applied, constructs a coherent interpretation of its sensory world.

The system is able to learn *data-efficiently*, from a tiny amount of data, because of the strong prior knowledge built into the system. This prior knowledge is a set of general, domain-independent constraints on the types of rule that can be constructed. These constraints are taken directly from Kant's *Analytic of Principles*: constraints that must be satisfied by any agent who seeks to reinterpret his sensory perturbations as a coherent whole, unified in the medium of time.

References

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

Corapi, D., A. Russo, and E. Lupu. 2010. Inductive logic programming as abductive search. In: ICLP (Technical Communications), 54–63.

Corapi, D., A. Russo, and E. Lupu. 2012. Inductive logic programming in answer set programming. In: *Inductive Logic Programming*, 91–97. Heidelberg/New York: Springer.

Frege, G., P. Geach, and M. Black. 1980. 'Über Sinn und Bedeutung', in Zeitschrift für Philosophie und philosophische Kritik, Translated as 'On Sense and Reference' by M. Black in Translations from the Philosophical Writings, 100: 25–50. Oxford: Blackwell, third edition.

Gelfond, M., and V. Lifschitz. 1988. International Conference on Logic Programming. The stable model semantics for logic programming. In: *ICLP/SLP*, vol. 88, 1070–1080.

Goodman, N.D., J.B. Tenenbaum, J. Feldman, and T.L. Griffiths. 2008. A rational analysis of rule-based concept learning. *Cognitive Science* 32(1): 108–154.

Graves, A., et al. 2012. Supervised sequence labelling with recurrent neural networks, vol. 385. University of Toronto, Springer.

Haugeland, J. 1990. The intentionality all-stars. *Philosophical Perspectives* 4: 383–427.

Hernandez-Orallo, J., and N. Minaya-Collado. 1998. Engineering of Intelligent Systems, A formal definition of intelligence based on an intensional variant of algorithmic complexity. In: *Proceedings of International Symposium of Engineering of Intelligent Systems (EIS98)*, February 11–13, 146–163.

Hofstadter, D.R. 2008. *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. New York: Basic Books.

Hofstadter, D.R., M. Mitchell, 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.

Hutter, M. 2007. On universal prediction and Bayesian confirmation. *Theoretical Computer Science* 384(1): 33–48.

Jordan, C., and L. Kaiser. 2013. Learning programs as logical queries. In: *The ICALP 2013 Satellite Workshop on Learning Theory and Complexity*, (ICALP is the "International Colloquium on Automata, Languages and Programming").

Kant, I. 1781. *Critique of Pure Reason*, Trans. P Guyer. Cambridge University Press.

Kowalski, R., and M. Sergot. 1989. A logic-based calculus of events. In: *Foundations of Knowledge Base Management*, 23–55. Berlin: Springer.

Lake, B.M., R. Salakhutdinov, and J.B. Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science* 350(6266): 1332–1338.

Law, M., A. Russo, and K. Broda. 2014. Inductive learning of answer set programs. In: *Logics in Artificial Intelligence*, 311–325. Cham: Springer.

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

McCarthy, J. 1963. Situations, actions, and causal laws. Technical Report, DTIC Document.

Meredith, M.J.E. 1986. Seek-whence: A model of pattern perception. Technical Report, Indiana University, Bloomington (USA).

Mitchell, M. 1993. Analogy-making as perception: A computer model. Cambridge: MIT Press.

Muggleton, S.H., D. Lin, and A. Tamaddoni-Nezhad. 2015. Meta-interpretive learning of higher-order dyadic datalog: Predicate invention revisited. *Machine Learning* 100(1): 49–73.

Reiter, R. 1980. A logic for default reasoning. *Artificial Intelligence* 13(1): 81–132.

Sellars, W. 1968. Science and metaphysics: Variations on Kantian themes. Ridgeview Publishing Company, Springer.

Shanahan, M. 2005. Perception as abduction: Turning sensor data into meaningful representation. *Cognitive Science* 29(1): 103–134.

Sloman, A. 2008. Kantian philosophy of mathematics and young robots. In: *International Conference on Intelligent Computer Mathematics*, 558–573. Berlin/Heidelberg: Springer.

Tenenbaum, J.B. 2000. Rules and similarity in concept learning. *Advances in Neural Information Processing Systems* 12: 59–65.

Waxman, W. 2013. *Kant's Anatomy of the Intelligent Mind*. Oxford: Oxford University Press.

Wittgenstein, L. 1958. *The Blue and Brown Books*. Oxford: Blackwell.