FREE WILL—EVEN FOR ROBOTS

John McCarthy
Computer Science Department
Stanford University
Stanford, CA 94305, USA
jmc@cs.stanford.edu
http://www-formal.stanford.edu/jmc/

2000 Feb 14, 4:12 p.m.

I can, but I won't.[1]

Abstract

Human free will is a product of evolution and contributes to the success of the human animal. Useful robots will also require free will of a similar kind, and we will have to design it into them.

Free will is not an all-or-nothing thing. Some agents have more free will, or free will of different kinds, than others, and we will try to analyze this phenomenon. Our objectives are primarily technological, i.e. to study what aspects of free will can make robots more useful, and we will not try to console those who find determinism distressing. We distinguish between having choices and being conscious of these choices; both are important, even for robots, and consciousness of choices requires more structure in the agent than just having choices and is important for robots. Consciousness of free will is therefore not just an epiphenomenon of structure serving other purposes.

Free will does not require a very complex system. Young children and rather simple computer systems can represent internally *'I can, but I won't'* and behave accordingly.

Naturally I hope this detailed *design stance* (Dennett 1978) will help understand human free will. It takes the *compatibilist* philosophical position.

There may be some readers interested in what the paper says about human free will and who are put off by logical formulas. The formulas are not important for the arguments about human free will; they are present for people contemplating AI systems using mathematical logic. They can skip the formulas, but the coherence of what remains is not absolutely guaranteed.

1 Introduction—two aspects of free will

Free will, both in humans and in computer programs has two aspects—the *external* aspect and the *introspective* aspect.

The external aspect is the set of results that an agent P can achieve, i.e. what it *can* do in a situation s,

$$Poss(P, s) = \{x | Can(P, x, s)\}. \tag{1}$$

Thus in the present situation, I can find my drink. In one sense I can climb on the roof of my house and jump off. In another sense I can't. (The different senses of *can* will be discussed in Section 3.1). In a certain position, a chess program can checkmate its opponent and can also move into a position leading to the opponent giving checkmate. What is x in $Can(P, x, s)$? In English it usually has the grammatical form of an action, but in the interesting cases it is not an elementary action like those treated in situation calculus. Thus we have 'I can go to Australia', 'I can make a million dollars', 'I can get a new house'. Often the what is to be achieved is a fluent, e.g. the state of having a new house.

In the most important case, $Poss(P, s)$ depends only on the causal position of P in the world and not on the internal structure of P.

The introspective aspect involves the agent P's knowledge of $Poss(P, s)$, i.e. its knowledge of what it can achieve. Here is where the human sense of free will comes in. It depends on P having an internal structure that allows certain aspects of its current state to be interpreted as expressing knowledge. I know I can find my drink. In a simple chess position I would know I could give checkmate in three, because the chess problem column in the newspaper said so, although I mightn't yet have been able to figure out how.

Some present computer programs, e.g. chess programs, have an extensive $Poss(P, s)$. However, their knowledge of $Poss(P, s)$ as a set is very limited. Indeed it is too limited for optimal functionality, and robots' knowledge of their possibilities need to be made more like that of humans. For example, a robot may conclude that in the present situation it has too limited a set of possibilities. It may then undertake to ensure that in future similar situations it will have more choices.

1.1 Preliminary philosophical remarks

Consider a machine, e.g. a computer program, that is entirely deterministic, i.e. is completely specified and contains no random element. A major question for philosophers is whether a human is deterministic in the above sense. If the answer is yes, then we must either regard the human as having no free will or regard free will as compatible with determinism. Some philosophers, called *compatibilists*, e.g. Daniel Dennett (Dennett 1984), take this view, and regard a person to have free will if his actions are determined by his internal decision processes even if these processes themselves are deterministic.[2] My view is compatibilist, but I don't need to take a position on determinism itself.

AI depends on a compatibilist view, but having taken it, there is a lot to be learned about the specific forms of free will that can be designed. That is the subject of this article.

I don't discuss the aspects of free will related to assigning credit or blame for actions according to whether they were done freely. More generally, the considerations of this article are orthogonal to many studied by philosophers, but I think they apply to human free will nevertheless.

Specifically, East Germany did not deny its citizens the kind of free will that some hope to establish via quantum mechanics or chaos theory. It did deny its citizens choices in the sense discussed in this article.

Logical AI has some further philosophical presuppositions. These are discussed in (McCarthy 1999b).

2 Informal discussion

There are different kinds and levels of free will. An automobile has none, a chess program has a minimal kind of free will, and a human has a lot. Human-level AI systems, i.e. those that match or exceed human intelligence will need a lot more than present chess programs, and most likely will need almost as much as a human possesses, even to be useful servants.

Consider chess programs. What kinds of free will do they have and can they have? A usual chess program, given a position, generates a list of moves available in the position. It then goes down the list and tries the moves successively getting a score for each move. It chooses the move with the highest score (or perhaps the first move considered good enough to achieve a certain objective.)

That the program considers alternatives is our reason for ascribing to it a little free will, whereas we ascribe none to the automobile. How is the chess program's free will limited, and what more could we ask? Could further free will help make it a more effective program?

A human doesn't usually consider his choices sequentially, scoring each and comparing only the scores. The human compares the consequences of the different choices in detail. Would it help a chess program to do that? Human chess players do it.

Beyond that is considering the set $Legals(p)$ of legal moves in position p as an object. A human considers his set of choices and doesn't just consider each choice individually. A chess position is called 'cramped' if there are few non-disastrous moves, and it is considered useful to cramp the opponent's position even if one hasn't other reasons for considering the position bad for the opponent. Very likely, a program that could play as well as Deep Blue but doing 10^{-6} as much computation would need a more elaborate choice structure, i.e. more free will. For example, one fluent of chess positions, e.g. having an open file for a rook, can be regarded as giving a better position than another without assigning numerical values to positions.

3 The finite automaton model of free will and *can*

This section treats $Poss(P, s)$ for finite automata. Finite automata raise the question of what an agent *can* do in a sharp form. However, they are not a useful representation of an agent's introspective knowledge of what it can do.

To the extent that a person or machine *can* achieve any of different goals, that person or machine has free will. Our ideas on this show up most sharply considering systems of interacting discrete finite automata. These are as deterministic as you can get, which is why I chose them to illustrate free will.

The material of this section revises that in (McCarthy and Hayes 1969), section 2.4 entitled 'The automaton representation and the notion of *can*'.

Let S be a system of interacting discrete finite automata such as that shown in figure 1.

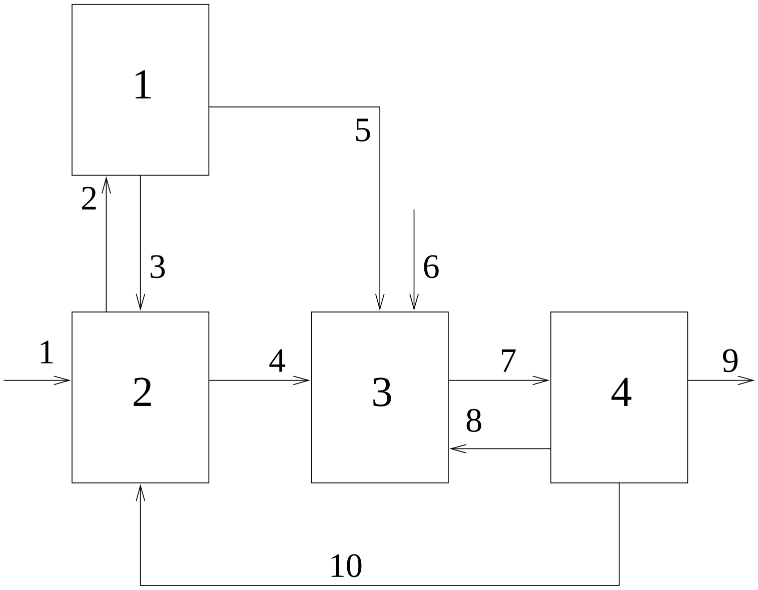

Figure 1: System S.

Each box represents a subautomaton and each line represents a signal. Time takes on integer values and the dynamic behavior of the whole automaton is given by the equations:

$$\begin{aligned}
a_1(t+1) &= A_1(a_1(t),\, s_2(t)) \\
a_2(t+1) &= A_2(a_2(t),\, s_1(t),\, s_3(t),\, s_{10}(t)) \\
a_3(t+1) &= A_3(a_3(t),\, s_4(t),\, s_5(t),\, s_6(t),\, s_8(t)) \\
a_4(t+1) &= A_4(a_4(t),\, s_7(t))
\end{aligned} \tag{2}$$

$$\begin{aligned}
s_2(t) &= S_2(a_2(t)) \\
s_3(t) &= S_3(a_1(t)) \\
s_4(t) &= S_4(a_2(t))
\end{aligned}$$

$$
\begin{aligned}
s_5(t) &= S_5(a_1(t)) \\
s_7(t) &= S_7(a_3(t)) \\
s_8(t) &= S_8(a_4(t)) \\
s_9(t) &= S_9(a_4(t)) \\
s_{10}(t) &= S_{10}(a_4(t))
\end{aligned}
\tag{3}
$$

The interpretation of these equations is that the state of any subautomaton at time $t+1$ is determined by its state at time t and by the signals received at time t. The value of a particular signal at time t is determined by the state at time t of the automaton from which it comes. Signals without a source subautomaton represent inputs from the outside and signals without a destination represent outputs.

Finite automata are the simplest examples of systems that interact over time. They are completely deterministic; if we know the initial states of all the automata and if we know the inputs as a function of time, the behavior of the system is completely determined by equations (2) and (3) for all future time.

The automaton representation consists in regarding the world as a system of interacting subautomata. For example, we might regard each person in the room as a subautomaton and the environment as consisting of one or more additional subautomata. As we shall see, this representation has many of the qualitative properties of interactions among things and persons. However, if we take the representation too seriously and attempt to represent particular interesting systems as systems of interacting automata, we encounter the following difficulties:

1. The number of states required in the subautomata is very large, for example $2^{10^{10}}$, if we try to represent a person's knowledge. Automata this large have to be represented by systems of equations or by computer programs, or in some other way that does not involve mentioning states individually. In Section 4 we'll represent them partially, by sentences of logic.

2. Geometric information is hard to represent. Consider, for example, the location of a multi-jointed object such as a person or a matter of even more difficulty—the shape of a lump of clay.

3. The system of fixed interconnections is inadequate. Since a person may handle any object in the room, an adequate automaton representation would require signal lines connecting him with every object.

4. The most serious objection, however, is that (in the terminology of (McCarthy and Hayes 1969)) the automaton representation is epistemologi-

cally inadequate. Namely, we do not ever know a person well enough to list his internal states. The kind of information we do have about him needs to be expressed in some other way.

Nevertheless, we may use the automaton representation for concepts of *can, causes,* useful kinds of counterfactual statements ('If another car had come over the hill when you passed just now, there would have been a head-on collision'). See (Costello and McCarthy 1999).

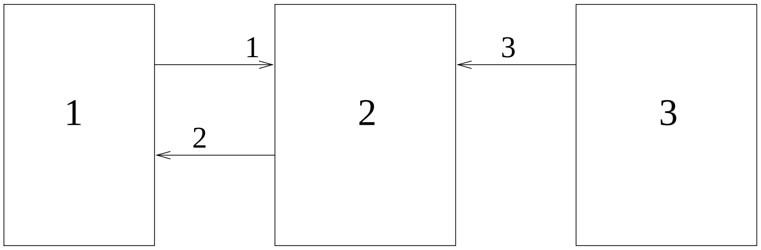

Figure 2: Another system S.

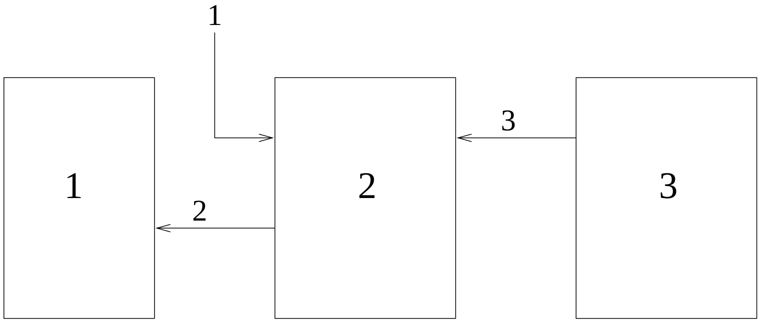

Figure 3: System S_1.

Let us consider the notion of *can*. Let S be a system of subautomata without external inputs such as that of figure 2. Let p be one of the subautomata, and suppose that there are m signal lines coming out of p. What p can do is defined in terms of a new system S_p, which is obtained from the system S by disconnecting the m signal lines coming from p and replacing them by m external input lines to the system. In figure 2, subautomaton 1 has one output, and in the system S_1 (figure 3) this is replaced by an external input. The new system S_p always has the same set of states as the system S. Now let π be a condition on the state such as, 'a_2 is even' or '$a_2 = a_3$'. (In the applications π may be a condition like 'The box is under the bananas'.)

We shall write

$$can(p, \pi, s)$$

which is read, 'The subautomaton p *can* bring about the condition π in the situation s' if there is a sequence of outputs from the automaton S_p that will eventually put S into a state a' that satisfies $\pi(a')$. In other words, in determining what p can achieve, we consider the effects of sequences of its actions, quite apart from the conditions that determine what it actually will do.

Here's an example based on figure 2. In order to write formulas conveniently, we use natural numberss for the values of the states of the subautomata and the signals.

$$\begin{array}{rcl}
a_1(t+1) & = & a_1(t) + s_2(t) \\
a_2(t+1) & = & a_2(t) + s_1(t) + 2s_3(t) \\
a_3(t+1) & = & \textbf{if } a_3(t) = 0 \textbf{ then } 0 \textbf{ else } a_3(t) + 1 \\
s_1(t) & = & \textbf{if } a_1(t) = 0 \textbf{ then } 2 \textbf{ else } 1 \\
s_2(t) & = & 1 \\
s_3(t) & = & \textbf{if } a_3(t) = 0 \textbf{ then } 0 \textbf{ else } 1.
\end{array} \tag{4}$$

Consider the initial state of S to be one in which all the subautomata are in state 0. We have the following propositions:

1. Subautomaton 2 *will* never be in state 1. [It starts in state 0 and goes to state 2 at time 1. After that it can never decrease.]

2. Subautomaton 1 *can* put Subautomaton 2 in state 1 but won't. [If Subautomaton 1 emitted 1 at time 0 instead of 2, Subautomaton 2 would go to state 1.]

3. Subautomaton 3 *cannot* put Subautomaton 2 in state 1. [The output from Subautomaton 1 suffices to put Subautomaton 2 in state 1 at time 1, after which it can never decrease.]

We claim that this notion of *can* is, to a first approximation, the appropriate one for a robot to use internally in deciding what to do by reasoning. We also claim that it corresponds in many cases to the common sense notion of *can* used in everyday speech.

In the first place, suppose we have a computer program that decides what to do by reasoning. Then its output is determined by the decisions it makes in the reasoning process. It does not know (has not computed) in advance what it will do, and, therefore, it is appropriate that it considers that it can do anything that can be achieved by some sequence of its outputs. Commonsense reasoning seems to operate in the same way.

The above rather simple notion of *can* requires some elaboration, both to represent adequately the commonsense notion and for practical purposes in the reasoning program.

First, suppose that the system of automata admits external inputs. There are two ways of defining *can* in this case. One way is to assert $can(p, \pi, s)$ if p can achieve π regardless of what signals appear on the external inputs. Thus, we require the existence of a sequence of outputs of p that achieves the goal regardless of the sequence of external inputs to the system. Note that, in this definition of *can*, we are not requiring that p have any way of knowing what the external inputs were. An alternative definition requires the outputs to depend on the inputs of p. This is equivalent to saying that p can achieve a goal, provided the goal would be achieved for arbitrary inputs by some automaton put in place of p. With either of these definitions *can* becomes a function of the place of the subautomaton in the system rather than of the subautomaton itself. Both of these treatments are likely to be useful, and so we shall call the first concept *cana* and the second *canb*.

3.1 Representing a person by a system of subautomata

The idea that what a person can do depends on his position rather than on his characteristics is somewhat counter-intuitive. This impression can be mitigated as follows: Imagine the person to be made up of several subautomata; the output of the outer subautomaton is the motion of the joints. If we break the connection to the world at that point we can answer questions like, 'Can he fit through a given hole?' We shall get some counter-intuitive answers, however, such as that he can run at top speed for an hour or can jump over a building, since these are sequences of motions of his joints that would achieve these results.

The next step, however, is to consider a subautomaton that receives the nerve impulses from the spinal cord and transmits them to the muscles. If we break at the input to this automaton, we shall no longer say that he can jump over a building or run long at top speed since the limitations of the muscles will be taken into account. We shall, however, say that he can ride a unicycle since appropriate nerve signals would achieve this result.

The notion of *can* corresponding to the intuitive notion in the largest number of cases might be obtained by hypothesizing an *organ of will,* which makes decisions to do things and transmits these decisions to the main part of the brain that tries to carry them out and contains all the knowledge of particular facts.[3] If we make the break at this point we shall be able to say that so-and-so cannot dial the President's secret and private telephone number because he does not know it, even though if the question were asked could he dial that particular number, the answer would be yes. However, even this break would not give the statement, 'I cannot go without saying goodbye, because this would hurt the child's feelings'.

On the basis of these examples, one might try to postulate a sequence of narrower and narrower notions of *can* terminating in a notion according to which a person can do only what he actually does. This extreme notion would then be superfluous. Actually, one should not look for a single best notion of *can*; each of the above-mentioned notions is useful and is actually used in some circumstances. Sometimes, more than one notion is used in a single sentence, when two different levels of constraint are mentioned.

Nondeterministic systems as approximations to deterministic systems are discussed in (McCarthy 1999a). For now we'll settle for an example involving a chess program. It can be reasoned about at various levels. Superhuman Martians can compute what it will do by looking at the initial electronic state and following the electronics. Someone with less computational power can interpret the program on another computer knowing the program and the position and determine the move that will be made. A mere human chess player may be reduced to saying that certain moves are excluded as obviously disastrous but be unable to decide which of (say) two moves the program will make. The chess player's model is a nondeterministic approximation to the program.

3.2 Causality

Besides its use in explicating the notion of *can*, the automaton representation of the world is very suited for illustrating notions of causality. For, we may say that subautomaton p caused the condition π in state s, if changing the output of p would prevent π. In fact the whole idea of a system of interacting automata is mainly a formalization of the commonsense notion of causality.

The automaton representation can be used to explicate certain counterfactual conditional sentences. For example, we have the sentence, 'If another car had come over the hill when you just passed, there would have been a head-on collision'. We can imagine an automaton representation in which whether a car came over the hill is one of the outputs of a traffic subautomaton. (Costello and McCarthy 1999) discusses useful counterfactuals, like the above that are imbedded in a description of a situation and have consequences. One use is that they permit learning from an experience you didn't quite have and would rather not have.

3.3 Good analyses into subautomata

In the foregoing we have taken the representation of the situation as a system of interacting subautomata for granted. Indeed if you want to take them for granted you can skip this section.

However, a given overall automaton system might be represented as a system of interacting subautomata in a number of ways, and different representations might yield different results about what a given subautomaton can achieve, what would have happened if some subautomaton had acted differently, or what caused what. Indeed, in a different representation, the same or corresponding subautomata might not be identifiable. Therefore, these notions depend on the representation chosen.

For example, suppose a pair of Martians observe the situation in a room. One Martian analyzes it as a collection of interacting people as we do, but the second Martian groups all the heads together into one subautomaton and all the bodies into another.[4] How is the first Martian to convince the second that his representation is to be preferred? Roughly speaking, he would argue that the interaction between the heads and bodies of the same person is closer than the interaction between the different heads, and so more of an analysis has been achieved from 'the primordial muddle' with the conventional representation. He will be especially convincing when he points out that when the meeting is over the heads will stop interacting with each other, but will continue to interact with their respective bodies.

We can express this kind of argument formally in terms of automata as follows: Suppose we have an autonomous automaton A, i.e. an automaton without inputs, and let it have k states. Further, let m and n be two integers such that $mn \geq k$. Now label k points of an m-by-n array with the states of A. This can be done in $\binom{mn}{k}!$ ways. For each of these ways we have a representation of the automaton A as a system of an m-state automaton B interacting with an n-state automaton C. Namely, corresponding to each row of the array we have a state of B and to each column a state of C. The signals are in 1–1 correspondence with the states themselves; thus each subautomaton has just as many values of its output as it has states.

Now it may happen that two of these signals are equivalent in their effect on the other subautomaton, and we use this equivalence relation to form equivalence classes of signals. We may then regard the equivalence classes as the signals themselves. Suppose then that there are now r signals from B to C and s signals from C to B. We ask how small r and s can be taken in general compared to m and n. The answer may be obtained by counting the number of inequivalent automata with k states and comparing it with the number of systems of two automata with m and n states respectively and r and s signals going in the respective directions. The result is not worth working out in detail, but tells us that only a few of the k state automata admit such a decomposition with r and s small compared to m and n. Therefore, if an automaton happens to admit such a decomposition it is very unusual for it to admit a second such decomposition that is not equivalent to the first with respect to some renaming of states. Applying this argument to the real world, we may say that it is overwhelmingly probable that our customary decomposition of the world automaton into separate people and things has a unique, objective and usually preferred status. Therefore, the notions of *can*, of causality, and of counterfactual associated with this decomposition also have a preferred status.

These considerations are similar to those used by Shannon, (Shannon 1938) to find lower bounds on the number of relay contacts required on the average to realize a boolean function.

An automaton can do various things. However, the automaton model proposed so far does not involve consciousness of the choices available. This requires that the automata be given a mental structure in which facts are represented by sentences. This is better done in a more sophisticated model than finite automata. We start on it in the next section.

4 Formalism for introspective free will

The previous section concerned only external free will, and it isn't convenient to represent knowledge by the states of subautomata of a reasoning automaton. (McCarthy 1979) has a more extensive formalization of knowing what and knowing that.

The situation calculus, (McCarthy and Hayes 1969) and (Shanahan 1997), offers a better formalism for a robot to represent facts about its own possibilities.

4.1 A minimal example of introspective free will

The following statement by Suppes (Suppes 1994) provides a good excuse for beginning with a very simple example of introspective free will.

> There are, it seems to me, two central principles that should govern our account of free will. The first is that small causes can produce large effects. The second is that random phenomena are maximally complex, and *it is complexity that is phenomenologically in many human actions that are not constrained but satisfy ordinary human notions of being free actions.*[my emphasis]

I don't agree that complexity is essentially involved so here's a minimal example that expresses, 'I can, but I won't'.

Because the agent is reasoning about its own actions, as is common in situation calculus formalization, the agent is not explicitly represented. Making the agent explicit offers no difficulties.

If an action a is possible in a situation s, then the situation $Result(a, s)$ that results from performing the action is achievable.

$$Possible(a, s) \rightarrow Can(Result(a, s), s) \tag{5}$$

If a situation $Result(a, s)$ is achievable and every other situation that is achievable is less good, then the action a should be done.

$$\begin{aligned} &Can(Result(a, s), s) \wedge (\forall s')(Can(s', s) \rightarrow s' <_{good} Result(a, s)) \\ &\qquad \rightarrow Should(a, s) \end{aligned} \tag{6}$$

Here $<_{good}$ means 'not so good as'.

Actions leading to situations inferior to what can be achieved won't be done.

$$\begin{aligned} &Can(Result(a, s), s) \wedge Result(a', s) <_{good} Result(a, s) \\ &\qquad \rightarrow \neg WillDo(a', s) \end{aligned} \tag{7}$$

This is reasonably close to formalizing 'It can, but it won't' except for not taking into account the distinction between 'but' and 'and'. As truth functions, 'but' and 'and' are equivalent. Uttering 'p but q' is a different *speech act* from uttering 'p and q', but this article is not the place to discuss the difference.

4.2 Representing more about an agent's capability

Here are some examples of introspective free will and some considerations. They need to be represented in logic so that a robot could use them to learn from its past and plan its future.

1. Did I make the wrong decision just now? Can I reverse it?

2. 'Yesterday I could have made my reservation and got a cheap fare.'

3. 'Next year I can apply to any university in the country. I don't need to make up my mind now.'.

4. 'If I haven't studied calculus, I will be unable to take differential equations.'

5. 'If I learn to program computers, I will have more choice of occupation.'

6. 'It is better to have an increased set of choices.'

7. 'I am not allowed to harm human beings.' Asimov imagined his three laws of robotics, of which this is one, as built into his imaginary positronic brains. In his numerous science fiction stories, the robots treated them as though engraved on tablets and requiring interpretation. This is necessary, because the robots did have to imagine their choices and their consequences.

8. Some of a person's behavior is controlled by reflexes and other automatic mechanisms. We rightly regard reflexive actions as not being deliberate and are always trying to get better control of them.

    - The coach helps the baseball player analyze how he swings at the ball and helps him improve the reflexive actions involved.
    - I'm a sucker for knight forks and for redheads and need to think more in such chess and social situations.

9. In the introduction I wrote

    $$Poss(P, s) = \{x | can(P, x, s)\}. \tag{8}$$

    What kind of an entity is x? In situation calculus, the simplest operand of *can* is a situation as treated above, but also we can consider an action itself or a propositional fluent. A propositional fluent p is a predicate taking a situation argument, and an agent can reason that it can (or cannot) bring about a future situation in which p holds.

(McCarthy 1996) includes an extensive discussion of what consciousness, including consciousness of self, will be required for robots.

5 Conclusions

1. Human level AI requires the ability of the agent to reason about its past, present, future and hypothetical choices.

2. What an agent can do is determined by its environment rather than by its internal structure.

3. Having choices is usefully distinguished from the higher capability of knowing about them.

4. What people can do and know about what they can do is similar to what robots can do and know.

AI needs a more developed formal theory of free will, i.e. the structures of choice a robot can have and what it can usefully know about them.

6 Acknowledgments

I thank Eyal Amir, Daniel Dennett, and Patrick Suppes for useful discussions.

This work was supported by the DARPA High Performance Knowledge Bases program and by the AFOSR New World Vistas program under AFOSR grant AF F49620-97-1-0207.

References

Costello, T., and J. McCarthy. 1999. Useful Counterfactuals[5]. *Electronic Transactions on Artificial Intelligence.* submitted 1999 July.

Dennett, D. 1978. *Brainstorms: Philosophical Essays on Mind and Psychology.* Cambridge: Bradford Books/MIT Press.

Dennett, D. 1984. *Elbow room : the varieties of free will worth wanting.* MIT Press.

McCarthy, J. 1979. First order theories of individual concepts and propositions. In D. Michie (Ed.), *Machine Intelligence*, Vol. 9. Edinburgh: Edinburgh University Press. Reprinted in (McCarthy 1990).

McCarthy, J. 1990. *Formalizing Common Sense: Papers by John McCarthy.* Ablex Publishing Corporation.

McCarthy, J. 1996. Making Robots Conscious of their Mental States[6]. In S. Muggleton (Ed.), *Machine Intelligence 15.* Oxford University Press. to appear in 2000. The web version is improved from that presented at Machine Intelligence 15 in 1995.

McCarthy, J. 1999a. Logical theories with approximate concepts—draft[7]. *submitted but web only for now.*

McCarthy, J. 1999b. Philosophical and Scientific Presuppositions of Logical Ai[8]. In H. J. Levesque and F. Pirri (Eds.), *Logical Foundations for Cognitive Agents: Contributions in Honor of Ray Reiter*, 72–78. Springer-Verlag.

McCarthy, J., and P. J. Hayes. 1969. Some Philosophical Problems from the Standpoint of Artificial Intelligence[9]. In B. Meltzer and D. Michie (Eds.), *Machine Intelligence 4*, 463–502. Edinburgh University Press.

Shanahan, M. 1997. *Solving the Frame Problem, a mathematical investigation of the common sense law of inertia.* M.I.T. Press.

Shannon, C. E. 1938. A symbolic analysis of relay and switching circuits. *Transactions American Institute of Electrical Engineers* 57:713–723. I think this is the article in which Shannon showed that most boolean functions require many relay contacts.

Suppes, P. 1994. Voluntary motion, biological computation and free will. *Midwest Studies in Philosophy* XIX:452–467.

7 Notes

[1] Sarah McCarthy, at age 4, personal communication.

[2] Some people ask whether making the system probabilistic or quantum mechanical or classical chaotic makes a difference in the matter of free will. I agree with those who say it doesn't.

[3] The idea of an organ of will cannot be given a precise definition, which has caused philosophers and psychologists to denounce as senseless ideas that separate will from intellect. However, it may be a useful *approximate concept* in the sense of (McCarthy 1999a). It presumably won't correspond to a specific part of the brain.

[4] An inhabitant of momentum space might regard the Fourier components of the distribution of matter as the separate interacting subautomata.

[5] http://www-formal.stanford.edu/jmc/counterfactuals.html

[6] http://www-formal.stanford.edu/jmc/consciousness.html

[7] http://www-formal.stanford.edu/jmc/approximate.html

[8] http://www-formal.stanford.edu/jmc/phil2.html

[9] http://www-formal.stanford.edu/jmc/mcchay69.html