Published in L. Magnani and R. Dossena (Eds.) *Computing, Philosophy and Cognition* (2005), London: King's College Publications, pp. 349–359.

mark.sprevak@ed.ac.uk

The frame problem and the treatment of prediction

Mark Sprevak
University of Edinburgh

12 July 2004

The frame problem is a problem in artificial intelligence that a number of philosophers have claimed has philosophical relevance. The structure of this paper is as follows: (1) An account of the frame problem is given; (2) The frame problem is distinguished from related problems; (3) The main strategies for dealing with the frame problem are outlined; (4) A difference between commonsense reasoning and prediction using a scientific theory is argued for; (5) Some implications for the computational theory of mind are discussed.

1 Introduction

The frame problem is a problem in artificial intelligence that was first described in McCarthy and Hayes (1969). As a problem in artificial intelligence, it has been extremely difficult to solve. A number of philosophers, including Dennett (1987), Fodor (1987), Glymour (1987), and Haugeland (1987) have suggested that the frame problem is either indicative of a new problem in philosophy, or has important connections to existing problems. Unfortunately, no one can agree what those connections are, or what the frame problem is. In this paper, I argue for two things. First, I argue for a view of what the frame problem is. Second, I argue that there is at least one sense in which the frame problem is relevant to philosophy: the frame problem provides a precise way of discriminating commonsense reasoning from prediction using scientific theory. Some people already believe that these two forms

of reasoning are distinct. However, even for these people, it has proved remarkably difficult to say exactly where the difference lies. I believe that the frame problem can help.

2 What is the frame problem?

The frame problem concerns how to represent a complex changing world. In artificial intelligence (AI), the standard way to represent a changing world is to use time-slices or situations. Time-slices represent what is true in the world at a particular moment in time; for example, they represent where a thing is located at an instant, or what the temperature is at an instant. Changes and events are represented as functions between time-slices. Applying an event to a particular time-slice yields another time-slice that represents the state of the world that the system thinks would be the result of that event. These functions or relations between time-slices are called the 'laws of motion': they describe how the agent thinks the world changes over time. This approach seems sensible, but it quickly runs into problems.

Consider how the functions that represent events are specified. For each object or state of affairs in the world one needs to specify how its state before the event relates to its state after the event. For example, imagine a function that represents the action of the agent moving through a door into another room. This function maps the agent's location before the event to a new location after the event. Part of the function would be specified as:

> IF the agent is in room R1 in situation S
> AND IF the door D from R1 to R2 is open in situation S
> THEN the agent is in R2 in situation goThrough(D, S)

This specification accounts for the *position* of the agent—it maps the position of the agent before the event to a new position after the event—but what about the rest of the world, how does the application of the goThrough function affect that? Most of the world will be unaffected by the agent going through the door. The agent's hair will remain the same colour. Paris will still be the capital of France, the walls of the room will still be the same, and so on. How does the system know this? The answer is that it doesn't: that information is not deductively entailed by the rule about position. The system has to be told about these other properties. How does one tell it? One way is to explicitly specify the information, for example:

> IF the agent has hair colour C in situation S
> THEN the agent has hair colour C in situation goThrough(D, S)

However, this specification needs to be repeated for nearly every property and relation in situation S. A huge number of no-change rules are needed in order for the system to know that moving rooms will not dramatically change the world. These no-change rules are called 'frame axioms'. The frame problem is that AI systems seem to need a lot of frame axioms. If there are 100 actions and 500 instantaneous facts, then the system will need up to 50,000 frame axioms. Worse, as the system learns more about the world, the number of frame axioms will come to dwarf everything else it knows.

The frame problem is the problem of getting a system to infer that the world remains largely the same before and after an action: it is the problem of getting the system to infer a large number of obvious non-changes. More precisely, the frame problem is the problem of getting the system to infer that the world remains the same *unless it has good reason for supposing otherwise*. It is worth clarifying a few things that the frame problem is not:

1. *Computational complexity*. The frame problem is not a point about computational complexity. Even if one had an infinitely fast machine the frame problem would still be present. It is representational systems that suffer from the frame problem, not individual machines. An infinitely machine has to be programmed. *We* have to decide what representational system to give it. If the representational system requires a vast number of no-change frame axioms, then we would still be stuck with the task of having to explicitly specify those axioms. In an even moderately complex world, this looks like an impossible task.

2. *Infallibility*. Humans make mistakes all the time; we are not infallible with our model of a changing world, so why should we expect our AI system to do better? Infallibility is not in question in the frame problem. The requirement is *not* that an AI system be infallible, but that it be reasonably reliable and robust. A system is reasonably reliable and robust if: (1) the system is right more often than not; and (2) the system does not fail catastrophically when changes that we would consider quite minor are made to the world. This specification is not precise, but it is clear that current AI systems fail to satisfy it.

3. *Hume's problem of induction*. Hume's problem of induction concerns how to justify our inductive inferences. The problem is to find a non-circular reason for trusting our inductive practices in the future. This is not the frame problem. The frame is not concerned with how to justify inferential systems, it is concerned with how to *construct* such systems. While Hume is interested

in whether our inductive beliefs are justified, AI researchers are interested in developing systems that simply possess those beliefs.

4. *The problem of knowing when to stop making inferences*. Real-world AI systems have to stop making inferences at some point and act. This is a familiar problem in decision theory: when does the expected cost of acquiring further information relevant to a decision exceed the expected value of that information? This is a problem of interest to many AI researchers, but it is not the frame problem.

5. *The problem of acquiring inductive beliefs (abduction)*. Abduction concerns how one generates inductive hypotheses from a finite set of observations. (This problem is distinct from Hume's problem of *justifying* the generated hypotheses). The problem of abduction is of great interest to AI researchers, but it is not the frame problem. For the purposes of the frame problem, we are happy to lend our AI system as much of our hard-won inductive knowledge as possible. The system is not required to infer its inductive beliefs from scratch.

6. *The problem of belief revision*. The problem of belief revision is how a system should react to new information. The problem of belief revision is distinct from the frame problem. Systems that suffer from the frame problem need not suffer from the belief revision problem. An example from Hayes (1987): Imagine a program which has to plan a sequence of movements of a robot arm across a table crowded with objects. Let the program have a complete representation of this world and its dynamics so that no new information will come its way. This program will have the frame problem in spades, but no belief revision problem.

7. *The ceteris paribus problem*. Problems associated with *ceteris paribus* clauses are the problems most commonly conflated with the frame problem. What I am calling the *ceteris paribus* problem is referred to in the AI literature as the *qualification problem*. The qualification problem is how to specify inductive inferences in such a way that they are defeasible. Inductive inferences only hold provided certain blocking circumstances do not obtain. There can be an unlimited number of blocking circumstances. We cannot specify all such circumstances, much less check that they do not obtain. As epistemic agents, what we tend to do is assume that they do not obtain, and then if a good reason to the contrary comes along, we consider our inference defeated. The qualification problem concerns how to formalise this feature of our reasoning. More specifically, the qualification problem concerns how to formalise the inference rule: 'If X is true and all else is equal, then conclude Y'.

The frame problem is also concerned with formalising inductive inference rules, but it is a different problem. The frame problem concerns the consequent, rather than the antecedent of inferences. It concerns how to formalise: 'Conclude Y *and* conclude that all else is unchanged unless there is good reason to suppose otherwise'.

(a) If *X* & …then *Y* (*Ceteris paribus problem*)

(b) If *X* then *Y* & … (*Frame problem*)

The *ceteris paribus* problem concerns how to formalise schema (a) in a way that does not require an explicit specification of the antecedents. The frame problem concerns how to formalise schema (b) in a way that does not require an explicit specification of the consequents. The *ceteris paribus* problem and the frame problem concern formalising inductive inference rules, but they concern different aspects of those rules.

To summarise: the frame problem is the problem of getting a system to infer that the world remains the same unless it already has a good reason to suppose otherwise.

3 Approaches to solutions

There is an enormous amount of literature on solving the frame problem, but no agreed solution. Rather than summarise all the approaches, I will describe one popular logic-based approach to solving the frame problem.

One might think of the frame problem as the problem of getting extra entailments out of limited assumptions. A tempting way to solve this problem is therefore to strengthen one's inferential system to get the extra entailments out. The aim is to build something like the following principle into an inferential system: 'Conclude that something stays the same, unless there is good reason to think otherwise'. Unfortunately, this principle is remarkably hard to formalise.

First, try a simple formalisation. Formalise the principle as: 'If there is no explicit rule for a property, then assume that property stays the same'. This formalisation has the virtue of being easy to implement in the original situation calculus. Does it solve the frame problem? Let us see how it works.

Imagine an AI system that reasons about a world containing coloured blocks. Suppose that this AI system, like the one described above, has axioms predicting what happens when a given events occurs. For example, the system may have axioms predicting what happens if a block is moved. Moving a block changes its location, so one of the axioms may be:

IF the location of block B is L in situation S
THEN the location of block B is L+D in situation moveBlock(B, D)

This may be just one of many axioms specifying how the moveBlock event affects the world. However, unlike the system described above, this system does not need a moveBlock axiom for *every* property in situation S. The system can use its no-change rule—'If there is no explicit rule for a property, then assume that property stays the same'—to deduce the state of other properties for itself. For example, it unnecessary to explicitly tell the system that the block will be the same colour before and after a move, or that Paris will be the capital of France, or that the walls of the room will be the same colour. The system can infer these consequences for itself. Therefore, this system does not need the vast number of frame axioms that our first system needed. Does it thereby solve the frame problem? Unfortunately, it does not.

Imagine that the world contains a blue spray-can that continuously sprays paint against a wall. Suppose that the system knows all about spray-cans, paint, and what paint can do to coloured blocks. Now suppose that the system is queried on what would happen if a red block were moved between the spray-can and the wall such that it is in the path of the blue paint. What will the system predict? The system will reason as follows. First, it will follow its explicit rules for moving blocks, such as its rule for how moving affects location. Then, the system will consider the properties for which changes are not explicitly specified. It will infer the state of these properties using its no-change rule: those properties will be marked as unchanged, since they lack explicit rules specifying otherwise.

The system will therefore make the following predictions. It will correctly predict that the position of the block changes when it is moved, but it will incorrectly predict that the red block will still be red when moved into the path of the blue paint. This is disappointing. We would have liked our system to make a correct prediction. But what makes this situation intolerable is that the system makes this prediction *despite the fact that it knows all about spray-cans, blue paint, and what they can do to coloured blocks*. The problem is not that the system does not know about spray-cans and paint, but that it cannot bring this information to bear to defeat its no-change rule. The no-change rule is too inflexible, it is not sensitive to being overridden in appropriate ways. What we would like is for the no-change rule to be defeated *when the system has good reason to think that a property will change*. This would be a better implementation of the original principle that we were trying to capture.

The failed formalisation we have just tried was: 'If there is no explicit rule for a property, then assume that property stays the same'. There are many ways to improve on it. Here are three popular example. First: 'Assume that a property stays the same, unless believing so results in a contradictory belief set.' Second:

'Assume that a property stays the same, so long as it is not believed that it changes.' Third: 'Assume that the maximal set of properties stay the same, consistent with the system's knowledge of the situation.' These three principles are associated with default logics, autoepistemic logics, and circumspection logics respectively.[1] All three formalisations produce different kinds of logic. None gives the same results, and none is as well-behaved as the original situational calculus. More importantly, none of the logics give results that match up with our intuitive understanding of the no-change principle. There are counterexamples, like the blue spray-can example, on which any given formalisation performs catastrophically badly. There seems to be no silver logical bullet to solve the frame problem.[2]

4 The frame problem and scientific theory

The purpose of the preceding section was to show that the frame problem is hard. In this section, I will argue that the frame problem provides a precise way in which commonsense reasoning can be distinguished from prediction using scientific theory.

The frame problem's most obvious application is in commonsense reasoning. Does the frame problem also affect prediction using scientific theory? The standard way of treating prediction using scientific theory is with the DN-model.[3] According to the DN-model, making a prediction is simple: a prediction deductively follows from a scientific theory conjoined with a numerical description of the current situation. This is familiar to anyone who has used a scientific theory: one has a theory, one plugs in the numbers, and the prediction deductively pops out. The DN-model has been criticised in other contexts, but its treatment of prediction has proved resilient. The main critics of the deductive model of prediction have been the DN-model's creator Carl Hempel (1988), and Nancy Cartwright (1983). Their criticism is that real scientific theories only entail their predictions *ceteris paribus* and the DN-model has difficulty modelling *ceteris paribus* clauses.

The DN-model also has difficulty modelling frame-type consequents. This is why the frame problem is so hard. If the DN-model could easily model frame-type

1. See Gabbay, Hogger and Robinson (1994) and Ginsberg (1987).

2. For a survey of problems with recent approaches to the frame problem see Morgenstern (1996). For recent logical approaches that hold promise for the future, see Levesque et al. (1997); Morgenstern (1996); Reiter (1991); Shanahan (1997).

3. Strictly speaking, the DN-model is a model of explanation, not prediction. However, the DN-model includes a deductive model of prediction. For the purpose of this paper, I focus only on this component of the DN-model. There is no special name for the component, so I will refer to it as the 'DN-model'. It is primarily this component that Cartwright targets in her criticism of the DN-model (Cartwright 1983, Ch. 7, 8).

consequents, then the frame problem would be easy, since formal AI models of inference are usually just variations on the DN-model. Therefore, the DN-model of prediction has difficulty with both *ceteris paribus* clauses and frame-problem-type consequents. Both features of inductive inference are clearly important. Does this mean that the frame problem, like the *ceteris paribus* problem, causes trouble for the DN-model of scientific inference? Does the frame problem demonstrate problems with the DN-model that *add* to those of Hempel and Cartwright?

I do not think so. The frame problem does not affect scientific inference. Consider an astronomical theory for predicting eclipses. Such a theory will almost certainly have *ceteris paribus* clauses: the theory will predict eclipses *provided* other planets do not upset the orbits, and so on. We would expect this feature to cause difficulties for a DN-type model of prediction using that theory. Does the frame problem introduce additional difficulties? It does not. What one wants from a theory of eclipses are times of eclipses, and this is exactly what the theory provides. One does not require that the theory *also* predict the non-change of other aspects of the world into the bargain. One does not ask for a *total* future world state from a scientific theory, only the state of certain properties—a partial future world state. We can fill in the other properties for ourselves using cognitive resources *outside* the scientific theory. In this sense, scientific theories are like specialised instruments for predicting the values of particular properties. They do a different job from commonsense reasoning, which fills in the rest of the world-state. The frame problem only bites if one tries to model this commonsense background in a DN-type way. Therefore, the frame problem affects commonsense reasoning, but not prediction using scientific theory.

The fact that models of these two forms of reasoning face different problems suggests that they are different forms of reasoning. Consider the grounds on which one justifies that two things are the same or different. A common reason for saying that two things are different is that accounts of those two things face different problems. One might claim that the physiological nature of horses is different to that of cows because, if one tries to give a physiological theory of horses, that theory faces significantly different problems from a physiological theory of cows. In contrast, a physiological theory of Betsy the cow does not face significantly different problems from a physiological theory of Daisy the cow (assuming Betsy and Daisy are normal cows). At the appropriate level of abstraction, different problems indicate different things. We have seen that at the level of abstraction at issue, accounts of commonsense reasoning faces different problems from accounts of scientific inference. Accounts of the commonsense reasoning face the frame problem, while accounts of prediction using scientific theory do not. This difference is not trivial because, as we saw, the frame problem is extremely hard to solve. Furthermore, the difference is not shown up by other epistemological problems—for example, the *ceteris paribus* problem affects both commonsense reasoning and prediction

using scientific theory. Therefore, the frame problem appears to provide at least a *prima facie* reason for saying that prediction using commonsense reasoning and prediction using scientific theory are different.

Whether this claim is ultimately correct is a delicate issue. One might feel that there is too much at stake for the issue to be settled on just one problem. However, even if one resists the claim above—that the two forms of reasoning are distinct—one still has to accept the claim that *accounts* of the two forms of reasoning face different problems. An account of commonsense reasoning faces different demands from an account of prediction using scientific theory: the former faces the frame problem while the latter does not. I wish to draw two implications of this claim.

First, since accounts of the two forms of reasoning faces different demands, one should be careful when proposing a model of scientific inference not to try to do too much. In particular, one should be careful not to try to implicitly model commonsense reasoning too. A good account of one need not be a good account of the other. The point is general but it suggests a way of defending the DN-model against Hempel and Cartwright's criticism. Hempel and Cartwright attack the DN-model for its inability to cope with *ceteris paribus* clauses. But perhaps *ceteris paribus* clauses, like frame-problem-type consequents, are better seen as part of the commonsense component of reasoning. A scientific theory may DN-entail its prediction and then our commonsense reasoning decides whether that theory applies in that circumstance. On this view, a scientific theory is like an instrument for predicting values of properties in a straightforward DN-type way, but we decide on the basis of our commonsense reasoning whether to use that instrument, just as we decide on the basis of commonsense reasoning what the state of the rest of the world is likely to be. A two-tier approach to inference has already been forced on us by the frame problem. It does not seem unreasonable to apply this distinction to *ceteris paribus* clauses.

The second implication is as follows. Since accounts of the two forms of reasoning face different problems, treating commonsense reasoning as a kind of scientific inference may be a mistake. The computational theory of mind models commonsense reasoning in exactly this way: as a kind of scientific theory. If the two forms of reasoning are distinct, or require radically different treatments, then there is something seriously wrong with this approach. This seems to be borne out both in Fodor (2000)'s comments that the frame problem creates serious trouble for the computational theory of mind, and in the difficulty that AI researchers have had in creating successful commonsense reasoning systems. The computational theory of mind works well for automated scientific reasoning systems, but it quickly breaks down when asked to produce anything commonsensical. This should be unsurprising given that the DN-model that works for scientific inference was not

designed to deal problems like the frame problem that face models of commonsense reasoning.

5 Conclusion

The frame problem provides a diagnostic difference between commonsense reasoning and prediction using scientific theory: the frame problem affects one but not the other. This difference can be taken in two ways. It can be taken either as evidence that the two forms of reasoning are essentially different; or, it can taken in a weaker way, as demonstrating that accounts of the two forms of reasoning face different demands. Either way, the difference has consequences for both the computational theory of mind and the treatment of prediction in scientific theories.

References

Cartwright, N. 1983. *How the Laws of Physics Lie.* Oxford: Oxford University Press.

Dennett, D. C. 1987. 'Cognitive Wheels: The Frame Problem of AI'. In *The Robot's Dilemma,* edited by Z. W. Pylyshyn, 41–64. Norwood, NJ.: Ablex.

Fodor, J. A. 1987. 'Modules, Frames, Fridgeons, Sleeping Dogs and the Music of the Spheres'. In *The Robot's Dilemma,* edited by Z. W. Pylyshyn, 139–150. Norwood, NJ.: Ablex.

———. 2000. *The Mind Doesn't Work That Way.* Cambridge, MA: MIT Press.

Gabbay, D., C. Hogger and J. Robinson, eds. 1994. *Handbook of Logic in Artificial Intelligence and Logic Programming.* Oxford: Oxford University Press.

Ginsberg, M., ed. 1987. *Readings in Nonmonotonic Reasoning.* San Francisco, CA: Morgan Kauffman.

Glymour, C. 1987. 'Android Epistemology and the Frame Problem: Comments on Dennett's 'Cognitive Wheels''. In *The Robot's Dilemma,* edited by Z. W. Pylyshyn, 65–76. Norwood, NJ.: Ablex.

Haugeland, J. 1987. 'An Overview of the Frame Problem'. In *The Robot's Dilemma,* edited by Z. W. Pylyshyn, 77–95. Norwood, NJ.: Ablex.

Hayes, P. J. 1987. 'What the Frame Problem Is and Isn't'. In *The Robot's Dilemma,* edited by Z. W. Pylyshyn, 123–138. Norwood, NJ.: Ablex.

Hempel, C. G. 1988. 'Provisos: A Problem Concerning the Inferential Function of Scientific Theories'. *Erkenntnis* 28:147–164.

Levesque, H., R. Reiter, Y. Lesperance, F. Lin and R. Scherl. 1997. 'Golog: A logic programming language for dynamic domains'. *Journal of Logic Programming* 31:59–84.

McCarthy, J., and P. J. Hayes. 1969. 'Some philosophical problems from the standpoint of artificial intelligence'. In *Machine Intelligence 4,* edited by B. Meltzer and D. Michie, 463–502. Edinburgh: Edinburgh University Press.

Morgenstern, L. 1996. 'The problem with solutions to the frame problem'. In *The Robot's Dilemma Revisited,* edited by K. Ford and Z. W. Pylyshyn, 99–133. Norwood, NJ: Ablex.

Reiter, R. 1991. 'The frame problem in the situation calculus: A simple solution (sometimes) and a completeness result for goal regression'. In *Artificial intelligence and mathematical theory of computation: Papers in honor of John McCarthy,* edited by V. L. Lifschitz, 359–380. Boston, MA: Academic Press.

Shanahan, M. 1997. *Solving the Frame Problem.* Cambridge, MA: Bradford Books, MIT Press.