Jurek Gryz

The frame problem in artificial intelligence and philosophy

Filozofia Nauki 21/2, 15-30

2013

Filozofia Nauki
Rok XXI, 2013, Nr 2(82)

Jarek Gryz

The Frame Problem in Artificial Intelligence and Philosophy

1. INTRODUCTION

The field of Artificial Intelligence (AI) started very soon after the first American digital computer ENIAC became operational in 1945. In 1956, key figures working in this area in the US (Herbert Simon, Allen Newell, Marvin Minsky, and John McCarthy among others) met for a two-month workshop in Dartmouth which was to become the official birthplace of the field. By that time, some impressive results had already been achieved. Newell and Simon had a reasoning program, the Logic Theorist, that was able to prove most of the theorems in Chapter 2 of Russell and Whitehead's *Principia Mathematica* (McCorduck 2004). Gelernter's Geometry Theorem Prover (Gelernter 1959) and Samuel's checkers program (Samuel 1959) matched human skills in mathematics and the game of checkers. Within a few years, first expert systems, to be applied in other areas of human expertise, have been built. MYCIN (Shortliffe 1976), developed to diagnose blood infections, was able to perform considerably better than junior doctors. These early successes led, understandably, to overblown expectations. As early as 1958, Simon boasted: "The simplest way [we] can summarize the situation is to say that there are now in the world machines that think, that learn, and that create" (Simon 1958: 8) and a few years later: "Machines will be capable, within twenty years, of doing any work that a man can do" (Simon 1965: 96).

The first signs of difficulties appeared when the solutions developed for small and well-defined domains were applied in real world scenarios. One problem was purely computational: many of the AI problems are intractable (Garey, Johnson 1979). To find the shortest route through n cities, one needs to search through all $n!$ possible

paths connecting them. When *n* is small, this can be done almost instantenously, when *n* is more realistic (say 50), the problem cannot be solved by any computer in our lifetime. Most of the AI algorithms were tested in so-called *microworlds*, limited domains which contained few objects and few possible actions (an example of such a microworld for manipulating blocks is shown in Figure 1). Solutions developed for microworlds simply did not scale up for more realistic domains.

The second difficulty facing AI turned out to be even more intricate. Proving theorems, playing checkers, or even diagnosing infections did not require knowledge outside of the domain in question. As soon as the problems become open-ended (writing a funny story, holding a conversation, or even grocery shopping) commonsense knowledge becomes indispensable. Even a seemingly simple task of language comprehension can be insurmountable as it may require large amount of sophisticated knowledge. Consider the following two sentences (Winograd 1972: 33):

> The police refused to give the students a permit to demonstrate because they feared violence.
>
> The police refused to give the students a permit to demonstrate because they advocated revolution.

What does the word "they" refer to in each of the two sentences? The grammatical structure of the two sentences is identical. Yet we have no difficulty in correctly understanding who "they" are in each case, *because* we know who is more likely to fear violence or to advocate revolution.

Once the importance of commonsense knowledge has been recognized, research in AI proceeded in two directions. First, such knowledge had to be gathered, formalized, and stored to be later used by a computer. In 1984, a large project under the name of CYC has been initiated by the Microelectronics and Computer Technology Corporation (MCC) in Texas. The goal of the project was to build a database containing a significant portion of commonsense knowledge possessed by a contemporary human being. The plan was to enter by hand around 100 million assertions. At any one time as many as 30 people were simultaneously entering data. They analyzed newspaper articles, encyclopedia entries, advertisements, etc. sentence by sentence. The idea was not to create yet another encyclopedia, but to encode the knowledge anyone needs to have before he can understand an encyclopedia. The types of entries are trivial by our human standards: water flows downhill, putting a hand into fire hurts, you can't walk through a closed door, yet they are necessary for any future thinking machine to understand. By the end of the first six years of the project over one million assertions have been entered manually into the database, but soon after that the project's life at MCC came to an end.[1] A database of raw facts is useless un-

[1] Given the huge resources put into the project and its rather unimpressive results, CYC has often been harshly criticized "[it is] one of the most controversial endeavors of the artificial intelligence history" (Bertino, Piero, Zarri 2001).

less a machine has a mechanism for applying this knowledge to solve real problems. Thus, a second important direction of AI research at that time went into formalizing commonsense reasoning in which planning played a key role. But these efforts very soon hit a roadblock, the Frame Problem.

2. THE FRAME PROBLEM IN ARTIFICIAL INTELLIGENCE

2.1. Planning

The *technical* or *logical* frame problem (as opposed to the philosophical one) was first discussed by McCarthy and Hayes (1969) in the context of situation calculus. The situation calculus was an instance of first-order logic designed to formalize commonsense reasoning. Its primary use has been to formalize planning tasks to be undertaken by a robot. The ontology of the situation calculus consists of situations, fluents, and actions. Situations are snapshots of the world at a particular time. A fluent is something that changes over time; boolean fluents are called states. Consider a world of blocks shown in Figure 1. To indicate that block `B` is on block `A` we write: `On(B,A)`; to indicate that there is nothing on block `B`, we write `Clear(B)`. To relate fluents and situations we write `Holds(On(B,A), S1)`, meaning that block `B` is on block `A` in situation `S1`. Thus, the complete description of the situation in Figure 1 can be specified as follows:

Holds(On(B,A), S1)
Holds(Clear(B), S1)
Holds(Clear(C), S1)
Holds(Clear(D), S1)

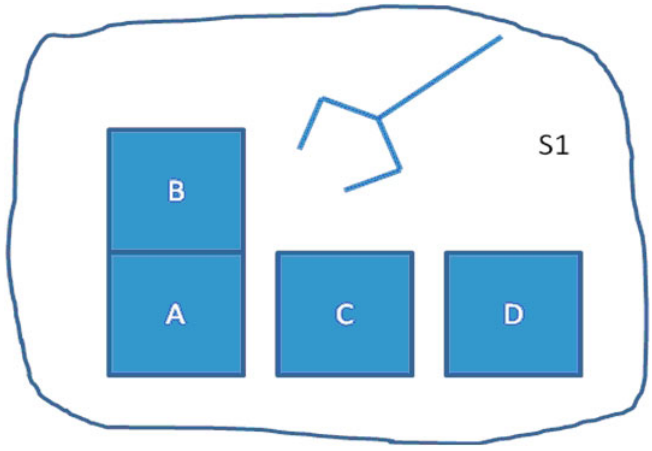

Figure 1. World of blocks in situation S1

Actions bring about new situations by changing fluents (states) in previous situations. Let us assume that there are only two possible actions in our block world: `Stack(X,Y)` and `Unstack(X,Y)` representing stacking and unstacking `X` on (from) `Y` respectively. The effect of the action `Stack(X,Y)` performed in situation `S1` can be described as:

Holds(On(X,Y), Result(Stack(X,Y), S1))

The complete description of the two actions is specified via axioms of the form:

(Holds(Clear(X),S) ∧ Holds(Clear(Y),S) ∧ X≠Y) → Holds(On(X,Y), Result(Stack(X,Y),S))

(Holds(On(X,Y),S) ∧ Holds(Clear(X),S)) → Holds(Clear(Y), Result(Unstack(X,Y),S))

in which all variables are universally quantified. The effects of a sequence of actions $A_1,\ldots, A_n$ (rather than a single action) can be defined inductively in a straightforward way:

Result(($A_1,\ldots, A_n$),S) = Result(A_n, Result(($A_1,\ldots, A_{n-1}$),S))

Now, consider a planning task in which a robot, given the initial situation `S1` (as shown in Figure 1) has to come up with a plan (that is, a sequence of actions α) to reach a goal where `A` is on `C` and `D` is on `A`, that is:

Holds(On(A,C), Result(α,S1))
Holds(On(D,A), Result(α,S1))

A straightforward solution to this task is the sequence of actions shown in Figure 2. Let us specify the first few steps in the 'proof' that leads to the required solution:

1. Holds(On(B,A),S1)
2. Holds(Clear(B),S1)
3. Holds(Clear(A),S2), where S2=Result(Unstack(B,A),S1)
4. Holds(Clear(C),S2)
5. Holds(On(A,C),S3), where S3=Result(Stack(A,C),S2)

…

Unfortunately, this is not a valid proof. The first three statements are true: (1) and (2) were part of the initial specification of `S1` and (3) follows from the action axiom for `Unstack` and the initial specifications. Statement (4) seems obviously true (unstacking `A` from `B` does not affect any other blocks), yet it is neither a description of `S1`, nor can it be derived by means of any action axioms. It is not enough to know that `C` is clear in the initial state, but that `C` remains also clear in the state that results from the `Unstack(B,A)` action, that is, that the following is true:

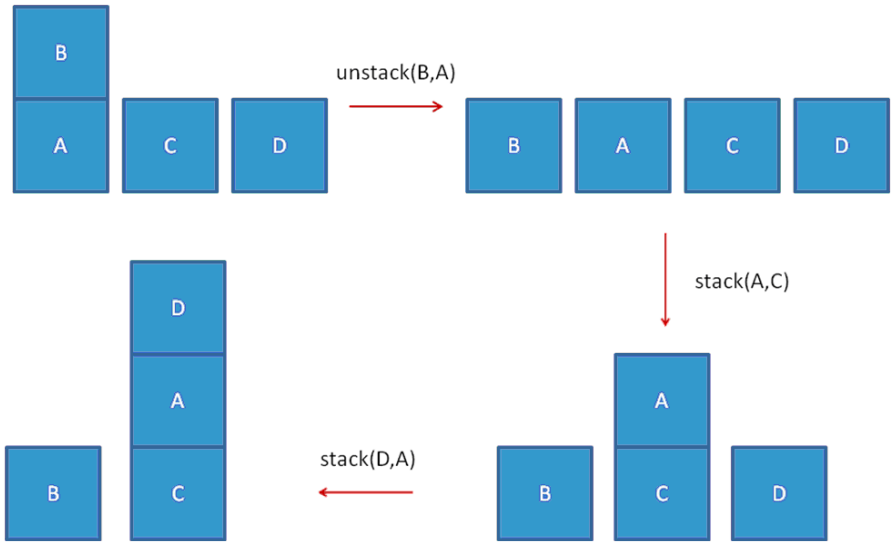

Figure 2. A plan to have A on C and D on A

Holds(Clear(C), Result(Unstack(B,A),S1))

Again, this step is perfectly natural in commonsense reasoning, yet it is not derivable as a matter of logic. McCarthy and Hayes (1969) suggested a straightforward way of handling this problem by adding formulas such as the one above (they called them *frame axioms*) for every fluent and every action. But the number of such axioms can be enormous: for *m* actions and *n* fluents one needs *m* × n axioms. For sufficiently large *m* and *n* (in practical scenarios they *will* be large) looking through all such axioms would dramatically degrade performance of any system. It is also counterintuitive. If the goal of AI is to build a system that resembles human mode of reasoning, than this is certainly not correct. When we pick up `B` from `A` we do not remind ourselves that the distance of `A` from the Sun has not changed, neither did its temperature or color or any other properties. Last, but not least, such axioms are context-dependent. In a scenario where removing `B` from `A` causes `D` to be put on `C`, the axiom specified above is false.

When an action is performed, some things change and some do not. How can we say in a concise way which things do not change when a particular action is performed? This is, in a nutshell, the classical frame problem.

2.2. Nonmonotonic Logic

Soon after the frame problem has been identified, a number of approaches have been proposed to solve it. Most of them extended classical logic to handle the prob-

lem, but there were also solutions using statistics and connectionist models.[2] Among the logical approaches, non-monotonic logic (Horty 2001) offered a particularly original and interesting solution. *Default logic* (Reiter 1980) and *circumscription* (McCarthy 1980) were the most prominent proposals to solve the frame problem using non-monotonic logic. We will discuss default logic in detail.[3]

The consequence relation of a classical logic is monotonic: if a formula p is a consequence of a set of formulas S, then p is also a consequence of $S \cup \{r\}$, for an arbitrary formula r. In other words, the set of conclusions we can draw from the premises grows monotonically with an addition of new formulas to the premises. In particular, a proof of a formula cannot be invalidated by *adding* a new formula to the derivation. But this is not the case in common sense reasoning. We tend to learn about the relationships in the world by making sweeping generalizations, such as *all swans are white* or *all birds fly.* And then, when we see a black swan or learn about ostriches, we retract or qualify our previous claims. In this sense, common sense reasoning is non-monotonic: a conclusion set need not grow monotonically with the premise set. If we could formalize this type of common sense reasoning, then we might be able to solve the frame problem. We could state succinctly that no properties change as a result of our actions except the ones we explicitly specify. The list of exceptions can grow in time, yet the *no-change* claim (modulo the exceptions) remains in place. This is the approach taken in default logic.

In addition to standard rules of inference, default logic adds *default rules*. A default rule is a triple: $\frac{\alpha:\beta}{\gamma}$, where α is a prerequisite, γ is a consequent, and β is the justification. The meaning of the rule is: if α has been already established, one can add γ to the set of conclusions assuming that this set is consistent with β. A *default theory* is a pair $\Delta = \langle \mathbf{W}, \mathbf{D} \rangle$, in which **W** is a set of ordinary formulas and **D** is a set of default rules. Consider our previous example *all birds fly.* This can be represented as a default rule, such that $\mathbf{D_1} = \{ \frac{B(x):F(x)}{F(x)} \}$ with the meaning *if x is a bird, then x flies unless we have information to the contrary.* If we now learn that Tweety is an ostrich (hence does not fly), that is, $\mathbf{W_1} = \{\neg F(\text{Tweety})\}$, we cannot draw a conclusion that Tweety flies.

To accommodate new rules of inference, the standard concept of logical consequence had to be modified. The conclusion set $\mathcal{E}$ associated with a default theory $\Delta = \langle \mathbf{W}, \mathbf{D} \rangle$ is called an extension and is defined as follows: $\mathcal{E} = \bigcup_{i=1}^{\infty} \mathcal{E}_i$, where:

$$\mathcal{E}_0 = \mathbf{W},$$

$$\mathcal{E}_{i+1} = Th(\mathcal{E}_i) \cup \{C \mid \frac{A:B}{C} \in \mathbf{D}, A \in Th(\mathcal{E}_i), \neg B \notin \mathcal{E}\}$$

[2] An overview of the solutions to the frame problem is provided in (Kamermans, Schmits 2004).

[3] We follow Horty's (2001) presentation of default logic.

and Th($\mathcal{E}_i$) is a set of standard logical consequences of $\mathcal{E}_i$. The idea behind this definition is that we first conjecture a candidate extension for a theory, $\mathcal{E}$, and then using this candidate define a sequence of approximations to some conclusion set. If this approximating sequence has $\mathcal{E}$ as its limit, $\mathcal{E}$ is indeed an extension of the default theory.[4] Note that the extension of Δ = <**W, D**> in which **D** is empty is simply Th(**W**).

One may argue that the concept of an extension in default logic seems more natural than the notion of consequence in classical logic. A non-empty set of formulas in classical logic will have a non-empty set of consequences. Not so in default logic. Consider the following example. Let $\mathbf{W_2}=\{\alpha, \beta\rightarrow\neg\gamma\}$ and $\mathbf{D_2}=\{\frac{\alpha:\gamma}{\beta}\}$. If this theory has an extension, then either $\neg\gamma \in \mathcal{E}$ or $\neg\gamma \notin \mathcal{E}$. Let us assume that $\neg\gamma \in \mathcal{E}$. Since $\neg\gamma$ is not in $\mathcal{E}_0$ and the default rule cannot be applied (since we assumed that $\neg\gamma \in \mathcal{E}$), $\neg\gamma$ cannot be added to any of $\mathcal{E}_i$, hence cannot be in $\mathcal{E}$. A contradiction. Assume the opposite then, that is, $\neg\gamma \notin \mathcal{E}$. The default rule is applicable now, hence β is added to $\mathcal{E}_1$. By the implication $\beta\rightarrow\neg\gamma$, $\neg\gamma$ is added to $\mathcal{E}_2$ and since $\mathcal{E}$ is a union of $\mathcal{E}_0, \mathcal{E}_1, \mathcal{E}_2$..., $\neg\gamma \in \mathcal{E}$. A contradiction again. An incoherent default theory has no extensions. An inconsistent set of propositions in classical logic implies everything. Arguably, the former is more intuitive than the latter.

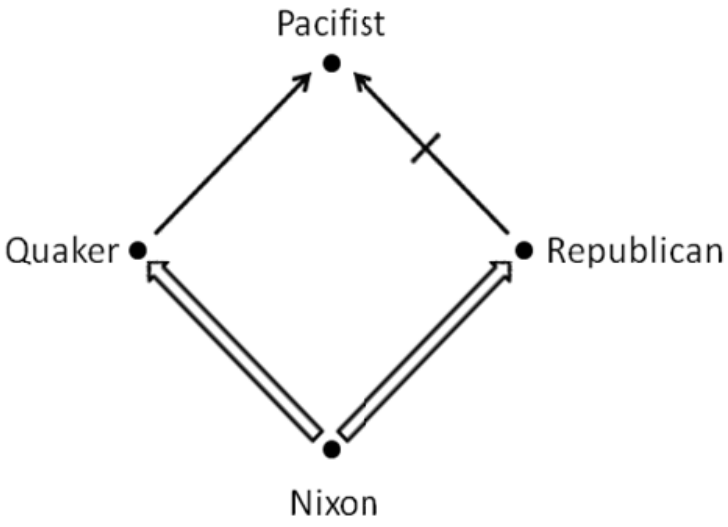

Figure 3. Nixon Diamond

Even more interestingly, a default theory can have multiple extensions. A famous example (Reiter, Criscuolo 1981) of this case encoded as an inheritance network in

[4] This definition of an extension was presented in (Horty 2001) and is more intuitive that the 'official' definition of Reiter's.

Figure 3 is called the Nixon Diamond (double lines indicate logical implication, single lines indicate default rules, crossed line indicates negation). The facts represented by this diagram are: Nixon is a Republican, Nixon is a Quaker, Republicans tend not to be pacifists and Quakers tend to be pacifists. As a default theory these facts can be stated as: $\mathbf{W_3}$={Q(nixon), R(nixon)} and $\mathbf{D_3}$={$\frac{R(x):P(x)}{P(x)}, \frac{R(X):\neg P(x)}{\neg P(x)}$}. This theory has two extensions, one with P(nixon) and one with ¬P(nixon). Both conclusions are equally valid, yet they cannot be both entertained at the same time. Again, this seems like a natural description of commonsense reasoning which cannot be captured in classical logic.

Finally, it is straightforward to encode by a single default rule the possibly numerous frame axioms. The following rule schema:

$$\frac{\text{Holds(f, s): Holds(f, Result}(\alpha, \text{S))}}{\text{Holds(f, Result}(\alpha, \text{S))}}$$

says that whenever a fact `f` holds in situation `s`, then — if it is consistent to assume that `f` still holds after the performance of action α — we should conclude by default that `f` still holds after the performance of action α. A reader can verify that the missing assumption in our planning proof of Section 1, that is:

Holds(Clear(C), Result(Unstack(B,A),S1))

can now be derived using the initial facts about the microworld together with the above default rule.

Does the default logic solve the frame problem? It certainly solves this particular instance of the problem, but it turns out that it leads to some very counterintuitive results in other cases. The following example shown in Figure 4 illustrates the *multiple extension problem*.

Hermann belongs to Pennsylvania Dutch (Amish) speakers who actually speak German. Since he was born in Pennsylvania, he was born in the USA. There are also two default rules here. One states that Pennsylvania Dutch speakers tend to be born in Pennsylvania, the other one states that German speakers tend *not* to be born in the USA. Common sense would tell us that Hermann was born in the USA. The fact that he speaks German is incidental and should not lead to the conclusion that he was not born in the USA (in fact, his speaking German is a *result* of him being a Pennsylvania Dutch, hence likely to be born in the USA). According to default logic, however, *both* extensions (one with Hermann born in the USA and another not born in the USA) are *equally* valid. What was considered an advantage in the case of Nixon Diamond (both extensions seemed reasonable) is clearly a flaw here. Default logic does not distinguish between extensions as it has no tools to do that. To find an intended extension, that is, to draw correct conclusions from arbitrary sets of facts we need the ability to identify information relevant to a particular context. Relevance, however, is not something that can be easily codified in logic.

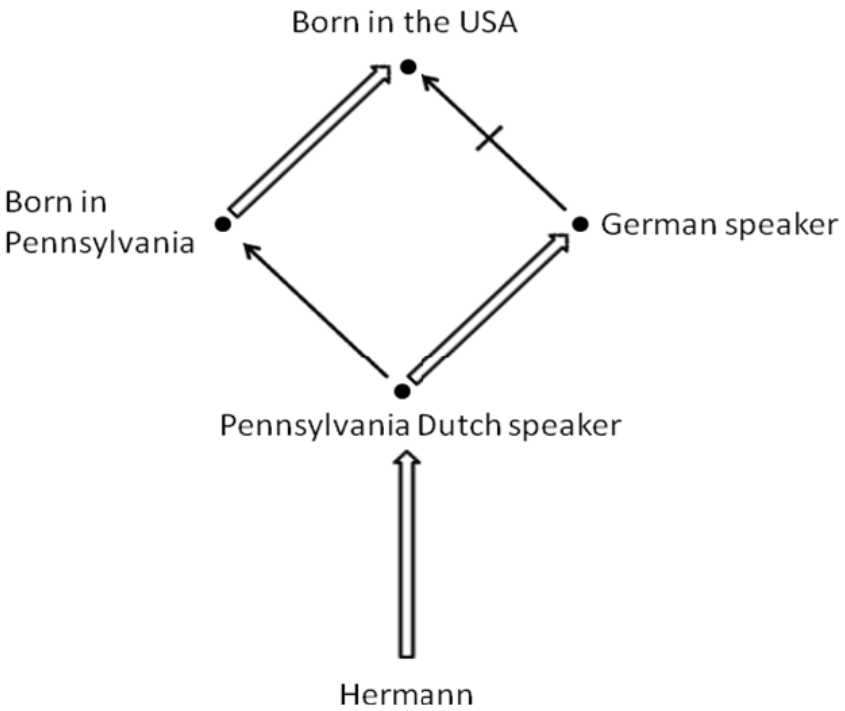

Figure 4. An illustration of the multiple extension problem

Circumscription, another non-monotonic logic approach to solving the frame problem, fared no better than default logic. A few years after McCarthy presented his solution, Hanks and McDermott (1987) offered a convincing scenario, called the Yale Shooting Problem, showing that the proposed solution leads to an anomaly. The situation they describe is as follows. We have an unloaded gun, which we then load, wait a while and fire it at our best friend Fred. Formalizing this scenario in logic and using circumscription to conclude what the outcome is, leads to two different solutions. According to one of them, Fred dies (as expected), according to another one (equally valid) gun mysteriously unloads itself and Fred survives. The idea of circumscription was to minimize changes that may occur in the world as a result of actions. In both solutions, the same number of default assumptions is violated: in the first one, Fred's staying alive, in the second one, gun's staying loaded. Common sense tells us that Fred should be dead, yet the facts and the particular logic employed here support with the same force both conclusions. We are faced again with the multiple extension problem.

A number of approaches have been offered to solve the Yale Shooting Problem in an attempt to save the solution to the frame problem.[5] None of them was fully suc-

[5] Morgenstern (1996) provides an overview of many such solutions.

cessful. Many people, even within AI community, think that "the solutions to the frame problem have all failed in at least one respect" (Morgenstern 1996). Even those who believe that the problem is "more-or-less solved" (Shanahan 1999), point out to difficulties that still lie ahead of any fully satisfactory solution (concurrent actions, actions with non-deterministic effects, continuous change, etc). But it was in philosophy where the frame problem took an entirely new dimension. Rather than trying to solve the problem, the philosophers asked themselves a different question: why is the problem so hard? We look at some of the answers in the next section.

3. THE FRAME PROBLEM IN PHILOSOPHY

The frame problem has been with us for over 40 years now. Dozens, if not hundreds of research papers and a number of conferences were devoted exclusively to this topic (Pylyshyn 1987), (Ford, Pylyshyn 1996). It has sparked heated debates about AI methodology. It has led some researchers to declare the logic-based approach to commonsense reasoning a failure. Why has a single technical problem become a source of so much frustration, even disillusionment with AI? After all, there are many unsolved problems in other sciences (for example, "Does P = NP?" in computer science), yet they do not threaten the existence of an entire discipline. There may be good reasons, however, why the frame problem is unique (Morgenstern 1996).

- The problem does not seem to be *that* difficult. It is straightforward to state that most things do not change as time passes. Yet logic cannot capture this rather obvious observation.
- The problem is central to AI. A solution is necessary for planning and it is hard to see how one can build an intelligent machine without the ability to plan. It is thus quite different from problems like "P = NP?" which are interesting and important for computer science, yet not crucial for its progress.
- So much work has been put into solving the problem and so little progress has been made. A number of solutions have been proposed to solve it, only to be shown, one by one, to be incomplete in one aspect or another. There is a feeling that these solutions have not touched the heart of the problem.

If the problem is so easy to state, yet so hard to solve, then, perhaps, it is just a symptom of a different, more general problem and it is *that* problem that we need to solve first. Or it is a series of closely interconnected problems involved, and "there would be little point in 'solving' the frame problem if it meant 'unsolving' some other problem" (Janlert 1996). This is where the problem was taken up by philosophers and given a new interpretation. Some of their opinions may seem radical (at least in the eyes of the AI community[6]), but they shed interesting light on AI and cognitive science in general. We discuss in detail three such views.

[6] Patrick Hayes went as far as to say about one of them: "Fodor doesn't know the frame problem from a bunch of bananas" (Hayes 1987: 132).

3.1. Dennett's Epistemological Holism

Dennett readily admits at the outset that he appropriates the name "frame problem" to cover more than just the narrow technical problem defined by McCarthy and Hayes. According to Dennett, the problem arises from our widely held assumptions about the nature of intelligence and it is "a new, deep epistemological problem — accessible in principle but unnoticed by generations of philosophers — brought to light by the novel methods of AI, and still far from being solved" (Dennett 1987: 42). To illustrate his understanding of the problem, Dennett presents the following scenario (Dennett 1987: 41–42):

> Once upon a time there was a robot, named R1 […]. One day its designers arranged for it to learn that its spare battery […] was locked in a room with a time bomb set to go off soon. R1 located the room […] and formulated the plan to rescue the battery. There was a wagon in the room, and the battery was on the wagon, and R1 hypothesized that a certain action […] 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, but did not realize that pulling the wagon would bring the bomb out along with the battery.

R1 did not pay attention to the ramifications of his actions. Thus, the designers decided that R1 must be redesigned to be able to recognize all the implications of his actions.

> They called their next model, the robot-deducer, R1D1. They placed R1D1 in much the same predicament that R1 succumbed to, and […] 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 would cause its wheels to turn more revolutions that there were wheels in the wagon — when the bomb exploded.

This time, R1D1 was equipped with all necessary knowledge to perform the action correctly, yet it failed to reach the proper conclusion in reasonable amount of time. This, one might say, is the computational aspect of the frame problem. The obvious way to avoid it is to appeal to the notion of relevance. Only certain properties are relevant in the context of any given action and we can confine the deduction only to those.

> 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 implications'.

But what is relevant in this case? Relevance clearly depends on a context. If the battery happened to be on the floor, then the fact that the bomb was on a wagon would be irrelevant to the action of picking up the battery and removing it from the room. If the temperature outside the room were –100°C (which would freeze the bomb), then pulling the wagon out of the room would have been a safe course of action. In this case, the temperature — to which the robot paid no attention before — would be suddenly relevant. A moment of thought should suffice to see that specify-

ing what propositions are relevant to what context is hopeless. Contexts are not independent from one another, one needs to appeal to a larger context to determine significance of elements in a narrower context (for example, to recognize two dots as eyes in a picture, one must have already recognized the context as a face). But then, as Dreyfus put it: "if each context can be recognized only in terms of features selected as relevant and interpreted in a broader context, the AI worker is faced with a regress of contexts" (Dreyfus 1992: 189).

So how do humans do it? Most of it we probably learn from experience, some of it may be innate. But the crucial question for AI is *how* this knowledge is stored and processed. In great majority of our actions we do not consciously think through our plans. In fact, the relevant knowledge rises to the surface *only* when we make mistakes and need to reconsider our course of action or plan an entirely novel action. But even in these cases introspection is not of much use. We do not ponder upon facts one by one, we are somehow capable to pay attention only to the relevant ones. Also, our information processing is so fast that it cannot involve drawing thousands of interim conclusions. We operate on a portion of our knowledge at any moment and that portion cannot be chosen by exhaustive consideration.

A great discovery of AI is the observation that a robot (a computer) is the fabled *tabula rasa*. For it to operate in a real world, it must have all the information explicitly specified and then organized in such a way that only the relevant portion of it is used when it is time to act. So far we have no idea how to do that. The philosophical frame problem, according to Dennett, is this: "How is it possible for holistic, open-ended, context-sensitive relevance to be captured by a set of propositional, language like representation of the sort used in classical AI?" (Shanahan 1999).

3.2. Fodor's Metaphysical Relativism

Fodor takes the frame problem one step deeper into the philosophical thicket. Most of the solutions of the classical frame problem follow the "sleeping dog" strategy (McDermott 1987), that is, the assumption that most events leave most of the facts untouched. Both default logic with its default assumptions as well as circumscription, which minimizes the changes resulting from actions, follow that strategy. In fact, the assumption of metaphysical inertia gives rise to the frame problem in the first place as we are trying to formalize the fact that *most* things do not change. Fodor challenges this assumption. To say that most properties do not change as a result of actions, one must make certain ontological commitments. The properties we assign to the objects in the world must indeed be inert to make a sweeping claim that *most* of them do not change with time. But what justifies this metaphysical assumption? To make his point more vivid, Fodor invents a new property of physical particles, the property of *fridgeon*: x is a fridgeon iff x is a particle at t and Fodor's fridge is on at t. Now, every time Fodor turns on or off his fridge, he changes the state of

every physical particle in the universe. The assumption of metaphysical inertia no longer holds and the sleeping dog strategy (widely accepted in AI) will no longer solve the frame problem.

The conclusions Fodor draws from this observation are far-reaching. The frame problem is empty unless we have the right ontology to go with it. Clearly, properties such as fridgeon should not be a part of this ontology because they do not represent computationally relevant facts. So how do we choose the right ontology? Fodor draws an analogy with science where the principle of scientific conservatism (similar to the sleeping dog strategy) is obeyed: alter the minimum possible amount of prior theory to accommodate new data. Just as we do not know how to build the simplest canonical notation for scientific theories to respect that principle, we do not know what canonical representation of the world is appropriate for AI. The frame problem is an indication of that difficulty.

It seems, however, that Fodor's argument applies as much to commonsense reasoning as it does to AI. First of all, one needs not to invent new "kooky" concepts to notice that some properties are just irrelevant when it comes to reason about everyday actions. A perfectly "natural" property of *location,* when applied to elementary particles, changes even more often than the property of *fridgeon.* We ignore it precisely because it is *irrelevant* for pretty much any of our actions. Perhaps the question of choosing the right ontology is just the question of choosing relevant properties within any (reasonable) ontology. In that sense, Fodor's argument can be reduced to Dennett's problem of relevance.

Fodor was not the only one to bring the issue of metaphysics into the discussion of the frame problem. Janlert (Janlert 1987) believes that McCarthy's notion of situation is an inappropriate basis for a metaphysics of change and the frame problem is just a problem of finding the right metaphysics. Fetzer (1991) took Fodor's analogy with science one step further and argued that the frame problem is an instance of the problem of induction.[7] This view has been rather controversial, however (see Dennett 1987: 52 or Morgenstern 1996: 110).

3.3. Dreyfus's Methodological Nihilism

Dreyfus has been an ardent critic of AI since 1960s. His views were so extreme that hardly anyone within AI community took him seriously at that time. Yet many of Dreyfus's arguments were sound and his predictions about the future of the field have been mostly correct. The frame problem did not feature large in these arguments, because — contrary to Dennett or Fodor — Dreyfus did not think of it as a real philosophical problem. It is a pseudoproblem spawned by a misguided research program. It may very well be insoluble, but we do not need to worry about it.

[7] Indeed, Fodor's concept of *fridgeon* serves similar purpose as Goodman's *grue* in his discussion of the new riddle of induction (Goodman 1954).

Dreyfus's criticism of AI touched the very heart of the discipline: its philosophical foundations. Although AI researchers never discussed these foundations explicitly, they "had taken over Hobbes's claim that reasoning was calculating, Descartes' mental representations, Leibniz's idea of "universal characteristic" — a set of primitives in which all knowledge could be expressed, Kant's claim that concepts were rules, Frege's formalization of such rules, and Wittgenstein's postulation of logical atoms in his Tractatus" (Dreyfus 2007: 1137). These assumptions were behind the famous *physical symbol system hypothesis* endorsed by Newell and Simon[8]: "A physical symbol system has the necessary and sufficient means for general intelligent action" (Simon, Newell 1976: 116). According to this hypothesis, human thinking is just symbol manipulation and any machine that implements the same rules for symbol manipulation can be intelligent.

Dreyfus questioned not only the physical symbol system hypothesis, but also several other implicit tenets shared by AI community. He organized them into the following four assumptions (Dreyfus 1992):

> Biological assumption: the brain is a symbol-manipulating device like a digital computer.
>
> Psychological assumption: the mind is a symbol-manipulating device like a digital computer.
>
> Epistemological assumption: intelligent behavior can be formalized and thus reproduced by a machine.
>
> Ontological assumption: the world consists of independent, discrete facts.

Dreyfus drew arguments against these assumptions from philosophy of Heidegger and Merleau-Ponty. He pointed out three aspects of intelligent behavior almost entirely ignored by AI. First, most of our everyday activities rely on skills, that is *knowledge how*, rather than on conscious deliberation over facts, that is *knowledge that*. Second, the body is an integral part of intelligent behaviour: our ability to respond to new situations involves a situated, material body. Third, human behavior is not goal-oriented but value-oriented and as such is always dependent on context or situation. Dreyfus argues convincingly that these aspects of human intelligence cannot be reproduced, or even simulated, within traditional AI.

Where does the frame problem fit in Dreyfus's attack on AI? Well, the persistence of the problem is yet another indication that there is something deeply flawed with AI methodology. Notice that we, humans, never face this problem in our own behavior. It was only discovered when logic-based AI attempted to build a system that was purportedly modeled on human intelligence. The frame problem is not a real problem. It is an aberration created and kept alive by the wrong model of human in-

[8] In fact, this is still an implicit assumption of most AI research today.

telligence. Change the model and the problem will go away. Ptolemaic cosmology puzzled for centuries over the mystery of Venus's phases — the problem disappeared with the advent of the heliocentric model. AI needs its own Copernican revolution.

4. CONCLUSIONS

The *frame problem* clearly means different things to different people. Besides the three philosophical perspectives presented here, one should mention its fame reaching psychology (*it is an instance of a symbol grounding problem* (Harnad 1993)), philosophy of science (*the frame problem as a diagnostic tool to distinguish commonsense reasoning from prediction using scientific theory* (Sprevak 2005)) or even ethics (*the frame problem shows that human moral normativity cannot be systematized by exceptionless general principles* (Horgan, Timmons 2009)). Interestingly, the discussion of the problem subsided within the AI community. It is hard to believe that everyone has accepted Shanahan's claim that the problem is "more-or-less solved" and moved on. If that were the case, where are the thinking machines that this problem has prevented us from building? A more likely explanation is that after so many failed attempts researchers lost faith that the problem can be solved by logic. Indeed, since 1980s there has been movement to break away from the traditional logic-based methodology of AI, most notably by Rodney Brooks at MIT. Brooks took to heart Dreyfus's arguments and attempted to build robots following a different methodological paradigm (Brooks 1991) that avoids running into the frame problem. The success of this project has been rather limited,[9] but perhaps the only way to overcome the frame problem is to avoid it rather than solve it.

REFERENCES

Bertino E., Catania B., Zarri G. P. (2001), *Intelligent Database Systems*, Addison-Wesley.

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

Dennett D. (1987), *Cognitive Wheels. The Frame Problem of AI* [in:] Pylyshyn 1987: 41–64.

Dennett D. (1994), *Consciousness in Human and Robot Minds* [in:] *Cognition, Computation and Consciousness*, M. Ito, Y. Miyashita, E. T. Rolls (eds.), Oxford: Oxford University Press.

Dreyfus H. L. (1992), *What Computers Still Can't Do*, Cambridge (MA): MIT Press.

Dreyfus H. L. (2007), *Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian*, „Artificial Intelligence" 171(18), 1137–1160.

Fetzer J. (1991), *The Frame Problem. Artificial Intelligence Meets David Hume* [in:] *Reasoning Agents in a Dynamic World. The Frame Problem*, K. M. Ford, P. J. Hayes (eds.), Greenwich (CT): JAI Press.

Ford K. M., Pylyshyn Z. W. (eds.) (1996), *The Robot's Dilemma Revisited. The Frame Problem in Artificial Intelligence*, Westport (CT): Ablex Publishing Corporation.

Garey M. R., Johnson D. S. (1979), *Computers and Intractability*, New York: W. H. Freeman.

[9] This time it was Dennett who made some overoptimistic predictions (Dennett 1994).

Gelernter H. (1959), *Realization of a Geometry-Theorem Proving Machine* [in:] *Proceedings of an International Conference on Information Processing*, Paris: UNESCO House, 273–282.
Goodman N. (1954), *Fact, Fiction, and Forecast*, Cambridge (MA): Harvard University Press.
Hanks S., McDermott D. (1987), *Nonmonotonic Logic and Temporal Projection*, „Artificial Intelligence" 33, 379–412.
Harnad S. (1993), *Problems, Problems. The Frame Problem as a Symptom of the Symbol Grounding Problem*, „Psycholoquy" 4(34).
Horgan T., Timmons M. (2009), *What Does the Frame Problem Tell Us About Moral Normativity?*, „Ethical Theory and Moral Practice" 12(1), 25–51.
Horty J. F. (2001), *Nonmonotonic Logic* [in:] *The Blackwell Guide to Philosophical Logic*, L. Goble (ed.), Oxford: Blackwell Publishers, 336–361.
Hayes P. (1987), *What the Frame Problem Is and Isn't* [in:] Pylyshyn 1987: 123–138.
Janlert L.-E. (1987), *Modelling Change — The Frame Problem* [in:] Pylyshyn 1987: 1–40.
Janlert L.-E. (1996), *The Frame Problem: Freedom or Stability? With Pictures We Can Have Both* [in:] Ford, Pylyshyn 1996: 35–48.
Kamermans M., Schmits T. (2004), *The History of the Frame Problem*, University of Amsterdam.
McCarthy J., Hayes P. (1969), *Some Philosophical Problems from the Standpoint of Artificial Intelligence* [in:] *Machine Intelligence*, B. Meltzer, D. Mitchie (eds.), Edinburgh: Edinburgh University Press, 463–502.
McCarthy J. (1980), *Circumscription — A Form of Non-Monotonic Reasoning*, „Artificial Intelligence" 13, 27–39.
McCorduck P. (2004), *Machines Who Think*, 2nd ed., Natick (MA): A. K. Peters.
McDermott D. (1987), *We've Been Framed. Or, Why AI Is Innocent of the Frame Problem* [in:] Pylyshyn 1987: 11–122.
Morgenstern L. (1996), *The Problem with Solutions to the Frame Problem* [in:] Ford, Pylyshyn 1996: 9–133.
Pylyshyn Z. W. (ed.) (1987), *The Robot's Dilemma. The Frame Problem in Artificial Intelligence*, Norwood (NJ): Ablex Publishing Corporation.
Reiter R., Criscuolo G. (1981), *On Interacting Defaults* [in:] *Proceedings of the International Joint Conference on Artificial Intelligence*, 27–276.
Reiter R. (1980), *A Logic for Default Reasoning*, „Artificial Intelligence" 13, 8–132.
Samuel A. L. (1959), *Some Studies in Machine Learning Using the Game of Checkers*, „IBM Journal of Research and Development" 3(3), 21–229.
Shanahan M. (1999), *The Frame Problem* [in:] *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/.
Shortliffe E. H. (1976), *Computer-Based Medical Consultations: MYCIN*, New York, Elsevier/North-Holland.
Simon H. A. (1965), *The Shape of Automation. For Man and Management*, New York: Harper & Row.
Simon H. A., Newell A. (1958), *Heuristic Problem Solving. The Next Advance in Operations Research*, „Operations Research" 6, 1–10.
Simon H. A., Newell A. (1976), *Computer Science as Empirical Inquiry. Symbols and Search*, „Communications of the ACM" 19(3), 11–126.
Sprevak M. (2005), *The Frame Problem and the Treatment of Prediction* [in:] *Computing, Philosophy and Cognition*, L. Magnani, R. Dossena (eds.), London: King's College Publications, 349–359.
Winograd T. A. (1972), *Understanding Natural Language*, New York: Academic Press.