PHILOSOPHICAL TRANSACTIONS A

rsta.royalsocietypublishing.org

Opinion piece

Cite this article: Reed C. 2018 How should we regulate artificial intelligence? *Phil. Trans. R. Soc. A* **376**: 20170360.
http://dx.doi.org/10.1098/rsta.2017.0360

Accepted: 29 April 2018

One contribution of 13 to a discussion meeting issue 'The growing ubiquity of algorithms in society: implications, impacts and innovations'.

Subject Areas:
artificial intelligence

Keywords:
artificial intelligence, machine learning, law, regulation, transparency

Author for correspondence:
Chris Reed
e-mail: chris.reed@qmul.ac.uk

How should we regulate artificial intelligence?

Chris Reed

Centre for Commercial Law Studies, School of Law, Queen Mary University of London, London, UK

CR, 0000-0002-9124-5489

Using artificial intelligence (AI) technology to replace human decision-making will inevitably create new risks whose consequences are unforeseeable. This naturally leads to calls for regulation, but I argue that it is too early to attempt a general system of AI regulation. Instead, we should work incrementally within the existing legal and regulatory schemes which allocate responsibility, and therefore liability, to persons. Where AI clearly creates risks which current law and regulation cannot deal with adequately, then new regulation will be needed. But in most cases, the current system can work effectively if the producers of AI technology can provide sufficient transparency in explaining how AI decisions are made. Transparency *ex post* can often be achieved through retrospective analysis of the technology's operations, and will be sufficient if the main goal is to compensate victims of incorrect decisions. *Ex ante* transparency is more challenging, and can limit the use of some AI technologies such as neural networks. It should only be demanded by regulation where the AI presents risks to fundamental rights, or where society needs reassuring that the technology can safely be used. Masterly inactivity in regulation is likely to achieve a better long-term solution than a rush to regulate in ignorance.

This article is part of a discussion meeting issue 'The growing ubiquity of algorithms in society: implications, impacts and innovations'.

1. Introduction

It is hardly surprising that there has been a sudden interest in regulating artificial intelligence (AI). AI technology has moved from the research laboratory to become part of our daily lives with remarkable speed.

THE ROYAL SOCIETY PUBLISHING

We have seen the first fatal accident involving an autonomous vehicle [1,2], AI applications are analysing images to detect potentially cancerous cells [3] and numerous other implementations are in place or in the pipeline.

The introduction of AI technologies creates societal risks. Although AI technologies aim to augment or replace human decision-making, leading to fewer wrong decisions, there is no doubt that AI will still get it wrong sometimes. And the ways in which AI gets it wrong are likely to be very different from the ways in which a human would make mistakes. This feels dangerous to society. We want to know the kinds of risks we are running, and purely statistical arguments that AI makes us safer are not convincing to the wider population.

Good regulation would improve our perception of safety, and also our perception that humans remain in control. It could also mitigate any new risks which the use of AI creates. But bad regulation risks stifling the development and implementation of useful AI solutions, perhaps even without improving safety and control. Thus, we need to understand what regulation can and cannot do so that we can shape it appropriately. It is also important that those who produce and use AI technologies are actually able to comply with regulation, and that regulation does not stifle worthwhile advances in the technology. Outside specifically regulated sectors, the general approach of law and regulation is that innovation is freely permitted, but that those responsible must bear the consequences if that innovation causes certain types of harm. If our existing law and regulation can deal with AI innovation in that way, no immediate change is needed. The argument, if one exists, for requiring *all* those who adopt an AI technology to demonstrate that it achieves a higher standard of performance and reliability than other innovations has not yet been made out.

2. The problem

Fundamentally, the problem which regulation must seek to solve is that of controlling undesirable risks. For any truly useful AI technology, there is likely to be empirical evidence that it is more cost-effective and, ideally, more accurate at making decisions than the human-based solution it replaces. But that evidence will be based on comparison with the human-based solution, whose deficiencies are currently tolerated by society. An AI-based solution will have its own deficiencies, and these will be less acceptable if they produce wrong answers where a human would have decided correctly. Regulation ought therefore to focus on any new risks which the AI solution presents, recognizing that some of these risks will be as yet unknown.

Some commentators are so alarmed by the prospect of unknown risks that they have proposed the establishment of a general regulator for AI [4]. But, there are three strong arguments against introducing new, generally applicable legal and regulatory obligations at this moment.

First, any regulatory body needs a defined field of operation, and a set of overriding principles on the basis of which it will devise and apply regulation. Those principles will be based on mitigating the risks to society which the regulated activity creates. Until the risks of AI are known, at least to some degree, this is not achievable. Regulation cannot control unknown risks, and devising a regulatory mandate on the basis of speculative risks seems unlikely to produce successful results.

Second, lawmakers are generally unsuccessful at prospective regulation, particularly in technology fields. The history of legislating prospectively for the digital technologies is one of almost complete failure [5].

Finally, and most importantly, a regulatory regime which aimed to deal with *all* uses of AI technology would be impossibly wide in scope. The range of potential applications is far too diverse, and it would be foolish to apply the same regulatory regime to autonomous vehicles as to smart refrigerators which order groceries based on consumption patterns. Probably, there is no plausible, let alone compelling, reason to regulate smart refrigerators at all. A regulatory project of this kind would risk becoming a project to regulate all aspects of human life.

The better strategy is to approach the problem incrementally. Some of the risks likely to be posed by AI technology are already apparent, and legal or regulatory action can be taken now to deal with them. Others will make themselves known as the technology becomes more widely used and can be dealt with in the same way. At some point, it will become apparent whether specific regulation is needed, and if so the scope and focus of that regulation will be possible to devise. But at present, we are some distance away from that point.

This should not be taken as a call for complete inaction. Working to understand the risks posed by AI developments and the possible legal and regulatory responses to them would be highly desirable, and this could be undertaken by establishing a new, non-regulatory body for that purpose [6] or by extending the remit and funding of existing agencies [7]. Properly informed analysis on these matters will help to ensure that the current law and regulation is applied most appropriately, identify when regulation is in fact needed, and decrease the likelihood of unnecessary or ineffective regulation in the future.

To illustrate the argument that existing law and regulation is largely able to cope with the immediate problems posed by AI, this article examines two potential risks (the infringement of fundamental rights, and the causing of loss or damage as the result of an AI decision) and one possible legal change (requiring transparency about the AI's decision-making process). The main focus is on areas of activity where there is no existing regulatory regime which concerns itself with the decision-making process of actors, and thus AI technology can be adopted without regulatory approval or oversight. In already regulated sectors, such as medical treatment and aviation, sector-specific regulatory change will of course be needed at the time of, and quite possibly in advance of, the adoption of AI technologies.

3. Identifiable risks to fundamental rights

Most countries have a range of laws which protect fundamental rights. Those most obviously threatened by AI decisions are the rights not to be discriminated against on the grounds of race, sex, sexual orientation, religion, etc. AI risks breaching these rights because of the difference in perspective between laws which protect human rights and machine learning technologies. The law takes the perspective of the individual as its starting point—was that individual treated fairly and reasonably in the light of the fundamental rights they possess under the law? By contrast, machine learning technologies examine past decisions, and from those decisions considered in aggregate, they develop the model for making their future decisions.

This latter approach can easily lead to a violation of fundamental rights because machine learning-based AI is trained on real-life examples of previous decisions. This training will embed any decision-making which infringes those rights in real life, rather than the ideal behaviour which is described in the law ([8], paras 47–49). For example, the practice of motor insurers granting insurance policies to women on more favourable terms than to men, which is objectively correct because the statistical evidence is clear that women present a lower risk, has been held unlawful on the ground of sex discrimination by the Court of Justice of the European Union [9]. More recently, potential for infringement of the right to a fair trial by using an AI tool to assist in sentencing has been recognized in US legal proceedings [10]. The Supreme Court of Wisconsin considered the issue at length and issued detailed guidelines for future cases, the most important of which is that fundamental rights will inevitably be infringed if decisions are made solely on the basis of a machine learning technology which cannot give a satisfactory account of the process by which its recommendations were produced ([10], paras 93–97). The fact that the tool's predictions were shown to be accurate enough in aggregate to justify its recommendations was an essential prerequisite for using the tool at all ([10], paras 87–92), but did not overcome the obligation to give individual consideration to the decision which thus required the tool to explain its reasoning.

The obvious regulatory response is to require each AI tool which has the potential to infringe fundamental rights to be able to explain the reasoning leading to the tool's decisions. This might be workable for those AI tools where the risk is obvious, such as tools which assist in recruitment

decisions [11], but for many, the risk of their making infringing decisions will not be recognized until an actual infringement occurs. It would, therefore, make sense to delay regulation if the risk can be mitigated in other ways.

The current focus of the law places the obligation to avoid human rights infringement on the person who is ultimately responsible for the decision, and not on the technology producer. This is likely to be sufficient in the short term to deal with this risk, because that person is potentially subject to liability claims from those whose rights are infringed. That potential liability incentivizes at least some of them to choose AI technologies which are able to provide decision-making explanations. This is because failure to provide satisfactory answers makes it likely that the liability claim would succeed. Thus, unless the liability risk can safely be carried by the user (e.g. if individual claims are likely to be small and manageable in aggregate) or covered by insurance, producers of machine learning technology will be incentivized by this to build explanation into their technologies so as to meet user demand. It therefore seems safe to delay regulation until the real magnitude of the risk becomes clearer, at which point it should also be easier to draw a line between regulated and unregulated activities.

4. Identifiable risks to the legal allocation of responsibility

Every society needs a scheme to allocate responsibility, and thus liability, for activities which cause loss or damage to others. It does so in respect of persons (including companies), not in respect of technologies. This is because the levers which the law uses to control or influence behaviour, such as awarding compensation or imposing fines or imprisonment, can only be applied to persons. Law and regulation which appears to be about technology is in fact always about the behaviour of persons responsible for that technology.

Some areas where AI technologies are likely to prove useful have tailored responsibility schemes, for instance, nuclear power [12], and here it will be up to the legislator or regulator to modify the scheme appropriately. But most activities fall within the general scheme of liability law.

In a minority of cases, the person responsible for the activity is made strictly liable, which means that they must compensate those who suffer loss or damage, regardless of whether the person responsible was careless when undertaking the activity. Strict liability under English law has been a piecemeal response to the recognition of dangerous activities or states of affairs against the consequences of which the person responsible is required to indemnify the remainder of society. Examples include keeping animals of a dangerous species, or keeping animals of a non-dangerous species which have characteristics which make them unusually dangerous compared to others of the same species [13]; flying an aircraft, where the owner is liable for all personal injury or property damage caused in flight or while taking off or landing [14]; and accumulating on a person's land something which is a non-natural use of the land, and which is likely to cause damage if it escapes from the land [15].

There is also a general category of strict product liability, under which the manufacturer or importer of a product is liable for personal injury or property damage caused by a defect in that product ([16], §2(1)), but products incorporating AI which have been tested and shown to perform better than their predecessor versions are unlikely to fall within the definition of 'defective'. Additionally, there is a defence that the defect is such that a reasonable producer would not, in the current state of the art in that industry, have discovered it ([16], §4(1)(e)), which in many cases involving AI would prevent the strict liability from arising. Thus, product liability is unlikely to be much help to, for example, victims of autonomous vehicle accidents.

It is worth noting that these strict liability regimes are not usually as strict as they at first appear. Often liability can be displaced by proving lack of fault, as in the case of product liability, or at least requires that the defendant ought to have appreciated the existence of the risk which came to pass. Fault is deeply embedded in liability law.

In most cases, the applicable responsibility regime will be the law of negligence, which is explicitly based on fault. This imposes liability for loss or damage caused by carelessness. The

law requires that a duty of care be owed, that the duty was broken and that the loss or damage was a foreseeable consequence (i.e. that the breach of duty caused the loss) [17]. Although the application of the law can be complicated for marginal cases, in general, it is quite easily applied to human decision-making. Thus, medical practitioners are required to be as careful as other reasonable medical practitioners, drivers are compared to reasonable drivers and so on. Carelessness is premised on how far a risk ought to have been foreseen and guarded against, and that is a question which the courts can answer if a human makes the decision which causes the loss or damage.

If we take the example of autonomous vehicles, it is immediately apparent that the nature of negligent driving changes. We can no longer ask whether the person driving was careless, because no person is driving. Thus, the only sensible questions which the law of negligence can ask are: How was the accident caused? and, Was the person who was responsible for the cause careless? These are both much harder to answer for AI technology.

Causation is a problem because the AI driving technology works together with the vehicle's control systems and sensors. So, we will need analysis of vehicle logs, and quite possibly simulation of the incident, in order to discover the cause of an accident. This will be more expensive, and require scarce expertise, compared to a collision between human-driven vehicles.

Even if a clear cause can be identified, allocating responsibility for that cause will be hard. Although the manufacturer of the vehicle owes a duty of care to that segment of the public at large which is put at risk when the vehicle is operating, that duty is only to take reasonable care in the design and construction of the vehicle. So, if the self-driving technology is produced by a third party, then the manufacturer's duty is only to take reasonable care when deciding to use the technology and in integrating it safely into the workings of the vehicle.

The production of the AI technology itself may also involve split responsibilities. There are at a minimum four components to this technology: the element which controls the operation of the vehicle; that which makes decisions about the driving; the element which learns about driving; and the data on which the learning element is trained and improves itself. Any one or more of these might be developed by different entities. Because the technology will perform directly in a way which might cause loss or damage, without human intervention, it is likely that the law of negligence will treat these producers in the same way as those who manufacture components for vehicles, and thus impose a duty of care on them. However, each will be entitled to rely on the others complying with their duty of care (in the absence of reasons to suspect that they have not done so). The likely scenario in any negligence claim is that each will argue that the loss or damage was caused by one or more of the others being negligent, and not by its own carelessness. Investigating how each of these components contributed to the accident, and how careful each producer has been, is likely to make negligence litigation involving self-driving vehicles vastly more uncertain and expensive than that which requires only determination of the level of care taken by a human driver.

The UK is likely to resolve this problem for autonomous vehicles by introducing what is, in effect, strict liability on the insurers or owners of such vehicles [18]. This is unproblematic because all road vehicles already require liability insurance; although premiums are initially predicted to be high [19], they will likely fall once the true risks are established by the insurance industry, and of course there will be a big saving in the costs of litigation. But, most fields of human activity which might make use of AI will not be subject to strict liability, so there the challenges to assessing negligence liability will remain.

If the machine learning technology does not cause loss or damage directly, but instead assists humans in deciding how to act, then determining negligence liability becomes more difficult. A good example here is a machine learning technology which assists in medical diagnosis and treatment by providing diagnosis advice to medical professionals. Certainly, the producers of that technology owe a duty of care to those professionals, but do they owe any duty to patients? English law has a long-standing policy that those who give advice will only owe a duty of care in negligence to that subset of society which is entitled to rely on the advice [20]. In the leading case

on this question [21], the auditors of a company had approved its accounts, and in reliance on those accounts the claimants invested money in the company. The accounts were defective and overvalued the shares, causing the investors financial losses. The House of Lords held that the auditors only owed a duty of care to those to whom they had undertaken a responsibility to act carefully in their audit activities. That group was restricted to the shareholders of the company, for whom the audit had been produced, and did not extend to external persons who were considering making an investment in the company.

If the law were applied in this way, we are faced with the situation where the medical professional has probably discharged her duty of care because the technology is normally so accurate in its advice that it is reasonable to rely on it, while the producer of the technology owes no duty of care at all to patients. Thus, even if the patient's loss were caused by carelessness on the part of the technology producer, there would still be no liability on the producer to compensate the patient.

Underlying this uncertainty about the law is a wider policy question. Where human decisions might cause harm, responsibility is shared widely between all those human decision-makers. But if the human decision-making is undertaken by AI, then responsibility is concentrated on a very small group of technology producers. Is it fair and just to place responsibility on such a small group? And more practically, would the risk of liability potentially drive that group of producers out of business, thus losing the wider benefits of adopting the technology? This problem was recognized in the early years of the commercial Internet where the law held Internet service providers jointly liable for content posted by users, and the policy response was to introduce wide-ranging liability immunities to prevent the risk of liability claims from stifling the development of the technology [22].

In the short term, AI technology will be introduced gradually, and this gives time for lawmakers to appreciate the problem and consider ways to deal with it, but in the long term, the law of negligence will likely become an inappropriate mechanism for assigning responsibility and liability in the case of AI technologies which are in widespread use.

5. Transparency as an interim solution?

In both the examples discussed above, explanations of decision-making play a fundamental role. The person who has an obligation to preserve fundamental rights needs to know how an AI technology makes its decisions, in order to be sure that relying on it will not result in a breach. To assess whether there should be negligence liability for an AI-based decision, the courts need to be told how the AI made its decision. So, requiring transparency about the workings of AI might be a suitable interim solution to some of the legal problems, and has already been recommended as a tool for regulation [6,7,23,24]. Transparency, if it could be achieved, would help to explain the options available to the technology and the choices it made between them ([25], paras 47–49). This alone would be of great assistance in helping to resolve responsibility and liability questions. Ideally, transparency would go further and explain how and why the technology made these choices, i.e. its reasoning.

But if the law were to demand transparency, it would need to define what was meant by the term. And, transparency for AI decision-making is a complex concept. Zarsky [26] demonstrates that transparency can take a range of meanings and, more importantly, be justified for a number of different purposes. These latter include: as an incentive to technology producers and users to modify or improve the technology in response to pressure from those subject to its decisions; to enable suggestions for improvement to be made by external experts; to give subjects notice about how information about them is processed and, possibly, to enable them to object or seek an alternative decision-maker. In the field of liability, not examined by Zarsky, it enables responsibility for decision-making failures to be assigned appropriately. It is these purposes and their justifications which determine what information should be provided about the AI, and to whom, it should be disclosed.

6. Transparency is in the eye of the (non-) beholder

Technical systems whose workings are not understandable by humans are often described as 'black box' systems. It is of fundamental importance that the law recognizes that a system might be a 'black box' to one person but not to another. For example, the producer of machine learning-based AI might be able to explain how and why it reaches its decisions, while to the user of the technology, these matters will be unknowable.

This distinction is important because the application of the law often depends on what a human knew, or ought to have known, at the time the potential liability arose. In most cases, all that the user of the technology knows is that he is ignorant of its workings, and that it is *de facto* a 'black box'. His only options are to rely on its decisions or reject them.

A producer of AI technology is in a different position, however. The law will ask what a technology producer knew or ought to have known in advance, for example through the process of testing and evaluating the technology. It will also ask what can be discovered after the event if the technology fails to make a correct decision.

And, a regulator who is deciding, for example, whether to licence some AI technology for general use will depend on the producer to provide some, but not all, of the information the producer can generate, but more importantly will need the producer to explain the meaning of that information and its implication for use of the technology. This is what Zarsky [26] terms 'interpretability'; disclosing an algorithm, for example, will probably only convey meaning to a technologist working in the field, so any explanation will need to translate what the AI is doing in terms which are meaningful to the recipient of the explanation. Such an explanation might even make no attempt to explain the algorithm at any level of abstraction, if doing so would not achieve the purpose for which the explanation is required. Wachter *et al.* [27] point out that what they describe as 'counterfactuals', examples of changes to the facts which would have produced a different decision by the AI, can be more illuminating to some recipients if the aim is to enable those recipients to understand why a decision was made.

Any legal requirement to incorporate transparency into AI needs to take account of these differences in perspective. Merely demanding transparency is meaningless without the context of the human who requires transparency. It is therefore essential to define a requirement for transparency in terms of the human who needs to understand the decision-making processes of the AI in question.

We also need to ask what purpose the understanding is to serve. The UK House of Lords Artificial Intelligence Committee has recently suggested that in order to achieve trust in AI tools

> it is not acceptable to deploy any artificial intelligence system which could have a substantial impact on an individual's life, unless it can generate a full and satisfactory explanation for the decisions it will take. ([7], para 105)

How full that explanation needs to be should depend on the functions performed by the tool and the needs of those who receive the explanation. For some tools, a satisfactory explanation might quite reasonably be short on detail.

7. Transparency: *ex ante* and *ex post*

There is also an important distinction to be made between *ex ante* transparency, where the decision-making process can be explained in advance of the AI being used, and *ex post* transparency, where the decision-making process is not known in advance but can be discovered retrospectively by testing the AI's performance in the same circumstances. Any law mandating transparency needs to make it clear which kind of transparency is required.

And the law needs also to recognize that the kinds of transparency which are achievable vary between types of AI. AI technology which implements a decision tree makes it theoretically

possible to explain the logical reasoning which led to the decision by the technology. But for all but the simplest systems, the decision tree will not be a product of the human mind, but rather will have been induced by a machine learning algorithm from a mass of data. The logic of this decision tree may well be too detailed and complicated for the human mind to understand fully, what Burrell describes as

> opacity that stems from the mismatch between mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of semantic interpretation [28].

Transparency, for the law's purposes, must mean an interpretable explanation of the kind which those humans who need the explanation can understand. This means users of technology, judges and regulators, and indeed the general population in some instances.

By contrast, where the machine learning technology's decision-making element comprises a neural network, or some similar technology, it will be difficult and perhaps impossible to provide any explanation at all. A legal requirement for transparency could thus effectively amount to a prohibition on using neural networks, and lawmakers need to understand this.

Of course, there are circumstances where such a prohibition might be appropriate. For example, researchers into the risk of death of pneumonia patients [29] found that neural net models were more accurate at making predictions than rule-based models. But, they also found that the rule-based models contained at least one incorrect rule that asthmatics were at lower risk of dying than other patients. This was because the differences in treatment given to asthmatics were not adequately captured in the machine learning dataset. The researchers concluded that neural nets could not safely be used here because of the risk that other errors might become embedded in a neural net system but be undetectable, and thus put patients at unnecessary risk. In this arena, it would clearly be appropriate for regulation to require *ex ante* transparency and thus prevent the use of neural nets.

By contrast, the FLARM system[1] uses AI technology to reduce mid-air collisions between light aircraft but provides no *ex ante* transparency about its workings because it is proprietary. However, it offers a technological solution for mitigating a known risk, where there was previously no solution, and has been demonstrated to be very effective. FLARM might have introduced new and unknown risks, but these are outweighed by the reductions in deaths which it has achieved. It would seem perverse to prohibit its use on grounds of lack of transparency.

From a legal perspective, it is unfortunate that machine learning development has, naturally, focused on the quality of learning and decisions rather than on providing accounts of how decisions are justified [25]. But, there is a growing body of research into how explanation facilities can be incorporated into machine learning [30], and also research into how explanations for the reasoning of 'black box' technologies can be reverse engineered retrospectively [31–33]. Additionally, if the data used as input to a decision are captured and logged, the technology should produce the same decision if that data are used as input again, which might allow an expert witness in litigation to observe its operation and gain some understanding of the relevant decision. Developments in all these areas may help resolve the transparency difficulty over time.

8. When should the law demand transparency?

In the absence of a wide range of real-life examples of AI in widespread use, it is difficult to identify any fundamental principles which could be used to determine whether a transparency obligation should be imposed, and if so, what kind of transparency. Some tentative starting points for such a discussion are proposed here.

[1]https://flarm.com/technology/traffic-collision-warning/.

(a) Complete lack of any transparency

Ex ante or *ex post*, as is currently the case for some AI based on neural networks. This should be legally acceptable where the AI produces benefits to society overall and the loss to individuals is minor and compensatable. An example might be AI controlling a domestic central heating system, where the only risk to the householder is that the use of AI results in higher bills than before it was installed. In cases like this, the law's policy decision might reasonably be to maintain the current legal position, which is that the householder bears this risk, sharing it with the supplier and producer of the technology via the existing law on liability for defective products and services. The situation seems no different from the introduction of any other new, but non-AI, technology.

It is conceivable that new uses of non-transparent AI might create risks of loss which is not easily compensatable under the current law, but it is too early to predict what form these might take, and whether some special liability scheme might need to be devised to cope.

(b) *Ex post* transparency only

This should be legally acceptable where the AI produces benefits to society overall and loss to individuals is legally compensatable (i.e. monetary damages will be accepted by society as an adequate remedy). This is effectively the current position for personal injury caused by motor accidents. Society accepts that some injuries are inevitable, and that human decision-making when driving is unpredictable, and also accepts that monetary compensation is the best that can be achieved while still permitting the societal benefits from motor travel.

Only the producers of an AI technology are likely to be able to provide *ex post* transparency, which means that any transparency regulation should focus on their systems of training and testing, recordkeeping and data retention. If this information is not recorded and preserved, *ex post* transparency is impossible.

Ex post transparency is certainly needed if the existing law of negligence is to be used as the mechanism for deciding liability, because without such transparency, there would be no way of deciding whether the accident resulted from a lack of care by any person. In practice, though, the cost and difficulty of obtaining this evidence (particularly if the AI producer is located in a different country) will make using the existing law more difficult and expensive. The law of negligence will be capable of evolving to deal with this problem over time, in the same way, it has done for other new technologies, and so it would seem reasonable to limit regulation to identifiably high-risk activities involving AI.

(c) *Ex ante* transparency

Because *ex ante* transparency is difficult to achieve, and many useful AI technologies may be unable to provide it, the law should be cautious about demanding it. There are, though, two justifications for doing so which are likely to override most objections.

The first is where the AI technology creates the risk that fundamental rights are breached. *Ex post* transparency is not good enough here because the contravention will still have occurred, and the law's objective is to prevent such contraventions rather than merely to compensate them after the event. If a fundamental right can be contravened on payment of compensation, it is not really fundamental. So, producers and users of AI which clearly create such risks might reasonably be required to use only technologies which can explain the decision-making process in advance. Technologies which are not obviously risky will still need to provide *ex post* transparency because, if the use of AI does lead to breaches, its producers and users will need to take steps to ensure they do not recur. There are already legal incentives for technology producers to build transparency mechanisms into technologies which might threaten fundamental rights, most notably Article 22 of the EU General Data Processing Regulation [34], which grants data subjects a limited right not to be subject to 'a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her', and

Articles 13 and 14, which provide that when profiling takes place using personal data the data subject has a right to 'meaningful information about the logic used' [35,36].

The second is where implementation of an AI technology will be unacceptable to society at large without a convincing explanation of how the technology takes particular decisions. Autonomous vehicles are likely to be such a technology—purely statistical assurances about safety might convince some readers of this journal, but the wider population requires a narrative explanation. This explains the continued fascination of the 'trolley problem', which in this context asks that the technology should explain the 'choice' of victims which has to be made when it is inevitable that a self-driving vehicle will crash [37–39]. Even if the vehicle is not making something which would be recognized as a choice in human terms, society is unlikely to be content with nothing more than an explanation that the vehicle is less likely to crash than if driven by a human. Although, as noted above, there is no legal objection to lack of *ex ante* transparency for such technology, law embodies the social settlement, not merely abstract principle. Thus, lawmakers may need to impose requirements for *ex ante* transparency to achieve social consensus even if they are not, from a theoretical legal perspective, necessary.

Lawmakers need to recognize, though, that any *ex ante* transparency regulation will reduce the ability of AI to improve in use via machine learning. AI which is required to provide *ex ante* transparency cannot evolve its decision-making through learning, but instead will be limited to capturing use data and uploading it to the producer's training set. The AI can then be trained on these data, and in due course an improved version released, but this is inevitably slower than evolution through learning in use. *Ex ante* transparency regulation might also prevent the use of AI incorporating neural networks, because it is usually not possible to explain *ex ante* (if at all) the reasoning through which the neural network reached its decision.

9. Masterly inactivity?

The analysis in this paper suggests that some form of regulation will be needed for some uses of AI. But does that mean that we need to regulate now?

I argue that the answer is a qualified 'No'. Responsibility for autonomous vehicles is clearly problematic, and the uncertainty about the current application of the law is likely to inhibit their adoption unless the position is clarified, as the UK and other lawmakers are currently doing. The use of technology in medicine is already regulated by the profession, and that regulation will certainly be adapted piecemeal as new AI technologies come into use. There are probably other high-risk uses of AI which will demand some level of legal and regulatory change. But, all these areas are likely to be regulated already, as is the case for road vehicles and medicine, so the existence of current regulation might provide a useful guideline about where to focus the immediate regulatory effort.

So far as regulating the rest of life is concerned, I have attempted to show that transparency will be enough to allow the current legal and regulatory regime to produce at least adequate answers. Because that regime also provides sufficient incentives for users to demand and producers to develop transparency of AI decision-making, there is no need to panic. A 'wait and see' approach is likely to produce better long-term results than hurried regulation based on, at best, a very partial understanding of what needs to be regulated.

Data accessibility. This article has no additional data.

Competing interests. I declare I have no competing interests.

Funding. This work emanates from the Microsoft Cloud Computing Research Centre, a collaboration between the Cloud Legal Project, Centre for Commercial Law Studies, Queen Mary University of London and the Department of Computer Science and Technology, University of Cambridge, which is generously supported by a Microsoft research donation.

Acknowledgements. The author is grateful to members of the Microsoft Cloud Computing Research Centre team for helpful comments, though the views here are entirely those of the author. The foundational research on which this article builds [40] was also conducted by Elizabeth Kennedy and Sara Nogueira Silva whose contribution is gratefully acknowledged.

References

1. Greenemeier L. 2016 Deadly Tesla crash exposes confusion over automated driving. *Scientific American* 8 July. See http://www.scientificamerican.com/article/deadly-tesla-crash-exposes-confusion-over-automated-driving/.
2. Tesla Motors statement. 2016 30 June 2016. See https://www.teslamotors.com/en_GB/blog/tragic-loss.
3. Al-shamasneh ARM, Obaidellah UHB. 2017 Artificial intelligence techniques for cancer detection and classification: review study. *Eur. Sci. J.* **13**, 342–370. (doi:10.19044/esj.2016.v13n3p342)
4. European Parliament Committee on Legal Affairs. 2017 *REPORT with recommendations to the Commission on Civil Law Rules on Robotics* (*2015/2103*(*INL*)) A8-0005/2017 27 January.
5. Reed C. 2010 How to make bad law: lessons from cyberspace. *Mod. Law Rev.* **73**, 903–932. (doi:10.1111/j.1468-2230.2010.00838.x)
6. British Academy & Royal Society. 2017 Data management and use: governance in the 21st century.
7. UK House of Lords Artificial Intelligence Committee. 2018 *AI in the UK: ready, willing and able?* (HL Paper 100).
8. UK House of Commons Science and Technology Committee. 2016 *Robotics and artificial intelligence* (HC 145 12 October).
9. *Association Belge des Consommateurs Test-Achats ASBL v Conseil des Ministres* (Case C-236/09, 1 March 2011).
10. State of Wisconsin v Loomis 2016 WI 68.
11. Chien C-F, Chien L-F. 2008 Data mining to improve personnel selection and enhance human capital: a case study in high-technology industry. *Expert Syst. Appl.* **34**, 280–290. (doi:10.1016/j.eswa.2006.09.003)
12. UK Nuclear Installations (Liability for Damage) Order 2016 No. 562.
13. UK Animals Act 1971 s 2.
14. UK Civil Aviation Act 1982 s 76(2).
15. Rylands v Fletcher 1868 UKHL 1
16. UK Consumer Protection Act 1987
17. Donoghue v Stevenson 1932 AC 562.
18. UK Automated and Electric Vehicles Bill 2017-19 s 2.
19. Paton G. 2017 Driverless cars will attract hefty insurance premium. *The Times*, 9 February 2017.
20. Morgan J. 2006 The rise and fall of the general duty of care. *Prof. Neglig.* **22**, 206–224.
21. Caparo v Dickman 1990 2 AC 605.
22. Lemley MA. 2007 Rationalizing internet safe harbors. *J. Telecomm. High Technol. Law* **6**, 101–119.
23. UK Department for Transport: Centre for Connected and Autonomous Vehicles. 2016 *Pathway to driverless cars: proposals to support advanced driver assistance systems and automated vehicle technologies*. London, UK: Department for Transport.
24. US Department of Transportation/NHTSA. 2016 *Federal automated vehicles policy – accelerating the next revolution in road safety*. Washington, DC: US Department of Transportation.
25. UK House of Commons Science and Technology Committee. 2016 *Robotics and artificial intelligence* (HC 145 12). London, UK: Science and Technology Committee.
26. Zarsky T. 2013 Transparent predictions. *Univ. Ill. Law Rev.* **4**, 1503.
27. Wachter S, Mittelstadt B, Russell C. In press. Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harv. J. Law Technol.* (doi:10.2139/ssrn.3063289)
28. Burrell, J. 2016 How the machine 'thinks': understanding opacity in machine learning algorithms. *Big Data Soc.* **3**, 1–12. (doi:10.1177/2053951715622512)
29. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Eldahad N. 2015 Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In *Proc. 21th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD '15, Sydney, Australia, 10–13 August*, pp. 1721–1730. New York, NY: ACM. (doi:10.1145/2783258.2788613)
30. Darlington K. 2013 Aspects of intelligent systems explanation. *Univ. J. Control Autom.* **1**, 40–51. (doi:10.13189/ujca.2013.010204)

31. Robnik-Sikonja M, Kononenko I. 2008 Explaining classifications for individual instances. *IEEE Trans. Knowl. Data Eng.* **20**, 589–600. (doi:10.1109/TKDE.2007.190734)
32. Štrumbelj E, Kononenko I. 2008 Towards a model independent method for explaining classification for individual instances. In *Data warehousing and knowledge discovery* (eds I-Y Song, J Eder, TM Nguyen), pp. 1–12. Berlin, Germany: Springer. (doi:10.1007/978-3-540-85836-2_26)
33. Baehrens D, Schroeter T, Harmeling S, Kawanabe M, Hansen K, Müller K-R. 2010 How to explain individual classification decisions. *J. Mach. Learn. Res.* **11**, 1803–1831.
34. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L119/1, 4 May 2016.
35. Goodman B, Seth Flaxman S. 2016 European union regulations on algorithmic decision-making and a 'right to explanation'. In *ICML Workshop on Human Interpretability in Machine Learning* (*WHI 2016*), *New York, NY, 23 June*. (https://arxiv.org/abs/1606.08813)
36. Karaminou D, Millard C. 2016 Machine learning with personal data. In *MCCRC Symp., Cambridge, UK, 8–9 September*. See https://ssrn.com/abstract=2865811.
37. Bonnefon J-F, Shariff A, Rahwan I. 2016 The social dilemma of autonomous vehicles. *Science* **352**, 1573. (doi:10.1126/science.aaf2654)
38. Why self-driving cars must be programmed to kill. *MIT Technology Review* 22 October 2015. See https://www.technologyreview.com/s/542626/why-self-driving-cars-must-be-programmed-to-kill/.
39. Larry Greenemeier L. 2016 Driverless cars will face moral dilemmas. *Scientific American* 23 June. See http://www.scientificamerican.com/article/driverless-cars-will-face-moral-dilemmas/.
40. Reed C, Kennedy E, Silva SN. 2016 Responsibility, autonomy and accountability: legal liability for machine learning, Queen Mary University of London, School of Law Legal Studies Research Paper No. 243/2016. See https://ssrn.com/abstract=2853462.