The Frame Problem: An AI Fairy Tale

KEVIN B. KORB
School of Computer Science and Software Engineering, Monash University, Clayton, Victoria 3168, Australia (email: korb@cs.monash.edu.au)

Abstract. I analyze the frame problem and its relation to other epistemological problems for artificial intelligence, such as the problem of induction, the qualification problem and the "general" AI problem. I dispute the claim that extensions to logic (default logic and circumscriptive logic) will ever offer a viable way out of the problem. In the discussion it will become clear that the original frame problem is really a fairy tale: as originally presented, and as tools for its solution are circumscribed by Pat Hayes, the problem is entertaining, but incapable of resolution. The solution to the frame problem becomes available, and even apparent, when we remove artificial restrictions on its treatment and understand the interrelation between the frame problem and the many other problems for artificial epistemology. I present the solution to the frame problem: an adequate theory and method for the machine induction of causal structure. Whereas this solution is clearly satisfactory in principle, and in practice real progress has been made in recent years in its application, its ultimate implementation is in prospect only for future generations of AI researchers.

Key words: artificial intelligence, frame problem, causal induction, machine learning, logicism, Bayesian learning, MML

1. Introduction

Once upon a time there was a little robot named Goldie. She lived in a world very different from yours and mine. It was called Blocks World. One day her designers (that's what they liked to call themselves) told her that they wanted her to bring them the *blue diamond.* They told her the blue diamond was in one of the three red boxes, either the little one, the medium-sized one or the great big one. (The designers were always telling Goldie things. Come to think of it, that was the only way she ever came to learn anything – except for the simplest of observational facts.) Goldie had a good model of her world (gratis her designers, of course), so she trundled over to the little tiny red box and peeked inside. The box was empty. Next she trundled over to the great big red box and just was able to raise her sensors over the edge. It was empty too. Finally she went over to the medium-sized box and picked it up (no need to look inside that one!). Suddenly, one of the designers asked "Is the blue diamond in the medium-sized box?" Now Goldie had no problem figuring out that the medium-sized box referred to was the red one, which she was holding. And she had no problem figuring out that she was holding it, since she had picked it up. And, as we saw, she knew it was the very same box that *used* to hold the blue diamond. But after picking up the box did it *still* hold the blue diamond? It turns out Goldie could not answer this question by ratiocination



Minds and Machines **8:** 317–351, 1998.
© 1998 *Kluwer Academic Publishers. Printed in the Netherlands.*

alone, for, when all was said and done, her head was as empty as anything in Blocks World.

This is McCarthy and Hayes's (1969) original frame problem: given some event or action and rules or laws specifying the resultant changes in some situation, how can an AI system deal with the *non-changes?* Stated thus it may appear to be a minor technical problem for AI, and, indeed, it has long been Hayes's principle claim about the Frame Problem that it *is* a technical problem for AI – albeit a major rather than minor problem (e.g., Hayes, 1987, p. 126). In a related response, Drew McDermott has claimed that it is a *problem* only for the narrow logicist approaches to artificial intelligence within which it arose and can be dismissed as of no import for broader approaches to AI (McDermott, 1987a). I claim contrariwise that the frame problem is a major problem for any artificial epistemology (AI) and more generally for any kind of epistemology. The reason it is such a general problem is precisely that Hayes is radically wrong in his view that the frame problem can be contained within the narrow technical framework within which he originally helped to introduce it.

The frame problem arose specifically within the context of McCarthy and Hayes's Situation Calculus, which was an attempt to formalize in first-order logic the common-sense reasoning that might be required of robots operating in the everyday world – which might be taken as a working definition of "logicism" in AI. [1] Briefiy (and crudely), the Situation Calculus introduces the following concepts:

Situations Situations are complete descriptions of the world at a particular time. They should be sufficient for predicting whatever may be predictable about the world at the next time step. Notationally, we shall write the set of all possible situations as *Sit* and use $s, s', \ldots \in Sit$ to represent particular situations.

Objects Objects are whatever it may be that we wish to reason about, people, things, etc. We shall designate them as $x, y, z \in Object$.

Fluents Fluents are functions whose domains are *Sit* or n-tuples one of whose elements is a member of *Sit.* There are two subspecies of fluents of interest:

Propositional fluents (predications): These are functions from *Sit* to *T (True)* and *F (False).* E.g.,

$$F : Sit \rightarrow \{T, F\}$$

$$F' : Sit \times Object \rightarrow \{T, F\}$$

Situational fluents (events, actions): These are functions from situations to new situations. E.g.,

$$F : Sit \rightarrow Sit$$

$$F' : Sit \times Object \rightarrow Sit$$

To illustrate the use of such a calculus, I now introduce the "Logic of Goings" (that is, as in going someplace). To reason logically about Goings we need some axioms – rules that are assumed generally true about Goings – and some initial conditions describing some situation about which we are to reason. For example:

Axioms (Al) *Animate(s, x)* ⊃ *At(Go(s, x, y), x, y)*

Initial Conditions (I1) *Animate(here&now, Fred)*

(I2) *Go(here&now, Fred, school)*

The axiom (actually, axiom schema) (Al) is to be read as asserting that if something is animate in a situation, then in the new situation obtained by going to *y* that thing will be "at" *y*. Given this and the initial conditions, we can logically infer that, after the initial conditions have been satisfied, Fred (an object) will be at school in the situation *Go(here&now, Fred, school),* which we might call "*there & then".*

Logicism has had quite a run within the AI community. It was the predominant theory (religion?) within the AI community generally from 1960 to 1980, more recently losing its grip due to successful competition from neural networks (Rumelhart and McClelland, 1986), genetic algorithms (Goldberg, 1989) and information-theoretic machine learning (Wallace and Freeman, 1987; Quinlan, 1993). Indeed, the ability to perform logical inferences, and even the fantasy ability to perform infinitely many logical inferences in finite time, has often been held to be the quintessential condition for achieving AI's goal, intelligence. Christopher Cherniak (1986) has pointed out in some detail, in his *Minimal Rationality,* how such criteria render the goal of AI unattainable in principle. [2] Despite these difficulties, logicism continues to dominate in the area of "knowledge representation", that is, efforts to figure out how to encode knowledge and belief in artificial systems and related concerns about how to reason with and about beliefs (e.g., defeasible reasoning, belief revision). [3] Thus, we find David Israel saying: "I am urging that we may want to think of the *first-order predicate calculus as a universal machine-language for Knowledge Representation"* (1985, p. 445; his emphasis). The basic idea of logicism is that something like the Logic of Goings can be completed – the ellipsis after axiom (AI) can be filled in – so that not only can we use it to reason about such simple situations and problems as Fred's, but we can use the completed theory to reason about *anyone's* goings *anytime* to *anywhere.* Perhaps the result will be unimaginably complex, but given any single aspect or fact about going places – what impedes arrival, what changes on the way, etc. – we can encode some axiom to represent it. Hence, by a kind of imaginary induction, we can imagine *completing* the Logic of Goings, eventually. Complexity alone is not much of an argument against this possibility: our brains are unimaginably complex as well, yet they exist and work. Having accepted the possibility of a complete and consistent Logic of Goings, we can now perform an imaginary super-induction: we can equally well imagine a Logic of Comings, a Logic of Stayings, a Logic of Eatings, a Logic

of Drinkings, etc. If we can agree that we can imagine reaching the end of this list, then we will have agreed that *any* situation in which common-sense reasoning might be required can be axiomatized, at least in principle; hence, a robot that implements the Situation Calculus can cope with all such situations. Hence, the logicist vision of an artificial intelligence is realizable. QED.

This may be a logical argument, but it is far from satisfactory.

Perhaps the most obvious difficulty that intrudes is that the whole picture is wrong: the image invoked is one of the human scientist-engineer stuffing axiomatic theories inside the little robot's big head. But if we are to take the claimed goal of AI seriously – the production of an artificial *intelligence* – we shall have to take seriously also the most prominent features of intelligence, what most prominently demarcates intelligent from unintelligent behavior, at least. And one feature that at least since Thorndike (1911) has been taken to most prominently differentiate the intelligent from the unintelligent is the ability to learn, to adapt behavior to new and challenging environments. If a robot, placed in an unanticipated environment, is unable to learn the simplest things about that environment and adjust to its simplest requirements, then that robot is not intelligent. But the fulfillment of the logicist enterprise envisioned above was of just such a monster: one whose head is full of pre-ordained theories, driven by the universal logical crank of *modus ponens*, with no ability whatsoever to adapt to new surrounds. This, I submit, is a disturbed vision of artificial intelligence. As we shall see, the precise role of *learning* for machine intelligence is an issue that refuses to go away.

It may be objected, with some justification, that some logic-based learning systems have shown some promise, such as inducive logic programming (e.g., Muggleton, 1992); I shall briefly consider the claims of inductive logic programming in §9 below. It may also be objected, and again with some justification, that the logicism I here deplore is a fanatic's religion and that most traditional AI researchers are prepared to accommodate deviations from the rigors of its edicts. For example, the information-theoretic classification trees and graphs of Oliver, Dowe and Wallace (1992) and Quinlan (1993) are not only recognized by traditional AI as a significant contribution to machine learning, but are often regarded as themselves within the tradition of symbolic AI, if not logicism proper. [4] Such tolerance is both laudable and desirable, particularly in view of the aggressiveness with which many of the disputes in the philosophy of AI are conducted. The classification of information-theoretic methods in AI as symbolic is not, however, particularly accurate: those methods in *principle* employ real-valued quantities, which are *in principle* strictly excluded by symbolicist AI. For example, their use violates Newell and Simon's stricture, recorded in the exegisis of their "Physical Symbol System Hypothesis" (1976), that all possible states of any intelligent system be designatable by representations of the system. [5] Consistent with this ideology, although not with the relevant facts of the matter, Simon has recently opined that "numbers seldom enter at all into" the computational modeling of cognition (Simon, 1994). My treatment of the frame problem is unapologetically directed

against the logicism which gave rise to its expression (a worthy achievement) and which continues to attempt to circumscribe its expression and possible resolution to logicist methods (which I think less worthy).

Returning to the remarkable super-induction to the truth of logicism, the less obvious difficulty with it just is the frame problem. Recall that we could infer of Fred that *At (there&then, Fred, school).* But what about all those things that *do not* change when Fred goes to school? What can we infer about Fred's hair color? sex? home address? These are emphatically *not* logical consequences of the initial conditions describing Fred's school-going situation plus the kind of axioms for a Logic of Goings described above. The latter just tell us those things which *change* when Fred goes somewhere, not what does not change. So, a purely logical system for common-sense reasoning needs to represent the latter explicitly, in what are called "frame axioms" (hence the name "frame problem"). For example, we could add as frame axioms:

Frame Axioms

$$\text{(Fl)} Hair-color(s, x, y) \supset Hair-color(Go(s, x, z), x, y)$$

$$\text{(F2)} Sex(s, x, y) \supset Sex(Go(s, x, z), x, y)$$

$$\text{(F3)} Home-address(s, x, y) \supset Home-address(Go(s, x, z)x, y)$$

Unfortunately, the enumeration of frame axioms is evidently hopeless: there is no end to the list of properties which are not affected by an action or event. Were the listing of frame axioms tractable, there would have been no frame problem. So, the frame problem can be considered to be: how to deal with non-effects of an event or action in a tractable way.

If AI is ever going to succeed, then the frame problem must be resolved – or dissolved – if for no other reason, because our best natural model for intelligent systems, humans, do not have such a problem. Of course, humans make mistakes about both the effects and non-effects of their actions and other events in the world, but put in Goldie's situation no normal human with a modest amount of worldly experience would be left baffled by the question of whether the diamond was still in the box. Whatever solution we may propose for our AI systems need not be perfect, but does need to work in ordinary circumstances, where moderately intelligent humans succeed.

2. Dennett's Fairy Story

Everyone who has written about the frame problem has some fairy tales to tell about it. Daniel Dennett has presented one of the more effective fairy tales for the frame problem – effective in raising relevant issues about the frame problem and

effective in provoking vigorous, if not always insightful, debates between philosophers and AI researchers on the subject. Here it is, again a problem for the likes of Goldie, but hardly for you or me (Dennett, 1984):

> Once upon a time there was a robot, named R1 by its creators. Its only task was to fend for itself. One day its designers arranged for it to learn that its spare battery, its precious energy supply, was locked in a room with a time bomb set to go off soon. R1 located the room, and the key to the door, and formulated a plan to rescue its battery. There was a wagon in the room, and the battery was on the wagon, and R1 hypothesized that a certain action which it called PULLOUT(WAGON,ROOM) would result in the battery being removed from the room. Straightway it acted, and did succeed in getting the battery out of the room before the bomb went off. Unfortunately, however, the bomb was also on the wagon. R1 *knew* that the bomb was on the wagon in the room, but did not realize that pulling the wagon would bring the bomb out along with the battery. Poor R1 had missed that obvious implication of its planned act.
>
> Back to the drawing board. "The solution is obvious," said the designers. "Our next robot must be made to recognize not just the intended implications of its acts, but also the implications about their side effects, by deducing these implications from the descriptions it uses in formulating its plans." They called their next model, the robot-deducer, R1D1. They placed R1D1 in much the same predicament that R1 had succumbed to, and as it too hit upon the idea of PULLOUT(WAGON,ROOM) it began, as designed, to consider the implications of such a course of action. It had just finished deducing that pulling the wagon out of the room would not change the color of the room's walls, and was embarking on a proof of the further implication that pulling the wagon out would cause its wheels to turn more revolutions than there were wheels on the wagon–when the bomb exploded.
>
> Back to the drawing board. "We must teach it the difference between relevant implications and irrelevant implications," said the designers, "and teach it to ignore the irrelevant ones." So they developed a method of tagging implications as either relevant or irrelevant to the project at hand, and installed the method in their next model, the robot-relevant-deducer, or R2D1 for short. When they subjected R2D1 to the test that had so unequivocally selected its ancestors for extinction, they were surprised to see it sitting, Hamlet-like, outside the room containing the ticking bomb, the native hue of its resolution sicklied o'er with the pale case of thought, as Shakespeare (and more recently Fodor) has aptly put it. "Do something!" they yelled at it. "I am," it retorted, "I'm busily ignoring some thousands of implications I have determined to be irrelevant. Just as soon as I find an irrelevant implication, I put it on the list of those I must ignore, and . . . " the bomb went off.

The question, of course, is what we ought to do for R2D2.

3. The Frame Problem Solved I: Default Logic

The idea for a solution that most people hit upon right away, and appears surely to be right, is that humans cope with the non-changes mostly by *ignoring* them. When we go to school, we simply fail to contemplate the host of non-changes occurring to our hair color, sex, homes, etc. We contemplate such things only when forced to, by someone's question or unusual circumstances. Hence, we can approach the problem as one of presumptive truth – that the normally unaffected properties are unaffected in any specific instance – without *attending* to them at all, unless there is a special circumstance that intervenes to affect the property. What we want then is an approach which allows our AI systems to ignore the non-changes ordinarily, but supports reasoning about them, and even retracting the conclusion of no change, when required – what has been called the "sleeping dog" strategy. [6]

There have been a number of attempts to formalize this intuition and fold it into the logicist program. Default logic is one of them.

Raymond Reiter's idea in introducing default logic (1978, 1980) was to extend logic by introducing new rules of inference which support non-valid modes of inference, in particular a kind of *defeasible* inference where conclusions could come to be retracted, but not because we discover that a premise has failed (which is the only ground for retraction supported by ordinary logics), instead because a defeater arises, we for example *observe* that the conclusion is false, even though the premises may remain clearly true. This kind of inference has come to be known as nonmonotonic because the chain of inferences may not always lead to a monotonic increase in the corpus of belief, some conclusions being withdrawn. One of the more common examples is that of the Tweety. Given that we know of Tweety only that it is a bird, we would ordinarily (if given occasion to) conclude that Tweety can fly. However, if we came to know that Tweety is a penguin, we would retract that claim, and we would do so without abandoning our premise that Tweety is a bird – or even the implicit premise that birds generally are capable of flight.

Reiter proposes that we add inference rules of the form:

$$\frac{A : B}{C}$$

Such a rule supports the defeasible inference of C given knowledge of A, so long as B is consistent with what we know. [7] More precisely, if A (conventionally called the *prerequisite)* is a member of K (where K is the corpus of knowledge / belief) and if B (the *justification)* is consistent with K, then defeasibly add C (the *consequent)* to K. If the above rule is named 'd', we can use 'PRE(d)', 'JUS(d)', and 'CON(d)' to refer respectively to its prerequisite, justification, and consequent; i.e., PRE(d) = A, JUS(d) = B, and CON(d) = C.

Successive use of default rules, employing intermediate default conclusions as the prerequisites of further rules, can characterize the default import of a state of belief K. We can introduce a formal concept of a default proof based upon this intuitive idea of using a sequence of default rules (cf. Reiter, 1980, p. 99).

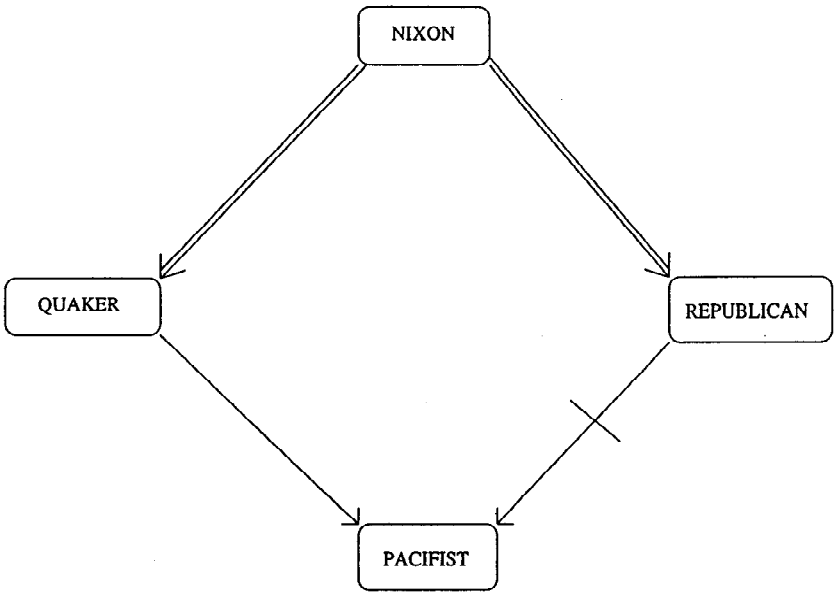


Figure 1. The Nixon Diamond

The resulting set of beliefs K′, obtained by some consistent application of default inferences, is called an *extension* of K.

Returning to the case of Tweety, the default rule for that inference would be:

$$\frac{Bird(Tweety) : Flies(Tweety)}{Flies(Tweety)}$$

This kind of rule is called a *normal* default because the consequent and justification are identical (i.e., JUS(d) = CON(d)); such rules can be abbreviated in the fashion Bird $\rightarrow$ Flies.

One difficulty which has been raised for default logic is that it very naturally gives rise to multiple extensions; that is, it frequently occurs that the default implications of our beliefs are just ambiguous – we cannot tell which of two consistent extensions of our beliefs which are mutually incompatible ought to be accepted. The "Nixon diamond" (referring to Richard M. Nixon) gives rise to a natural case of ambiguity. This is an inference net, where each '$\Rightarrow$' arc represents a universal generalization, '$\rightarrow$' represents a default rule, and '$\not\Rightarrow$' (or '$\nrightarrow$') represents the denial of a universal generalization (or, of a default rule) [8] – in this case, we have the default rule Republican(x) $\rightarrow$ $\neg$ Pacifist(x). Here we have two potential chains of inference, one leading to Pacifist(a) and one to $\neg$ Pacifist(a), where 'a' designates Nixon. [9] This means that we can use default logic to get two competing, incompatible extensions of K: K′ = {Nixon(a), Quaker(a), Republican(a), Pacifist(a)}* and K" = {Nixon(a), Quaker(a), Republican(a), $\neg$ Pacifist(a)}*. But the existence of multiple extensions here is not really a problem. It is true that *we* have grounds for

preferring one conclusion over the other, but those are *additional grounds.* Were we confronted with only the information that Nixon is a Republican and a Quaker, then the response that the information is ambiguous with respect to pacifism seems quite legitimate. In other words, it is a plausible interpretation of default logic that only sentences in the intersection of extensions to the knowledge base are to be considered valid default conclusions. [10]

3.1. DEFAULT LOGIC MEETS THE FRAME PROBLEM

The intuitive notion behind turning to default logic as a fix for the frame problem is that instead of invoking explicit reasoning about various unchanging properties we can let default rules cover such properties. We would not have to execute the default rules unless the situation demanded that we examine the default commitments of our explicit beliefs. Furthermore, if we can find rules that govern the acceptability of default rules, then we need not attempt the Herculean task of providing all possibly relevant default rules in advance – we could construct and use them on the fly, as needed. In other words, we could automate the learning of relevant default rules and invoke such learning only when the dogs have been awakened.

After moving Dennett's wagon, what we want of any sensible reasoner is that it: (1) believe that the battery has moved; (2) *not* believe (or even think about the possibility) that the bomb was moved – *unless* a defeater rears its head. [11] Some default rules that apparently achieve this are:

$$(D1) \quad Moved - to(t_0, x, y) \rightarrow At(t_1, x, y)$$

$$(D2) \quad At(t_o, x, y) \rightarrow At(t_1, x, y)$$

$$(D3) \quad Supported - by(t_0, x, y) \,\&\, Moved - to(t_0, y, z) \rightarrow At(t_1, x, z)$$

This, together with initial conditions describing the placements of the battery and bomb, etc. will lead to the appropriate beliefs requested above. As we noted, the adequacy of this answer is subject to our ability to introduce new defaults, and to introduce new qualifications on their application, as exceptions or difficulties may arise. Thus, we may need further defaults to deal with cases where the battery, although supported by the wagon, is tied to the door knob, *unless* the rope is badly frayed, *except* when the wagon's surface has been greased, ... Assuming for the sake of argument, that all difficulties with specifying a semantics for default rules and methods of learning them have been resolved, then we seem to have a resolution of the frame problem. At any rate, we have a resolution of at least some *instances* of the frame problem, such as Dennett's wagon problem and Goldie's problem with the diamond. We have converted the problem raised by the Situation Calculus – how to generate and cope with infinitely many frame axioms – into a problem that at least *might* be tractable – how to understand and learn default rules – all within the logicist framework.

Unfortunately, this view turns out to be far too optimistic. The first, and in my opinion decisive, difficulty is that there is no logicist theory in the offing for how

to understand and generate default rules, as Reiter confesses (1987, p. 183). The only *promising candidate* ever put forward for interpreting default rules is a *probabilistic* semantics, and the learning of contingent probabilistic relations is nothing any purely logical system can rationally aspire to. This objection I have already put elsewhere (Korb, 1995), so I turn to another. Although multiple extensions in the Nixon diamond appeared to be innocuous, because there is a good argument that given only such a modest set of beliefs we *should* be indecisive about Nixon's warlike nature, there are other cases that are not at all innocuous. In Hanks and McDermott's (1986) "Yale Shooting Problem" in particular, we have a case with multiple extensions when it is manifest that we *should not.*

> Suppose first (t_0) that Fred is alive and Goldie is, well, Goldie the robot – we need not decide whether she is alive or not. Goldie picks up a loaded gun and aims it at Fred and fires (t_1). Is Fred alive at t_2?

This is not much of a challenge. At least, not if we presume that the gun, its charge, aim and distance are such as to nearly preclude survival. A plausible encoding then of the inferential relations we use to answer the question, putting them into default rules, is:

$$(Dl) \quad Loaded(t, x) \rightarrow Loaded(t', x)$$
$$(D2) \quad Alive(t, x) \rightarrow Alive(t', x)$$
$$(D3) \quad Loaded(t, x) \,\&\, Fired-at(t, x, y) \,\&\, Alive(t, y) \rightarrow \neg Alive(t', y)$$

The first two rules allow us to retain the right context from the beginning of the story to the crucial stage when Goldie fires the gun; that is, they assert that normally (by default) loaded weapons remain loaded and live people remain alive (the hyphen, as in t', is meant to indicate the next successive time step). If we interpret "*Fired - at*" in the strong sense suggested above, we can employ default rules in the the following way, to recover the ordinary story we want. Certainly, each of the defaults is plausible: they correspond to conclusions we would all normally make in isolation from defeating conditions. Thus, we get the following sequence of default conclusions:

$$Loaded(t_1, gun) \text{ (using } D1)$$
$$Alive(t_1, Fred) \text{ (using } D2)$$
$$\neg Alive(t_2, Fred) \text{ (using } D3)$$

Unfortunately, as Hanks and McDermott point out, we can equally well obtain a different sequence of default conclusions:

$$Loaded(t_1, gun) \text{ (using } D1)$$
$$Alive(t_1, Fred) \text{ (using } D2)$$
$$Alive(t_2, Fred) \text{ (using } D2)$$

How does this last step square with the rest of the story? Of course, we cannot have that Fred is both alive and dead, so if we stick with the latter story we have to deny the first. But relying upon the defaults available there is no difficulty accepting the implausible second story: there is no more reason to use D3 at the end than there is to use D2 twice. Perhaps in this alternative story when Goldie fires the gun, instead of the gun discharging, it unloads itself in a non-standard way, leaving Fred alive, the gun unloaded, and the bullet lying on the floor. In short, what we have with these defaults is a case of multiple extensions and no way, within default logic alone, to resolve the issue.

Some have tried to employ a syntactical rule of maximal specificity to force such things, noting say that the antecedent of D3 is more detailed than that of D2. However, such reasoning is notoriously language-specific. For example, we could defeat such reasoning by introducing a single primitive predicate whose semantics are equivalent to those of the antecedent of D3, replacing D3 by, for example:

$$(\text{D3}')\, Twiddled(t, y) \rightarrow \neg Alive(t', y)$$

The problem here does not seem to be one of specificity alone, even if we consider *semantic* specificity: we can identify all kinds of additional properties of Fred and incorporate them into the antecedent of D2 (in order to defeat the greater semantic specificity of D3), all without making sense of the second story.

In contrast with the Nixon case, it is not so plausible here to say that the ambiguity is normal, that is, that the problem is simply that D1-D3 failed to capture the knowledge that we humans in fact employ to solve the problem, but could be readily extended with further defaults to do so. Certainly the defaults fail to capture the knowledge we employ, but it is hard to see what *explicit* knowledge encoded in further symbolic default rules would make the right difference. Humans, knowing what the defaults D1-D3 are *supposed* to encode already, do not encounter this difficulty in deciding upon an "extension". It does seem intuitively plausible that humans are relying upon the information D3 is supposed to encode (our knowledge of guns) being more specific than the information that D2 is supposed to represent (our knowledge of the more general tendency of life to persist). But the concept of specificity that intuitively seems relevant here is that of Reichenbach (1949, p. 374) and Salmon (1971): the maximal reference class (of, for example, incidents of guns firing) which is statistically *homogeneous* – within which no subclass determined by properties measurable prior to knowledge of the outcome (e.g., death) would reveal a probability of that outcome distinct from that of the original reference class.[12] But such considerations are not purely symbolic, they are *essentially* quantitative, and so out of bounds for any purely symbolicist or logicist research program. In any case, at a minimum, the default logic research program has failed thus far to produce any effective means for mimicking human judgments in such matters as the mysterious shooting at Yale.

In short, default logic may be usable for reasoning without the frame problem intruding in some cases. However, there are two major and substantial restrictions

upon its utility: it is lacking an accepted semantics for its rules, which it must have for a satisfactory general solution to the frame problem, since it needs to be augmented by a method for learning defaults as the situation demands; it fails to support useful reasoning whenever multiple defaults lead to contradictory conclusions, when nothing within default logic alone provides a means to arbitrate between them. This latter situation, unfortunately, appears to be very common: every exception case to a default which is not itself universal will lead to conflicting defaults. For example, insects generally fly, but ants generally do not; of an ant default logic must remain silent, however, since *some* ants do fly. Humans have no problem concluding of a randomly selected ant that it (probably) does not fly.

4. The Problem of Induction and the Frame Problem

David Hume's problem with induction (1748) was that we fail to *perceive* causal relations: we perceive both smoking and cancer, but *not* that smoking causes cancer; we perceive a boy throwing a baseball and a window breaking, but not that throwing the ball *caused* the window to break. So why, asked Hume, do we assert that some X has caused some Y? Because we have been exposed to a pattern of regular association between the two, that is *we* become habituated to first X, then Y – it is a fact about us, not the world, that we acquire an expectation about Y given X, which we then honor by claiming that the world incorporates a necessary connection between the two. In reality, there is no necessary connection of the kind, we can *imagine* X followed by anything whatsoever: fire by freezing, eating by hunger, death by life (this one being more common than the others, for obscure reasons). If we decide these will not follow, that is because (?) of our experiences rather than because of logical reasoning. In short, there is no *empirical* difference between: (A) Whenever X occurs Y occurs; and (B) Whenever X occurs Y occurs necessarily. We have no way to distinguish these cases observationally, so there is no substantive, non-psychological difference.

In consequence, no matter how many conjunctions of X and Y we may experience, there is no justification for concluding after the next X that Y will follow. The past provides no grounds for belief about the future. There is no such thing as *rational* induction from the data.

Recently, machine learning researchers have rediscovered Hume. Thus, Schaffer (1994) has repackaged Hume's problem with induction as the "Conservation Law" of induction (cf. also Wolpert's "No Free Lunch" theorem, 1995): that inductive generalization is a "zero-sum enterprise – for every performance gain in some subclass of learning situations there is an equal and opposite effect in others." To take a simple example of what he has in mind, suppose we adopted the inductive rule: If you've seen every instance but one of some class A and each instance has property B, then predict that the last instance is also a B. This rule has very limited application. It applies only to finite domains and, even then, only when we've nearly exhausted the domain observationally. Nevertheless, even within this

limited scope of application, according to Schaffer, it fails to perform any better than flipping a coin. The reason is that for each logically possible case where the rule works – because the world is such that all A's are B's – there is an equal and opposite logically possible world where the rule fails – because the final instance is not-B. What we have, in other words, is that for every possible world that supports the utility of some inductive rule there is an equal and opposite possible world that undermines that inductive rule. Unless we have some *a priori reason* to believe we are in one possible world rather than another, we can have no reason to prefer one inductive rule over another.[13] Of course, if we cannot have reasons to prefer one inductive rule over another, then we shall not have any good reason to prefer one induction over another, as rules can be invented to support any induction. Hence, we are once again driven to the conclusion that humans and their artificial inventions, such as robots and scientific theories, are guided by nothing more rational than their passions and urges and have no claim to having a superior, rational approach to understanding the world.

One possible view is that the frame problem *just is* the problem of induction. In fact, this is James Fetzer's (1991a) view of the frame problem. The problem of induction is the problem of how to form reliable expectations about the unknown; the frame problem is the problem of how to form reliable expectations about the consequences of events and actions. Since the world of events, their precursors and successors, just is the world, the two problems are the same. Dennett (1984) has a simple argument that this is wrong. Imagine an act of God which gives an agent all and only reliable beliefs about the (some) future. The agent then will have all of the frame axioms by fiat. But the agent might still have a "bad case of the frames": it must still *avoid* reasoning with those frame axioms when solving problems about the future that are not affected by them. The frame problem is as much about *using* justified beliefs as it is about *finding* them.

5. The Many Frame Problems

Pat Hayes is even more emphatic in rejecting the "problem of induction" interpretation of the frame problem. The frame problem does not have to do with predictions of *changes* – of the effects of our actions and events – but instead how we deal with, appropriately represent and reason with, the *non-changes* (1991, p. 72). Indeed, the frame problem is not about justification or inference at all, but is simply a problem about representation.[14] Although in Dennett's wagon tale R1D1 and R2D1 wandered off into the haze either thinking or furiously not-thinking about irrelevancies, that too is not specifically the frame problem – that is the relevance problem. Also, the point that frame axioms can be overruled in specific cases by exceptions, that the battery may have been tied to a doorknob, etc., is again not the frame problem; it is the qualification problem. For one thing, most rules (including non-frame axioms) are similarly subject to qualification; for another, we would have a problem with representing infinitely many framing conditions even if they

were not subject to qualification. Hayes also points out (1987) that dealing with complex causal stories is again not the frame problem: the frame problem arises in the simplest circumstances. Thus, he notes, if he were to pick up a glass of water, we know right away the answer to the question, will the table follow? Finally (that is, this is the last of the many problems Hayes distinguishes from the frame problem that I will report), Hayes points out that the general AI problem – "GAIP" – is not the frame problem: that problem is the all-embracing one of how to build a real artificial intelligence. It is very clear that without solving the frame problem we will not solve GAIP, yet the solution of the frame problem is a specific precondition to a solution of GAIP, and not the very same thing. This paragraph, I believe, adequately sums up much of Pat Hayes's contribution to the discussion of the frame problem since its introduction.[15]

If Hayes is right, even in part, then default logic was never very promising as an instrument for solving the frame problem. Default logic, if it provides anything, provides a mechanism for dealing with qualifications; it provides a generic mechanism for defeasible reasoning, without providing any guidance for selecting among the many available defaults, and therefore without any guidance in how to avoid using frame rules when they are not relevant to a problem.

6. The Frame Problem Solved II: Circumscription

Circumscription, introduced by John McCarthy (1980), has seemed to many a more promising approach to handling the frame problem, and again one which is true to logicist impulses. McCarthy's idea was to introduce a second-order axiom schema which could be used to let properties persist unless something specifically overrode that persistence – a more direct attempt to implement the sleeping dog strategy. Somewhat in line with Grice's rules of implicature (1975), we should presume that an action affects a property only if we are directly told that it does; that is, predicates which report *changes* should be minimized as much as possible, while retaining consistency with whatever is explicitly reported. We can implement this in the following way: Given a conjunction of first-order axioms A in which the predicate we want to minimize, P, appears, the circumscription of P in A(P) is:

$$A(P) \wedge \forall \Phi \neg (A(\Phi) \wedge \Phi \subset P)$$

where $\Phi \subset$ P is true if and only if

$$\forall x[\Phi(x) \supset P(x)] \wedge \neg \forall (x)[P(x) \supset \Phi(x)]$$

Given the truth of the circumscription sentence, P must be minimal, for if P were any larger than what A(·) forces, some smaller Φ would violate the circumscription.

For the Tweety example, let P be Abnormal – in particular, the flightless kind of abnormality that some bird types have. Let all that we know about birds be expressed by the sentence A(P):

$$\forall x[(\neg P(x) \wedge Bird(x)) \supset Flies(x)] \wedge \forall x[(Penguin(x) \vee Kiwi(x)) \supset P(x)]$$

In other words, we know that normal birds fly and that penguins and kiwis are not normal. The circumscription of P relative to A($\cdot$) is:

$$A(P) \wedge \forall \Phi \neg (A(\Phi) \wedge \Phi \subset P)$$

which is equivalent to

$$A(P) \wedge \forall x[P(x) \equiv (Penguin(x) \wedge Kiwi(x))]$$

Since circumscription entails that all the Abnormal cases are those which we explicitly know of, this is sometimes called "predicate completion". The circumscriptive approach to nonmonotonicity supposes that when we need to retract a defeasible conclusion, such as that Tweety can fly, we do so because we are confronted with some new explicit evidence contradicting the conclusion, such as that Tweety is a penguin.

As with default logic, there has been considerable effort to specify conditions and rules which will allow circumscription to support all and only nonmonotonic inferences that humans find sensible. Similarly, much effort has been expended to get the circumscriptive solution to the frame problem just right. I will not concern myself with the nuances of different formulations of circumscription, for the question before us is whether circumscription is even aimed in the general vicinity of the frame problem. To examine this question I present Lifschitz's (1987) version of a circumscriptive resolution of the frame problem and, in particular, how Lifschitz applies his formalism to the Yale Shooting Problem.

Let $a, a' \ldots$ range over actions, $s, s', \ldots$ over situations and $f, f', \ldots$ over states (properties of situations). Causal structure is supposed to be described by direct causal links between states, which are expressed in the form *causes (action, state*$_1$, *state*$_2$) – i.e., if *state*$_1$ *(s),* then *state*$_2$ (*action(s)).* For the Yale Shooting Problem we adopt the following domain axioms:

(A1) *causes*(*load*, *unloaded*, *loaded*)

(A2) *causes*(*shoot*, *loaded*, *unloaded*)

(A3) *causes*(*shoot*, *alive*,¬*alive*)

Lifschitz introduces.a concept of feasibility: an action is feasible if, in a situation, it is possible to carry it out. This requires that any precondition for the action obtain, which we may write *precond(action, state).* For the shooting problem Lifschitz introduces the following precondition:

(A4) *precond*(*shoot*, *loaded*)

Lifschitz now introduces some definitional statements (schemata) about feasibility and what it is for an action to affect a state (property) in a situation:

$$(D1) \quad feasible(a, s) \equiv_{df} \forall f[precond(a, f) \supset f(s)]$$

That is, an action is feasible if and only if all of its preconditions are true.

$$(D2)\ \textit{affects}(a, f, s) \equiv_{df} feasible(a, s) \wedge \exists f' causes(a, f, f')$$

That is, an action will affect a state if and only if the action is feasible and there is some new state caused by that action in the situation. The next schema Lifschitz calls the "prediction axiom", since it requires that feasible causes have their effect:

$$(D3)[\ \textit{feasible}(a, s) \wedge causes(a, f, f')] \supset [f'(a(s)) \equiv f(s)]$$

Finally, Lifschitz's frame axiom:

$$(D4)\ \neg affects(a, f, s) \supset [f(a(s)) \equiv f(s)]$$

Given all of this apparatus, we can now circumscribe both *causes* and *precond* – i.e., we complete them so that the explicitly required causes and preconditions in Yale Shooting Problem are the *only* causes and preconditions. This results in the following predicate completions:

$$\begin{aligned} \forall aff'[causes(a, f, f') \equiv\ & (a = load \wedge f = unloaded \wedge f' = loaded) \\ & \vee (a = shoot \wedge f = loaded \wedge f' = unloaded) \\ & \vee (a = shoot \wedge f = alive \wedge f' = \neg alive)] \\ \forall af[precond(a, f) \equiv\ & (a = shoot \wedge f = loaded)] \end{aligned}$$

From all of the above, and the initial conditions *unloaded(s)* $\wedge$ *alive(s)* we can simply *deduce:* $\neg$ *alive(shoot(load(s))).* Since this is a deduction – a monotonic inference – there is no chance of also deducing *alive(shoot(load(s)))* (unless the axioms are inconsistent!). The only way to block this deduction is to introduce new preconditions. Hence, we are able to infer what we need (whether a positive prediction or a non-change) without infinitely many frame axioms. Therefore, the frame problem is dead.

7. The Frame Problem Undead

The notion that we have solved the frame problem by introducing a formal mechanism for computing predicate completions seems to me rather far fetched. There are substantive questions about the generic axioms. Both D3 and D4 require features of determinism, D3 implying that feasible actions always produce their effects and D4 implying that there are no uncaused changes. D4 additionally implies that at each "time step" in the world only a single causal action or event can occur. The whole approach behind the Situation Calculus and these offshoots seems to assume a discrete time universe in which parallel events either do not occur or, if they do, they do not interact. Whatever universe this is, it is not ours. [16]

But setting aside these metaphysical difficulties – granting the ontology of the Situation Calculus, the mechanism of circumscription and the dubious generic

axioms D1-D4 – what Lifschitz has done is to wheel in just the right domain-specific axioms A1-A4 to obtain the result he desired. If it counts as a solution to the frame problem to find any mechanism X which produces sentences and for any specific instance of the frame problem to find some input A for X such that X produces the desired sentence, then there are *infinitely* many *easy* solutions to the frame problem. For example, anybody's word processor will produce the output ¬ *alive(shoot(load(s)))*, if it is encouraged in the right way. Similarly, any of the RxDx series of robots could have solved the frame problem if only Dennett had had the wit to pose the problem correctly – for example, by telling the robot the right plan in the first place.

If you think this misses the point, then you had better be prepared to answer the question: Where did the domain axioms come from? One thing is certain: if you substitute the following axioms for A2 and A3 above, Lifschitz's circumscriptive machinery will get us back the same old wrong answer to the Yale Shooting Problem:

$$(A2') \; causes(shoot, loaded, bullet - lying - on - floor)$$

$$(A3') \; causes(shoot, alive, alive - and - laughing - at - you)$$

8. The Frame Problem Solved III: McDermott

McDermott brings much the same criticism to circumscription, but claims then that the frame problem really is just solved by wheeling in any old X that can spit out solutions to particular problems. McDermott correctly points out that the circumscriptive machinery is *epiphenomenal* (1987b, p. 111): since the axioms must be found and cared for by humans, circumscriptive logics simply provide the shell of a solution. We are always going to need humans to hold the hands of our robots. McDermott claims that the frame problem, while real enough for logicist attempts at AI, is not real for AI generally. In particular, he claims that "no working AI program has ever been bothered at all by the frame problem" (1987a, p. 116). Here he did not mean to be relying upon the fact that there are no working artificial intelligences, he meant rather that programming up solutions to "frame problems" is nearly as trivial as using a wordprocessor.

Contrary to all the pessimistic moaning of AI researchers, writing algorithms which avoid the frame problem is just a breeze. He gives the example: writing an algorithm for a robot to pull an object is trivial! Figure 2 presents McDermott's algorithm, *sans* frame problem (1987a, p. 118).

McDermott's solution certainly works! *If* the world is sufficiently cooperative. Similarly, the following solution works for Dennett's problem: put R2D2 to sleep; remove the battery from the room; turn R2D2 back on. But what will happen to McDermott's algorithm in an uncooperative world? What if . . . there is an obstacle

in front of tuggee? Oops! Or, I put a hand grenade in the Tugger's hand? Oops! Or, what if the various steps described by "find everything connected", "figure out the forces", "find the weak links", turn out to be non-trivial? The latter steps are little better specified in McDermott's description than Lifschitz's remarkable step 1: get the right axioms.

All of the steps that have been missing in all of the solutions so far have had one thing in common: find the right default rules; find the right domain axioms; find the right links and connections. What is missing in every solution to the frame problem is a theory of how to *find out* about the world, a theory of machine induction. Pat Hayes is no doubt right in formally distinguishing the frame problem from the general AI problem from the problem of learning about the world. But that just is not the important point; the important point is that we can solve any one of these problems only by solving them all.

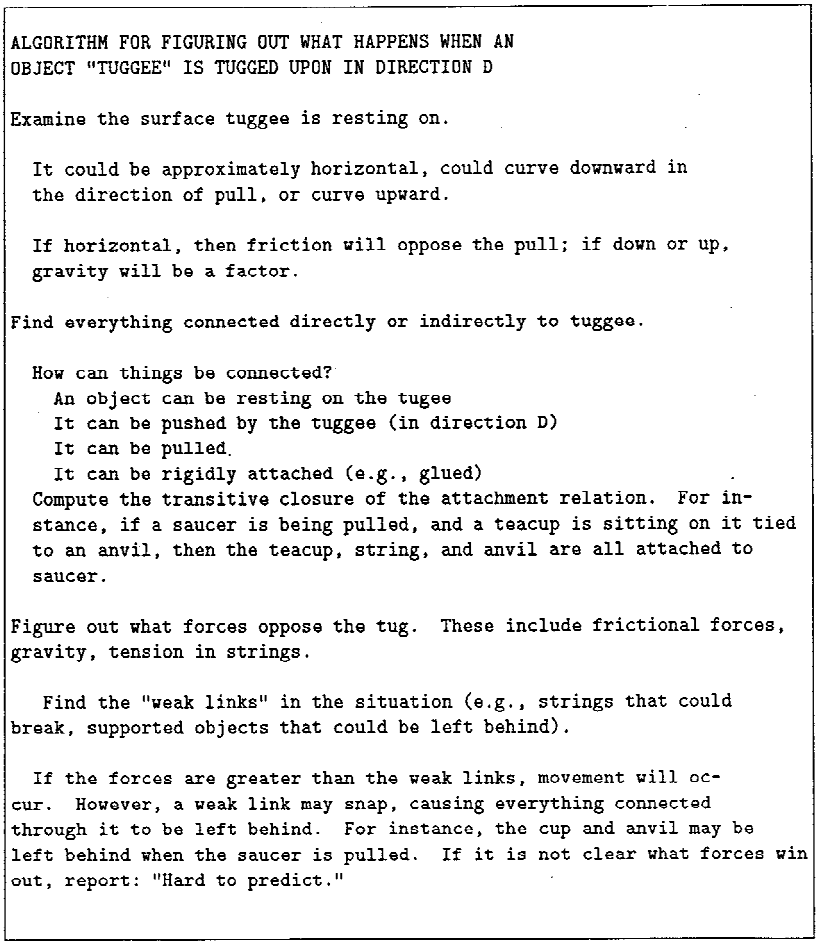
9. The Real Solution to the Frame Problem

There is a solution to the frame problem. Consider Dennett's instance of the frame problem: R2D2 must work out (plan) how to remove its battery from a room, *without* worrying about wall color and so on, *while* worrying about rope, grease and so on but *only if* they actually are *causally relevant* to the outcome. I now present two solutions to this problem.

9.1. SOLUTION 1: CAUSAL MICROMODELS

In the first solution I do not worry about what is physically and epistemically possible – it is a fantasy solution, if you like, that nevertheless leads the way to a real solution. What we do is build scale models of the Dennett's room inside R2D2's head.[17] These models are superb models of the real room about which R2D2 must reason; in fact, they are so good they are *causally isomorphic.* to the real room. That is, there is a one-to-one mapping from Dennett's room to each scale model that exactly preserves the causal structure of the real room in the scale model. Thus, if there is a bomb on the wagon in the real room, there is a scaled-down bomb on the wagon in each scale model; if there is a rope tying the battery to a door knob in the room, so too is there a scaled-down rope tying the scaled-down battery to a micro-knob in the scale model; etc. The requirement here is specific and *exhaustive:* for example, the rope will fray and break *if and only if* the micro-rope will do the same in the same circumstances. (We shall also need to assume that time is scaled down, so that R2D2 has some opportunity to explore the micro-models before losing its battery.)

Given an open-ended supply of such models (and micro-bomb resistant chambers in which to put them), R2D2 can solve its problem with ordinary AI search techniques: it can search the state space obtained by applying various operators to its scale models, until it finds a combination of actions that works. Each time a



```
ALGORITHM FOR FIGURING OUT WHAT HAPPENS WHEN AN
OBJECT "TUGGEE" IS TUGGED UPON IN DIRECTION D

Examine the surface tuggee is resting on.

  It could be approximately horizontal, could curve downward in
  the direction of pull, or curve upward.

  If horizontal, then friction will oppose the pull; if down or up,
  gravity will be a factor.

Find everything connected directly or indirectly to tuggee.

  How can things be connected?
    An object can be resting on the tugee
    It can be pushed by the tuggee (in direction D)
    It can be pulled.
    It can be rigidly attached (e.g., glued)
  Compute the transitive closure of the attachment relation.  For in-
  stance, if a saucer is being pulled, and a teacup is sitting on it tied
  to an anvil, then the teacup, string, and anvil are all attached to
  saucer.

Figure out what forces oppose the tug.  These include frictional forces,
gravity, tension in strings.

   Find the "weak links" in the situation (e.g., strings that could
break, supported objects that could be left behind).

  If the forces are greater than the weak links, movement will oc-
cur.  However, a weak link may snap, causing everything connected
through it to be left behind.  For instance, the cup and anvil may be
left behind when the saucer is pulled.  If it is not clear what forces win
out, report: "Hard to predict."
```

Figure 2. McDermott's Algorithm

scale model blows up the micro-battery R2D2 knows it has a losing sequence of operators and can scratch that plan. Assuming that there is a solution, R2D2 will find it (and whether it finds it in time or not is an ordinary time complexity issue about its search technique).

In short, given a magical supply of magical micro-rooms, the Frame Problem is no more. If there is no rope tied to the battery, then R2D2 need not – and will not – "think" about the possibility; all of the scale models it manipulates ("all of the situations it thinks about") will lack such a rope. R2D2 will think about the color of the wall if and only if the color of the wall actually figures in a positive or negative sequence of actions which R2D2 tests. When operating in the real world – or such a close surrogate for the real world – the sleeping dog strategy is automatic; it is trivial that properties surface in "thought" only if they surface in reality.

Dennett has a ready answer to my "solution"; he would say, "Korb, you're a cheat! You're using *cognitive* wheels to get your robot rolling." A wheel, according to Dennett's definition, is a biologically implausible, but highly effective, mode of transport. A cognitive wheel is cheating at AI–that is, it is the attempt to substitute some arcane and artificial technology for cognition in the place of something that ought to be biologically plausible.

In response, I confess to cheating! There are at least two distinct approaches to AI: *descriptive* AI aims at building computational mechanisms that simulate human or animal cognition by implementing true theories of their natural cognitive processing mechanisms (this is what Herbert Simon calls "computational psychology"). By contrast, I am more directly concerned with *normative* AI: the attempt to build computational mechanisms that at the behavioral level demonstrate intelligence and at the micro-level do anything that works. Since in normative AI it is not a goal to simulate anything biological (although if something biological works that would surely be taken as a promising approach to anybody's AI), there can be no such thing as a "cognitive wheel".

Of course, it remains true that for anybody's AI proposed mechanisms must actually be capable of implementation. It remains a drawback to my solution that it invokes magic. We do not know how to program magic.

9.2. SOLUTION 2: CAUSAL INDUCTION

Hence we must move on to the next, and last, solution. This one is the real thing. First, it is open to implementation, as far as we know; in fact, it is being implemented now in various research laboratories around the world, having become a hot research topic in AI in the last few years. Second, it retains all the important features of the scale-model solution; in particular, the whole idea of these research projects is to produce computational models of physical systems which mimic their causal characteristics as closely as possible. The research programs I am referring to have as their stated goal *Learning Causal Structure* or *Learning Bayesian Networks.* This research has exploded in popularity because of recent work on graphical representations of probability distributions has made the use of probabilities for "reasoning under uncertainty" tractable in many realistic cases. As a result, there is great interest in using Bayesian networks as a primary representation for expert systems (Heckerman et al., 1995). But going further than that there is great interest in automating the learning of Bayesian networks in order to allow intelligent systems to adapt to change, deal with massive amounts of information that humans have not codified (in so called "knowledge discovery in data bases"), and to support greater autonomy for artificial systems. [18]

But what has causal learning to do with the frame problem? Everything. The causal theories (Bayesian networks) that are learned are, ideally, causally isomorphic to the physical systems about which the AI system is to reason. Such an ideal is generally no more obtainable for an artificial system than it is for a hu-

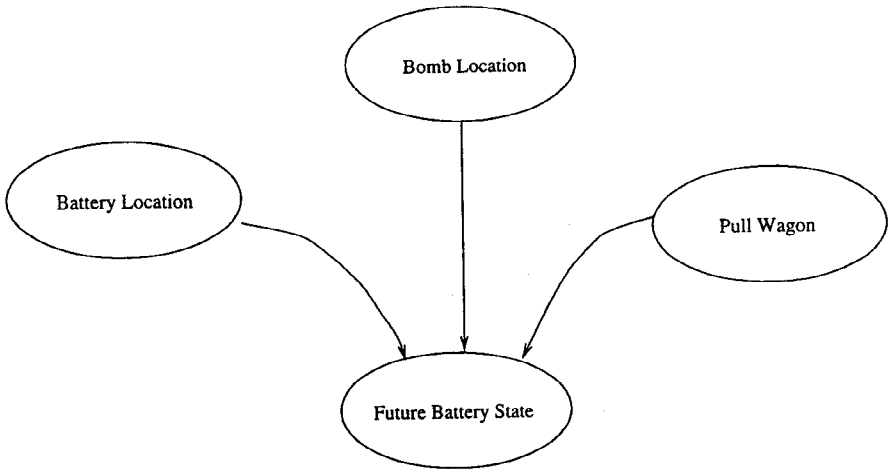


Figure 3. Wagon model

man scientist, so we can substitute a more realistic goal: learning a theory that is causally homomorphic to the physical system under study. This means that some states or properties that are distinct in the physical system will be collapsed into a single state or property in the learned theory, or perhaps go entirely unrepresented. Another way of putting this is: when R2D2 enters Dennett's room and observes its features it will construct a causal model which it will subsequently manipulate in order to test various plans (in lieu of ideal scale models). Thus, R2D2 may adopt the causal model outlined in Figure 3. Whether this model suffices for the problem depends upon the learning that has gone into building the model (e.g., whether the probabilistic relations incorporated in the model, which are not shown in Figure 3, are accurate) and on the extent to which the model structurally matches the problem. For example, if a rope ties the bomb to the wagon, this will need to be reflected in the model, perhaps by an additional variable.

The empirical background to learning such causal structures includes the joint measurement of observable variables and so their covariations. Such data accumulated have been used since Sewall Wright (1934) to confirm or disconfirm causal models that might explain them; and more recently have been applied in computer programs to the same effect by Glymour and company at CMU (e.g., Spirtes et al., 1993) and by us (e.g., Wallace et al., 1996). The empirical data may also be supplemented by experimental interventions, as discussed by Spirtes et al. (1993).

Two ways in which the learned causal models might fail to capture all of the variables relevant to a problem are: by failing to generate a hypothetical (hidden, latent) variable which is operative in the real system; and by incorrectly selecting among the variables measured, missing the relevance of one or more of them. For the causal induction solution to the frame problem to be realized, it is not necessary that automated causal induction never succumb to such difficulties, but

it is necessary that our programs do not immediately or easily succumb to them. Latent variable discovery is the hard problem here, however we have some ground for optimism that it is a problem that can be solved. Humans routinely learn about theoretical entities and even come to manipulate them in their problem solving. For some examples: atomic nuclei, gravity, chemical bonds, human personality characteristics, Mendelian genes. And there are various approaches in statistics and machine learning for postulating latent variables and testing for their effects on measured variables. Thus, latent variable discovery has already been done with some modest success, although research into it seems to be in its early stages (see again Spirtes et al., 1993, and also Wallace and Freeman, 1992, and Wallace, 1995). Variable selection, by contrast, is already done in all causal discovery algorithms and is represented in the explicit causal relations which they attribute to pairs of variables. Finally, another way in which the causal model applied to a reasoning problem may fail to include relevant variables is in a failure of the process whereby a previously learned model is matched to the reasoning problem. But such pattern matching is comparatively well understood and I will assume in application to the frame problem it introduces no new issues.

In short, the causal induction solution to the frame problem is as follows. Learned causal models have the right properties to solve the frame problem in practical cases, because their behavior mimics that of isomorphic micro-models, to the extent that their deviations from isomorphism are small or unimportant to the variables of practical interest. The automated learning of causal models has begun as an area of serious research, with initial success. The obstacles to learning causal models that are sufficiently close to the real systems being reasoned about – in particular, learning latent variable structure and variable selection – are also under attack, with some initial success. Hence, although the frame problem is not yet solved in practice, it does not appear to offer any theoretical barrier to this approach to generating an artificial intelligence.

9.3. THE BAYESIAN INDUCTION OF CAUSAL MODELS

Selecting a precise *method* for automating the learning of causal models is not a part of the causal induction solution to the frame problem, so long as some such method exists. Nevertheless, I shall outline a Bayesian method for causal learning in order to provide a clearer idea of what form the solution takes.

The particular method we have been using for causal discovery employs an information–theoretic evaluation metric for causal models given sample data. This metric, Minimum Message Length (MML), was first formulated by Solomonoff (1964) and first applied by Wallace and Boulton (1968). The fundamental principle in MML induction is to find that hypothesis which leads to the shortest message encoding both the hypothesis and the data to be explained. Shannon's measure of information (Shannon and Weaver, 1949) contained in a message m is $I(m) = -\log_2 P(m)$. For Shannon's information measure to correspond to the

actual lengths of messages in any language, it is necessary that the method of encoding hypotheses and data in that language be *efficient,* i.e., that the encoding method minimizes entropy in the communication of messages using that language. In other words, we must design an encoding procedure so that the probabilities of messages will be respected, so they can be read back directly from the lengths of their encodings via Shannon's equation.

The possibility of applying MML induction depends upon finding an efficient encoding for the application. If we in fact have a prior probability function available for a domain of inquiry (as subjectivist Bayesians assert that we always do), then we can *guarantee* efficiency: given known probabilities $P(h_i)$ for each hypothesis in the hypothesis space and for the likelihoods $P(e_j|h_i)$ for each outcome in an observational or experimental outcome space per hypothesis, we can construct an encoding technique which enforces efficiency. The methods for doing so are straightforward applications of coding theory, and are described, for example, by Oliver and Hand (1995). In domains where we do not have prior probabilities to guide our encoding (or where, at any rate, explicit priors are not obviously available), we can nevertheless invent natural coding procedures which incorporate basic common sense – in particular, which encode simpler hypotheses more compactly and which penalize precisely specified parameters over inexactly specified parameters – and which are efficient with respect to *some* probability distribution (i.e., which are probabilistically coherent).

Given efficiency, MML induction can be proven correct from the Bayesian point of view, as a trivial corollary to Bayesian conditionalization; hence, MML is properly described as a Bayesian method of induction.

PROOF. We begin with Bayes' Theorem:

$$P(h_i|e) = \frac{P(e|h_i)P(h_i)}{Pe)}$$

Applying the functions used in Shannon's information measure:

$$log_2 P(h_i|e) = log_2 P(e|h_i) + log_2 P(h_i) - log_2 P(e)$$
$$-log_2 P(h_i|e) = -log_2 P(e|h_2) - log_2 P(h_2) + log_2 P(e)$$

By the total probability theorem:

$$P(e) = \sum_j P(e|h_j)P(h_j)$$

So, the value of $P(e)$ is the same for all hypotheses in the hypothesis space. This implies that $P(e)$ is irrelevant to the *ranking* of hypotheses by posterior probability; i.e., $-log_2 P(e|h_i) - log_2 P(h_i)$ differs from $-log_2 P(h_i|e)$ by a constant. MML induction obliges us to find the hypothesis (h_i) that minimizes $I(h_i) + I(e|h_i)$. Since $I(h_i) + I(e|h_i) = -log_2 P(e|h_i) - log_2 P(h_i)$, this proves that MML induction produces that hypothesis having the greatest posterior probability. QED.

It is important to note that MML induction deals naturally – *intrinsically* – with noise in the data. Indeed, MML induction just is a principle for playing a balancing act between order and disorder. Consider a set of k sample points in a polynomial curve-fitting exercise. If we allow in the hypothesis space $(k-1)$ th-order polynomials for fitting the k data points, then there will be hypotheses available which fully imply each data point in the sample – we can have a hypothesis which overfits the sample. For such a hypothesis $I(e|h) = 0$ (i.e., the likelihood is 1), which is an advantage in the MML measure; however, the hypothesis will be highly complex and improbable so that $I(h)$ will be large, typically on the order of $I(e)$ in fact. Between the extremes of an empty hypothesis (where $I(h) \approx 0$) and an overfitting hypothesis (where $I(e|h) \approx 0$) typically lies a minimum point which MML induction aims at finding.

In the case of causal induction, MML evaluation plays the tradeoff between complexity of the hypothesis and its fit to the data by selecting or rejecting direct causal relations between variables according to their value in accommodating the data and by rendering the parameters associated with causal influences with greater or less precision. For a general account of MML see Georgeff and Wallace (1984) and Wallace and Freeman (1987); for MML causal induction see Korb and Wallace (1997), Wallace, Korb and Dai (1996) and Wallace and Korb (forthcoming).

9.4. REMARKS ON THE CAUSAL INDUCTION SOLUTION

This solution to the frame problem incorporates some of the suggestions mentioned previously. First, specificity is an essential ingredient to the solution. The idealized goal of the solution is to discover and use a causally isomorphic model. As that is not normally possible, the approximative substitute goal is to find a good causal homomorphism of the real causal system. This means using some model which merges or ignores the less important causal properties for the problem at hand while retaining those which are more important. ("Importance" here is presumably to be understood as the variation in expected utility which inclusion or exclusion of the variable will have in the judgments required by the problem context.) By identifying the variables chosen for the model with greater specificity the system will be able to discover more exact causal relationships between the variables and so, for example, be able to make more precise predictions. But accuracy, in addition to precision, is important, so the system is simultaneously obliged to aim for more general variables, for which it will have more statistical evidence (larger samples) available from which to estimate their causal relationships with each other. In other words, we need models which incorporate Reichenbach and Salmon's maximal homogeneous reference classes. The trade-off involved in capturing maximality (i.e., generality, so as to obtain the most reliable statistics) within homogeneity (i.e., specificity, so as to respond to all relevant statistical relations) is one which is natural to the Bayesian inductive methods outlined above. [19]

The causal induction solution to the frame problem also essentially incorporates the idea that focusing attention on a subset of possibly relevant variables helps to resolve the frame problem in practical cases (e.g., Nutter, 1991). In particular, the variable selection implied by the Bayesian inductive criterion above provides a focus of attention on a subset of variables, which we may hope constitute a tractable model for the problem at hand.

Finally, although Hayes was right to say that the concept of relevance is not itself relevant to the *introduction* of the frame problem, the causal induction solution shows that Dennett was also right that relevance is pertinent to its *solution.* An appropriate computational treatment of relevance is required at two points: in the discovery of direct causal relations between pairs of variables, when irrelevant, unconnected variables must be left unconnected; and when matching a learned causal model to a practical problem, when irrelevant, unmatched variables must be left behind.

9.5. OBJECTIONS

I shall briefly review some likely objections to my causal induction solution to the frame problem. Of course, Dennett's "cognitive wheels" objection could be raised again: it is implausible (or, minimally, unclear) that humans use the Bayesian methods adumbrated above to learn about their world. Cognitive psychologists are beginning to investigate just how humans do learn causal structure (cf. Sperber, Premack and Premack, 1995), and it is possible that our methods approximate Bayesian induction (certainly, there *has* been success in characterizing human science in Bayesian terms – see Howson and Urbach, 1993). But again, of course, we needn't claim to be doing cognitive psychology when we construct our AI systems – we may be doing research in a different branch of cognitive science and engineering.

Another possible objection is that real causal systems are typically not closed: they are perturbed by the influences of variables that are "out of focus." On the one hand, these external influences are explicitly represented in the causal models as "unexplained variance" in predicted variables, so the lack of closure is acknowledged. On the other hand, once we have abandoned causal isomorphism as a realistic goal, we are free to accept imperfections in the model, so long as the model remains practically useful for improving an artificial agent's understanding of, predictions about, and interventions in the physical system of interest. It is no one's precondition of an artificial intelligence that perfection be achieved.

Finally, it might be objected that bringing to bear the apparatus of Bayesian induction, or any other statistical inference method, is unnecessary since symbolic learning methods, which are inherently simpler and cleaner, suffice for our artificial learning needs. Inductive logic programming, in particular, has been widely acclaimed as a promising symbolic approach to machine learning. I have not the space (nor the time!) to produce a full rebuttal to this claim. As a Bayesian, I

note that my own prior probability that some essentially symbolic learning method shall one day be shown sufficient for artificial agents learning about significant chunks of the real world is near (but not at) zero. The most obvious and pervasive obstacle to such an unlikely achievement is that of measurement noise: real-world data confronting both humans and artifacts almost invariably incorporate missing values and mismeasured values. Bayesian induction, as explained above, accommodates measurement noise *intrinsically.* The same can hardly be said of symbolic learning algorithms. Tom Mitchell's (1977) version-space learning, twenty years ago acclaimed as the epitome of symbolic learning methods, simply did not acknowledge the possibility of noisy data. Mitchell (1978) subsequently extended the algorithm to incorporate noise in the data – by the hideously (exponentially) complex device of requiring that all possible subsets of the data be considered as potentially in error. The "scientific discovery" research of Langley et al. (1987) encountered exactly the same problem. The initial versions of BACON required predigested data, with all noise removed. Subsequent versions (e.g., Langley et al., 1990) allowed predigested noise parameters to be entered by the user. No version of BACON (or its direct descendants) has successfully recognized and accommodated measurement noise independent of human intervention. Symbolicism and logicism are, on principle, blind to the statistical features of data and of inductive learning problems; it should come as no surprise that the learning algorithms produced by these research programs turn out to be severely handicapped in the real world.

Inductive logic programming (ILP) again is handicapped in precisely the same way. The fundamental idea behind ILP is that, since resolution [20] provides a complete system for first-order deduction and since deduction is the generation of particular statements from generalizations whereas induction is the inverse generation of general statements from particulars (cf. Muggleton, 1990, pp. 13–14),[21] we may obtain induction by reversing resolution. When a set of data requires explanation, an ILP algorithm will search for a hypothesis (or set of hypotheses) which, together with any background knowledge provided and initial conditions, deductively implies the data under consideration. (This will be familiar to many as a form of Peircean abduction!) Since in the general case one can find infinitely many hypotheses which imply a consistent set of data, there is some concern within the ILP research program to find methods to constrain the search, such as employing simplicity biases and reversing only specialized forms of resolution. A different problem, which the ILP community has by and large ignored (the majority of published examples of successes for ILP learning being examples of artificial learning problems, such as learning the grammar of an invented language), is that real-world data are often inconsistent and almost always incomplete and inaccurate. Finding a set of hypotheses that deductively imply the truth of inaccurate observations is, of course, a way of guaranteeing the failure of our inductions.

I do not wish my criticism of ILP, or philosophical logics in AI, etc., to be misread as the claim that these techniques are without any value. ILP might well be combined with statistical or information-theoretic techniques (or otherwise made

to approximate their use), while relaxing the requirement of strict deducibility for the data, so as to cope with noise in the data; indeed, Quinlan has applied some of his information-theoretic measures to ILP in his program FOIL (Quinlan, 1990). Again, default or nonmonotonic logics may well find application in some reasoning problems. What I am claiming is that the ideology which has underwritten and driven much of the research in these applications of and extensions to logic – an ideology intolerant of any admixture of non-symbolic ingredients – is itself bankrupt. Artificial intelligence without quantitative machine learning is hopeless.

10. The Causal Induction Hypothesis

I can sum up my position in a modest proposal, the *Causal Induction Hypothesis* (CIH): the ability to learn the causal structure of physical systems is the necessary and sufficient condition of artificial intelligence. In asserting this I mean to invoke the same variety of qualifications that Newell and Simon (1976) imposed for their Physical Symbol System Hypothesis (PSSH), such as that the intelligences must be embodied (physical), that they have a semantics (e.g., be causally related in the right way to the objects of their representations), that the sufficiency is subject to the causal learning being supplemented in the "right" (unspecified) ways[22] and especially that the Causal Induction Hypothesis is empirical and testable – although I can give motivating arguments, I can not offer proofs.

As motivation, first I shall argue that causal learning is necessary and sufficient for solving the frame problem. The case for sufficiency has just been made in the prior section: if causal learning can yield models which are causally homomorphic to the physical systems reasoned about (and the homomorphism does not lose important information about the states or variables of particular interest to the problem at hand), then the addition of ordinary AI search techniques solves the frame problem. But causal learning is also *necessary* for the solution of the frame problem. In order to avoid irrelevant inferences about a changing world (which is very nearly Hayes's own (1987) characterization of the frame problem) it is necessary to restrict inferences to those which are relevant to worldly changes. This can only be done by guiding one's inferences by what is relevant to the changes in the world at issue, and that means guiding them by the causal structure of that world. Hence, it is necessary that an autonomous agent be able to learn what that causal structure is.

The next step in the argument is that solutions to the frame problem are solutions to GAIP – the problem of producing a real artificial intelligence (note that this is *not* to say that the two *problems* are the same, but that their *solutions* are equivalent).[23] I have already made clear that a solution to the frame problem is necessary for an artificial intelligence. Solving the frame problem is also sufficient for a real artificial intelligence, because an appropriately robust (and supplemented) causal learner (which is equivalent to a system that solves the frame problem, per above) is sufficient for a real artificial intelligence. Why? If pressed here, I shall have to wave

my hands. I wave my hands in the general direction of what plausible criteria for intelligence might look like. Certainly some minimal criteria for intelligence would include: the ability to come to terms with an earth-like environment and adapt to noncatastrophic changes in the environment (which a robust causal learner should be able to do); to plan simple sequences of actions to achieve basic goals, such as saving one's battery (which an appropriately supplemented causal learner can do); to be able to understand such assertions as that Fred has fired a gun, including the presumption that Fred had a gun and the expectation that the next step in the story will not have the bullet lying innocently on the floor (a good causal model for gun-firing will lead to both); etc. (These criteria, by the way, correspond respectively to major branches of AI research: machine learning, planning, and natural language understanding.)

Although I can give motivating (if hardly compelling) arguments now on behalf of CIH, it is subject to the empirical test of whether research in automating the induction of causal models will ultimately lead to programs which can learn enough usable causal theories to be able negotiate in practice the kinds of difficulties listed above. It is possible that causal learning will fail to progress, that it will reach hurdles that we cannot surmount. There is no specific reason to believe that will happen, as progress has been considerable now that serious attention has turned to the subject (for examples of such progress, see: Heckerman and Geiger, 1995; Spirtes et al., 1993; Wallace et al., 1996). Even so, if causal learning stalls, this will provide evidence against CIH only if artificial intelligence does not then also stall.

Pat Hayes argues that causal learning *must* stall and so, given that there is a solution to the frame problem (i.e., there is an "existence proof " in that humans are not subject to it), CIH must be false. The reason Hayes offers that causal learning is impossible as a solution is that causal structure is "unimaginably complex . . . . Since one of the things which actions can do is alter the pattern of causal connections," causal laws must be indexed by what actions have been applied, which leads to a recursive frame problem (1987, p. 130). It would certainly be news to physical and biological scientists, however, that in addition to disturbing measurements and upsetting experiments their actions can actually *change the physical and biological laws that they are studying!* There is no support for such a claim in any biophysical science – including quantum mechanics (where measurement by an observer is an action that is subsumed by quantum mechanics). It is true that in the psychosocial sciences things are apparently different: placebo effects and self-fulfilling prophecies are common. But it is at least a viable thesis that the laws of psychosocial behavior which can be manipulated by human intervention are not themselves fundamental, but are instead heuristic, and may be replaced by more fundamental laws in (a perhaps distant) future – laws which incorporate such "interventions". The fact that placebo effects, for example, are themselves quite regular and reproducible supports this supposition (see the excellent paper by Roger Buck (1963) on self-fulfilling prophecies and their import for social science).

In principle, then, the final solution to the frame problem is this: given a surrogate for the physical system, a model that is causally homomorphic to it and in particular which incorporates all the causal features relevant to a particular task, manipulate the surrogate in order to test various possible solutions to the problem. This final solution must wait upon the solution to the problem of learning relevant causal homomorphisms to be realizable in practice. The latter problem is difficult and has only just recently become the object of serious research in artificial intelligence, partly because of the long-dominant and jealous religion of logicism. Nevertheless, its difficulties do not appear to be impenetrable or interminable, except perhaps to those – both enemies of and enthusiasts for AI – whose time scale for progress in AI is set without reference to the other human sciences.

11. Conclusion

The frame problem is many things to many people: When Patrick Hayes (1987) and Vladimir Lifschitz (1987) look at the frame problem, they see a technical problem to be overcome by a new philosophical logic; when Walter Freeman (1992) looks at it, he sees the problem of understanding reasoning as a dynamic process; when James Fetzer (1991a, 1991b, 1993) looks at it, he sees the problem of induction; Stevan Harnad (1993) sees the symbol-grounding problem; Hubert and Stuart Dreyfus (1987) see a failure to accept holistic representation. In fact, if I am right, the solution to the frame problem touches upon every issue facing artificial intelligence, so it is easy enough for people to see the problem in whatever light they prefer. But what is central to any solution is automating the understanding of the world's causal structure – and this, short of divine (designer) intervention, is tantamount to automating the learning of causal structure.

This brings the frame problem into some relation with the problem of induction: it seems that to solve the frame problem we must be able to answer Hume's challenge–we must be able to justify some inductions over others, despite Hayes's denial. And trying to answer this challenge lands us in a dilemma: we either justify rules of induction inductively, which is manifestly circular, or we justify rules of induction deductively, which appears to be impossible.

But the necessity of meeting Hume's challenge is illusory. [24] The problem of induction is closely analogous to the frame problem. One aspect they share is that – whatever our difficulties in coming to terms with them, formulating theories about them, or inventing algorithms to cope with them – humans are not impeded by these problems in their day-today activities. Humans are no more struck dumb by the question of whether or not the sun will "rise" tomorrow than they are by the question whether or not the table will rise with the glass of water. If there are behavioral criteria for intelligence, and if handling ordinary inductions and conditioning one's behavior thereby is an aspect of such criteria, then the task at hand is just to formulate theories of how to meet these criteria and to generate the algorithms which can test the theories. What we need are theories and algorithms

descriptive of how humans (and other animals) *in fact* perform their inductions, for if we are not prepared to be satisfied with a machine which can do as well but not better, then we are demanding something nonsensical (at least for an *initial* artificial intelligence).

In other words, we need a solution to the *descriptive problem of induction* – how humans perform their inductions – rather than the (possibly unanswerable) prescriptive problem of induction – how to justify the inductions we perform. The descriptive problem of induction will be unanswerable only if human behavior is in fact unintelligible, a proposition which we need not (yet) concede.

Of course, as I previously insisted, artificial intelligence need not restrict itself to descriptive, naturalistic approaches. We may entertain behavioral criteria for intelligent inductions without constraining ourselves to human methods for achieving them. In other words, a normative AI may employ non-naturalistic (e.g., Bayesian or scientific) methods of induction which also satisfy our behavioral criteria, without resolving all foundational queries about epistemology. [25]

In the end, then, if we are successful, we shall have an understanding of how we make inductions – or, at any rate, how inductions useful in our world might be made – and we shall have working automata to prove it. We shall not likely have any theory or automata for performing useful inductions in *arbitrary* worlds. Whereas the prescriptive problem of induction appears to be an enduring problem, it also appears to be one which impedes nothing in any of the sciences, whether natural or artificial.[26]

Notes

[1] Two critiques of logicism are McDermott (1987c) and Korb (1995).

[2] The "ideal rationality" that Dietrich and Fields (1996) criticize as unattainable appears to be at least a close relative of logicism.

[3] Perhaps this dominance contributes fuel to recent disputes sparked by so called anti-representationalists: Rodney Brooks (1991), at least, in his remarkably titled "Intelligence without Representation" is more charitably read as decrying symbolic representations rather than (as he actually writes) *all* representations, since his own "Creatures" maintain non-symbolic representations. In particular, they retain information about the environment (i.e., intentional representations) by circulating information through an electrical delay system – in a way highly reminiscent of how generic mercury-tube memory systems worked in early electronic computers.

[4] Note that the logicism (and symbolicism) at issue is *not* the AI tradition that Gillies (1996) identifies as the "Turing tradition" and as including Turing, Michie and Quinlan. Turing himself, as Gillies discusses elsewhere (Gillies, 1990), was an advocate of the utility of probabilistic reasoning; similarly, Quinlan (1993) directly incorporates probabilistic reasoning into his machine learning programs (in its dual form of an information measure). Logicism in AI is largely an American invention, represented by Simon, Newell, McCarthy and Hayes, among others.

[5] This is expressed in their condition 2 of a physical symbol system: "There exist expressions [i.e., symbol structures] that designate every process of which the machine is capable." Of course, in (a possible) practice this condition may be satisfied by a quantitative AI, accepting *approximate* digital representations of real-valued quantities. But it is the in principle exclusion of real-valued quantities

which explains the actual practice of much of "traditional" AI, a practice which is highly biased against quantitative reasoning.

[6] This describes J. Terry Nutter's (1991) solution as well, despite her claims to the contrary. Her suggested approach is to employ a concept of attention, focused by a graphical representation of context, which would simply avert one's gaze from irrelevant non-change. But Nutter has to admit (1991, p. 186) "the primary untouched issue here is how our system decides what initial nodes and arc types to put into the context." This untouched issue is tantamount to the question of how to leave the sleeping dogs lying. (My thanks to Adam Frey for pointing this out).

[7] Since consistency is not decidable in first-order logic, this implies there will be no algorithm for default inference; however, heuristic searches for inconsistency can be substituted for a formally correct algorithm.

[8] In this context the denial of $A \rightarrow B$ means $A \rightarrow \neg B$, in contrast to the negation, which would be $\neg (A \rightarrow B)$

[9] Of course, 'Nixon is a Republican' is not a universal generalization, but to let the example go through we could replace 'Nixon' by some uniquely identifying description of Nixon and get a universal statement.

[10] Parts of this section were reproduced from Korb (1995) by permission of the publisher, Taylor and Francis.

[11] The bomb being on the wagon is just such a defeater, of course.

[12] See Korb (1998) for a discussion of homogeneity and its importance for the understanding of causality.

[13] Similar difficulties have been explored by philosophers of science previously. Thus, Salmon examined a *counter*-inductive rule as an alternative to Reichenbach's "straight rule" for enumerative induction (Salmon, 1967); Carnap's system of inductive logic (1962) was plagued by the arbitrariness of the selection of a fundamental measure, which could yield an arbitrary rule for induction.

[14] But this claim is clearly an error, probably due to a common tendency within AI to treat inferential issues under the label of "knowledge representation". The frame problem has *both* inferential and representational aspects – it requires us to avoid explicitly describing (representing) facts about unimportant non-changes, but of course it *therefore* requires us to avoid inferences which would lead us to reason about such non-changes, whatever our representational approach. Everyone who has written seriously about the frame problem, including Hayes, has written about inferential issues, and that is no accident.

[15] Except that it does not adequately convey his vehemence and belligerence, which is probably just as well.

[16] I note that McDermott (1987a, pp. 115f) also objects to the discrete universe assumption.

[17] Previously suggested by Haugeland (1987).

[18] AI systems that learn and use causal models to cope with the world might plausibly be construed to have implemented the "causally-structured programming" that Fetzer (1991b) called for and that Ford and Hayes (1991a, p. xiii) claimed not to be able to understand.

[19] Note that the objection raised in §3 to the use of homogeneous reference classes was not that they were irrelevant to the frame problem, but that their use is inconsistent with logicist principles, which were there under review.

[20] An inference rule well suited to automated theorem proving, introduced, by Robinson, 1965.

[21] This latter is the common, but mistaken, view of deduction and induction in artificial intelligence. Deduction is not the derivation of particulars from general statements, nor is induction the generation of general statements from particulars. For some trivial counterexamples: one may deduce particulars from particulars; one may also induce particulars from particulars, as in Carnap's "instance confirmation" yielding a prediction about the *next* observation of a kind (Carnap, 1962).

[22] *Sufficiency* here must be understood as a technical term introduced by Newell and Simon: since the necessary ingredients to achieve (real) sufficiency are unspecified and open-ended, the technical term is vacuous, having only propaganda value.

[23] In other words, the problem descriptions are not synonymous – they differ intensionally. If you care to identify problems with their solution sets, then of course my claim amounts to the problems being the same – *extensionally* (and even so qualified by the previous note).

[24] It is in this regard that I most clearly differ from Fetzer on the frame problem. Since, on my account, the frame problem is open to a practical solution, but one which does not even address Hume's prescriptive problem of justifying induction, it follows that the frame problem and Hume's problem of induction are not the same.

[25] This distinction between a naturalistic and a non-naturalistic approach to resolving the descriptive problem of induction roughly corresponds to the distinction Hayes (1991; and Hayes and Ford, 1993) draws between automating common-sense reasoning and automating scientific reasoning, if we strip his conception of its logicist presuppositions. Hayes argues that we should be concerned only with automating common-sense reasoning on the grounds that that is what humans do (i.e., it is an existence proof of its possibility); however, this neglects the fact that humans also do scientific reasoning and provides no argument that automating the latter will not serve the needs of intelligence quite as well.

[26] I wish to acknowledge helpful comments by Chris Wallace, Jon Oliver, an anonymous referee and the participants at a seminar at the Center for the Philosophy of Science at the University of Pittsburgh. This work was supported in part by ARC grant A49330662.

References

Brooks, R. (1991), 'Intelligence without Representation', *Artificial Intelligence,* 47, pp 139–159.

Brown, F.M. (ed.) (1987), *The Frame Problem in Artificial Intelligence,* Morgan Kaufmann.

Buck, R.C. (1963), 'Do Reflexive Predictions Pose Special Problems for the Social Scientist?', *Philosophy of Science* , 30, pp 359–374.

Carnap, R. (1962), *Logical Foundations of Probability,* second edition, University of Chicago.

Cherniak, C. (1986), *Minimal Rationality,* Cambridge, Mass.: MIT Press.

Dennett, D. (1984), 'Cognitive Wheels: The Frame Problem of Artificial Intelligence', in C. Hookaway (ed.) *Minds, Machines and Evolution,* Cambridge University. Reprinted in Pylyshyn (1987).

Dietrich, E. and Fields, C. (1996), 'The Role of the Frame Problem in Fodor's Modularity Thesis: A Case Study of Rationalist Cognitive Science', in K. M. Ford and Z. W Pylyshyn, (eds.) *The Robot's Dilemma Revisited:* Norwood, N. J. : Ablex, pp 9–24.

Dreyfus, H. and Dreyfus, S. (1987), 'How to Stop Worrying about the Frame Problem even though it's Computationally Insoluble', in Z. W. Pylyshyn (ed), *The Robot's Dilemma* Norwood, N. J.: Ablex, pp. 95–111.

Fetzer, J. (1991b), 'The Frame Problem: Artificial Intelligence Meets David Hume', In K.. M. Ford and P. J Hayes (eds.) *Reasoning Agents in a Dynamic World* Greenwich, CT: JAI Press pp. 55-69.

Fetzer, J. (1991b), 'Artificial Intelligence Meets David Hume: A Response to Patrick Hayes', in K. M Ford and P. J Hayes (eds.) *Reasoning Agents in a Dynamic World* Greenwich, CT: JAI Press pp. 77–85.

Fetzer, J. (1993), 'Philosophy Unframed', *Psycoloquy* 4, (33).

Ford, K.M. and Hayes, P.J. (eds.) (1991a), *Reasoning Agents in a Dynamic World: The Frame Problem*. Greenwich, CT: JAI Press.

Ford, K.M. and Hayes, P.J. (1991b), 'Framing the Problem', in K. M Ford and P. J Hayes (eds.) *Reasoning Agents in a Dynamic World* Greenwich, CT: JAI Press pp. ix–xiv.

Ford, K.M. and Pylyshyn, Z.W. (1996), *The Robot's Dilemma Revisited*, Norwood, N.J.: Ablex.

Freeman, W. (1992), 'Framing Is a Dynamic Process', *Psycoloquy,* 3(62).
Georgeff, M.P. and Wallace, C.S. (1984) 'A General Selection Criterion for Inductive Inference', *European Conference on Artificial Intelligence,* 6, pp. 473–482.
Gillies, D. (1990), 'The Turing-Good Weight of Evidence Function and Popper's Measure of the Severity of a Test', *British Journal for the Philosophy of Science* 41, pp 143–146.
Gillies, D. (1996), *Artificial Intelligence and Scientific Method*, Oxford: Oxford University Press.
Goldberg, D.E. (1989), *Genetic Algorithms in Search, Optimization, and Machine Learning)*, Addison-Wesley.
Grice, H.P. (1975), 'Logic and Conversation', in P. Cole and J.P. Morgan (eds.) *Syntax and Semantics, Volume 3: Speech Acts*, pp. 41–58) New York: Seminar Press.
Hanks, S. and D. McDermott (1986), 'Default Reasoning, Nonmonotonic Logic, and the Frame Problem', in T. Kehler, S. Rosenschein, R. Filman and P. Patel-Schneider (eds.) *Proceedings of the Fifth National Conference on Artificial Intelligence,* pp. 328–333, Morgan Kaufmann.
Harnad, S. (1993), 'Problems, Problems: The Frame Problem as a Symptom of the Symbol Grounding Problem', *Psycoloquy* 4(34). http://cogsci.ecs.soton.ac.uk/harnard/psyc.html
Haugeland, J. (1987), 'An Overview of the Frame Problem', in Z.W. Pylyshyn *The Robot's Dilemma* Norwood, N. J.: Ablex, pp. 77–93.
Hayes, P. (1987), 'What the Frame Problem Is and Isn't', in Z. W. Pylyshyn (ed), *The Robot's Dilemma* Norwood, N. J.: Ablex, pp. 123-137.
Hayes, P. (1991), 'Commentary on 'The Frame Problem: Artificial Intelligence Meets David Hume', in in K. M Ford and P. J Hayes (eds.) *Reasoning Agents in a Dynamic World* Greenwich, CT: JAI Press, pp. 71–76.
Hayes, P. and Ford, K. M. (1993), 'Modeling our Adaptive Intelligence, Not God's', *Psycoloquy,* 4 (42). http://cogsci.ecs.soton.ac.uk/harnard/psyc.html
Heckerman, D. and Geiger, D. (1995), 'Likelihoods and Priors for Learning Bayesian Networks', *Neural Information Processing Systems 1995 Workshop on Learning in Bayesian Networks and Other Graphical Models, December 1995, Veil, Colorado.*
Heckerman, D., Mamdani, A. and Wellman, M. P. (1995), 'Real-world Applications of Bayesian Networks', *Communications of the ACM* 38 (March), p 24–26.
Howson, C. and Urbach, P. (1993), *Scientific Reasoning: The Bayesian Approach,* second edition, Chicago: Open Court.
Hume, D. (1748/1975) *An Enquiry Concerning Human Understanding*, ed. L. A. Selby-Bigge, revised by P. H. Nidditch, Oxford University.
Israel, D. (1985), 'A Short Companion to the Naive Physics Manifesto', in J.R. Hobbs and R.C. Moore (eds.) *Formal Theories of the Commonsense World*, Norwood, N. J. : Ablex. pp. 427–447.
Korb, K.B. (1995), 'Inductive Learning and Defeasible Inference', *Journal of Experimental and Theoretical Artificial Intelligence* 7, pp. 291–324.
Korb, K.B. (1998), 'Probabilistic Causal Structure', forthcoming in H. Sankey (ed.), *Causation and Natural Laws: Australasian Studies in the History and Philosophy of Science,* Kluwer.
Korb, K.B. and Wallace, C. S. (1997) 'In Search of the Philosopher's Stone: Remarks on Humphreys and Freedman's Critique of Causal Discovery', *British Journal for the Philosophy of Science* 48, pp. 543–553.
Langley, P. Simon, H. A. and Bradshaw, G. L. (1990), ' Heuristics for Empirical Discovery,' in Shavlik and Diettrich (eds.), *Readings in Machine Learning,* Morgan Kaufmann.
Langley, P. Simon, H. A. Bradshaw, G. L. and Zytow, J. M. (1987), *Scientific Discovery,* MIT.
McCarthy, J. and Hayes, P. J. (1969), 'Some Philosophical Problems from the Standpoint of Artificial Intelligence', in B. Meltzer and D. Michie (eds.), *Machine Intelligence* 4, pp. 463–502, Edinburgh University Press. Reprinted in M.L. Ginsberg (ed.) *Readings in Nonmonotonic Reasoning,* pp. 26–45, Morgan Kaufmann.
McDermott, D. (1987a), 'We've Been Framed: Or, Why AI Is Innocent of the Frame Problem', in Z. W. Pylyshyn (ed), *The Robot's Dilemma* Norwood, N. J.: Ablex, pp. 113–122.

McDermott, D. (1987b), 'AI, Logic and the Frame Problem', in F. M. Brown *The Frame Problem in Artificial Intelligence*, Morgan Kaufmann, pp. 105–118.
McDermott, D. (1987c), 'A Critique of Pure Reason', *Computational Intelligence* 3, pp. 151–160.
Mitchell, T.M. (1977), 'Version Spaces: An Approach to Rule Revision during Rule Induction', *5th International Joint Conference on Artificial Intelligence*, pp. 305–310.
Mitchell, T.M. (1978), *Version Spaces: An Approach to Concept Learning.* Technical Report STAN–CS–78–711, Computer Science Dept., Stanford University (Ph.D. dissertation).
Muggleton, 5. (1990), *Inductive Acquisition of Expert Knowledge* Workingham, England: Addison-Wesley (Turing Institute Press).
Muggleton, S. (ed.) (1992), *Inductive Logic Programming,* Academic Press.
Muggleton, S. and Feng, C. (1992), 'Efficient Induction of Logic Programs', in S Muggleton (ed.), *Inductive Logic Programming,)* Academic Press. pp. 281–298.
Nutter, J.T. (1991), 'Focus of Attention, Context, and the Frame Problem', in in K. M. Ford and P. J. Hayes (eds.) *Reasoning Agents in a Dynamic World* Greenwich, CT: JAI Press. pp. 171–188.
Oliver, J. Dowe, D. and Wallace, C. S. (1992), 'Inferring Decision Graphs using the Minimum Message Length Principle', *Proceedings of the 1992 Australian Joint Conference on Artificial Intelligence, Hobart,* Tasmania, pp. 361–367, World Scientific.
Oliver, J. and Hand, D. (1995) *Introduction to Minimum Encoding Inference.* Technical Report 95/205, Dept. of Computer Science, Monash University.
Pylyshyn, Z.W. (ed.) (1987), *The Robot's Dilemma: The Frame Problem in Artificial Intelligence* Norwood, N.J.: Ablex.
Quinlan, J.R. (1990), 'Learning Logical Definitions from Relations', *Machine Learning* 5, pp 239–266.
Quinlan, J. R. (1993), C4.5: *Programs for Machine Learning,* Morgan Kaufmann.
Reichenbach, H. (1949), *The Theory of Probability,* second edition, translated by E. H. Hutton and M. Reichenbach, Berkeley: University of California.
Reiter, R. (1978), 'On Reasoning by Default', *Proceedings of Theoretical Issues in Natural Language Processing,* Urbana, Illinois, pp. 210–218.
Reiter, R. (1980), 'A Logic for Default Reasoning', *Artificial Intelligence* 13, pp. 81–132.
Reiter, R., (1987), 'Nonmonotonic Reasoning', *Annual Review of Computer Science* 2, pp. 147–186.
Robinson, J.A. (1965), 'A Machine-oriented Logic Based on the Resolution Principle', *Journal of the ACM* 12, pp. 23–41.
Rumelhart, D. E. and McClelland, J. (1986), *Parallel Distributed Processing, volume 1,* MIT.
Salmon, W.C. (1967), *The Foundations of Scientific Inference*, University of Pittsburgh.
Salmon, W. C. (1971), *Statistical Explanation and Statistical Relevance,* University of Pittsburgh.
Schaffer, C. (1994), 'A Conservation Law for Generalization Performance', *Proceedings of the 1994 International Conference on Machine Learning,* Morgan Kaufmann.
Shannon, C. E. and Weaver, W. (1949) *The Mathematical Theory of Communication,* University of Illinois.
Simon, H. (1994), 'Literary Criticism: A Cognitive Approach', in G. Guzeldere and S. Franchi (eds.), *Stanford Humanities Review Special Supplement: Bridging the Gap,* pp. 1–26.
Solomonoff, R. (1964) 'A Formal Theory of Inductive Inference, I and II,' *Information and Control* 7, 1–22 and 224–254.
Sperber, D. Premack, D. and Premack, A. J. (eds.), (1995), *Causal Cognition: A Multidisciplinary Debate*, Oxford: Clarendon Press.
Spirtes, P., Glymour, C. and Schemes, R. (1993), *Causation, Prediction, and Search*, New York: Springer Verlag.
Thorndike, E. L. (1911), *Animal Intelligence*, New York: Macmillan.
Wallace, C. S. (1995) *Multiple Factor Analysis by MML Estimation.* Technical Report 95/218, Dept. Computer Science, Monash University.

Wallace C. S. and Boulton D. M. (1968) 'An Information Measure for Classification' *Computer Journal* 11, 185–194.

Wallace, C. S. and Freeman, P. R. (1987), 'Estimation and Inference by Compact Coding', *Journal of the Royal Statistical Society, Series B,* 49, pp 240–252.

Wallace, C. S. and Freeman, P. R. (1992), ' Single-factor analysis by minimum message length estimation', *Journal of the Royal Statistical Society, Series B,* 54, pp. 195–2O9.

Wallace, C. S. and Korb, K. B. (forthcoming), 'A Study of Causal Discovery by MML Sampling,' forthcoming in M. Slater (ed.), *Causal Models and Intelligent Data Analysis,* Springer Verlag.

Wallace, C. S. Korb, K. B. and Dai, H. (1996), 'Causal Discovery via MML', in L. Saitta (ed.), *Proceedings of the 13th International Conference on Machine Learning,* Morgan Kaufmann, pp. 516–524.

Wolpert, D. H. and McReady, W. G. (1995), *No Free Lunch Theorems for Search.* TR 95–02–010 Santa Fe Institute. http://www.santafe.edu/sfi/publications/95wplist.html

Wright, S. (1934) 'The Method of Path Coefficients', *Annals of Mathematical Statistics* 5, pp. 161–215.