SIMPLE DETERMINISTIC FREE WILL

John McCarthy

from May 16, 2002 until November 6, 2005

Abstract

A common feature of free will is that a person has choices among alternative actions and chooses the action with the apparently most preferred consequences. In a determinist theory, the mechanism that makes the choice among the alternatives is determinist. The sensation of free will comes from the fact that the mechanism that generates the choices uses a non-determinist theory as a computational device and that the stage in which the choices have been identified is introspectable. The present formalism is based on work in artificial intelligence (AI).

We present a theory of *simple deterministic free will* (**SDFW**) in a deterministic world. The theory splits the mechanism that determines action into two parts. The first part computes possible actions and their consequences. Then the second part decides which action is most preferable and does it.

We formalize SDFW by two sentences in *situation calculus*, a mathematical logical theory often used in AI. The situation calculus formalization makes the notion of free will technical. According to this notion, almost no animal behavior exhibits free will, because exercising free will involves considering the consequences of alternative actions. A major advantage of our notion of free will is that whether an animal does have free will may be determinable by experiment. Some computer programs, e.g. chess programs, exhibit SDFW. Almost all do not. At least SDFW seems to be required for effective chess performance and also for human-level AI.

Many features usually considered as properties of free will are omitted in SDFW. That's what makes it simple. The criterion for whether an entity uses SDFW is not behavioristic but is expressed in terms of the internal structure of the entity.

1 The Informal theory

Let the course of events, including events in my brain (or yours or his or its) be deterministic. It seems to many people that there is no place for free will. Even our thoughts are determined.

However, if we examine closely how a human brain (or chess program) deterministically makes decisions, free will (or imitation free will if your philosophy forbids calling it real free will) must come back in. Some deterministic processes consider alternative actions and their consequences and choose the actions they think have the most preferred consequences. This deterministic decision process uses a nondeterministic theory to present the set of available actions and the consequences of each of them.

When a person, animal, or machine reacts directly to a situation rather than comparing the consequences of alternative actions, free will is not involved. So far as I can see, no animals consider the consequences of alternative actions; hence they don't have free will. Others think that apes sometimes do compare consequences. A relevant experiment is suggested in section 7. Using free will is too slow in many situations, and training and practice often have the purpose of replacing comparison of consequences by automatic reaction to a situation.

We believe this simple theory covers the most basic phenomenon of human free will. We'll call it *simple deterministic free will* and abbreviate it SDFW. Robots with human-level intelligence will also require at least this much free will in order to be useful.

Beyond having free will, some systems are conscious of having free will and communicate about it. If asked to tell what it is doing, humans or some machine will tell about their choices for action and say that they intend to determine which action leads to the best consequence. Such a report, whether given externally or contemplated internally, constitutes the human sensation and the human report of free will. SDFW does not require consciousness of having free will or the ability to communicate about it. That's what's simple about SDFW. Thinking about one's free will requires theoretical structure beyond or above SDFW. So will considering actions as praiseworthy or blameworthy. SDFW also doesn't treat game theoretic situations in which probabilistic mixed strategies are appropriate.

In AI research one must treat simple cases of phenomena, e.g. intentional behavior, because full generality is beyond the state of the art. Many philosophers are inclined to only consider the general phenomenon, but this limits what can be accomplished. I recommend to them the AI approach of doing the simplest cases first.

2 Situation calculus formulas for SDFW

Artificial intelligence requires expressing this phenomenon formally, and we'll do it here in the mathematical logical language of situation calculus. Situation calculus is described in [MH69], [Sha97], [Rei01], and in the extended form used here, in [McC02]. Richmond Thomason in [Tho03] compares situation calculus to theories of action in the philosophical literature. As usually presented, situation calculus is a non-deterministic theory. The equation

$$s' = Result(e, s)$$

asserts that s' is the situation that results when event e occurs in the situation s. Since there may be many different events that can occur in s, and the theory of the function *Result* does not say which occurs, the theory is non-deterministic. Some AI jargon refers to it as a theory with branching time rather than linear time. Actions are a special case of events, but most AI work discusses only actions.

Usually, there are some preconditions for the event to occur, and then we have the formula

$$Precond(e, s) \rightarrow s' = Result(e, s).$$

[McC02] proposes adding a formula $Occurs(e, s)$ to the language that can be used to assert that the event e occurs in situation s. We have

$$Occurs(e, s) \rightarrow (Next(s) = Result(e, s)).$$

Adding occurrence axioms, which assert that certain actions occur, makes a theory more deterministic by specifying that certain events occur in situations satisfying specified conditions. In general the theory will remain partly non-deterministic, but if there are occurrence axioms specifying what events occur in all situations, then the theory becomes deterministic, i.e. has linear time.

We can now give a situation calculus theory for SDFW illustrating the role of a non-deterministic theory in determining what will deterministically happen, i.e. by saying what choice a person or machine will make.

In these formulas, lower case terms denote variables and capitalized terms denote constants. Suppose that *actor* has a choice of just two actions $a1$ and $a2$ that he may perform in situation s. We want to say that the event $Does(\text{actor}, a1)$ or $Does(\text{actor}, a2)$ occurs in s according to which of $Result(Does(\text{actor}, a1), s)$ or $Result(Does(\text{actor}, a2), s)$ *actor* prefers.

The formulas that assert that a person (actor) will do the action that he, she or it thinks results in the better situation for him are

$$Occurs(Does(actor, Choose(actor, a1, a2, s), s), s), \tag{1}$$

(1)
and

$$\begin{aligned} &Choose(actor, a1, a2, s) = \\ &\textbf{if } Prefers(actor, Result(a1, s), Result(a2, s)) \\ &\textbf{then } a1 \textbf{ else } a2. \end{aligned} \tag{2}$$

Adding (2) makes the theory determinist by specifying which choice us made.[1]

Here $\text{Prefers}(actor, s1, s2)$ is to be understood as asserting that *actor* prefers $s1$ to $s2$.

Here's a non-deterministic theory of greedy John.

$$\begin{aligned} &Result(A1, S0) = S1, \\ &Result(A2, S0) = S2, \\ &Wealth(John, S1) = \$2.0 \times 10^6, \\ &Wealth(John, S2) = \$1.0 \times 10^6, \\ &(\forall s\ s')(Wealth(John, s) > Wealth(John, s') \\ &\qquad \to \text{Prefers}(John, s, s'). \end{aligned} \tag{3}$$

As we see, greedy John has a choice of at least two actions in situation $S0$ and prefers a situation in which he has greater wealth to one in which he has lesser wealth.

From equations 1-3 we can infer

$$Occurs(Does(John, A1, S0)). \tag{4}$$

[1] (2) uses a conditional expression. **if** p **then** a **else** b has the value a if the proposition p is true and otherwise has the value b. The theory of conditional expressions is discussed in [McC63]. Conditional expressions are used in the Lisp, Algol 60, Algol 68, and Scheme programming languages.

For simplicity, we have omitted the axioms asserting that $A1$ and $A2$ are exactly the actions available and the nonmonotonic reasoning used to derive the conclusion.

Here Prefers($actor, s1, s2$) is to be understood as asserting that *actor* prefers $s1$ to $s2$. I used just two actions to keep the formula for *Choose* short. Having more actions or even making *Result* probabilistic or quantum would not change the nature of SDFW. A substantial theory of Prefers is beyond the scope of this article.

This illustrates the role of the non-deterministic theory of *Result* within a deterministic theory of what occurs. (1) includes the non-deterministic of *Result* used to compute which action leads to the better situation. (2) is the deterministic part that tells which action occurs.

We make four claims.

1. Effective AI systems, e.g. robots, will require identifying and reasoning about their choices once they get beyond what can be achieved with situation-action rules. Chess programs always have.

2. The above theory captures the most basic feature of human free will.

3. $Result(a1, s)$ and $Result(a2, s)$, as they are computed by the agent, are not full states of the world but elements of some theoretical space of approximate situations the agent uses in making its decisions. [McC00] has a discussion of approximate entities. Part of the problem of building human-level AI lies in inventing what kind of entity $Result(a, s)$ shall be taken to be.

4. Whether a human or an animal uses simple free will in a type of situation is subject to experimental investigation—as discussed in section 7.

Formulas (1) and (2) illustrate *person* making a choice. They don't say anything about *person* knowing it has choices or preferring situations in which it has more choices. SDFW is therefore a partial theory that requires extension when we need to account for these phenomena.

3 A generalization of SDFW

We can generalize SDFW by applying preferences to actions rather than to the situations resulting from actions. The formulas then become

$$Occurs(Does(actor, Choose\text{-}action(actor, a1, a2, s), s), s) \qquad (5)$$

(1) and

$$\begin{array}{ll}\mathit{Choose\text{-}action}(actor, a1, a2, s) = & \\ \textbf{if } \mathit{Prefers\text{-}action}(actor, a1, a2, s) & \quad (6)\\ \textbf{then } a1 \textbf{ else } a2. & \end{array}$$

(5) and (6) obviously generalize (1) and (2), because the earlier case is obtained by writing

$$\mathit{Prefers\text{-}action}(a1, a2, s) \equiv \mathit{Prefers}(\mathit{Result}(a, s), \mathit{Result}(a2, s)). \qquad (7)$$

I am doubtful about the generalization, because I don't see how to represent commonsense preferences between actions except in terms of preferring one resulting situation to another.

4 Knowledge of one's free will and wanting more or fewer choices

This section is less worked out than basic SDFW and not axiomatized. That's why it was best to start simple.

Here are some examples of it being good to have more choices.

"I'll take my car to work today rather than bicycling so I can shop on the way home if I want to."

"If you learn mathematics, you will more choice of scientific occupations".

"The more money I have, the more models of car I can choose from."

"If I escape from Cuba, I will have more choice of what to read, what I can say or write, and where to travel."

We want to say that situation $s1$ is at least as free as situation $s2$, written $s1 \geq_{freedom} s2$, if every fluent achievable by a single action from $s2$ is achievable from $s1$. Just as with equation (1), we can say that *person* chooses an action that leads to more freedom at the next situation.

$$\begin{array}{ll} s1 \geq_{freedom} s2 & \\ \equiv & \\ (\forall f)((\exists a)(\mathit{Holds}(f, \mathit{Result}(\mathit{Does}(\text{person}, a), s2))) & \quad (8)\\ \qquad\qquad \rightarrow & \\ (\exists a)(\mathit{Holds}(f, \mathit{Result}(\mathit{Does}(\text{person}, a), s1)))). & \end{array}$$

Here f ranges over fluents. Having more choices is usually preferred. However, one sometimes wants fewer choices. Burning one's bridges, nailing the flag to the mast, and promising to love until death do us part are examples of actions that reduce choices. The conditions under which this occurs are too difficult for me to formalize at present. They can involve fearing that one's preferences in the future might be different from one's present preferences for future actions or that making a commitment about one's future actions confers a present benefit.

5 Philosophical issues

The formalism of this paper takes sides in several philosophical controversies.

1. It considers determinism and free will, as experienced and observed by humans, as compatible. This is in accordance with the views of Locke and Hume.

2. It takes a third person point of view, i.e. considers the free will of others and not just the free will of the observer.

3. It breaks the phenomenon of free will into parts and considers the simplest systems first—in contrast to approaches that demand that all complications be understood before anything can be said. In this it resembles the approaches to belief and other intentional states discussed in [Den71], [Den78], and [McC79]. Starting with simple systems is the practice in AI, because only what is understood can be implemented in computer programs.

It seems to me that formulas (1) and (2) expressing the use of the branching time $Result(e, s)$ function in determining what events occur make the philosophical ideas definite. Thus we can see which modifications of the notions are compatible with (1) and (2), and which require different axioms.

The process of deciding what to do often involves considering a pruned set of actions which eliminate those that have obviously bad consequences. The remaining actions are those that one *can* do. When someone refers to a pruned action, one sometimes gets the reply, "Oh, I could do that, but I really can't, because ...."

6 Praise and blame

We have maintained that the basic notion of free will is the same for humans, animals and robots. Praising or blaming agents for their actions is an advanced notion requiring more structure, e.g. including good or bad actions or outcomes. Blaming or praising humans requires taking human peculiarities, not shared with agents in general, e.g. robots, into account.

Consider the verdict: *"Not guilty by reason of insanity"* as applied to a person with schizophrenia. Schizophrenia is basically a disease of the chemistry of the blood and nervous system. At a higher level of abstraction, it is regarded as a disease in which certain kinds of thoughts enter and dominate consciousness. A patient's belief that the CIA has planted a radio in his brain is relieved by medicines that change blood chemistry. If the patient's belief caused him to kill someone whom he imagined to be a CIA agent, he would be found not guilty by reason of insanity. If we wanted robots susceptible to schizophrenia, we would have to program in something like schizophrenia, and it would be a complicated and unmotivated undertaking—unmotivated by anything but the goal of imitating human schizophrenia. The older McNachten criterion, "unable to understand the nature and consequences of his acts", uses essentially the criteria of the present article for assessing the presence or absence of free will.

I don't know if all praise or blame for robots is artificial; the matter requires more thought. Verbally one might praise a robot as way of getting it to do more of the same.

7 A possible experiment with apes

Here's a *gedanken* experiment aimed at determining whether apes (or other animals) have free will in the sense of this article. The criterion is whether they consider the consequences of alternate actions.

The ape can move a lever either to the left or the right. The lever causes a prize to be pushed off a shelf, either to the left or the right. The goody then hits a baffle and is deflected either to the ape in control of the lever or to a rival ape. On each trial, the baffle is set by the experimenter. The whole apparatus is visible to the ape, so it can see the consequences of each choice.

The free will involves the ape having two choices and being able to determine the consequences of each choice.

There is a possibility that the ape can win without determining the consequences of the possible actions. It may just learn a rule relating the position of the baffle and the action that will get the prize. Maybe we wouldn't be able to tell whether the ape predicted the consequences or not.

We can elaborate the experiment to obviate this difficulty. Let there be a sequence of (say) six baffles that are put in a randomly selected configuration by the experimenter or his program at each trial. Each baffle deflects the prize one way or the other according to how it is set. If the ape can mentally follow the prize as it would bounce from baffle to baffle, it will succeed. However, there are 64 combinations of baffle positions. If a training set of (say) 32 combinations permits the ape to do the remaining 32 without further trial and error, it would be reasonable to conclude that the ape can predict the effects of the successive bounces.

I hope someone who works with apes will try this or a similar experiment.

Frogs are simpler than apes. Suppose a frog sees two flies and can stick out its tongue to capture one or the other. My prejudice is that the frog doesn't consider the consequences of capturing each of the two flies but reacts directly to its sensory inputs. My prejudice might be refuted by a physiological experiment.

Suppose first that frogs can taste flies, i.e. when a frog has a fly in its mouth, an area of the frog's brain becomes active in a way that depends on the kind of fly. Suppose further that when a frog sees a fly, this area becomes active, perhaps weakly, in the same way as when the frog has the fly in its mouth. We can interpret this as the frog imagining the taste of the fly that it sees. Now further suppose that when the frog sees two flies, it successively imagines their tastes and chooses one or the other in a consistent way depending on the taste. If all this were demonstrated, I would give up my prejudice that frogs don't have SDFW.

8 Comparison with Dennett's ideas

Daniel Dennett [Den03] writes about *The evolution of freedom.* I agree with him that free will is a result of evolution. It may be based on a more basic ability to predict something about what future will result from the occurrence of certain events including actions. He compares *determinism* and *inevitability*, and makes definitions so that in a deterministic world, not all events that occur are inevitable. He considers that freedom evolves in such a way as to make more and more events *evitable*, especially events that are bad for the organism.

Dennett's ideas and those of this paper are in the same direction and somewhat overlap. I think SDFW is simpler, catches the intuitive concepts of freedom and free will better, and are of more potential utility in AI.

Consider a species of animal with eyes but without a blink reflex. Every so often the animal will be hit in the eye and suffer an injured cornea. Now suppose the species evolves a blink reflex. Getting hit in the eye is now often evitable in Dennett's sense. However, it is not an exercise of free will in my sense.[2] On the other hand, deciding whether or not to go through some bushes where there was a danger of getting hit in the eye on the basis of weighing the advantages against the dangers would be an exercise of free will in my sense. It would also be an evitability in Dennett's sense.

Evitability assumes that there is a normal course of events some of which may be avoided, e.g. that getting hit in they eye is normal and is avoided by the blink reflex. My notion of free will does not involve this, because the choice between actions $a1$ and $a2$ is symmetric. It is interesting to ask when there are normal events that can sometimes be avoided.

The converse of an evitability is an opportunity. Both depend on a distinction between an action and non-action. In the case of non-action, nature takes its course.

9 Summary and remarks

A system operating only with situation-action rules in which an action in a situation is determined directly from the characteristics of the situation does not involve free will. Much human action and almost all animal action reacts directly to the present situation and does not involve anticipating the consequences of alternative actions.

One of the effects of practicing an action is to remove deliberate choice from the computation and to respond immediately to the stimulus. This is often, but not always, appropriate.

Human free will, i.e. considering the consequences of action, is surely the product of evolution.

[2] Dennett (email of 2003 Feb 27) tells me that the blink reflex involves no significant free will in his sense

Do animals, even apes, ever make decisions based on comparing anticipated consequences? Almost always no. Thus when a frog sees a fly and flicks out its tongue to catch it, the frog is not comparing the consequences of catching the fly with the consequences of not catching the fly.

One computer scientist claims that dogs (at least his dog) consider the consequences of alternate actions. I'll bet the proposition can be tested, but I don't yet see how.

According to Dennett (phone conversation), some recent experiments suggest that apes sometimes consider the consequences of alternate actions. If so, they have free will in the sense of this article.

If not even apes ordinarily compare consequences, maybe apes can be trained to do it.

Chess programs do compare the consequences of various moves, and so have free will in the sense of this article. Present programs are not conscious of their free will, however. [McC96] discusses what consciousness computer programs need.

People and chess programs carry thinking about choice beyond the first level. Thus "If I make this move, my opponent (or nature regarded as an opponent) will have the following choices, each of which will give me further choices." Examining such trees of possibilities is an aspect of free will in the world, but the simplest form of free will in a deterministic world does not involve branching more than once.

Daniel Dennett [Den78] and [Den03] argue that a system having free will depends on it being complex. I don't agree, and it would be interesting to design the simplest possible system exhibiting deterministic free will. A program for tic-tac-toe is simpler than a chess program, but the usual program does consider choices.

However, the number of possible tic-tac-toe positions is small enough so that one could make a program with the same external behavior that just looked up each position in a table to determine its move. Such a program would not have SDFW. Likewise, Ken Thompson has built chess programs for end games with five or fewer pieces on the board that use table lookup rather than look-ahead. See [Tho86]. Thus whether a system has SDFW depends on its structure and not just on its behavior. Beyond 5 pieces, direct lookup in chess is infeasible, and all present chess programs for the full game use look-ahead, i.e. they consider alternatives for themselves and their opponents. I'll conjecture that successful chess programs must have at least SDFW. This is not the only matter in which quantitative considerations make a philosophical difference. Thus whether the translation of a text is indeterminate depends on the length of the text.

Simpler systems than tic-tac-toe programs with SDFW are readily constructed. The theory of greedy John formalized by (3) may be about as simple as possible and still involves free will.

Essential to having any kind of free will is knowledge of one's choices of action and choosing among them. In many environments, animals with at least SDFW are more likely to survive than those without it. This seems to be why human free will evolved. When and how it evolved, as with other questions about evolution, won't be easy to answer.

Gary Drescher [Dre91] contrasts situation-action laws with what he calls the *prediction-value paradigm*. His prediction-value paradigm corresponds approximately to the deterministic free will discussed in this article.

I thank Drescher for showing me his forthcoming [Dre06]. His notion of *choice system* corresponds pretty well to SDFW, although it is imbedded in a more elaborate context.

This article benefited from discussions with Johan van Benthem, Daniel Dennett, Gary Drescher, and Jon Perry. The work was partly supported by the Defense Advanced Research Projects Agency (DARPA).

References

[Den71] Daniel C. Dennett. Intentional systems. *The Journal of Philosophy*, 68(4):87–106, 1971.

[Den78] Daniel Dennett. *Brainstorms: Philosophical Essays on Mind and Psychology.* Bradford Books/MIT Press, Cambridge, 1978.

[Den03] Daniel Dennett. *Freedom Evolves.* Viking, 2003.

[Dre91] Gary Drescher. *Made up minds: a constructivist approach to artificial intelligence.* MIT Press, 1991. Q335D724.

[Dre06] Gary Drescher. *Good and real: Paradoxes from physics to ethics.* M.I.T. Press, forthcoming, 2006.

[McC63] John McCarthy. A Basis for a Mathematical Theory of Computation[3]. In P. Braffort and D. Hirschberg, editors, *Computer Pro-*

[3] http://www-formal.stanford.edu/jmc/basis.html

gramming and Formal Systems, pages 33–70. North-Holland, Amsterdam, 1963.

[McC79] John McCarthy. Ascribing mental qualities to machines[4]. In Martin Ringle, editor, *Philosophical Perspectives in Artificial Intelligence.* Harvester Press, 1979. Reprinted in [McC90].

[McC90] John McCarthy. *Formalizing Common Sense: Papers by John McCarthy.* Ablex Publishing Corporation, 1990.

[McC96] John McCarthy. Making Robots Conscious of their Mental States[5]. In Stephen Muggleton, editor, *Machine Intelligence 15.* Oxford University Press, 1996. Appeared in 2000. The web version is improved from that presented at Machine Intelligence 15 in 1995.

[McC00] John McCarthy. Approximate objects and approximate theories[6]. In Anthony G. Cohn, Fausto Giunchiglia, and Bart Selman, editors, *KR2000: Principles of Knowledge Representation and Reasoning,Proceedings of the Seventh International conference*, pages 519–526. Morgan-Kaufman, 2000.

[McC02] John McCarthy. Actions and other events in situation calculus[7]. In B. Selman A.G. Cohn, F. Giunchiglia, editor, *Principles of knowledge representation and reasoning: Proceedings of the eighth international conference (KR2002).* Morgan-Kaufmann, 2002.

[MH69] John McCarthy and Patrick J. Hayes. Some Philosophical Problems from the Standpoint of Artificial Intelligence[8]. In B. Meltzer and D. Michie, editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, 1969. Reprinted in [McC90].

[Rei01] Raymond Reiter. *Knowledge in Action.* M.I.T. Press, 2001.

[Sha97] Murray Shanahan. *Solving the Frame Problem, a mathematical investigation of the common sense law of inertia.* M.I.T. Press, 1997.

[4] http://www-formal.stanford.edu/jmc/ascribing.html

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

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

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

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

[Tho86] K. Thompson. Retrograde analysis of certain endgames. *ICCA (International Computer Chess Association) Journal*, 9(3):131–139, 1986.

[Tho03] Richmond Thomason. Logic and artificial intelligence. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* 2003. http://plato.stanford.edu/archives/fall2003/entries/logic-ai/.

/@steam.stanford.edu:/u/ftp/jmc/freewill2.tex: begun Thu May 16, 2002, latexed November 6, 2005