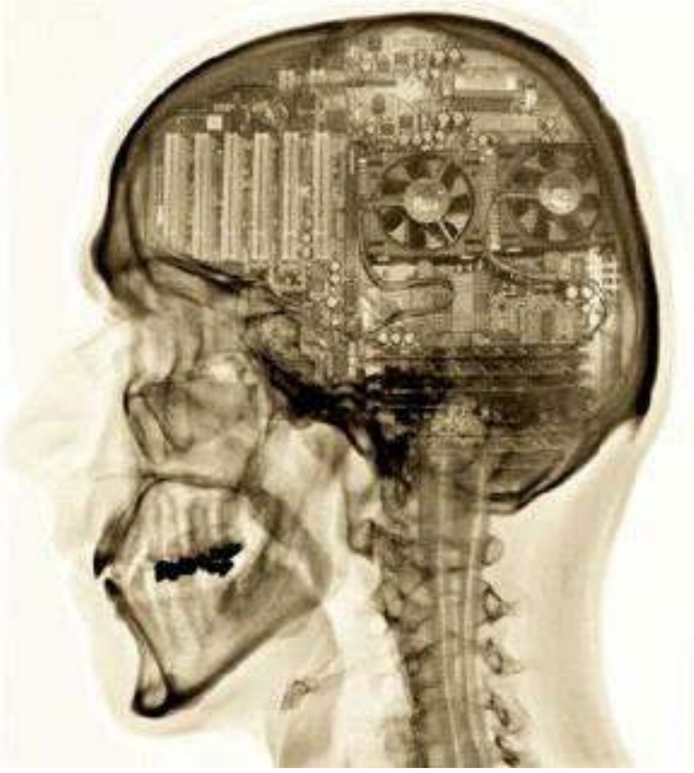
Is Regulation of Artificial Intelligence Possible? - h+ Media

authors John Danaher



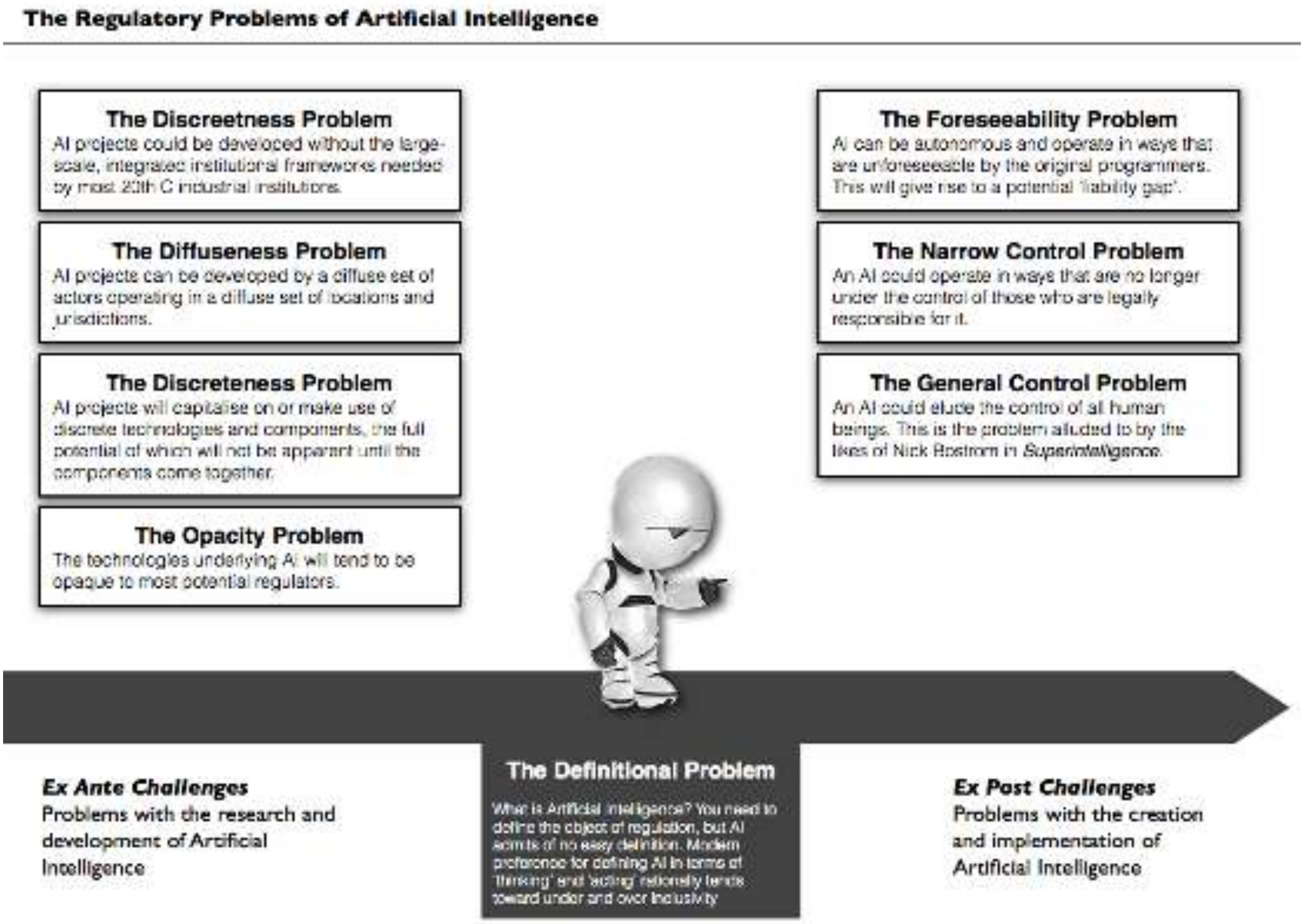
The halcyon days of the mid-20th century, when researchers at the (in?)famous Dartmouth summer school on AI dreamed of creating the first intelligent machine, seem so far away. Worries about the societal impacts of artificial intelligence (AI) are on the rise. Recent pronouncements from tech gurus like Elon Musk and Bill Gates have taken on a dramatically dystopian edge. They suggest that the proliferation and advance of AI could pose a existential threat to the human race.

Despite these worries, debates about the proper role of government regulation of AI have generally been lacking. There are a number of explanations for this: law is nearly always playing catch-up when it comes to technological advances; there is a decidedly anti-government libertarian bent to some of the leading thinkers and developers of AI; and the technology itself would seem to elude traditional regulatory structures.

Fortunately, the gap in the existing literature is starting to be filled. One recent addition to it comes in the shape of Matthew Scherer's article 'Regulating Artificial Intelligence Systems'. Among the many things that this article does well is that it develops the case for thinking that AI is (and will be) exceptionally difficult to regulate, whilst at the same time trying to develop a concrete proposal for some form of appropriate regulation.

In this post, I want to consider Scherer's case for thinking that AI is (and will be) exceptionally difficult to regulate. That case consists of three main arguments: (i) the definitional argument; (ii) the ex post argument and (iii) the ex ante argument. These arguments give rise to eight specific regulatory problems (illustrated below). Let's address in each in turn.

(Note: I won't be considering whether the risks from AI are worth taking seriously in this post, nor will I be considering the general philosophical-political question of whether regulation is a good thing or a bad thing; I'll be assuming that it has some value, however minimal that may be)



The Regulatory Problems of Artificial Intelligence

The Discreetness Problem
AI projects could be developed without the large-scale, integrated institutional frameworks needed by most 20th C industrial institutions.

The Diffuseness Problem
AI projects can be developed by a diffuse set of actors operating in a diffuse set of locations and jurisdictions.

The Discreteness Problem
AI projects will capitalise on or make use of discrete technologies and components, the full potential of which will not be apparent until the components come together.

The Opacity Problem
The technologies underlying AI will tend to be opaque to most potential regulators.

The Foreseeability Problem
AI can be autonomous and operate in ways that are unforeseeable by the original programmers. This will give rise to a potential 'liability gap'.

The Narrow Control Problem
An AI could operate in ways that are no longer under the control of those who are legally responsible for it.

The General Control Problem
An AI could elude the control of all human beings. This is the problem alluded to by the likes of Nick Bostrom in *Superintelligence*.

Ex Ante Challenges
Problems with the research and development of Artificial Intelligence

The Definitional Problem
What is Artificial intelligence? You need to define the object of regulation, but AI admits of no easy definition. Modern preference for defining AI in terms of 'thinking' and 'acting' rationally tends toward under and over inclusivity

Ex Post Challenges
Problems with the creation and implementation of Artificial Intelligence

1. The Definitional Argument

Scherer's first argument focuses on the difficulty of defining AI. Scherer argues that an effective regulatory system needs to have some clear definition of what is being regulated. The problem is that the term 'artificial intelligence' admits of no easy definition. Consequently, and although Scherer does not express it in this manner, it seems like the following argument is compelling:

- (1) If we cannot adequately define what it is that we are regulating, then the construction of an effective regulatory system will be difficult.
- (2) We cannot adequately define 'artificial intelligence'.
- (3) Therefore, the construction of an effective regulatory system for AI will be difficult.

Scherer spends most of his time looking at premise (2). He argues that there is no widely-accepted definition of an artificially intelligent system, and that the definitions that have been offered would be unhelpful in practice. To illustrate the point, he appeals to the definitions offered in Russell and Norvig's leading textbook on artificial intelligence. These authors note that definitions of AI tend to fit into one of four major categories: (i) *thinking like a human*, i.e. AI systems are ones that adopt similar thought processes to human beings; (ii) *acting like a human*, i.e. AI systems are ones that are behaviourally equivalent to human beings; (iii) *thinking rationally*, i.e. AI systems are ones that have goals and reason their way toward achieving those goals; (iv) *acting rationally*, i.e. AI systems are ones that act in a manner that can be described as goal-directed and goal-achieving. There are further distinctions then depending on whether the AI system is narrow/weak (i.e. focused on one task) or broad/strong (i.e. focused on many). Scherer argues that none of these definitions is satisfactory from a regulatory standpoint.

Thinking and acting like a human was a popular way of defining AI in the early days. Indeed, the pioneering paper in the field — Alan Turing's 'Computing Machinery and Intelligence' — adopts an 'acting like a human' definition of AI. But that popularity has now waned. This is for several reasons, chief among them being the fact that designing systems that try to mimic human cognitive processes, or that are behaviourally indistinguishable from humans, is not very productive when it comes to building actual systems. The classic example of this being the development of chess-playing computers. These systems do not play chess, or think about chess, in a human-like way; but they are now better at chess than any human being. If we adopted a thinking/acting like a human definition for regulatory purposes, we would miss many of these AI systems. Since these systems are the ones that could pose the largest public risk, this wouldn't be very useful.

Thinking and acting rationally is a more popular approach to AI definition nowadays. These definitions focus on whether the system can achieve a goal in narrow or broad domains (i.e. is the system capable of optimising a value function). But they too have their problems. Scherer argues that thinking rationally definitions are problematic because thinking in a goal-directed manner often assumes, colloquially, that the system doing the thinking has mental states like desires and intentions. It is very difficult to say whether an AI system has such mental states. At the very least, this seems like a philosophical question that legal regulators would be ill-equipped to address (not that philosophers are much better equipped). Acting rationally definitions might seem more promising, but they tend to be both under and over-inclusive. They tend to be over-inclusive insofar as virtually any machine can be said to act in a goal directed manner (Scherer gives the example of a simple stamping machine). They tend to be under-inclusive insofar as systems that act irrationally may pose an even greater risk to the public and hence warrant much closer regulatory scrutiny.

I think Scherer is right to highlight these definitional problems, but I wonder how serious they are. Regulatory architectures are made possible by law, and law is expressed in the vague and imprecise medium of language, but problems of vagueness and imprecision are everywhere in law and that doesn't prove an insuperable bar to regulation. We regulate 'energy' and 'medicine' and 'transport', even though all these things are, to greater or lesser extent, vague.

This brings us back to premise (1). Everything hinges on what we deem to be an 'adequate' definition. If we are looking for a definition that gives us necessary and sufficient conditions for category membership, then we are probably looking for the wrong thing. If we are looking for something that covers most phenomena of interest and can be used to address the public risks associated with the technology, then there may be reason for more optimism. I tend to think we should offer vague and over-inclusive definitions in the legislation that establishes the regulatory system, and then leave it to the regulators to figure out what exactly deserves their scrutiny.

In fairness to him, Scherer admits that this argument is not a complete bar to regulation, and goes so far as to offer his own, admittedly circular, definition of an AI as any system that performs a task that, if it were performed by a human, would be said to require intelligence. I think that might be under-inclusive, but it is a start.

2. The Ex Post Argument: Liability Gaps and Control Problems

The terms 'ex post' and 'ex ante' are used frequently in legal scholarship. Their meanings will be apparent to anyone who has studied Latin or is familiar with the meanings of 'p.m.' and 'a.m.'. They mean, roughly and respectively, 'after the fact' and 'before the fact'. In this case, the 'fact' in question relates to the construction and implementation of an AI system. Scherer argues that regulatory problems arise both at the research and development of the AI (the ex ante phase) and once the AI is 'unleashed' into the world (the ex post phase). This might seem banal, but it is worth dividing up the regulatory challenges into these distinct phases just so as to get a clearer sense of the problems that might be out there.

We can start by looking at problems that arise once the AI is 'unleashed' into the world. It is, of course, very difficult to predict what these problems will be before the fact, but there are two general problems that putative regulators would need to be aware of.

The first is something we can call the 'foreseeability problem'. It highlights the problem that AI could pose for traditional standards for legal liability. Those traditional standards hold that if some harm is done to another person somebody else may be held liable for that harm provided that the harm in question was reasonably foreseeable (there's more to the legal standard than that, but that's all we need to know for now). For most industrial products, this legal standard is more than adequate: the manufacturer can be held responsible for all injuries that are reasonably foreseeable from use of the product. With AI things might be trickier. AI systems are often designed to be autonomous and to act in creative ways (i.e. ways that are not always reasonably foreseeable by the original designers and engineers).

Scherer gives the example of C-Path, a cancer pathology machine learning algorithm. C-Path found that certain characteristics of stroma (supportive tissue) around cancerous cells were better prognostic indicators of disease progression than actually cancerous cells. This surprised many cancer researchers. If autonomous creativity of this sort becomes common, then what the AI does may not be reasonably foreseeable and people may not have ready access to legal compensation if an AI program causes some injury or harm.

While it is worth thinking about this problem, I suspect that it is not particularly serious. The main reason for this is that 'reasonable foreseeability' standards of liability are not the only game in town. The law already provides from strict liability standards (i.e. liability in the absence of fault) and for vicarious liability (i.e. liability for actions performed by another agent). These forms of liability could be expanded to cover the 'liability gaps' that might arise from autonomous and creative AI.

The second ex post problem is the 'control problem'. This is the one that worries the likes of Elon Musk, Bill Gates and Nick Bostrom. It arises when an AI program acts in such a way that it is no longer capable of being controlled by its human makers. This can happen for a number of reasons. The most extreme reason would be that the AI is smarter and faster than the humans; less extreme reasons could include flawed programming and design. The loss of control can be particularly problematic when the interests of the AI and the programmers no longer align with one another. Scherer argues that there are two distinct control problems:

> **Local Control Problem**: Arises when a particular AI system can no longer be controlled by the humans who have been assigned legal responsibility for controlling that system.

> **Global Control Problem**: Arises when an AI can no longer be controlled by any humans.

Both of these control problems would present regulatory difficulties, but the latter would obviously be much more worrying than the former (assuming the AI is capable of doing serious harm).

I don't have too much to say about this since I agree that this is a problem. I also like this particular framing of the control problem insofar as it doesn't place too heavy an emphasis on the intelligence of an AI. The current furore about artificial superintelligence is philosophically interesting, but it can serve to obscure the fact that AI systems with much lower levels of ability could pose serious problems if they act outside the control of human beings (be that locally or globally).

3. The Ex Ante Argument: Discreetness, Diffuseness, Discreteness and Opacity

So much for the regulatory problems that arise after the creation and implementation of an AI system. What about the problems that arise during the research and development phase? Scherer argues that there are four such problems, each associated with the way in which AI research and development could leverage the infrastructure that has been created during the information technology age. In this sense, the regulatory problems posed by AI are not intrinsically different from the regulatory problems created by other systems of software development, but the stakes might be much higher.

The four problems are:

> **The Discreetness Problem**: AI research and development could take place using infrastructures that

> are not readily visible to the regulators. The idea here is that an AI program could be assembled online, using equipment that is readily available to most people, and using small teams of programmers and developers that are located in different areas. Many regulatory institutions are designed to deal with largescale industrial manufacturers and energy producers. These entities required huge capital investments and were often highly visible; creating institutions than can deal with less visible operators could prove tricky.

> **The Diffuseness Problem**: This is related to the preceding problem. It is the problem that arises when AI systems are developed using teams of researchers that are organisationally, geographically, and perhaps more importantly, jurisdictionally separate. Thus, for example, I could compile an AI program using researchers located in America, Europe, Asia and Africa. We need not form any coherent, legally recognisable organisation, and we could take advantage of our jurisdictional diffusion to evade regulation.

> **The Discreteness Problem**: AI projects could leverage many discrete, pre-existing hardware and software components, some of which will be proprietary (so-called 'off the shelf' components). The effects of bringing all these components together may not be fully appreciated until after the fact. (Not to be confused with the discreetness problem).

> **The Opacity Problem**: The way in which AI systems work may be much more opaque than previous technologies. This could be for a number of reasons. It could be because the systems are compiled from different components that are themselves subject to proprietary protection. Or it could be because the systems themselves are creative and autonomous, thus rendering them more difficult to reverse engineer. Again, this poses problems for regulators as there is a lack of clarity concerning the problems that may be posed by such systems and how those problems can be addressed.

Each of these problems looks to be serious and any regulatory system would need to deal with them. To my mind, the diffuseness and opacity problems are likely to be the most serious. The diffuseness problem suggests that there is a need for global coordination in relation to AI regulation, but past efforts at global coordination do not inspire confidence (e.g. climate change; nuclear proliferation). The opacity problem is also serious and likely to be compounded by the growing use of (and need for) AI in regulatory decision-making. I have written about this before.

Scherer, for his part, thinks that some of these problems may not be as serious as they first appear. For instance, he suggests that although discreetness is a possibility, it is still likely that AI research and development will be undertaken by largescale corporations or government bodies that are much more visible to potential regulators. Thus, from a regulatory standpoint, we should be thankful that big corporations like Google, Apple and Facebook are buying-up smaller scale AI developers. These bigger corporations are easier to regulate given existing regulatory institutional structures, though this must be balanced against the considerable lobbying power of such organisations.

Okay, that's it for this post. Hopefully, this gives you some sense of the problems that might arise with AI regulation. Scherer says much more about this topic in his paper, and develops his own preferred regulatory proposal. I hope to cover that in another post.

###

##

John Danaher is an academic with interests in the philosophy of technology, religion, ethics and law. John holds a PhD specialising in the philosophy of criminal law (specifically, criminal responsibility and game theory). He formerly was a lecturer in law at Keele University, interested in technology, ethics, philosophy and law. He is currently a lecturer at the National University of Ireland, Galway (starting July 2014).

He blogs at http://philosophicaldisquisitions.blogspot.com and can be found here: https://plus.google.com/112656369144630104923/posts

This article originally appeared on John's site here. Republished under creative commons license.