

## **Four Concepts of Freewill: Two of them incoherent**

**Aaron Sloman**

---

**Last updated: 11 Aug 2007: Freedom as permission**

The discussion below could be extended by pointing out that there is a fifth notion of freedom which refers to what you are free to do within a context of a game, a system of laws, a moral regime etc. This notion of freedom is close to the notion of permission. It is worth noting that the law may forbid something without enforcing that proscription. So many people constantly do what they are not free to do in this sense.

---

### **NOTE ON FORMATTING:**

Adjust the width of your browser window to make the lines the length you prefer.  
This web site does not attempt to impose restrictions on line length or font size.

---

*I have found power in the mysteries of thought.*

**Euripides, 438 B.C.**

---

### **NOTE**

These notes form an extension to my paper originally written around 1988 and revised a few times:

"How to dispose of the free will issue"

---

## **Four Concepts of Free-will (two good, two garbage)**

There are at least four different notions of freedom that drive philosophical and scientific discussions of free will. Two of them make sense and are not only consistent with determinism, but depend on the world being (largely) deterministic.

The other two notions of free will, one theological and one romantic, are incoherent, but worries about them cause people to argue that free will is inconsistent with determinism. People with such worries, may need therapy more than they need arguments, since arguments typically don't remove their worry, and, in my experience, they are not able to produce counter-arguments, only counter-assertions.

The four concepts are only briefly explained here.

### **1. Ordinary language -- I came of my own free will, nobody forced me.**

This is a widely used concept (or collection of concepts), at least in our culture. Its use is totally compatible with normal deterministic brain function. But there may be borderline/difficult cases (e.g. hypnosis, indoctrination).

In 'A plea for excuses', J.L. Austin (in 1956) brilliantly and amusingly discussed some of the ways in which this concept is commonly used in making excuses, e.g. by explaining why something was or was not done, and resisting accusations either of acting under duress or coercion, or of not acting under duress or coercion.

---

*"In this sort of way, the philosophical study of conduct can get off to a positive fresh start. But by the way, and more negatively, a number of traditional cruces or mistakes in this field can be resolved or removed. First among these comes the problem of Freedom. While it has been the tradition to present this as the 'positive' term requiring elucidation, there is little doubt that to say we acted 'freely' (in the philosopher's use, which is only faintly related to the everyday use) is to say only that we acted not un-freely, in one or another of the many heterogeneous ways of so acting (under duress, or what not). Like 'real', 'free' is only used to rule out the suggestion of some or all of its recognized antitheses. As 'truth' is not a name for a characteristic of assertions, so 'freedom' is not a name for a characteristic of actions, but the name of a dimension in which actions are assessed. In examining all the ways in which each action may not be 'free', i.e. the cases in which it will not do to say simply 'X did A', we may hope to dispose of the problem of Freedom."*  
(Austin, A Plea for Excuses)

Related points were made by around the same time by Isaiah Berlin in Two concepts of freedom: he distinguished positive freedom, i.e. 'freedom to' from negative freedom, 'freedom from'.

Neither Austin nor Berlin adopted or even knew about the design standpoint that AI-informed philosophers adopt.

---

This is the main notion that supports compatibilist theories of freewill: which claim that the existence of freewill is compatible with universal determinism.

An introduction to compatibilist theories of freewill can be found at [this philosophical web site](#). However that document claims that assuming that freewill is 'uniquely human' is a theory neutral assumption! I leave it to pet owners and robot designers to point out the silliness of that claim.

**NOTE: 8 Dec 2010:** I am pleased to report that that wording seems to have been revised, since I wrote the above.

## **2. Legal concept (close to 1): concerned with whether people can be held responsible for their actions in court.**

This legal concept grows out of the ordinary language concept, though courts generally attempt (with varying success) to be more precise in the concepts and forms of arguments that can be used in settling questions whose answers may determine whether someone does or does not get fined, imprisoned, or in barbaric cultures mutilated or put to death.

An example of legal attempts at clarification, including for example the relative importance of impaired ability to distinguish acceptable from unacceptable behaviour and impaired ability to resist impulses, can be found [at this web site on penal law](#).

The legal arguments are also totally compatible with normal deterministic brain function, and in fact depend (implicitly) on assumptions concerning normal and abnormal functioning of the brains of plaintiffs, as indicated by reference to 'disease' or 'mental defect'.

But as with the ordinary concept of doing something of your own free will, there may be borderline/difficult cases for the legal notion (e.g. hypnotism, duress, mental illnesses, etc.) where experts invoked by legal teams will not necessarily agree. The existence of indeterminate borderline cases is a feature of many useful concepts of everyday life, though attempts to import them into scientific theories often generate muddle and confusion.

## **3. Theological concept designed to let God off the hook:**

The third notion of freewill is used in the theological claim that the alleged all-powerful, all-knowing, all-good god who created everything cannot be blamed for all the evil stuff that occurs in that creating (and there is a lot of it!) because it arises from our free will which he also gave us.

This notion of freewill is an incoherent concept because nothing could possibly satisfy all its mutually contradictory requirements. It implies that we are to be blamed for the effects of our actions, because the world is sufficiently deterministic for the actions to have definite consequences, yet god cannot be so blamed. If god's creatures are not deterministic then that implies that god cannot know the consequences of what he is doing, which contradicts the claim that he is all-knowing. If he does know the consequences and they are evil and he takes no action to prevent the consequences (which he could do because he is all powerful) then he too is evil. (Read Hume for more detailed arguments: it's old stuff.)

It is also worth commenting that a god who does not interfere when his 'offspring' turn bad cannot be described as anything like a loving parent, who will typically do as much as possible to help an errant child mend his or her ways, without being accused of tampering with the child's free will. (But the story in A Clockwork Orange draws attention to the possibility of some moral disagreement regarding what means are acceptable for improving decision making. It's just a myth that every question that can be asked must have right and wrong answers. Sometimes we just have to make up our minds what to do.)

#### **4. Romantic concept: used by many amateur and many professional philosophers**

People don't like the idea of being determined and predictable in principle, no matter how difficult it is to predict their decisions in practice. So they wish they were not like that, yet would like to feel in control of and responsible for their decisions

Another incoherent concept -- because nothing could possibly satisfy all its mutually contradictory requirements.

(This is related to the discussion below of whether there are any true counterfactual conditionals in a deterministic universe.)

The whole idea of freewill driving the philosophical quest for an alternative to determinism is incoherent because nothing could satisfy it -- not even Quantum Mechanics: randomness has nothing to do with being free, as Hume and many others have pointed out. The main points were made long, long, ago by David Hume, and some of them before him by John Locke, and probably others before them. (I am no good at history.)

---

There are various key ideas that have to be grasped.

One is that within a complex portion of the universe, e.g. your left thumb, or your brain, or the whole of you, or a nation, or an ecosystem, there will be very many processes that can be described at different levels of abstraction: physical, chemical, physiological, and in some cases in terms of thoughts, percepts, desires, preferences, values, memories, capabilities, skills, and many kinds of information. In some cases there are also social and political processes. A mechanical clock has fewer levels of abstraction.

At any level of abstraction we can talk about how things might have been different from what they actually are. E.g. the clock now registers 13 minutes past 10, but it could have registered 23 minutes past 11 as far as the workings of the clock are concerned. They allow for that possibility, which is why we can regard some counterfactual conditionals as true and others false. (This point is analysed further below.)

In the case of a typical mechanical watch or clock, things are very rigidly determined, and it always goes through a fixed sequence of states. But other mechanisms allow more flexibility. E.g. many digital clocks and watches have buttons that allow the clock to be moved into a mode in which it can jump between states in a different order, e.g. increasing hours or minutes, or changing the day of week, month or year.

In particular, a digital watch allows the hour to be changed without changing the minutes, but old mechanical watches do not allow this, unless you dismantle the watch and either disengage the gears, or remove the hour hand and replace it in a new position (which may or may not be possible -- depending on how the spindle is shaped).

When the time on a digital watch is changed, the buttons in the watch have to be pressed by someone *outside* the watch! But suppose you built a system that had all sorts of internal controls that alter internal and external behaviour, where those controls can themselves be operated by other internal mechanisms. E.g. most digital watches cannot adjust their own time, but the recently available radio watches can detect that they are fast or slow and take remedial action. Maybe some of them can also check whether the time signal is coming from an authenticated source, and if

not ignore it. We can say of such a watch that it could have changed the time it is displaying, and it would have done so if there had been a discrepancy with the radio timing signal.

If there is such a control architecture, then, for many states of the system (at any level of abstraction) we may be able to say 'It could have been otherwise', and we may be able to say that unlike ordinary digital watches, but like more sophisticated digital watches, it has internal components that could have made the state different from what it is now.

In the case of humans, some animals, and future robots, some of those components (or states of components) play the role of desires, values, preferences, goals, attitudes, likes, dislikes, ambitions, fears, etc. and others contain beliefs about how the world is, about what can and cannot be done, about what the prerequisites and consequences of various changes are, about how various processes can be produced.

Analysing what desires, values, etc. are in terms of their functional role in a working system is a non-trivial task. A beginning is made in The Architectural Basis of Affective States and Processes (2005) by Sloman, Chrisley and Scheutz.

This sort of analysis adopts *the design stance*, which is also needed in order to understand products of evolution even though evolution is not a designer with explicit goals that its designs achieve. Instead it is a mechanism that manages to produce systems that actually achieve those goals even though nobody designed them to achieve those goals, and even though those goals were never explicitly formulated, by evolutionary mechanisms or anything else: they merely happen to be implicit in the processes of natural selection. (Unlike plants or animals bred for a specific purpose, whether by humans or anything else.)

The attempt to explain what beliefs, desires, values, preferences, goals, attitudes, likes, dislikes, ambitions, fears, etc. are within the framework of Dennett's *intentional stance* cannot succeed, because that (like the 'Knowledge Level' proposed by Newell) assumes that things with those mental states, processes, etc. are *rational* whereas humans and other animals obviously can have such states without being rational.

Many of those internal states will themselves have been produced previously by a combination of influences, including both previous choices made by the individual, states and processes in the environment, and, of course, the initial state of the system.

We can then distinguish two cases:

1. Most of your control decisions (selections between possible options, whether internal or external) made over a period of time are controlled by your own *internal* components -- some of which may be involved in determining which options are even considered, e.g. when you construct alternative plans and then choose between them,

as in fully deliberative systems.

and

2. Most of the selections, or at least most of the major ones that influence all the others, during that time period, are made by *external* influences, such as strong winds, avalanches, or other individuals pushing you around, making threats, or even tampering with your brain.

In the second case you can say that your behaviour was not free insofar as other objects or other people produced selections and actions that would not have been produced if you had not had the external influences.

Of course, there is no *sharp* dichotomy of cases, but a spectrum with many subtle discontinuities, like the difference between doing something by mistake and doing it by accident, or the difference between doing something intentionally and wanting to do it. (See J.L. Austin for more examples.)

In case 1. you can say that you were free insofar as it was *your* desires, beliefs, preferences, hopes, fears, values, plans, etc. that selected what actually happened from among the alternative possibilities.

Of course your desires, beliefs, hopes, etc. are themselves products of a large collection of influences over many different time scales, some going back over millions of years of evolution that made it possible for you, unlike a flea, to consider complex future options and choose between them. (The architectural requirements for that, which are presupposed by some AI researchers, but are not always made explicit, are summarised in this discussion of requirements for 'fully deliberative' systems.)

In this sense you are free to the extent that your decisions and actions are mainly determined by *your* desires, tastes, preferences, beliefs, hopes, fears, values, etc. rather than by other things.

What other kind of freedom could you wish for? Do you want to be able to change all your desires, tastes, etc. at any time? On what basis? Randomly? If they change randomly why call that freedom? If they are not random what should determine them other than other desires, tastes, preferences, etc.?

### **Worries about remote causes**

Some people worry that even though what they do is the result of their own desires, tastes, etc., these desires and tastes and the mechanisms through which they influence decisions and behaviour are themselves products of deterministic processes -- during our evolution, individual development, etc.

So what: those prior determinants of what you are made it possible for you to have kinds of freedom that clocks and most other types of animals do not have. If you wish to be free and do not wish to have deterministic mental mechanisms produced by prior deterministic influences, then you are asking for something that is incoherent --- it is logically impossible.

OK -- so you lack the freedom to have something incoherent. To that extent you are unfree. But that's not a result of living in a deterministic universe. If you were in a universe full of totally undetermined occurrences you would not be more free, just more unpredictable. Actually that may be an incoherent notion too, but that's a topic for another day.

### **Some real limitations on freedom**

Of course people may sometimes regret that they have certain desires and wish they had others -- a common consequence of religious indoctrination which sometimes persuades people that their healthy biological desires are wicked even if they don't act on them (a form of cruelty that should be strongly opposed, like other aspects of mind-binding cultures).

Another sort of case is a person who wishes he were were not addicted to gambling, or nicotine. In such cases changing to a new state can be very hard. *People like that really do have their freedom restricted*. Likewise someone who wants to be a concert pianist and plays beautifully at home but collapses with nerves on the stage has his freedom restricted, and may not be able to change easily. Sometimes a good teacher or therapist can help. Or sometimes the change just happens gradually after repeated attempts.

The fact that we *sometimes* have our freedom restricted does not imply that we *never* do anything of our own free will (in the every day or legal sense).

### **Internal freedom-restrictors**

As some of the previous examples illustrate, it's not only external influences that can interfere with your freedom. A virus infection can make you so weak that you are as incapable of doing certain things as you would be if somebody tied you up.

More subtly, there can be things in you, even things in your collection of desires, that you would prefer not to have. Examples are addictions. These may be viewed as buggy products of mechanisms that are highly desirable in other contexts. There are many complex designs that have bad side-effects that arise out of unintended interactions between parts, or between the mechanisms and certain sorts of environments.

It is very important for our well-being that physical processes can alter our desires, e.g. when your body is short of water you have mechanisms that detect that state and produce a desire to drink, and that's just as well or you'd die. Likewise, temperature sensors, pressure sensors, stress sensors, exhaustion sensors, can all twiddle the contents of your current collection of desires. That's not normally regarded as a loss of freedom, but a requirement for a long and healthy life.

Some desires, such as the addict's desire for a smoke or an alcoholic drink, or a shot of cocaine, or the addicted gambler's desire to make another attempt to win, are not part of the functioning of the system that is required for its well-being and, on the contrary, they can shorten life and cause much misery. But the fact that they are internal does not mean that the person who wishes to get rid of them can easily do so: they may be much harder to get rid of than a virus infection or a rope tying you to a tree.

There is much literature presenting more complex cases, in plays, stories and novels, and also in the clinical literature of psychiatrists and psychoanalysts. Philosophers have also written many words about these.

These internal causes are important for the first two concepts of freedom because they are relevant to what can and what cannot be given as an excuse or mitigating circumstance either in ordinary personal interactions or in the law courts. There may be no right or wrong answers to where the boundaries should be drawn: those are ethical questions that different cultures may answer differently.

This is not the place for a more complete discussion. My main point is that addictions, hypnotic suggestions, virus infections, uncontrollable desires to be violent produced by brain damage, are different from your evolutionary history, your normal biological needs, your personal hopes, likes, and ambitions, etc. Putting them all the same basket is one step towards an incoherent notion of freewill.

## **The importance of counterfactual conditionals**

Implicit throughout this discussion is the assumption that it makes sense to talk about 'what would have happened if...' E.g. if you had not felt hungry you would not have taken a detour to the sandwich bar.

This notion has been the subject of huge amounts of philosophical discussion, including discussion of whether that notion is consistent with determinism.

For a recent example see 'Counterfactuals and explanation' by Boris Kment in *Mind*, 115, 458, April 2006.

There is a lot more philosophical literature including some very complicated theories about possible worlds, causation, prediction, explanation, probability, and the meaning of 'if'. A small selection is referenced in Kment's paper.

In particular there *appears* to be a conflict between determinism and the first two notions of freedom, which can lead to fatalism, as follows.

Suppose you have a choice to make, e.g. between staying in bed to help yourself recover from an illness and going to work to give a lecture, which may make the illness worse, but will help your students who have an examination the next day. While you lie in bed thinking about the two options you assume that there really are two options and your weighing up the considerations will determine which of them is selected.

But if determinism is true then well before your final decision the whole state of the universe determines what your decision will be. Therefore your deliberations and your decision have no effect and the freedom to choose is illusory.

This is a fallacious conclusion because when the action is selected it is selected *as a result of* a process that *included* your deliberations: *If you had not gone through that process you might have performed a different action.*

Suppose you decide to go to work and give the lecture, because the examination is imminent and some students badly need last minute advice, whereas the harm you do to yourself by lecturing while ill will have more limited effects on you than failing an important exam will have on students. Then it is true that if you had not thought about the consequences of cancelling the lecture, or if you had forgotten that there was an examination the next day then you would have done something different, e.g. staying at home on the doctor's advice.

All of this implies that some of the features of the situation before you decided could have been different, so that it makes sense to ask what would have happened if they had been different. However, 'things could have been different' *is always relative to some context that determines a set of constraints*.

Obviously if the *only* such context you are willing to consider is 'everything being *exactly* as it was before you started trying to decide', then in that situation you would not have reached a different situation (unless it was a case where some random brain process, perhaps based on quantum noise, played a significant role: but, as I keep saying, randomness does not produce freedom -- there is more on that below).

But we are perfectly capable of thinking about contexts that are not so constrained. E.g. I happened to remember the examination, but I can still reason about what would have happened if I had not remembered it, because consideration of the examination played an important role in the decision. So I can truly say that if I had not remembered the examination I would have stayed at home.

Of course in that situation (my not remembering the examination) something else must have been different -- e.g. the state of my brain or the entry in my diary that caused me to remember the examination and take it into account, or some external distraction that stops me thinking of the examination. But I don't need to know, or explicitly describe, all the different possible states of the universe that might have left me taking the decision without considering the examination.

I don't need that because I just know from experience that there are lots of relatively small portions of the universe, like the digital watch, and my brain, that at any time, could have been in a different state without anything being badly broken: because the structure of those mechanisms allows for those alternative combinations of states, as the design of a digital watch allows me to make the hour change without the minutes changing, though normally that doesn't occur.

We also know that some of those possibilities would have had only very minor effects whereas others would have major effects, even if we can't always tell which is which. In highly structured mechanisms, such as digital watches and brains, the distinction may be much easier to make because the causal pathways are far more constrained. It's different for weather systems or ecosystems.

It may also be difficult to tell in brains which changes have big effects and (because they are amplified in chaotic, non-linear, feedback loops) and which have small effects (because homeostatic mechanisms stop small perturbations from propagating).

---

## **How randomness screws up counterfactual conditionals**

Much is made above of the fact that there are many true counterfactual conditional statements about what would or would not have happened if something had or had not occurred.

The person trying to avoid opprobrium or punishment might argue that he would not have stolen some medicine from a rich doctor if his impoverished neighbour had not been on the point of dying, or if he had had enough money to buy the medicines, or if the doctor had been willing to help.

However if an effect is produced by a random mechanism then it makes no sense to ask precisely what would have happened if something had been different: all that can be truly said is that *some* selection from the range of processes made possible by the structure of the device would have occurred, and that selection would have had consequences.

A different view is taken by some philosophers, e.g. Kment in the paper cited above, who argue that counterfactual conditionals regarding random devices can be true in cases where the counterfactual assumption has nothing to do with the operation of the device. E.g. if Fred selected the number 888 in a lottery, and a 'truly random' machine (if there can be such a thing) produces the winning number 889, then Kment and many others would say that the following is true.

(A) *If Fred had selected 889 then he would have won.*

However I believe this is one of those indeterminate cases where there are equally strong arguments for saying that that statement is not true unless the antecedent is expanded as follows:

(B) *If Fred had selected 889 and the machine had produced the same result as it actually produced, then Fred would have won.*

I suspect many people unwittingly treat the extra conjunct in the antecedent as presupposed by the question. But it does not have to be presupposed.

If the machine is random and we don't restrict 'what if' considerations to the cases where the result is stipulated to be the same then we have no idea whether Fred would have won.

(This amounts to the suggestion that attempts to make sense of counterfactuals in terms of possible worlds and 'closeness' relations should either exclude random devices whose operation makes a significant difference to what happens, or else say that closeness of possible worlds to the actual world must require all random devices with major consequences to produce the same effect as occurred in the actual world.)

I am not trying to prove that statement like (A) are false, merely to suggest that it isn't clear whether they can be true. This is linked to unclarity in the notion of randomness.

---

## **Different notions of "different levels"** (Added 1 Aug 2008)

Not all differences of level are relevant to the points being made here. A rectangular grid of dots on a sheet of paper can be described at different levels, e.g.

- Describing each dot and its location
- Specifying that there are rows of dots of particular lengths and particular distances apart, and columns of dots of particular lengths and particular distances apart.
- Specifying that there is a rectangular grid of dots.
- Specifying that the dots lie along diagonal lines.
- Specifying that the dots form a regular pattern.

However in previous cases we talked about different counterfactual conditionals being true at different levels because there are causal interactions at different levels, whereas these different descriptions of the dot grid do not specify different sets of causes and effects.

The important cases of levels of abstraction relevant to the free-will debate depend on there being virtual machines at different levels, in which causes and effects occur. For more on this see

<http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#bielefeld>

---

## On conceptual analysis

This is a very short position paper. I have referred to two coherent concepts, but that should not be taken to imply that those concepts are very precise, well understood, or easy to analyse. Both of them have had many books and articles written about them, and there are a lot of intricacies that I have ignored here. People who wish to go further will have to devote a lot of time to learning to do analytical philosophy: it is not easy.

A short but dense tutorial on some of the techniques of analytical philosophy can be found in [Chapter 4 of \*The Computer Revolution in Philosophy\*](#)

Gilbert Ryle coined a label for the results of the kind of analysis that describes collections of relationships, often unobvious and unnoticed relationships, between collections of concepts, namely 'logical geography'. I have tried to describe connections between that kind of analysis, and a deeper analysis that identifies aspects of reality that can support different logical geographies for the same 'terrain'. My provisional name for the second sort of analysis is 'logical topography'. The distinction is explained, and some misinterpretations of the distinction countered, in this draft paper:

[Two Notions Contrasted: 'Logical Geography' and 'Logical Topography'](#)  
[Variations on a theme by Gilbert Ryle:](#)  
[The logical topography of 'Logical Geography'](#).

---

My small additional contribution to what are essentially Hume's ideas was originally posted to an internet news group in 1988. I tried to show how we can better understand the ways in which humans and other animals or machines may have more or less freedom/autonomy/control/responsibility, depending on their [architecture](#). It can be found here:

<http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#8>

HOW TO DISPOSE OF THE FREE WILL ISSUE

available in various formats.

This was originally posted to comp.ai.philosophy around 1989. A slightly revised version appeared in *AISB Quarterly*, Winter 1992/3, Issue 82, pp.31-2.

In that note I tried to show how within the space of possible designs for more or less sophisticated information-processing systems, there are some with more opportunities to do things and some with fewer. Evolution produced a very wide variety of designs, some with more 'freedom' some with less. There is no reason to believe that humans are any kind of optimum, or that there is some coherent limiting case of perfect freedom.

Stan Franklin expounded and developed those ideas (with acknowledgements) in Chapter 2 of: his book: [Artificial Minds](#) MIT Press, 1995

Daniel Dennett has written at least two books on the subject. If you wish to see the arguments spelled out in much greater detail with more examples, often in an entertaining fashion, read his [Elbow Room](#) and [Freedom Evolves](#) (I have not yet read the latter, but he is always worth reading).

---

Maintained by [Aaron Sloman](#)  
[School of Computer Science](#)  
[The University of Birmingham](#)