Towards Machine Agency: a Philosophical and Technological Roadmap

J Storrs Hall

March 30, 2012

Abstract

A key question which will face fields of inquiry as diverse as philosophy, law, and even theology in the coming decades is that of machine agency. Simply put, at what point does the responsibility, legal or moral, for an act of a machine inhere in the machine itself instead of its (human) designers and builders? We examine in this context a selection of cognitive architectures for an artificial general intelligence, and compare them with various standards of autonomy, moral agency, and legal adulthood.

The Problem: Can a Machine be a Moral Agent?

> *The cause of thy sin is inevitably determined in heaven; this did Venus, or Saturn, or Mars: That man, forsooth, flesh and blood, and proud corruption, might be blameless; while the Creator and Ordainer of heaven and the stars is to bear the blame.*
>
> –Augustine

Currently it is more or less universally accepted that when a machine does something wrong, morally or legally, the fault lies with the machine's designers and builders, and not of the machine itself. The advance of robotics technology, however, is pushing us rapidly into a situation where this will not be so clearly true as it is today. There are several reasons for this trend:

- The first is primarily psychological. As physically humanoid robots more closely resemble humans in grace and skill of movement, and as natural language interfaces become more adept at informal dialogue, people will instinctively assign them a status closer and closer to "human." This position comes with moral agency *ex officio*.
- Progress in AI has produced, and will continue to produce, machines whose actions are less and less predictable by their designers. A simple foretaste of this is the chess-playing programs which can beat, and thus surprise, not only their programmers but the best human players.
- The trend to greater complexity and autonomy is exacerbated by the trend within AI to having an ever-greater learning component in the machine's design. The decisions made by a modern AI are more and more a product of its experience and training and to that extent less a direct product of its program code. The parallel between this and the roles of nature and nurture in the formation of human character and skill is more than metaphorical.
- At the same time, the bar is being lowered: advances in neuroscience have injected a mechanistic account of human decision-making into the public conversation, which has weakened to some extent the traditional model of human free will and accountability. The more that humans are seen as being "mere machines," the more, *eo ipso*, will mere machines be seen as human.

The notion of moral agency has long been tied, in philosophical thought, to that of free will. Legal and ethical codes have traditionally taken the intuitive, "folk psychology," concept of free will for granted. Yet philosophers have long understood that the concept becomes quite problematical, even paradoxical, under closer examination. In particular, there appears to be no ontological room between the deterministic causality of macroscopic physics and the pure randomness underlying quantum and statistical mechanics into which free will can fit. We cannot blame someone for an act which was an inevitable outcome of his physical structure; but neither can we blame him for one resulting from a roll of the molecular dice.

Determinism

'Tis impossible to admit any medium betwixt chance and necessity.

–David Hume

In the world of software, problem of determinism is brought into focus in a way which must illuminate the ontological paradox. Because all software is reducible, by construction, to finite chains of primitive Boolean operations, the output of a program is completely deterministic given its input. The fact that the input might include randomness in the form of measurement variance, or even a radioisotopic random number generator, is clearly irrelevant to the question of whether the program has free will, and thus, in the intuitive account, agency.

Suppose, as a *gedankenexperiment*, that the universe is simulated on a computer that is equivalent to a Turing machine. The laws of physics at the most microscopic level are the program code of the simulator. This simulation has the property that given an initial state, the entire history of the universe can be computed mechanically. Because it is digital, we can ignore issues such as sensitive dependence on initial conditions; we assume we know every bit in the initial state. Because the simulator is deterministic, there can only be as many trajectories of the history of the universe as there are initial states.

Now let us imagine an additional constraint on the simulation: suppose we require that the simulator is an invertible function. That is to say, that there is a reverse-time simulator that can run the universe backwards, just as deterministically as the forward one. This is equivalent to saying that information cannot be created or destroyed by the simulator, or to saying that the trajectories of states cannot branch or merge.

Now let us consider the sequence of states of the universe in such a trajectory. Because the inverse simulator is as deterministic as the forward one, it is equally valid to say that state n is caused by state $n+1$ as it is by state $n-1$. It is just as valid to say that the trajectory is determined by some state in the far future as by one in the far past. It is in fact just as correct to claim that both past and future are determined by a snapshot of the present.

Is it possible that such a microscopically reversible simulation could present an experience to its inhabitants in which causality only went in one direction? In which puddles never leapt up into overturned glasses which then righted themselves? In which photons from the walls never converged on a filament, generating electricity?

We can answer, with a high degree of certainty, that it is not only possible, but that it is actually true of our own universe. The laws of physics, as well as we understand them, are in fact microscopically reversible. Ironically, it is the very conservation of information at the microscale which implies the Second Law of thermodynamics at the macroscale. Indeed Hawking (1988) conjectures that our perception of the direction of time and causality is due to the direction of entropy increase (which is contingent on boundary conditions), rather than any property of the underlying microscopic mechanics. This situation is referred to as the Arrow of Time (or Loschmidt's) paradox.

This paradox stands as an example of the fact that arguing from microscopic properties to macroscopic ones is fraught with epistemological peril. Given the bi-directional nature of microscopic determinism, one might as readily argue that our actions are caused by their consequences. It seems preferable to allow that microscopic and macroscopic accounts of the universe are ontologically distinct; in particular, our intuitions do not properly map properties of one onto the other.

Given that our intuitions of determinism fit poorly with what we know of reality, and of course contradict our intuitions of freedom directly, it is at least reasonable to attempt an understanding from another point of view. We know that impenetrability, randomness, and monotonic entropy at the macroscopic scale can arise in systems that are completely deterministic, digital, and reversible at the microscopic scale. Can we not then look for a *macroscopic* computational property of such systems that we can reasonably identify with freedom?

Note in passing that the property in question must be at the highest architectural level of the system. This is because the implementation of the lower levels can vary completely and yet provide the same "virtual machine" to the higher ones. Different hardware can implement the same machine-language instruction set. The same high-level programming language can be implemented on different instruction-set architectures. The same operating system can be implemented in different languages. There are typically many levels in the "software stack," and the details of each level are independent of the other levels (which is the reason for building them that way).

An Operational Definition

Surveyor: *That's the good news. The bad news is that the trillions of synapses involved in this make it almost impossible for the other parts of the brain to figure out how these decisions are made. As far as their higher-level reasoning can tell, those decisions just happen – without any cause.*

Apprentice: *Is that what they refer to as "free will?"*

–Marvin Minsky

Various philosophers have staked out positions in the general area of compatibilism, accepting both the notions that the universe is microscopically deterministic and that there is something real that our intuitive notion of free will actually corresponds to.

The best such theory for application to machines comes from computer scientist Drew McDermott. It is based on a cognitive architecture that is squarely in the mainstream of artificial intelligence research, and yet it is general enough to be applicable to a broad range of architectures, including ultimately that of our own brains.

But before examining McDermott's model in detail, consider the desiderata a theory of free will must meet operationally, i.e. in order to be useful:

- It must be the case that in a situation where a reasonable observer would conclude that a human "had a choice" about an action, such an observer would conclude that the machine also had a choice, though it might differ in detail.

- There must some non-trivial difference between punishing a free machine versus a non-free one; the former should be reasonable in a way the latter is not.

- Other limits to agency should hold as well: the theory should explain, for example, why an agent is not responsible for acts whose consequences it does not understand, or outcomes it was powerless to prevent.

Now let us consider a robot which chooses its actions by means of a model – a computational simulation – of the world around it. The robot knows that if it drops a pound of lead to the floor, it will fall rapidly and hit with a thud. It knows because it can imagine – create in its simulation – such a lump and run the computational experiment. The robot similarly knows what will happen if it drops a china vase, or knocks over a glass of milk, or jumps off a cliff.

Note that in order for this "world model" to be useful to the robot, it must be deterministic: it must reliably predict the results of actions in the real world.

Besides the model, the robot needs a "utility function," a way to evaluate and compare states of the world as represented in the model. For example, scenarios including a puddle of milk on the floor might be scored lower than ones without.

As a part of the model of the world, the robot has a model of itself. In order to decide what action to perform in the real world, the robot runs computational experiments in which its model of itself performs a set of candidate actions in the model world. For each action, the resulting state is calculated, and the utility function of that state evaluated. Then the robot undertakes in the real world the action that resulted in the highest simulated utility.

This is in essence a standard "rational agent" cognitive architecture, but McDermott (2001) adds to it a remarkable observation. Consider what happens when the designers of the robot, having gone to the trouble of building a comprehensive world model, proceed to reuse the model as a resource for the robot's intuitive ontological judgements. To answer a question about a property of something in the world, the robot consults its model of the thing and reports the property as modelled.

In particular, the robot will perceive the dynamics of its world to be deterministic, since that's the way the model world works.

But the robot's model of itself cannot be deterministic. The rational agent algorithm requires that the self-model arbitrarily attempt a set of different actions from exactly the same initial condition. From the point of view of the part of the robot outside the model, its inner model of itself must have a "control knob" that selects an action – a knob that every other object in the deterministic world model lacks.

Operational Agency

> *What about free will? Here it is very difficult to say anything without saying something that will be contested by some philosopher.*

–Stanford Encyclopedia of Philosophy

Let us pass lightly over the question of whether McDermott's control knob represents "real free will." Many philosophers would refer to the above as an error theory, of why we think we have free will when in fact we don't.

Others might contend that, the concept being so problematical, the control knob represents our best description of what "real free will" actually is. For the present discussion, however, we can ask a more pragmatic question: whether it makes sense to use the rational agent cognitive architecture, hereinafter "rational," as a stand-in for "real free will" in moral and legal reasoning.

Let's consider the desiderata above:

- A reasonable observer would conclude that a rational robot would "have a choice" in the same kinds of situations as a human. A clear example of this is a chess-playing program, a prototypical rational agent. In the midst of a complex game, the move it will make is not at all obvious, and it is clearly weighing its alternatives – its choices – the same way a human would do.
- It makes no sense to kick your dryer if your clothes come out wrinkled. However, punishment makes sense for rational machines in exactly the same cases as it does for humans: credibly to threaten punishment for undesirable acts, and to exact it as an example and maintain credibility in the case of future ones. In the rational machine, unlike other machines, a credible threat of punishment (or reward) will be added to the calculated utility of the predicted outcome of an act. In other words, for the rational robot, *punishment changes behavior* and thus makes sense.
- Even though a robot is rational, if an act has consequences it couldn't predict, e.g. by virtue of being outside the world model's computational capacity, a threatened punishment for that consequence would not change its utility estimations and thus would not change its actions. The rationale for punishment, and thus agency, disappears.
- The question of outcomes that were beyond the robot's control brings us back face to face with the question of its being, on the microscopic level at least, completely deterministic, and thus in some sense incapable of having done anything other than it actually did. However, I shall claim that while superficially appealing, this objection is beside the point. The real question is whether punishment changes behavior. It is a perfectly reasonable to say that an outcome was "beyond the robot's control" if the outcome would have happened in spite of any conceivable threat of punishment or promise of reward. It is not reasonable to say that an outcome was beyond the robot's control if a reward of $0.02 would have changed the outcome.

We can clarify the final point by considering the purest possible form of punishment or reward for a rational robot. The point of punishment is to change the effective result of its utility calculation. Thus we could achieve the same effect simply by changing its utility function. Thus to state that something was beyond the robot's control seems equivalent to saying that the robot would not have changed the outcome under any possible utility function. This seems to map quite naturally on to a statement about a human to the effect that "he couldn't have changed it if he had wanted to."

Gazanniga (2011) points out that a key intuitive difference between humans (and animals such as dogs and horses) and machines is that when a human misbehaves, you punish it, whereas when a machine does, you fix it. On our present theory, however, it becomes clear that punishing and fixing are essentially the same: punishing is a clumsy, external way of modifying the utility function.

Furthermore, a closer analysis reveals that fixing – modifying the robot's utility function directly – is tantamount to punishment, in the sense that the robot would not "want" it to happen and would act if possible to avoid it.

Consider a robot in a situation with two alternatives: it can pick up a $5 bill or a $10 bill, but not both. Its utility function is simply the amount of money it has. It will choose to pick up the $10.

Suppose we want the robot to pick the $5 instead. We threaten to fine it $6 for picking the $10 bill. It will of course pick up the $5, and be better off than the net $4 resulting from the other choice.

Now suppose we give the robot the choice between being in the situation where it is free to choose unencumbered, and the one in which we impose the fine. It will pick the former, since in that situation it winds up with $10 and in the other, $5.

Suppose instead that we give the robot a choice between the unencumbered situation, and being "fixed" – having its utility function changed to prefer the $5 to the $10. It will choose the unencumbered situation for the same reason as before: it will gain $10 from that and only $5 from the other one.

It would be incorrect to think that the prospect of preferring the $5 *after* being "fixed" will make a difference to the first choice. The first choice is being made under the present utility function, which by stipulation was concerned with money only. In fact the logical form of the robot's reasoning is that of a two-player game, where the robot's first choice is its own move, and its second choice after possibly being "fixed," is the opponent's move. The rational robot will apply a standard minimax evaluation.

Learning and moral development

The existence of a learning capability in the robot's world model changes the dynamics of the third desideratum somewhat. The utility function needs to be modified to make a meta-judgement: how much effort to put into learning – improving the accuracy and reach of its world model – versus acting to increase its utility directly using the current model. Assuming this meta-evaluation itself has a learning component, i.e. the robot can learn from experience how much study is worth, it can make sense to punish the robot for unintended and unpredicted consequenses. In some sense, the robot is being punished not for doing the wrong thing, but for not knowing what would happen. Given the appropriate learning capabilities, the lack of foresight falls into the category of something the robot "could have done differently."

Finally, there is the question of "knowing the difference between right and wrong." There seem to be two general cases in the mainstream of thought. We punish children who do wrong to teach them the difference. We do not punish those who have such substantial mental disabilities that punishment would amount to cruelty without significant salutary effect. A strong case can be made that the dividing line follows the distinction in learning capability discussed above.

Summary

> *Freedom is only necessity understood.*
>
> –William James

It makes sense to treat a robot as a moral agent – to praise and reward or blame and punish it for acts good and bad – to the extent it has a cognitive architecture that allows such *predicted* consequences to influence its behavior. The rational agent cognitive architecture is one such. The robot must have:

- a world model with which it can predict the results of its actions
- a utility function which it uses to prefer one outcome over another
- a model of itself within the world model that is decoupled from internal causality
- the ability to incorporate promises of punishment or reward, or associations of past actions with punishments or rewards, into the causal model
- a control structure which weighs and chooses actions using the above

As with humans, children, and animals, the appropriateness of punishment or reward varies with what the recipient could be reasonably expected to anticipate or reasonably expected to be able to learn to anticipate.

Bibliography

Gazzaniga, Michael S. *Who's in Charge: Free Will and the Science of the Brain,* HarperCollins, 2011

Hawking, Stephen. *A Brief History of Time: from the Big Bang to Black Holes*, Bantam Dell, 1988

Hume, David. *An Enquiry Concerning Human Understanding*, Millar, 1748

McDermott, Drew V. *Mind and Mechanism*, Bradford, 2001

Wegner, Daniel M. *The Illusion of Conscious Will*, Bradford, 2002