A Pragmatic Theory of Induction

John Bell[1]

Abstract.

This paper develops a qualitative, logical, theory of induction. It begins with Hempel's attempt to produce a "purely syntactical" theory of confirmation and the demise of this attempt as a result of Goodman's paradox. Ideas from the informal, pragmatic, solutions to this paradox proposed by Goodman and Quine are then adopted, adapted and extended in order to produce a formal, pragmatic theory of induction. According to this theory, induction takes place in an evolving context of inference; which includes an evolving system of kinds and, typically, a background theory. The theory is illustrated by giving a formal solution of Goodman's paradox, and a further difficulty raised by Davidson is discussed.

1 INTRODUCTION

In 1943 Hempel proposed a "purely syntactical definition of confirmation", [10]. By "confirmation" he meant a relation between observation sentences and the conclusions which can reasonably be induced from them. Central to his definition is the idea of the development of a hypothesis for a finite class of individuals. Thus, for example, the observations $P(a) \wedge Q(a) \wedge \neg Q(b) \wedge \neg P(b)$ determine the class of individuals $\{a, b\}$, and the hypothesis $\forall x(P(x) \rightarrow Q(x))$ is confirmed by these observations because its development for $\{a, b\}$ is the sentence $(P(a) \rightarrow Q(a)) \wedge (P(b) \rightarrow Q(b))$ and this is entailed by the observations. This idea captures an essential feature of induction, as it requires that all of the stated evidence is taken into account. Thus, in particular, it is defeasible; if the evidence is extended to include the observations $P(c) \wedge \neg Q(c)$, then the development of the above hypothesis for $\{a, b, c\}$ is not entailed by the observations, and so the hypothesis is no longer confirmed by them. Hempel's definition of confirmation can thus be seen as a precursor of ideas such as the closed world assumption and circumscription. It can also be seen as a pragmatic, or context-dependent, account; where the *context of inference* consists of all of the stated evidence, and conclusions are only appropriate if it is assumed that this is all of the relevant evidence. As an example of the "material adequacy" of his definition, Hempel argued that it could be used as the basis for a solution to his "paradox of the ravens".

However, in 1955 Goodman [9] proposed another, devastating, paradox which lead to the abandonment of the attempt to produce a qualitative, logical, theory of induction, and efforts shifted to the development of quantitative, probabilistic, theories. Goodman asks us to suppose that all emeralds examined before some future time T (tomorrow, next Thursday, etc.) have been found to be green. Then these observations confirm the hypothesis that all emeralds are green, and hence the prediction that emeralds observed at or after T will be green. Now call an object "grue" at time t if it is green and t is before T or it is blue and t is T or later. Then the observations equally confirm the hypothesis that all emeralds are grue, and hence the prediction that emeralds observed at or after T will be blue.

Attempts to argue that the predicate 'grue' is illegitimate on syntactic grounds are futile. For example, Goodman points out [pp. 79-80] that it is useless to argue that 'grue' is more complex than 'green' because complexity is relative to a choice of language; thus, in a language ("Grublish"?) in which 'grue' and 'bleen' (blue before T, green thereafter) are primitive and 'green' is defined as being grue before T or bleen thereafter, 'green' is more complex than grue. Semantic arguments are also futile. For example, Barker and Achinstein [1] attempt to show that, unlike 'green', 'grue' is positional. In order to do so they introduce a certain Mr. Grue, arm him with easel and sketchpad, and set him the task of representing (now, at some time $t < T$) the colour of grass at T. As Mr. Grue believes that grass is grue, it seems that he should select blue pigment and thus his representation of the grass will be a different colour from its present colour. However Ullian [14] argues that what counts as a concept is entirely a matter of convention, and that instead of our concept 'colour', Mr. Grue could have the concept 'shmolour'; where an object has the shmolour grue at time t just in case it is green and t is before T or it is blue and t is T or later. He can thus proceed to paint the grass grue in the belief that it will be the very same shmolour at T. Ullian concludes: "One extension is as good as another for a class qua class, no matter how much (or how little) its description may cut across the boundaries of our ordinary classifications. Unless we privilege special classes—and logic alone cannot allow us to do this—there is no hope of distinguishing those extensions which may be taken as belonging to bona fide predicates", [pp. 388-9].

Goodman considers induction to be a special case of the problem of *projecting* from a given set of cases to others. A lawlike statement is one that is *projectible*, that is "capable from receiving confirmation from an instance of it" [9, p. 83]. And, Goodman suggests, a statement is projectible if all of the predicates occurring in it are projectible. The problem is thus to find a means of distinguishing between projectible predicates (such as 'green') and non-projectible predicates (such as 'grue'). He suggests that one way of doing so is *entrenchment*. Now, purely as a matter of historical fact, the predicate 'green' has been projected (has featured in projections) many more times than the predicate 'grue' has. Thus 'green' is better entrenched than 'grue', and consequently projections involving 'grue' can be ruled out when they conflict with projections involving 'green'. Goodman concludes that: "the roots of inductive validity are to be found in our use of language . . . the line between valid and invalid predictions (or inductions or projections) is drawn upon the basis of how the world is and has been described and anticipated in words", [pp. 120-1].

Quine [12] agrees, but seeks a more fundamental explanation in

[1] Department of Computer Science, Queen Mary, University of London, London E1 4NS, email: jb@dcs.qmul.ac.uk

the nature of human, and indeed animal, cognition. He suggests that we sort things into *kinds* on the basis of an innate sense of similarity: “A standard of similarity is in some sense innate. This point . . . is a commonplace of behavioral psychology. A response to a red circle, if it is rewarded, will be elicited again more readily by a pink ellipse than by a blue triangle, the red circle resembles the pink ellipse more than the blue triangle. Without some such prior spacing of qualities, we could never acquire a habit; all stimuli would be equally alike and equally different”, [p. 123].

This paper aims to revive “Hempel’s programme”, by proposing a logical theory of induction which is not susceptible to paradoxes such as Goodman’s. It develops a model-based pragmatic theory which draws on Goodman’s notion of entrenchment and Quine’s notion of kinds. According to the theory, induction takes place in an evolving context of inference, which includes an evolving system of kinds and, typically, a background theory. The theory is developed in Section 2. By way of illustrated by giving a formal solution to Goodman’s paradox in Section 3, and a further difficulty raised by Davidson is discussed in Section 4.

2 SIMILARITY, KINDS, AND INDUCTION

The theory of induction is defined in a language called the Second-order Temporal Calculus ($\mathcal{STC}$). This language will be introduced informally here, but a formal account can be found in [2].

$\mathcal{STC}$ is based on Kleene’s strong three-valued language [11] which, when interpreted epistemically, can be seen as providing a means for doing Classical reasoning with partial information. Accordingly, the truth conditions for the propositional fragment yield a Boolean truth value whenever possible. Thus the sentence $\neg\phi$ is true if ϕ is false, is false if ϕ is true, and is undefined otherwise. Similarly the sentence $\phi \wedge \psi$ is true if ϕ and ψ are both true, is false if either is false, and is undefined otherwise. Further connectives can be defined as in Classical logic; for example, $\phi \vee \psi$ is defined as $\neg(\neg\phi \wedge \neg\psi)$. In order to increase the expressiveness of Kleene’s language the undefined operator $?$ is added; thus the sentence $?\phi$ states that ϕ is undefined, that ϕ is neither true nor false. Then $\circ\phi$ can be defined as $?\phi \vee \phi$, $\bullet\phi$ as $?\phi \vee \neg\phi$, $\phi \rightarrow \psi$ as $\bullet\phi \vee \neg\bullet\psi$, and $\phi \equiv \psi$ as $(\neg\bullet\phi \wedge \neg\bullet\psi) \vee (\neg\circ\phi \wedge \neg\circ\psi) \vee (?\phi \wedge ?\psi)$. Thus $\circ\phi$ states that ϕ is not false, $\bullet\phi$ states that ϕ is not true, $\phi \rightarrow \psi$ is true if ψ is true or ϕ is not, and $\phi \equiv \psi$ states that ϕ and ψ are equivalent (are both true, or both false, or both undefined). The first-order extension follows Kleene. Atomic sentences are interpreted using partial functions; thus an atomic sentence of the form $r(u_1, \ldots, u_n)$, may be true, false or undefined in a model. And a universally quantified sentence $\forall v\mathcal{F}$ is true if every instance of $\mathcal{F}$ is true, false if one instance is false, and is undefined otherwise. As in Classical logic, $\exists v\mathcal{F}$ is defined as $\neg\forall v\neg\mathcal{F}$.

In order to represent time, a temporal index is added to each atomic sentence of the underlying language. Thus the first-order atomic sentence $r(u_1, \ldots, u_n)(t)$ should be understood as stating that the relation r holds for the objects denoted by the terms $u_1, \ldots, u_n$ at time point t. It is assumed that time consists of points and that it is discrete and linear.

Finally, the language is extended to the second-order; by adding second-order relations and permitting quantification over first-order relations. The interpretation of a second-order formula $\forall v\mathcal{F}$ proceeds in two stages: a first-order relation symbol (rather than a first-order relation) is substituted for each free occurrence of v in $\mathcal{F}$, this relation symbol is then interpreted in the process of interpreting the atomic formulas in $\mathcal{F}$ and relative to their temporal indices.

The second-order features of $\mathcal{STC}$ are used to define similarity, kinds and induction, and the axioms for these are listed in Table 1.

Table 1. Axioms for similarity, kinds and induction.

$$\forall S, x, y, z, t((\mathit{QSpace}(S)(t) \wedge S(x,y)(t) \wedge y \neq z) \rightarrow \neg S(x,z)(t)) \tag{1}$$

$$\begin{aligned}
&\forall P, S, t(\mathit{Similar1}(P,S)(t) \equiv \\
&\qquad (\exists t' \leq t\, P(x)(t') \\
&\qquad \wedge \exists y \forall x, t' \leq t(P(x)(t') \\
&\qquad\qquad \rightarrow (\mathit{QPred}(S,x,P)(t') \wedge S(x,y)(t')))))
\end{aligned} \tag{2}$$

$$\begin{aligned}
&\forall P, S, r, s, t(\mathit{Similar}(P,r)(t) \equiv \\
&\qquad (r = [S|s] \wedge \mathit{Similar1}(P,S)(t) \\
&\qquad \wedge (s = [\,] \vee \mathit{Similar}(P,s)(t))))
\end{aligned} \tag{3}$$

$$\forall P, t(\mathit{Kind}(P)(t) \equiv \exists r\, \mathit{Similar}(P,r)(t)) \tag{4}$$

$$\forall P, t((\mathit{Kind}(P)(t) \wedge \bullet \mathit{AffK}(P)(t)) \rightarrow \mathit{Kind}(P)(t+1)) \tag{5}$$

$$\begin{aligned}
&\forall P, Q, t(\mathit{Proj}(P,Q)(t) \equiv \\
&\qquad (\mathit{Kind}(P)(t) \wedge \mathit{Kind}(Q)(t) \\
&\qquad \wedge \forall x, t' \leq t(P(x)(t') \rightarrow Q(x)(t'))))
\end{aligned} \tag{6}$$

$$\forall P, Q, t((\mathit{Proj}(P,Q)(t) \wedge \bullet \mathit{AffP}(P,Q)(t)) \rightarrow \mathit{Proj}(P,Q)(t+1)) \tag{7}$$

Beginning with quality spaces, Axiom (1) states that no object has two different qualities (occupies two distinct regions in a quality space) simultaneously. Thus the second-order formula $\mathit{QSpace}(S)(t)$ states that S is a quality space at time t and the (first-order) formula $S(x,y)(t)$ states that object x has the quality in region y of S at t; for example, $\mathit{QSpace}(\mathit{Colour})(t)$ and $\mathit{Colour}(O,C)(t)$ state respectively that, at t, *Colour* is a quality space and that the colour of object O is C.

Similarity is defined by axioms (2) and (3). Axiom (2) states that all P’s, past and present, are similar with respect to quality space S (are *Similar1* in S) iff they all have (or had when P) the same quality in S. Thus similarity in this sense is a historical, intensional, notion, involving both the present and the past extensions of P. In the axiom, the second-order predicate *QPred* is used to relate predicates with the qualities that they predicate. The formula $\mathit{QPred}(S,x,P)(t)$ states that, at time t, the S of x is P (where S is a quality space, x is an object and P is a predicate); for example $\mathit{QPred}(\mathit{Colour}, x, \mathit{Green})(t)$ states that, at t, the colour of x is green. It is necessary to relate P and S in this way in view of the subsequent definition of kinds. Axiom (3) merely extends the definition of similarity to a non-empty list of qualities. Thus the P’s are *Similar* in respect r at time t iff r is a non-empty list of first-order relation symbols each of which is used to represent a quality space in which the P’s are *Similar1* at t. (Technically, r is a first-order term, $[\,]$ denotes the empty list and $[Q|s]$ denotes the list with head Q and tail s.)

Kinds are defined in terms of similarity. Thus Axiom (4) states that the predicate P is (denotes, constitutes) a kind at time t if all P’s are *Similar* in some respect r. If r contains a single relation symbol S, then P is said to be a *simple* kind (as the P’s have a single quality in common; are *Similar1* in S), otherwise P is said to be a *complex* kind. It is now clear that if the *QPred*-condition were dropped from Axiom (2), then Axiom (4) would permit unnatural kinds; for example, suppose that grue objects are all the same shape and that some objects are grue after T, then the proposed simplification of Axiom (2) would result in *Grue* being a kind because of the similarity in shape and despite the dissimilarity in colour.

As the extension of predicates may vary over time, kinds have natural histories. A kind becomes established or entrenched as a result of observed regularities and of its use in projections. However a kind can become *defunct* at time t if it remains a kind but has no new members at t; so extinct species can be regarded as permanently defunct kinds. A kind can also become *defective* at t if its members are no longer all alike in what was the defining respect of the kind; for example we consider that the kind grue becomes defective at T if is not defunct; as it then contains both green and blue objects.

The final three axioms concern the entrenchment of kinds and projection. Axiom (6) states that the relation between predicates P and Q is projectible at time t iff it is the subkind relation; thus *Proj*$(P, Q)(t)$ is true iff P and Q are both kinds at t and, the extension of P up to t is a subset of that of Q. The reason for projecting the subkind relation between predicates, rather than projecting predicates individually, as Goodman and Quine suggest, is discussed further in Section 4.

Axioms (5) and (7) are the speculative axioms of the theory, and are used to make predictions on the basis of the current context of inference. Axiom (5) is the entrenchment axiom. Intuitively, the second-order atom *AffK*$(P)(t)$ states that the kind denoted by the predicate P is affected at time t; that is, that there is reason to doubt its persistence beyond t. So the axiom states that if P is a kind at t and it is not true that P is affected at t, then P remains a kind at $t+1$. This axiom is intended to be interpreted pragmatically: its interpretation should take account of the current context of inference, and it should be interpreted positively (rather than contrapositively) whenever possible. Thus, given *Kind*$(P)(t)$, the *entrenchment assumption* $?$*AffK*$(P)(t)$ should be made, and the axiom used to conclude *Kind*$(P)(t+1)$ whenever it is consistent to do so.

Axiom (7) is the projection axiom. Again intuitively, the second-order atom *AffP*$(P, Q)(t)$ states that the projectibility of the subkind relation between P and Q is affected at t; that is, that there is reason to doubt that the relation can be projected beyond t. So the axiom states that if the subkind relation between P and Q holds t and it is not true that its projectibility is affected at t, then the relation holds at $t+1$. This axiom is also intended to be interpreted pragmatically; its interpretation should take the current context of inference into account, and it should be interpreted positively whenever possible. Thus, given *Proj*$(P, Q)(t)$, the *projection assumption* $?$*AffP*$(P, Q)(t)$ should be made and the axiom used to conclude *Proj*$(P, Q)(t+1)$ whenever doing so is consistent.

As Quine remarks, entrenchment, projection, and induction are intimately related: “We revise our standards of similarity or of natural kinds on the strength . . . of second-order inductions. New groupings, hypothetically adopted at the suggestion of a growing theory, prove favorable to inductions and so become “entrenched”. We newly establish the projectibility of some predicate, to our satisfaction, by successfully trying to project it. In induction, nothing succeeds like success”, [12, pp. 128-9]. This should be reflected in the formal theory, with inductive inferences arising from the combined effect of the entrenchment and projection axioms. However, if these axioms are to be combined, it is necessary to have a strategy for resolving conflicts between them. Conflicts arise when the consequents of a pair of instances of the axioms are inconsistent, thereby forcing a choice between the assumptions in their antecedents. Quine remarks that “every reasonable expectation depends on similarity” [12, p. 124], and a general principle seems to be that predictions which accord with the established system of kinds should be preferred to those which violate it. Thus a projection which further entrenches a kind, or which leaves a kind defunct, should be preferred to a projection which makes a kind defective. Consequently, preference should be given to entrenchment assumptions, with the effect that, in cases of conflict, applications of the entrenchment axiom should take precedence over applications of the projection axiom. The entrenchment axiom should thus play a significant role in constraining projection assumptions. Note that the proposed strategy for conflict resolution differs from the one proposed by Goodman in that it is based on similarity rather than the historical use of language. One upshot is that it can be used to resolve conflicts between two equally entrenched (in Goodman’s sense) predicates, as the formal treatment of his paradox in the next section shows.

The *theory of induction*, Θ_I, consists of the axioms $\{(1), \ldots, (7)\}$, and any theory which contains Θ_I will be called an *induction theory*. In order to enforce the intended interpretation of induction theories, a formal pragmatics is needed. As suggested, an induction theory should be interpreted chronologically: given the context of inference at time t, the axioms of Θ_I should be used in order to extend the context to time $t+1$. Moreover, as suggested, in case of conflict, preference should be given to entrenchment over projection. These effects can be realized by defining the class of preferred models of a given induction theory, in which the theory is interpreted as intended, and then defining an inductive entailment relation on the basis of these models; in short, by defining a preference logic [13].

So, let M and M' be $\mathcal{STC}$ models which differ only on the interpretation of first-order and second-order relations, and the relations *AffK* and *AffP*. Then M is *inductively preferred* (*I-preferred*) to M' (written $M \prec_I M'$) iff there is a time point t such that M and M' agree for any earlier time point and:

- at least one more first-order atom is defined (is either true or false) in M' at t, or
- M and M' agree on the interpretation of all of the above atoms and at least one more *AffK* atom is defined in M' at t, or
- M and M' agree on the interpretation of all of the above atoms and at least one more *AffP* atom is defined in M' at t, or
- M and M' agree on the interpretation of all of the above atoms and at least one more second-order atom is defined in M' at t.

A model M is said to be an *I-preferred model of a theory* Θ iff M is a model of Θ and there is no model M' of Θ such that $M' \prec_I M$. An induction theory Θ *inductively entails* (*I-entails*) a sentence ϕ (written $\Theta \approx\!\!\!\!|_I\, \phi$) iff all I-preferred models of Θ are also models of ϕ.

The preferred models of an induction theory are those in which defined atoms are minimized chronologically according to type. At each time point in such a model the present facts (the first-order literals which follow from the interpretation of the theory at earlier time points) are fixed before any inductive assumptions are made about the future. This has the desired effect that speculating about the future cannot change the present. Then inductive assumptions are made in the priority order defined; entrenchment assumptions before projection assumptions. Finally, any remaining second-order literals are minimized.

3 GOODMAN’S PARADOX

As an example of the theory at work, a formal solution to Goodman’s paradox is now developed. The solution is presented as an imaginary example of robot induction. It is assumed that the robot is equipped with a high-level vision system, such as the one developed by Ullman [15], and that the robot performs high-level symbolic reasoning in $\mathcal{STC}$. The details of the vision system are unimportant. All that is assumed is that this system is capable of classifying objects according to their colour and shape, and of reporting its conclusions to the high-level reasoning system using atomic sentences of $\mathcal{STC}$. Thus, specifically, it is assumed that the vision system employs the quality spaces *Colour* and *Shape*, and that its reports are sentences such as

$\mathit{Colour}(O, C_1)(T)$, which states that object O has colour C_1 at time T. The high-level reasoning system, henceforth simply "the robot", is thus assumed to have the axioms listed in Table 2.

Table 2. Axioms for Goodman's paradox.

$$\bigwedge \mathit{UNA}[C_1, C_2, S_1] \tag{8}$$

$$\forall t \geq 1(\mathit{QSpace}(\mathit{Colour})(t) \wedge \mathit{QSpace}(\mathit{Shape})(t)) \tag{9}$$

$$\forall x, t(\mathit{Green}(x)(t) \equiv \mathit{QPred}(\mathit{Colour}, x, \mathit{Green})(t) \equiv \mathit{Colour}(x, C_1)(t)) \tag{10}$$

$$\forall x, t(\mathit{Blue}(x)(t) \equiv \mathit{QPred}(\mathit{Colour}, x, \mathit{Blue})(t) \equiv \mathit{Colour}(x, C_2)(t)) \tag{11}$$

$$\forall x, t(\mathit{Grue}(x)(t) \equiv \mathit{QPred}(\mathit{Colour}, x, \mathit{Grue})(t) \equiv ((t < T \wedge \mathit{Green}(x)(t)) \vee (t \geq T \wedge \mathit{Blue}(x)(t)))) \tag{12}$$

$$\forall x, t(\mathit{Emerald}(x)(t) \equiv \mathit{QPred}(\mathit{Emerald}, x, \mathit{Shape})(t) \equiv \mathit{Shape}(x, S_1)(t)) \tag{13}$$

$$\mathit{UNA}[O_1, \ldots, O_T] \tag{14}$$

$$T > 1 \wedge \bigwedge_{i=1}^{T-1} (\mathit{Shape}(O_i, S_1)(i) \wedge \mathit{Colour}(O_i, C_1)(i)) \tag{15}$$

$$\mathit{Shape}(O_T, S_1)(T) \tag{16}$$

Axiom (8) states that the names C_1, etc., refer uniquely; the notation $\mathit{UNA}[u_1, \ldots, u_n]$ is adopted as a convenient abbreviation for the set $\{u_i \neq u_j : u_i, u_j \in \{u_1, \ldots, u_n\}$ and $1 \leq i < j \leq n\}$. Axiom (9) states that, from time 1 onwards, *Colour* and *Shape* are quality spaces. The next four axioms, represent the robot's use of language. Thus, for example, Axiom (10) represents the robot's use of the predicate *Green* for objects of colour C_1, and Axiom (12) represents its use of the predicate *Grue*. The final three axioms represent the robot's observations. Thus, in view of the earlier axioms, the robot observes a different green emerald at each time point between 1 and $T-1$, and a further emerald of unknown colour at time T. However, given the background theory $\{(8), \ldots, (13)\}$ and the observations $\{(14), \ldots, (16)\}$, the robot can use the theory of induction, Θ_I, to infer that the emerald observed at T is green. It can then conclude that all emeralds are green, and that it is not the case that all emeralds are grue.

Proposition 1. *Let* $\Theta_1 = \Theta_I \cup \{(8), \ldots (16)\}$. *Then:*

$$\Theta_1 \mathrel{|\!\!\approx}_I \forall x, t(\mathit{Emerald}(x)(t) \rightarrow \mathit{Green}(x)(t)) \wedge \neg\forall x, t(\mathit{Emerald}(x)(t) \rightarrow \mathit{Grue}(x)(t)).$$

Proof. In any I-preferred model M of Θ_1 it follows by the chronological minimization of (first- and second-order) atomic sentences that no atomic sentence with temporal index $t < 1$ is true. It is also clear that, at each successive time point i such that $1 \leq i \leq T-1$, each object O_i examined at i is classified as being green, grue, and an emerald at i (axioms $(10), (12), (13), (15)$). Moreover, it follows from the chronological minimization of first-order atoms that nothing else is established as having any of these properties at i. Consequently, as time progresses, the predicates *Green*, *Grue* and *Emerald* become entrenched as kinds (axioms $(2), (4), (10), (12), (13)$), and the following sentences are true in M:

$$\begin{aligned}&\mathit{Kind}(\mathit{Emerald})(T-1),\\ &\mathit{Kind}(Green)(T-1), \mathit{Kind}(\mathit{Grue})(T-1),\\ &\forall x, t \leq T-1(\mathit{Emerald}(x)(t) \rightarrow \mathit{Green}(x)(t)),\\ &\forall x, t \leq T-1(\mathit{Emerald}(x)(t) \rightarrow \mathit{Grue}(x)(t)).\end{aligned}$$

Up to time $T-1$ the pragmatic axioms (5) and (7) play no significant part in the reasoning; instances of these axioms with antecedents which refer to time points before $T-1$ have been trivially satisfied.

However a conflict arises at $T-1$ when it comes to predicting the colour of object O_T, which, by axioms (13) and (16), is an emerald. It follows from the displayed sentences and Axiom (6) that both $\mathit{Proj}(\mathit{Emerald}, \mathit{Green})(T-1)$ and $\mathit{Proj}(\mathit{Emerald}, \mathit{Grue})(T-1)$ are true in M. But, the projection assumptions $?\mathit{AffP}(\mathit{Emerald}, \mathit{Green})(T-1)$ and $?\mathit{AffP}(\mathit{Emerald}, \mathit{Grue})(T-1)$ cannot both be true in M. For then it would follow from axioms (6) and (7) that $\mathit{Green}(O_T)(T)$ and $\mathit{Grue}(O_T)(T)$ would be true in M, which would result in a contradiction. By Axiom (12), $\mathit{Blue}(O_T)(T)$ would be true, so, by Axiom (11) $\mathit{Colour}(O_T, C_2)(T)$ would be true. Axioms (1), (8) and (9) would give $\neg\mathit{Colour}(O_T, C_1)(T)$, which with Axiom (10) gives $\neg\mathit{Green}(O_T)(T)$.

Moreover, if $?\mathit{AffP}(\mathit{Emerald}, \mathit{Grue})(T-1)$ were true in M, then this assumption would result in the kind *Grue* becoming defective at T. For it would follow (as above) that $\mathit{Grue}(O_T)(T)$ would be true, with the result that (as above) $\mathit{Colour}(O_T, C_2)(T)$ would be true. But as (for instance) $\mathit{Green}(O_1)(1)$ is true, it follows from Axiom (10) that $\mathit{Colour}(O_1, C_1)(1)$ would be true. So it would follow from axioms (1), (8), (9), (10), (11) and (12) that $\neg\mathit{Colour}(O_1, C_2)(1)$, and $\neg\mathit{Colour}(O_T, C_1)(T)$ would be true. By chronological minimization of first-order atoms there would be no other y such that $\mathit{Colour}(O_1, y)(1)$ and $\mathit{Colour}(O_T, y)(T)$ were both true. Hence, by Axiom (2), $\neg\mathit{Similar1}(\mathit{Grue}, \mathit{Colour})(T)$ would be true. Moreover, it would follow by chronological minimization of second-order atoms that $\mathit{QPred}(S, O_T, \mathit{Grue})(T)$ would be undefined for any value of S other than *Colour*. So it would follow by Axiom (2) that $\neg\exists S\ \mathit{Similar1}(\mathit{Grue}, S)(T)$ would be true. Consequently, it would follow from Axiom (4) that $\neg\mathit{Kind}(\mathit{Grue})(T)$ would be true. So it would follow from the contrapositive of Axiom (5) that $\mathit{AffK}(\mathit{Grue})(T-1)$ would be true.

On the other hand $?\mathit{Aff}(\mathit{Emerald}, \mathit{Green})(T-1)$ can be true in M along with both $?\mathit{AffK}(\mathit{Green})(T-1)$ and $?\mathit{AffK}(\mathit{Grue})(T-1)$. In particular, $\mathit{Green}(O_T)(T)$ is consistent with $\mathit{Kind}(\mathit{Grue})(T)$, as *Grue* has no new members at T.

Now, as M is an I-preferred model, *AffK* atoms are minimized before *AffP* atoms at $T-1$. So the assumptions $?\mathit{AffK}(\mathit{Green})(T-1)$, $?\mathit{AffK}(\mathit{Grue})(T-1)$ and $?\mathit{AffP}(\mathit{Emerald}, \mathit{Green})(T-1)$ are true in M. So it follows (as above) that $\mathit{Green}(O_T)(T)$ is true in M, and consequently the following sentences are true in M:

$$\begin{aligned}&\mathit{Kind}(\mathit{Emerald})(T), \mathit{Kind}(Green)(T), \mathit{Kind}(\mathit{Grue})(T),\\ &\forall x(\mathit{Emerald}(x)(T) \rightarrow \mathit{Green}(x)(T)),\\ &\neg\forall x(\mathit{Emerald}(x)(T) \rightarrow \mathit{Grue}(x)(T)).\end{aligned}$$

It is clear from the last of these sentences that the second of the conjuncts to be proved is true. It is also clear now that the first conjunct to be proved is true; as, by chronological minimization, no emeralds are observed after T. □

The formal argument illustrates the need to maximize entrenchment assumptions (to minimize *AffK* atoms) before maximizing projection assumptions (minimizing *AffP* atoms) at each time point, thereby giving preference to the entrenchment axiomx over the projection axiom.

The paradox can be restated in formalized Grublish, with *Grue* and *Bleen* defined in terms of the *Colour* predicate, and *Green* defined in terms of *Grue*, *Bleen* and time point T. However, as induction depends on similarity and kinds rather than syntactic simplicity, the intended conclusions would still follow.

4 DAVIDSON'S DIFFICULTY

Davidson [3] asks us to consider the following hypothesis:

H_1 All emeroses are gred;

which states that "everything that is examined before t and is an emerald (or else is a rose) is green if examined before t (or else is red)" [p. 225]. He continues: "If H_1 is lawlike, it is a counterexample to Goodman's analysis . . . and one that would seem to cut pretty deep. Goodman's tests for deciding whether a statement is lawlike depend primarily on how well behaved its predicates are, taken one by one; thus for Goodman H_1 comes out doubly illegal. What H_1 suggests, however, is that it is a relation between the predicates that makes a statement lawlike, and it is not evident that this relation can be defined on the basis of the entrenchment of individual predicates", [pp. 225-6].

But is H_1 lawlike? If, Davidson suggests, we suppose that the following two hypotheses are true and lawlike:

H_2 All emeralds are green. H_3 All roses are red.

Then H_1 is true, and we have good reason to believe it.

Nevertheless, Goodman replies [8], it need not follow that H_1 is lawlike. The fact that H_1 is entailed by two hypotheses which are confirmed by their positive instances does not imply that H_1 is confirmed by its positive instances: " however true H_1 may be, it is unprojectible in that positive instances do not in general increase its credibility; emeralds found before t to be green do not confirm H_1", [p. 328]. However, Davidson counters: "The positive instances of H_1 are gred emeroses, and if they are examined before t they are also green emeralds examined before t. But green emeralds examined before t do not tell us anything about the colour of roses examined after t. Unfortunately, if this were a good argument, it would also show that H_2 is not lawlike, for the positive instances of H_2 examined before t would be nothing but gred emeroses examined before t; and what can they tell us about the colour of emeralds after t?", [p. 226].

Given the theory of induction developed in this paper, it seems that this dispute can be resolved as follows. Davidson is right in claiming that lawlikeness cannot be determined by considering the projectibility of its predicates "taken one by one", and that "it is a relation between the predicates that makes a statement lawlike". However, Goodman is right in claiming that H_1 is not lawlike because it "is unprojectible in that positive instances do not in general increase its credibility".

The idea of projecting the subkind relation in the projection axiom, Axiom (7), arose in response to Davidson's objection, and suggests the following definition: call a statement *lawlike* iff it is of the form $\forall x, t(P(x)(t) \to Q(x)(t))$, or is logically equivalent to a statement of this form, and $\forall t\mathit{Proj}(P, Q)(t)$ is true. The truth of a lawlike statement depends on the subset relation holding between its antecedent and consequent, its lawlikeness depends on the fact that the subset relation is also a subkind relation. So if we assume that the predicates 'gred' and 'emerose' are, like 'grue', both defined relative to some fixed future time point T, then H_1 is only lawlike if we suppose that no gred emeroses are examined at T or subsequently. If a gred emerose (a red rose) is observed at T, then the kind 'gred emerose' becomes doubly defective at T; because gred emeroses examined before T differ in both colour and shape from the gred emerose which is observed at T. So, while H_2 and H_3 are always projectible, H_1 would cease to be projectible at T and would be revealed as an accidental generalization. Thus it seems that positive instances of a hypothesis can only confirm it while the subkind relation holds.

5 CONCLUDING REMARKS

A great deal of work has been done on induction and machine learning; see, for example, [5, 7]. However, I believe that this is the first attempt to revive Hempel's programme and produce a logical account of induction which is not susceptible to Goodman's paradox. A more extensive treatment of this work is given in [2], including a discussion of the logical properties of the theory, in terms of Hempel's logical conditions of adequacy for confirmation [10] and Flach's rationality postulates for induction [4], and an extension to include common sense reasoning about change and inertia.

Recent work by Gärdenfors [6] develops Quine's notion of quality spaces into conceptual spaces, and proposes that natural kinds form convex regions in such spaces; although, as ever, Mr. Grue has his own idea of shmonvex regions, etc. The theory proposed here appears to complement Gärdenfors' work. While he develops richer definitions of kinds at what he calls the conceptual level (which roughly corresponds our robot's classifying colours and shapes), this paper has developed a theory of induction at what he calls the symbolic level.

This paper has dealt with representation. In future work the direct model-building implementation of the underlying theory of events [16] will be extended to the rest of the theory of induction, making actual robot induction of the kind envisaged here possible.

ACKNOWLEDGEMENTS

Earlier versions of this paper were presented at the ECAI 2000 workshop *Scientific Reasoning in Philosophy and AI*, and the Context 2001 workshop *Contexts in Logics*. I am grateful to everyone who has commented on it.

REFERENCES

[1] S.F. Barker and P. Achinstein (1960) On the new riddle of induction. *Philosophical Review* 69, pp. 511-22.

[2] J. Bell (2002) A Pragmatic Theory of Induction. Extended version of this paper. Available at www.dcs.qmul.ac.uk/˜jb.

[3] D. Davidson, Emeroses by Other Names. Note dated 1966, in: *Essays on Actions and Events*, Oxford Univ. Press, Oxford, 1980, pp. 225-227.

[4] P. Flach (2000) On the logic of hypothesis generation. In [5], pp. 89-106.

[5] P. Flach and A. Kakas, Eds. (2000) *Abduction and Induction: Essays on their relation and integration*. Kluwer, Amsterdam.

[6] P. Gärdenfors (2000) *Conceptual Spaces; The Geometry of Thought*, MIT Press, Cambridge, Mass.

[7] D. Gillies (1996) *Artificial Intelligence and Scientific Method*. Oxford University Press, Oxford.

[8] N. Goodman (1966) Comments. *Journal of Philosophy*, 63, pp. 328-31.

[9] N. Goodman (1983) *Fact, Fiction, and Forecast*, 4th Edition, Harvard University Press, Cambridge Mass. (1st Edition, 1955.)

[10] C.G. Hempel, A Purely Syntactical Definition of Confirmation, *Journal of Symbolic Logic*, vol. 8, 1943, pp. 122-143.

[11] S.C. Kleene (1952) *Introduction to Metamathematics*. North-Holland, Amsterdam.

[12] W.V. Quine (1969) Natural Kinds, in *Ontological Relativity and Other Essays*, Columbia University Press, New York, pp. 114-138.

[13] Y. Shoham (1988) *Reasoning About Change*, M.I.T. Press, Cambridge Mass.

[14] J.S. Ullian (1961) More on "grue" and grue. *Phil. Review* 70, pp. 386-9.

[15] S. Ullman (1996) *High-level Vision; Object Recognition and Visual Perception*, MIT Press, Cambridge, Mass.

[16] G. White, J. Bell and W. Hodges (1998) Building Models of Prediction Theories. Proc. KR'98, Morgan Kaufmann, San Francisco, pp. 557-568.