POLITICA & SOCIETÀ
2/2014, 265-292

ISSN 2240-7901


# Determinism, machine agency, and responsibility[1]

Peter M. Asaro

Abstract

There has been some recent interest in robot ethics and machine responsibility, as well as the legal frameworks for judging the agency of robots and machines in questions of legal responsibility and liability[2]. In no small measure this interest has revived some of the central questions of philosophy of the past few centuries. Among them: the nature of determinism and the deterministic nature of algorithms, the question of free will and its relation to morality, and the relation of punishment to free will and algorithmic decision-making. While these questions have not garnered much attention in recent decades, the increasing interest in machine ethics, and moral machines[3] has raised the questions anew[4]. Among the central questions here are: If machines are algorithmic, and thus deterministic, how can they be responsible, moral agents? What does it mean to punish an algorithmically, and presumably deterministic, machine? And is simple non-determinism, of the probabilistic sort, a sufficient basis for asserting agency or ascribing moral responsibility? These questions excite interest

*Peter M. Asaro, Assistant Professor and Director of Graduate Studies, School of Media Studies – The New School, 79 5th Avenue – New York - asarop@newschool.edu.*

[1] This paper was originally prepared as a response to J. Storrs Hall, «Towards Machine Agency: A Philosophical and Technological Roadmap» (http://robots.law.miami.edu/wp-content/uploads/2012/01/Hall-MachineAgencyLong.pdf) and was presented at the 2012 *We Robot Conference on Law and Robotics* at University of Miami. I have attempted to make its arguments more general, by addressing commonly held views of computation, determinism and responsibility expressed in that paper.

[2] See: W. Wallach, C. Allen, *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press, Oxford 2009; J. Storrs Hall, «Towards Machine Agency: A Philosophical and Technological Roadmap», presented at the 2012 *We Robot Conference*, University of Miami Law School. Downloaded from: http://robots.law.miami.edu/wp-content/uploads/2012/01/Hall-MachineAgencyLong.pdf.

[3] W. Wallach, C. Allen, *Moral Machines*, cit.

[4] J. Storrs Hall, «Towards Machine Agency», cit.

in this new field, much as they do among students in their first philosophy class. It is important, however, to realize that philosophers did not simply throw-up their hands and give up on these questions. While such questions are rarely completely resolved or definitively answered, great intellectual progress was made through their consideration. And while these insights were made in somewhat remote areas of philosophy, it warrants a fresh look to see how those insights might be brought to bear on a contemporary quandary, such as the moral status of machines and robots, in light of our best scientific and philosophical understanding of determinism and morality. That is the aim of this paper – to shed light on a contemporary problem by drawing upon some insights from the philosophy of science and moral philosophy of the past century. The first section of the paper will review some of the leading approaches to thinking about physical determinism, especially those that grew out of thinking about physics in the early 20th century. This is particularly interesting in regard to its relation to our understanding of information theory, and bears quite directly upon our understanding of computation and algorithms. In the second part of the paper, I will move from this understanding of determinism in computation to how it relates to our notions of free will in moral responsibility, and its relation to ascriptions of moral praise and blame. In light of these discussions, I will conclude with a consideration of how we might understand punishment in algorithmic systems, such as machine learning systems and the moral agency of robots, and how this might bear on legal liability and responsibility.

Keywords: determinism; machine agency; robots; responsibility; punishment.

## 1. Determinism

As with most discussions of determinism, the best place to start is with the work of Pierre-Simon Laplace, which is the most straightforward, if extreme, view of physical determinism. According to Laplacian determinism, everything in the universe is determined. Let us think of a simple physical system, like a billiard table with an assortment of billiard balls. If we know the state of the system (the positions of the balls, their mass, the friction of the felt, the positions of the holes, the level and evenness of the table, etc.) and we know the causal inputs to the system (the angle and force of the cue stick, the shape and friction of the cue tip, and where it strikes, etc.), then we can calculate precisely how the systems will evolve after the shot is taken – which balls will sink, and where the others will come to rest. The Laplacian notion of determinism is that the entire universe is like our billiard table at a fundamental level. That is, if we knew the current state of the entire system in complete detail, then we could, in principle, calculate

the next state and all subsequent states from that, assuming also that we know the laws of physics and apply them correctly.

There are two obvious problems with Laplacian determinism. The first is that human knowledge is limited in various ways, so that we cannot know the exact position and location and mass of every particle, and every energy state, in the universe. On its face, this is an epistemic limitation, however, and not a metaphysical one. The universe is just too big, too complicated, and constantly changing for us to get a proper inventory of its current state. We might also worry that our favored physical laws and equations are not quite right, and that they are only approximations or contain inaccuracies. And it would be very impractical for us to attempt to calculate the next state of the universe, given its size and complexity, even with large and powerful computers to automate the calculations. Laplace's response to this, however, is that while we may not be able to know exactly what the universe will do in the future, we might believe that there is an all-knowing god or even the universe itself acting as its own best representation, that in some sense knows or stores its own state, applies the correct laws of physics, and is thus perfectly physically determined.

The second problem is that if this view were true, it would seem that we should also be able to run the universe backwards. That is to say that if we always get exactly the same result when we run the system over and over again, then those physical processes should be reversible. *I.e.*, there are no one-way or irreversible processes. Just as the present state determines all of the future states, so the future states should determine the present and past states. Indeed, we should be able to determine the entire past and future of the universe from any one state of the universe. Laplace is credited with this formulation of physical determinism in a *Philosophical Essay on Probabilities* (1814).

While this conceptualization of the physical world is quite clear and compelling, it becomes problematic in light of a great deal of philosophical and scientific work of the last few centuries that has observed and recognized the irreversibility of physical processes in a very fundamental way, as in thermodynamics. For example, when we stir the cool milk into the hot coffee they mix together in temperature as well as color, and we cannot easily separate them by stirring in the other direction, which just mixes them more rather than unmixing them. Even if we were to manually pick each bit of milk back out of

the coffee, we would not also thereby restore the coffee to its original temperature, that would require adding some energy, and the process of picking out the milk would require much more energy and information than the process of stirring. The rather simple and common process of stirring is an irreversible process, and our best understanding of the physical laws governing it stand in opposition to our understanding of the Laplacean universe.

The problem of stirring milk into coffee is not just a clever counter example to the assumptions required of physical determinism. Indeed, much of our understanding of information as a physical phenomenon, with real physical properties and limitations, is rooted in our understanding of thermodynamics and entropy. Recent research into quantum computing has challenged some aspects of our understanding of information and determinism, but has not overcome these fundamental limitations, so I want to start with a brief overview of quantum uncertainty.

It is worth reviewing the history of science surrounding the Copenhagen Interpretation of quantum mechanics put forward in the early 20th Century. And for those readers who are not students of physics and not familiar with this history, you might be familiar with Isaac Einstein's oft-quoted statement: «God does not play dice». This quote was Einstein's response to the Copenhagen Interpretation. The Copenhagen Interpretation is a way of understanding the sets of equations that were used to explain quantum phenomena in the early part of the 20th century. Quantum phenomena are a variety of observed phenomena where the energy levels of different forms of radiation do not vary in intensity continuously, or smoothly, but in steps, or quanta. These phenomena are only observable at very small scales, such as atomic scales, and the size of the quanta varies with the type of energy, with high-energy radiation such as x-rays having much larger quanta than lower-energy radiation such as infra-red. The interpretation was developed by Niels Bohr and Werner Heisenberg, and you may also recognize this interpretation as central to the Heisenberg Uncertainty Principle[5].

[5] For a good introduction to determinism in contemporary physics, I recommend S. Hawking's lecture «Does God Play Dice?» (http://www.hawking.org.uk/does-god-play-dice.html).

The Copenhagen interpretation holds that the state of a particle, such as an electron, is a probabilistic state – not in any particular state or other – but is best understood as a wave function – a distribution of potential states. Now, it is possible to "collapse" the wave function and determine some aspect of its current state, but doing so necessarily precludes being able to determine other aspects of its state. This means that the wave function is best understood as a probability function of possible states that the atomic system is in, but one cannot precisely determine the various aspects of that state, where the electron is, or what its spin is, without perturbing the system. That is, you have to actively interfere with the system, destroying some aspects of its current state in order to determine others.

A challenging feature of this is that your interference in the system is destructive – you cannot determine every aspect simultaneously and destroy some potential information about the system in order to gain some other information about it.

In order to obtain knowledge of this system, in order to extract information about the state of the particle, we have to touch it, to interfere with it. This interference is with some form of radiation, which comes in different quanta, but the more accurate measurements come from higher-energy radiations, which are more destructive of other aspects (*e.g.* position vs. spin). Which is just to say that information is not some magical non-physical property. Rather, information has physical causal requirements, and implications, and there is a physical cost to obtaining information that puts physical limits on our knowledge about the universe at a fundamental level.

Thus, the Copenhagen Interpretation states that it is, in fact, fundamentally uncertain what state the system is in, it is in a state of uncertainty. This uncertainty is absolute but not complete. Rather, our uncertainty is bound by the wave function, and even though we cannot know absolutely everything about a particle, we can make reasonably accurate predictions about its behavior based on our probabilistic models. But this also means there are things we cannot predict, like butterflies flapping their wing and causing hurricanes.

Further, work on General Relativity and black holes has led to the acceptance by physicists that information can be lost in the universe. That is, when light or objects fall into a black hole, the information they contain is lost forever to the rest of the universe. While this

happens to all the information that passes beyond the event horizon of a macro-scale black hole, there may also be micro-scale black holes which pop in and out of existence for incredibly brief periods as part of particle-antiparticle interactions, which nonetheless eat up random bits of information from the universe. The fact that information can be lost means that there are physical processes that are, in principle, not reversible. We simply cannot reconstruct the information that falls into a black hole.

The debate then between Einstein on the one hand, and Heisenberg and Bohr on the other, is whether this interpretation of quantum phenomena is correct or not. There was agreement on the best equations for explaining the phenomena – there were real limits on the information that could be obtained about electron states. The debate was whether the universe itself is fundamentally metaphysically uncertain, or whether it is in fact fundamentally deterministic and we simply cannot know epistemically what state it is in. Bohr's response to Einstein's familiar quote was, famously: «Don't tell God what to do». The Copenhagen Interpretation is now widely accepted among physicists. Of course, scientists can and should continue to challenge accepted theories, and various attempts have been made to promote hidden variable theories, which postulate that there are other hidden variables of which we are unaware, of which if we knew we would be able to replace the wave function with a deterministic function. There have actually been experimental tests which appear to disprove the hidden variable theories, at least for local variables.

At this point in the debate, we have reached the edge of the known and perhaps the edge of the knowable. We can never know who is right, if it is beyond our ability to know. So we can go on debating this, but we do not know, and cannot know whether the universe itself is fundamentally deterministic or non-deterministic. But we do know that our knowledge, as human beings, of any natural system or even in any real system that we design, is going to be epistemically limited in the same way. So, even if metaphysical determinism is true, we will never know what state the universe is in or where it is going. Effectively, we must treat the universe as uncertain, and indeed we do this in our physical models and simulations. We must thus accept epistemic indeterminacy, at least, or reject accepted physics. We must also accept that the physical properties of

information will have powerful implications for what is practically possible to know or compute.

This has several implications for computational simulations, which are actually quite crucial to robotics and artificial intelligence, information theory and other fields like cryptography. Something that comes out of the Copenhagen Interpretation when applied to thermodynamics, which is very interesting in this regard, is the concept of entropy and its relation to uncertainty. Indeed, in information theory, information gets defined as negative entropy. As I mentioned earlier, a deterministic system ought to be able to be run backwards as well as forwards. But for a physical system to be reversible, it would mean that there was no entropy in that system, as entropy is not reversible. Entropy is where the energy states of the universe decay, where order breaks down into chaos, information gets noisy, and these processes are not reversible; entropy is not reversible, according to the second law of thermo-dynamics. But if Laplacian determinism were true then entropy should be reversible, and so we should be able to unstir the coffee and milk that we stirred up earlier. The reason we cannot unstir our coffee is because of the properties of information and energy. We need all of that information about how to separate the coffee and milk and we have to capture that information somehow, we don't have access to it. And even if we did, all of this takes extra energy from the universe, and we have still transformed the energy states of our coffee and milk particles along the way, so they are not really in the same state if we just suddenly separate them again.

Consider what would be required to obtain the necessary information to unstir two liquids. And to avoid relying on other physical properties, let's take a glass of hot water and a glass of cold water, and pour them together and stir them. How could we separate the two glasses again, and get a glass of hot water and a glass of cold water? Getting access to information about the mixed particles requires, on one hand, intervention which then changes the world in the case of the particle. More importantly, if we look at Claude Shannon's conception of information theory, information does not just sit around, it is conveyed and transmitted[6]. The measure of information is how

[6] C.E. Shannon, W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana Ill. 1948.

much of the structure gets transmitted from one point to another, when that information is constantly decaying due to entropy. Thus, Shannon developed the error correcting channel, so that when we get the information wrong, we have an additional error correcting channel that will allow us to correct that, to bring our representation of the universe back into the state that we desire – but we need that extra information, and that takes extra energy.

Yet, we have to get that information from somewhere, it is not *sui generis* – it does not create itself or transmit itself, as these operations cost energy. This is part of our fundamental understanding of the physics, and is tied up into the equation, $E = mc^2$. The question of information in physics is also tied up in the transmission of light and other forms of radiation, which are the carriers of information. Light travels at a determinate speed (at least relative to its own frame of reference), but it also costs (and carries) a determinate amount of information and we can calculate that, and it is called Bremermann's Limit[7]. This is essentially why cryptography works, which is that you need certain time to run calculations, but also that encoding, transmitting and calculating information requires certain amounts of energy. So if we think that we would like to perfectly model a system – just a part of the universe – and we want to know every single aspect of every single particle and every single component of that system. Then that means we have to touch all of those particles in some way or other, and collect radiating information about each part, and we have to represent those in some physical form, a physical symbol system or computer, and all of this involves energy. The more complicated the system is that we are trying to represent or model perfectly and completely, the more energy we need and more time it is going to take.

Information is not free, it is not outside of physical bounds. Information is physical, it is an integral aspect of a physical world. We are constrained by that, our models and simulations and computations are constrained by that. Which also means that if we wanted to, for instance, compute everything about the planet Earth in absolute physical detail, we would need a computer at least as big as the Earth

[7] W.R. Ashby, «Some Consequences of Bremermann's Limit for Information Processing Systems», in H. Oestreicher *et. al.* (eds.), *Cybernetic Problems in Bionics*, Gordon and Breach, New York 1968, pp. 69-76.

to do that[8]. So, we have these very practical constrains on our ability to actually enact or simulate a deterministic interpretation of the world computationally. Even if we believe that the world is deterministic and that we would like to collect "all" the information about it, and to then model and simulate it and run it through the computer to see what is going to happen, we would still need to have enough computing power to do that, we have to collect all the information to do that. And each of these are deeply problematic and physically impossible. Even if we get our quantum computer right, which is why quantum computing is very seductive, and are able to simultaneous compute over multiple probabilistic alternatives, we can still only do this for very simple systems.

This is something that is very important to remember about computational models and simulations. They are very useful, precisely because they *leave out* a great deal of information. What we are really doing, most or all of the time with computations, is approximating the world. We are constructing simplified models, for specific purposes and interests. Our best scientific models are not absolute models, but limited, human models. This is the nature of human knowledge, and information becomes knowledge by being highly selective in leaving out what is not important. So, while in some cases we could develop more sophisticated higher-resolution models, we could include more variables, we could include more of the initial states and conditions, and we could try to force our model to behave as a deterministic system, we know that most systems are more usefully and more accurately modeled probabilistically and non-deterministically. Indeed, even when we start modeling large deterministic systems, we find that they begin to have unpredictable results anyway. Given that we know fundamentally that the universe is uncertain, it is probably better to stick to non-deterministic models.

Additionally, there are issues about open and closed systems. Closed systems are those that we assume have no inputs or outputs

[8] According to Bremermann's Limit, the maximum computation possible from any piece of matter is $1.36 \times 10^{50}$ bits per second per kilogram. Thus, a computer with the mass of the entire Earth operating at the Bremermann's limit could perform approximately $10^{75}$ mathematical computations per second (http://en.wikipedia.org/wiki/Bremermann%27s_limit).

to a larger system, such as its environment. Of course, there are no truly closed systems, this is just a strategy for simplifying our models. So for a system to be causally deterministic, it also has to be a closed system, which means that we have to be able to bound the system, which means that we have to know the entire universe. Otherwise it is going to be influenced by something outside of itself and those influences have to get modeled or they are considered inputs that would then change the behavior of that system. As observers, whether we are looking at electrons or whether we are building models, we have to take responsibility for the definition of the systems. We ultimately must decide what the boundaries of a system are. We decide what its inputs and outputs are, what its states are, and so forth, and there are different ways to do that. That is the practice of representation, the practice of modeling and simulation.

So, if we look at the robot, and we want to model it and its behavior, there are various ways to do this. We can model it physically, we can try to look at its software, we can try to look at its behavior, we can run a robot simulator which models both its software and its structure in a physical simulation. There are different levels of analysis we might approach it at, and we have to choose a level of analysis that we want to model. We have to choose what the operating conditions of those levels of description, and the factors that we want to bracket off as external and outside the simulation. That is our responsibility as observers ad modelers; we affect in these ways the interpretations of the system and how it will behave, and whether we read that system as deterministic or non-deterministic, as probabilistic or non-probabilistic. We choose what variables to use, what they represent and how those variables operate in our descriptions of those systems.

Moreover, in simulating formal systems, such as a computational model, there is also maintenance. So, we might wish to say that a computer is a deterministic system, but this only sounds convincing when the computer is working properly. We actually have to design sophisticated circuits inside of the computer that constantly try to ensure that all the electrons are going to the right places, that memory registers with ones stay ones and zeros stay zeros. Because in reality all those electrons are in myriad probabilistic states, power supplies vary, cosmic radiation bombards silicone chips and the electrons misbehave with some frequency. Following Shannon's law, we have many error

correcting channels built into our circuits and chips, checksums and such-like that, to ensure that we are maintaining a correct representation. So while we describe computer logic in terms of platonic formal systems, the reality is that we are constantly channeling and disciplining material reality into performing as if it were a well-behaved formal and deterministic system.

And of course these systems crash a lot, bugs are induced, memories overflow. Many things happen inside of the physical computer that are not causally determined by the "deterministic" program. When we think about this in terms of models of the world, this is entropy, this is decay. Information is the preservation of form, of structure, so that is why you have to have an error correcting channels which again takes further information from the world, which requires additional energy, but this is the continual process of taming uncertainty and disciplining computing machinery that we go through to create computational simulations of so-called deterministic systems.

I hope the preceding discussion has convinced you that, while deeply appealing to many people, physical determinism is, according to the best available science, most likely false. And even if it were true we can never know that, not that it would make any difference in how we treat the universe in our scientific theories and computational models. While increasing the resolution of computational simulations can sometimes improve them, simulations actually derive much of their value from what they leave out of the simulation – by identifying only the essentials. And finally, that the notion of treating computers, the quintessential information processing machines, as deterministic systems is just such a simplification – a convenient way to think about them for practical purposes. The reality of physical computers is one of instability and indeterminacy, with engineers working at every turn to stabilize and discipline these machines into behaving as if they were following the programmed deterministic logic we desire. And thankfully, as our engineering improves this convenient fiction becomes easier to accept as our systems become more reliable.

My aim at reviewing all of this really comes down to releasing those who might be caught up in the illusion of determinism. And the conclusion I want the readers to draw for themselves is that we should not be terribly concerned about the world being deterministic. If you are still not convinced by these arguments, or you really believe

that the world is in fact deterministic and that it's just an epistemic boundary, or perhaps that we will have the quantum computers and quantum robots that will overcome these physical limitations of information processing computation today, or that we are on the verge of a scientific revolution that will revise the basic laws of physics, then there is nothing more I can say to convince you. But even if you believe this in your heart and accept that practical progress often requires proceeding based on practical simplifications, then you should still be convinced that science, engineering and philosophizing ought to proceed as if the world were not deterministic. And thus, when we talk about deterministic systems, we do not mean this literally, but only that we find useful and convenient to treat certain systems as if they behave deterministically for practical reasons.

It is with this in mind that we turn to consider what determinism has to do with free will, once we accept that there is no determinism. We can also examine what implications this has for understanding the relationship between causality, free will and moral responsibility.

## 2. Free will and determinism

The traditional problem of free will, as developed and debated for centuries beginning with medieval theologians and philosophers, is that one cannot be held responsible for acts for which one is causally determined to have done. There is a sense in which one must choose to act in order to be responsible for the act. This is reflected in criminal law, where there is a requirement for *mens rea*, the guilty mind, to accompany the guilty act. This mental choice or intent is what separates *e.g.* involuntary manslaughter from first degree murder.

Following Descartes in the modern era, as mechanistic theories of the mind and later neuroscience, which treated the mind/brain as itself a deterministic mechanical system, the problem was reborn in a new more powerful form. Our actions may be determined by our thoughts and choices, but those are in turn determined by other thoughts and mental states. If thoughts are just another physical cause and we are not free from our own thoughts which are determining our actions, perhaps we are not responsible for our actions even when we choose them.

Currently, the received view in philosophy regarding free will and determinism comes from Strawson's work in the paper «On Freedom and Resentment»[9]. In that paper, Strawson turns the ancient debate about free will and determinism on its head by saying that in fact the reason why we ascribe the moral agency is not because we make a judgment about the theoretical determinism or indeterminism of a system. Instead, the reverse is true, and we start with an ascription of moral agency and then begin constructing a model of agency and determinism in order to understand somebody's actions. We are constantly in this process of analogizing, using metaphors, and trying to understand a person's situation, their options and their thinking. We often try to put ourselves in the same situation to decide what we would have done if it had been us. And this is, in fact, how we decide whether somebody is responsible for their actions or not, whether they had agency, whether they could have done anything differently given their situation.

Making ascriptions of moral agency is not a formal process. It is very much a psychological process and depends on intuition and experience, but it also utilizes sense of whether this person is acting in good faith or bad faith, what were their intentions, goals and desires, and other aspects of their mental life, which become much more relevant for morality and law, as we will see.

It is not straightforward to see how this might apply to ascriptions of moral agency to machines. On the one hand, we might argue that to the extent that we can relate to them and their goals, then we can attempt to apply our usual techniques. But to the extent that computers and robots are alien or unrelatable, this may not work. Moreover, the fact that we program and control them, and thus sometimes have a unique sort of access to what they are thinking and why they are acting as they do, this may also influence our ascriptions in ways that do not conform well to the way we ascribe agency to humans or animals.

[9] P.F. Strawson, «Freedom and Resentment», *Proceedings of the British Academy*, n. 48 (1962), pp. 1-25.

## 3. Material agency

While I think it's useful to think about determinism, we should always keep in mind that this universe is probabilistic and we should not worry about it too much when it comes to moral agency. What we should worry about is how we are ascribing moral agency to the machines, people, and to other agents.

There are other theorists who have attempted to understand non-human agency and its role in shaping knowledge, society and technology. In the field of science and technology studies, there has been a long standing discourse on the nature of non-human agency. Bruno Latour describes assemblages of humans and non-humans working together in terms of networks of actors, where actors can be any element of the system that asserts agency – people, animals, machines, nature, scientific phenomena, *etc.*[10]. Latour specifically choose the term «actor» or «actant» to get away from the loaded conception of agency, because agency carries the weight of intention and intentionality. When the fisherman attempts to enroll the scallops in their political maneuverings, the scallops do not really intend much of anything, but they do become causal agents, and social actors (actants) in network assemblage that emerges and tries to stabilize itself. Thus the scallops «do things» or «fail to do things» apart from any intentions. So in a causal sense, they have agency, though in a moral or legal sense they do not. This motivates Andrew Pickering to use the term «material agency»[11]. And just as I described the efforts of engineers to discipline the computer into realizing a formal computational system, Pickering examines scientific and technological development as the disciplining of material agency into forms that can be reliably counted on to perform in predictable and useful ways.

Latour's infamous example is: «Guns don't kill people, people kill people», the slogan of the National Rifle Association. The implicit claim of the slogan is that it is the human agents that are responsible for killing, and not the material agents, the guns. Legally and morally,

[10] B. Latour, «On Technical Mediation», *Common Knowledge*, n. 3 (1994), 2, pp. 29-64.

[11] A. Pickering, *The Mangle of Practice: Time, Agency, and Science*, The University of Chicago Press, Chicago 1995.

this seems compelling, as we would not hold the gun responsible for who or what gets shot. Yet, it also seems to miss the point that guns are in fact quite dangerous. In his analysis Latour points out that it is not really a simple binary – that people or guns are responsible. In fact it is an assemblage, person+gun, that kills people. The person+gun is different to the naked person. When you put a gun into somebody's hand you have redefined your system, or have a new hybrid system. Clearly, a person+gun can be much more effective and efficient at killing than a person alone. And this has implications, for instance the police are going to behave very differently towards you if you are a person than if you are a person+gun, including they might shoot you. And so the gun has a real causal efficacy, in terms of what the assemblage is capable of, and as the element that changes the nature and interpretation of the overall system there is real agency there. The gun is having an effect on the world, even if it does not intend to, or has no intentions of its own at all. It has become part of a system and that system is capable of things that its components alone are not. There is some difficulty in identifying the source of agency in such socio-technical hybrids, and we still tend to ascribe agency primarily to the human agents.

One difficulty with agency is, of course, intentionality and so I want to add my little bit which is that people+robots+guns can also kill people. And as we add more elements to the system, in this case the robots, which have not only causal efficacy carrying and firing weapons, but also potentially decision-making – the traditional domain of intentionality. And beyond the capability of making decisions towards achieving a goal or plan, others have also suggested that these systems might learn, adapt, evolve and otherwise develop unpredictably or autonomously, and this might represent another level of agency or even intentionality. Before addressing this question directly, I want to turn for a moment to the other side of agency, which is that of responsibility and punishment.

## 4. Punishment

In thinking about ascribing moral agency and responsibility, it is helpful to consider why we punish people, when we find them respon-

sible for wrong-doing. The reasons turn out to be more complicated and interesting than we might assume. There are, in fact, multiple reasons for punishing people and they can even be at odds with each other. As a society, we do not even need to agree on the reasons for punishing people, and in practice the law and moral judgment often mixes and blends these reasons together. If we try to separate them out, the main classical elements are retribution, deterrence and reform.

The notion of punishment as a form of retributive justice has roots in ancient law, *e.g.* «An eye for an eye, a tooth for a tooth», but also has a modern formulation. *Retribution* is the idea that your violation of a law as an individual creates a debt to society. You have in some sense taken advantage of everyone else following the law without following it yourself and you have taken an extra privilege against society at large. Under the social contract, we enter into a law-bound society to gain certain protections from harm and we agree not to harm other members of society. For instance, I am not supposed to steal your stuff, and you are not supposed to steal my stuff. If I steal your stuff, I have clearly harmed you, but I have also violated the law. So it is not enough for me to simply return your stuff – that would be the straight liability or tort. Theft is not a tort, but a crime. For an act to be a crime there must be criminal intent, and the punishment goes beyond the monetary value of the theft because there is more that is due to the society whose laws have been broken. While there might be a fine imposed, the fine is paid to the state, not directly to the victims, and other punishments such as imprisonment do not directly benefit the state or the prisoner at all.

In the era of psychological behaviorism and social engineering, *deterrence* has emerged as a principle function of punishment, at least in the framing of many laws.

As a function of punishment that aims to prevent future crimes, deterrence works on two levels. The first is the individual causal level of deterrence when we imprison people or use capital punishment, and actually removing agents from society or prevent them from acting for a period of time. If you acted badly, we do not want you in society doing more bad things, so we are going to remove you form the society (*e.g.* exile, capital punishment, imprisonment). And this is more and more becoming the justification for a lot of extended sen-

tences and zero-tolerance policies in the current political rhetoric – to keep all the bad apples off the streets.

The other level is social psychological level of deterrence, wherein we aim to alter everybody's future actions by placing a negative cost on taking wrongful actions. Because we can recognize other people's mistakes and witness the punishment they receive for their transgressions, we are all meant to learn from the action and think twice before doing what they did. This was in part why punishments were often a public display – to demonstrate both that the state would catch you, and that the consequences would be bad – whether it was the pillory to the gallows. Confinement serves mostly the first aspect of deterrence, to prevent the individual from acting.

In the last few decades of social engineering, the most salient and discussed purpose for punishment is *reform*, wherein the intention is to change the character or behavior of the person who has done something wrong. Reform has multiple interpretations, depending on what we think is wrong with the person who does wrong, and thus upon on our moral theories. If we are utilitarians, we might think that the wrongdoer has miscalculated the values of certain actions, or failed to consider the costs to others and only considered the benefits they might receive. What we need to do in such cases is revise people's utility functions. For instance, we could impose monetary penalties and, in terms of economic decision making, the rational person will recognize the additional costs of getting caught and being punished, and thus avoid choosing illegal actions.

We could also apply virtue ethics here instead. In this case, our aim is to train people to internalize our moral and legal structures, and reform aims to provide the moral education that might have previously been lacking. This notion also has important implications if we think about automating law enforcement: Is it sufficient just to get people to obey the rules or do we want them actually to understand why those rules are there and to internalize those rules as members of society? And this begins to get at other notions of virtues, self-realization and autonomy, especially when we think about children, why we punish children, and how this differs from adult punishment. We want children to learn the specific lesson and not repeat their behavior, but we really want them to learn a deeper lesson as well. We want them to become a better person, and that is about internalizing the reasons

why the rules are in place and that they should not violate them even for a desired advantage, or when it is almost certain they can get away without being caught or punished in a particular situation.

And we could also approach reform as Kantians. In this case, we would want the wrongdoer to recognize their duties, to respect the rights of others and to internalize these duties and rights into their decision-making as they act in the world.

## 5. Robots, reinforcement learning and punishment

It is important to keep these different variations on the notion of reform and its distinction from deterrence and retribution, as we consider various proposals for the punishment of robots, such as that of Storrs Hall[12]. According to that proposal, we are meant to consider a robot that is capable of sophisticated decision-making and also capable of learning. We are further asked to consider that punishments of this system are meant primarily or exclusively as reforms of the system, aimed at improving its future performance and actions.

The first thing to note here is that a robot following a utility function is only deterministic in a qualified sense, not in the metaphysical sense discussed above. We can treat a system that implements a rational decision function as being logically deterministic and expect it to make the same decision given the same inputs. Apart from building in a randomizing function, we are not really dealing with «magical free will» in such cases[13].

In reinforcement learning the idea is that if the robot makes the wrong choice and you want it to make the right choice in the future, you need to change its decision structure. Typically what we want to do is to change the probabilities of making a certain decision or to change the values placed on the outcomes, so that we re-weight the decision process and the desired outcome becomes more likely or guaranteed next time. There are many technical problems with the actual implementation of such learning. Among these is the temporality problem. If you have a robot that has made a very complicated

[12] J. Storrs Hall, «Towards Machine Agency», cit.
[13] W. Wallach, C. Allen, *Moral Machines*, cit., pp. 59-63.

sequence of decisions, *e.g.* played a long game of chess and then lost. What do you change? How do you decide which move lost the game? Or what was the responsibility of each individual move in the overall sequence of moves that lost the game? How do you decide how you are going to re-weight that whole chain of decisions, based on one outcome at the end? This becomes a really difficult problem, computationally speaking.

With a well-constrained system like chess we can try to deal with that formally and we have additional information, such as looking at multiple games, looking at multiple alternatives for each position, things like that, to try to determine more accurately where revisions should take place. It is not a straightforward problem at all, it is a very complicated problem even in a formal closed system like a game of chess. And then when we add to the fact that chess is not a probabilistic game in the sense that every state is determined or deterministic and the other player is choosing moves based on different probabilities for expected outcomes. We could further try to model the probability functions and strategy of our opponent, as those diverge from our own model of an ideal player, which raises a whole new set of issues. As we start to consider our robot taking actions in a world with many agents, in which the options are not always clear, figuring out where our decisions might have gone wrong gets even more difficult. A related problem arises in the punishment of another type of non-human agent known – as corporations, as I will discuss below.

Things start to get really interesting when we consider what might happen if we programmed the robot to actually reflect on this problem for itself. Storrs Hall uses an interesting example to get at this. The examples involves training a robot assistant and raises the question of how we think about punishment and its relation to revising the utility functions that form its algorithmic decision-making. For the sake of clarity, here is his full description of the example:

> 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 or 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[14].

I think this equivocation of punishment and fixing is mistaken and reductive. It is reductive not only because it reduces punishment to a particularly narrow interpretation of reform, but also because it takes a very narrow interpretation of how we should think about representing decisions to change our own methods of making decisions. But it will take a bit to explain why I see it this way.

Storrs Hall's example corresponds closely to what are called Ulysses' problems in decision theory. The story of Ulysses is that he wanted to hear the song of the sirens but he knows that when you hear the song of the sirens, it is so seductive that you are going to steer your ship into the troubled waters and sink. So he tells all of his sailors to plug their ears with wax and tie him to the mast. And he also tells them that when they sail through, and he hears the sirens' song and begs his crew to untie him from the mast or listen to the song, they should ignore his requests and orders. In this situation Ulysses has a certain set of probabilities, values, desires, *e.g.* he does not want to crash his ship, but he does want to hear the sirens' song. But he also knows that in that future point in time, when he is listening to the sirens, he will be willing to crash his ship to get closer to them and their

[14] J. Storrs Hall, «Towards Machine Agency», cit. p. 4.

song. Thus, he knows now that he is not going to be rational at that moment in the future, that he will have a different utility function, and that the one he has now is better or more desirable in the long run than the one he will have then. And so he has himself tied to the mast to prevent himself from acting under the irrational utility function.

The Ulysses problem in decision theory deals with the meta-choices over different utility functions, and are fascinating for a number of reasons, especially with explanations of the decisions made around drug addiction. If you know you are a drug addict and you are going to make poor choices when you are under the influence of a drug, then what is your rationality towards decisions to take a drug or not, when you know that it's going to change your sets of values and your rational deliberation in a future point in time? But it also applies to education and learning. When we decide to pursue an education, such as a college degree, we do not really know what exactly we will learn or how the educational experience will change us and our utility functions. In this sense, going to college is just as irrational as taking addictive drugs. Of course, we have institutions and social values which aim to ensure the positive value of public education, and these are not the song of the sirens or mind-altering drugs. The point here is that it matters considerably how we interpret the situation in which we revise or decision structures. Not all revisions are good, nor are they all bad, and often we do not have sufficient means for judging these in advance, and less often we are able to choose them freely.

Generally, in reinforcement learning, what we are doing is very small tweaks to our model of the world, the utility functions. We can change the values for how we evaluate certain outcomes and we can make some outcomes more or less desirable. But we can also change probabilities – our expectations for how the world will behave – and this is what a lot of scientific understanding is about. Knowledge aims to develop better understanding, better estimates of the probabilities of certain outcomes in the world given certain conditions. In our uncertain and probabilistic universe this is very important.

There is another way to approach the revision of utility functions, pointed to by Storrs Halls' reference to the minimax solution, which includes strategy and risk aversion and is operative in multi-agent reasoning and game theory. We can have estimations not only about our utility functions, but also models of our opponent and their decision

structures. So we have to weigh our own uncertainties there but also our opponent's probability estimates of outcomes. But we have an additional uncertainty about whether we have an accurate representation of their probability estimate about outcomes and their aversion to risk. Do I really know when the Russians are going to launch their nuclear missiles or not? Are they using the same utility function that I would use if I were them? Or do I think they are fundamentally irrational in a particular way? This becomes very complicated, moreover, because I can have different relations to risk: I can be risk averse or I can be risk accepting. I may take a big risk for a big outcome. Or I may decide not to do it. And I can change that, I can change my risk aversion. When do I decide to change my own risk aversion? That is a whole other issue.

The suggestion in the passage above is that the robot can play this multi-agent game with itself, running its alternative utility functions against each other to choose the winner. But this is problematic. First, it is not clear what it means to play a zero-sum game with oneself. The challenge of game-theoretic problems such as the prisoner's dilemma is that one is not able to communicate with the other prisoner to establish cooperation. The successful tit-for-tat strategy in repeated games is essentially a mode of communicating the intention to cooperate. What would it mean to cooperate with oneself in such games? Moreover, the notion of communication points to the problem of access to information. How is the robot meant to access the information about its future self, the results of its future decisions and the influence of its learning on those decisions? For any non-trivial learning, it necessarily lacks access to that information.

And this brings us back to our earlier discussion of the reversibility of functions that is required by determinism. In this case, we can imagine that the procedure by which we revise our utility function is reversible, and thus can both be reversed in practice, but also modeled by us externally. And if we can model this externally, then we can also imagine a more sophisticated robot who can also model this, and thus choose for itself if it wants to revise its utility function or not. But in practice, such processes are not reversible.

For any sufficiently sophisticated learning algorithm, learning amounts to the encoding of experience as an inclination to behave in a certain way in the future and is a form of data compression. All

statistical reinforcement learning algorithms are essentially processes for generating compressed representations of behavioral outputs over sets of inputs. This process is what is known in data compression as a lossy process. That is, there is a loss of information in the process of encoding, making it irreversible in the sense that you cannot recover the full resolution of the original from the copy. The loss is desirable in the sense that the goal is to create a compact and efficient representation rather than store every possible input and treat it individually. But it also means that learning cannot be undone, at least after further learning has taken place. Once you have learned from two different experiences, it becomes difficult or impossible to separate which experiences were responsible for which aspects of your current representation or utility function. Given all of the historical states of the system, we might claim that this could be reconstructed and reversed. This is really the same problem as physical determinism – if we know the whole history of the universe and its states, it becomes deterministic and reversible. But the reality is that the learning process is entropic, and irreversible. And as much as we would like to track all of the states, inputs and outputs of a system, we can only do this for very simple or trivial systems. For any complex systems we very quickly run into Bremermann's Limit and require far more information than is possible or practical to manage. Indeed, the whole point of learning algorithms is to compress this sort of information.

## 6. Creating choices

There is a deeper problem here and this is where I think we have to think about robots and learning at a more fundamental level. It gets into the question of when we might start treating robots and machines as agents, who might be punishable. This is about changing their model of the world, on the one hand. So, to go back to this (visual), all we have been talking about so far is just changing numbers on this graph. Or changing my relationships to probability outcomes and risk in that graph, but the graph itself has not changed. I have neither added nor deleted any option from my set of choices. My interpretation of the world has only changed in terms of how I have estimated the objects in the world and their probabilities and their values. But I can change

the world itself, I can introduce new entities in the world, I can create new alternatives and options for choices that I can make and those are going to have their own probabilities, outcomes and values.

How do we do that? It's a very creative process and that is what happens when we are learning and when a child goes from a two year old that we would not want to ascribe full moral agency to, to being an eighteen year old that we do ascribe moral agency to. Just because they understand a lot more things about the world, their world is far more sophisticated. It is not just that we have adjusted some utility functions on the world of the two year old. And moreover, they create new alternatives, things that their parents, teachers and society did not foresee or provide. In creating these alternatives, if you are a Kantian, you are creating their own moral autonomy, deciding who you want to be by how you see the world and how you make choices. This is about creating a model of ourselves to the extent that we recognize the virtues, values or ethics that we are internalizing as a part of self-knowledge and self-discovery and assertions of our autonomy, of taking responsibility for our actions, instead of just conforming to a behavioral outcome, following the rules. We can follow the rules without believing in them, or we can follow the rules by believing in them and that is different, it is a different experience.

When it comes to punishment, the tricky bit is really intentionality. Let us go back to the question of material agency, and if the robot does something wrong, we will fix the robot. Is this a punishment, or is this just repairing the robot, and why is it important to distinguish these? It is helpful to think about the law of liability and torts here. There are cases where human action brings about damages, or where human inaction with regard to responsibilities or property can cause damages, but there is also the situation in which nature can do things. Bad things can just happen in the world and cause damage to property, like hurricanes, and nobody is responsible for that – we might call it an act of nature or an act of God. Now, if I own a robot, and a robot does something wrong, it's no longer an act of nature because I have this material and legal relation to the robot that makes me liable for its actions, but I am not necessarily intending for the robot to do all the things that it does. But torts and liability is quite capable of dealing with unintentional harms. And if the damages are intentional,

then we are talking about the guilt and culpability of the owner of the robot, and we move into criminal law. This is the Latin *mens rea*, which is that you have to have the guilty intention in order for the act to be criminal. There are some difficulties in the context of criminal negligence because the omission of acts can also be guilty and it is a bit peculiar that you can be guilty for not doing things. This notion of guilt goes back to duties and the understanding that you had a duty and when you fail to enact that duty.

One proposal, from Daniel Dennett, is that when a robot becomes so complicated that it is a better authority on its own internal states then we are, as external observers, then at that point we should treat it as its own agent. His argument is also epistemic in the sense that it becomes easier to model and predict the behavior of such a complex system by treating it as an agent[15]. I think this is an interesting proposal, but I do not think it quite gets us closer to the point of understanding punishment. Mostly because it does not bring us to thinking about the construction of new alternatives and, really, autonomy.

Even with complicated technologies that do no exhibit intention-like behaviors, we find ourselves in situations of great uncertainty. Right now I cannot figure out the internal state of my laptop computer and there are limits on even what the best engineers in the world, with great amounts of time, effort and energy, could actually figure out about the internal state of this machine and why it crashes occasionally, which it does. But that means that the system is, in some sense, the better reporter of its states than any of us are. It certainly has material agency because it messes me up a lot by crashing, but I am not willing to give it moral agency. I am not willing to say it is guilty of a crime or malice when it crashes.

We might also consider the second-order responsibility – that is my responsibility in relying on this machine which sometimes fails. If I am tasking a robot to do something, I need to have a reasonable expectation of probabilities predicting its actions and things like that, and that does make me responsible as its supervisor. If I am not being able to predict the behavior of something, then I should not be using it to do those things, but then we are just back to a straight kind of li-

[15] D. Dennett, *Freedom Evolves*, Viking Press, New York 2003.

ability, we are not really talking about the system having agency unless it goes on a frolic of its own, and that is a different question[16].

If we consider other aspects of punishment, retribution, then we are also concerned about the intention behind the actions and not just the consequences of the actions. Is it very provocative to think about what constitutes malice in a robot? This has to be something more than just the harm that is caused by a robot, but also a specific intention that its act causes harm. Such a robot would have to be able to represent the world, to represent the agents in the world, to represent itself in the world, and have a moral model of itself and other agents in the world. It would also have to recognize that if it does a certain act, it is violating its own morality by its own model, and recognize that it wants to do that anyway and chooses that at some level. This is very complicated to think about in terms of a robot, but that is the kind of direction we might be going.

We also have some other notions of how to punish non-human agency. Corporate punishment is the most obvious. Corporations are non-human agents, and if we look at John Coffee's work on corporate punishment, what you are really trying to do is to influence the decision making processes of corporations[17]. Ultimately you want to punish a corporation so that they do not do the same thing again. But it turns out to be very difficult to apply a punishment that effectively changes the decision process of a corporation because corporations are complicated distributed systems. Coffee actually uses the administrative behavior work of artificial intelligence pioneer Herbert Simon, for which he won the Nobel Prize in Economics[18]. If you want to influence organizational decision making externally, you have to know the structure of the organizational and you have to know how decisions are made on different levels of the organization, so that you can apply incentives and penalties that will influence the values of

[16] P. Asaro, «A Body to Kick, But Still No Soul to Damn: Legal Perspectives on Robotics», in P. Lin, K. Abney, G. Bekey (eds.), *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press, Cambridge MA 2011, pp. 169-186.

[17] J. Coffee, «No Soul to Damn, No Body to Kick: An Unscandalized Inquiry into the Problem of Corporate Punishment», *Michigan Law Review*, n. 79 (1981), 3, pp. 386-459.

[18] H. Simon, *Administrative Behavior: A Study of Decision-making Processes in Administrative Organizations*, Free Press, New York 1947.

those decisions and their utility functions there. But this can be very difficult to determine even for agents inside the organization, much less from outside the organization.

Practically, you can punish the whole corporation by affecting their bottom line, taking money from them, but that doesn't necessarily influence the people who made decisions, the managers, the board of trustees, the CEO, the various people who made those decisions that lead to the wrongful action. Specific individuals might get fired if held responsible by the corporation, or lose their own money if they own shares and those shares lose value. As a matter of legal recourse, unless those individuals have committed an explicit crime, it is difficult to hold them individually responsible, and they may be soon replaced by another person. So you are left with the rather blunt instrument of fining the whole company, and it is very hard to target these processes and the decision making. This is a problem for corporate decision making, and it is a problem for punishing robots too which is this question: how do you target, the thing you really want to change when the robot does something you do not want it to do again? At least with corporations, they exist essentially to make money, so taking money away from them actually hurts them, because that is their fundamental reason of existing.

Robots may not have something comparable to the profit motive. They may lack money or even the interest in acquiring it. Moreover, they may not have a fundamental reason for existing. Different robots may have different fundamental reasons for existing, depending how we try to construe that from their programing and operation. What does it really mean to punish them effectively, if they don't care about money? If they do not care about their physical presence or state? The famous statement «No body to kick, no soul to damn», is what the Lord Chancellor said about the corporations in Britain in 1800. While robots have a body, it is not clear that putting them in prison, or killing them, or ripping of their arm, or any other corporeal punishment is really going to change their minds about anything or effectively punish them if they do not care about their bodies. And if they do not really care about anything, how do you punish them at all?

In order to address the issues before us, and the way of going forward, the crucial things we need are theories of punishment, agency and responsibility that apply to these complicated systems, systems of

humans and machines working together. Legal and moral theory is focused on individuals. Legal theory deals with organizations, but it still treats these as individual as legal entities and it is always trying to find individual agents to hold responsible, which is why it is convenient to treat corporations as persons because the law cannot really deal with the complexity of an organization.

The consequence of this is that we end up with organizations and systems that are increasingly designed for irresponsibility. When we design a system, we have the choice to design it in a way that no individual has a clear responsibility for it. Even though that system might cause harms to others, or to the environment, yet nobody can be held responsible. We can look at the recent stock market collapse and real estate bubble and nobody goes to jail. Great deal of wealth transferred, everybody feels that it was a great injustice, except the people who received the wealth, but there is nobody to point the finger at, to say you broke a law, you violated the public trust, you did something illegal, you are responsible. No individual is responsible, but at a certain level the system is responsible for what happened and needs to be reformed.

Alternatively, we could design systems that insist on holding individuals responsible. Often, in large organizations this can result in scapegoating – punishing those who in fact had very little or no responsibility. It can also result in designating certain officials as being responsible, without necessarily providing them with the means to assert that responsibility.

Neither solution is very satisfying, and what is called for are a new set of alternatives. Theories in which responsibility and agency can be meaningfully designed and shared, so that large organizations of people and machines can produce desirable results and be held accountable and reformed when they fail to do so. And I think that it is something to keep in mind and that we need to address more generally in legal and moral theory.