How do we hold AI itself accountable? We can't.

Joanna Bryson

University of Bath, England, United Kingdom, European Union

Abstract

Artificial Intelligence (AI) is often presented to us as another race or gender of human that has growing superhuman capacities. It is natural therefore that many ask how we can integrate these new individuals into our society and our system of justice. Unfortunately, this presentation is entirely erroneous. Intelligence is an attribute of an agent, not an agent in itself, and artefacts with or without this attribute cannot be dissuaded by human justice. Human justice is uniquely designed for maintaining societies of organisms like ourselves.

Keywords: Artificial Intelligence, Legal Personhood, Suffering, Dissuasion, Recompense

This report is a short precise of a formal academic article on legal personhood for AI that I wrote with two leading law professors in legal personality, Tom Grant of Cambridge University, and Mihailis E. Diamantis of University of Iowa. Since they each had far more influence on the article than I did, I can sincerely and humbly say that that article is a great paper that I think everyone should for themselves. The title is Of, for, and by the people: the legal lacuna of synthetic persons, and it appeared open access (thanks to fees paid by the non-profit University of Bath to the for-profit publisher Springer) in the journal *Artificial Intelligence and Law* **25**(3):273291 in September of 2017 [1].

I recently received an email about that paper, and I repeat the letters we exchanged here.

First, the (anonymised) initial email:

"I'm writing to you in view of your article 'Of, For and by the people: the legal lacuna of synthetic Persons'. What are some of the mitigation measures that should be in place to ensure synthetic persons are legally accountable for their acts in case they are granted electronic personhood?"

Here is an extended version of my response:

Thank you for your interest in our work. I trust you have read the paper to which you refer? Since it is open access I hope you have no trouble getting a copy, but if you do have trouble let me know; I can even send a hard copy if necessary. The reason I ask is that the point of our article is that there is no way to ensure that a synthetic person can be held legally accountable. It does not matter whether you mean a 'synthetic person' to refer to a robot, or to the legal fiction that is used to make a corporation appear like a person. The only way to ensure that law is stable is to have a human be accountable for the actions of an artefact, and that same human be the one in control of the artefact's behaviour.

In this report, for clarity, the term *human* will always refer to a biological entity of the species *Homo sapiens*. However a *person* will be a person recognised as such by the law. Some humans are not person, because they are not competent to operate in the context of the law (e.g. infants or those with severe dementia), or because they are not recognised by the law as persons (something that might happen for example to a member of an ethnic minority living under an autocratic regime). But some non-human entities are legal persons, such as companies, and sometimes religious idols, including in one case a river. For all of these non-human entities, legal personhood is attributed (assigned) to the entity because it is legally convenient, and there is a sense in which justice can be upheld. Idols are only assigned personhood in that they are *moral patients*, that is, that they need to be protected as if they were a human. There are two reasons this makes sense for an idol:

- Real humans have been shown to suffer grievous harm when the idols do. This is partly because the idols are of great religious significance and therefore are part of both individual and community identity. The other part is the second problem:

- the idols are unique and irreplaceable. They are either ancient artefacts that require preservation, or as I mentioned in one case the artefact is a river, which can be said to be killed if the pollution in the river is so great that the life depending on the river is destroyed.

An AI system might be unique, but if so, that would be *a design decision.* All AI is by definition an attribute of an artefact, and if it is a digital artefact, any intelligence on it, for example its individual memories, can be backed up and stored. Whoever built the artefact could likewise choose to use mass

produced, perfectly replicable components. So unlike humans, rivers, or ancient religious artefacts, if an artefact with AI is unique that was a decision taken by a contemporary individual who could easily have made a different decision and protected the intelligent system they were building. What we recommend in our article is that all legal commercial products including AI should be manufactured not to be unique, if there is any concern that humans would suffer were they to lose access to the AI in that artefact.

It should be said first that not every legal system recognises idols (or even corporations) as legal persons, and second that I learned the above about idols from an excellent paper by Solaiman [2], which was also core to the arguments my colleagues and I made in our paper. Finally, it should also be said that the arguments I make below about why AI cannot be held accountable through this mechanism also apply with increasing frequency to corporations. "Shell companies" are those founded only too deceive the law and remove the threat of legal action from humans or companies that the humans in control really care about. I'm sure the janitors of a shell company goes bankrupt, but increasingly some actors are happy to (for example) build buildings with the sole purpose of having the project go bankrupt and thus serve for money laundering. They may also enjoy as a power move or benefit politically from removing whatever attribute of a city had previously been built on the location of the bankrupt building, but that's only tangentially relevant to the question at hand.

There are two necessary conditions for an entity to be a legal person.

1. First, that entity must be able to know about and be able to execute the law on their own behalf. This is why animals are not held to be legal persons, though note that we do routinely allow infirm humans (and in some countries, idols) to be represented by others.
2. Second, the penalties of law have to serve as dissuasion to the entity. This is where shell companies (as just described)—and AI—fall down.

Although many people think the purpose of the law is to compensate those who are wronged, what the law mostly does really is to maintain order by dissuading people from doing wrong in the first place by making it clear what the costs are for doing wrong. If they do do wrong, they are forced to pay those costs, with the hope that this more immediate experience of the dissuasion will stop them from doing it again, or sometimes they are just forever prevented from free action either by being jailed for life or executed. Of course, sometimes part of the dissuasion includes recompense to persons

wronged, for example the return of property, money, or even the granting of money to compensate for injury or time.

Humans are incredibly social beings. One consequence of that is that our society and self image has co-evolved with our sense of justice. So often people do *feel* compensated when they see someone else dissuaded. But having the murderer of your partner jailed or executed by no means brings your partner back to life. It is good for the victims that they can feel a sense of peace, and perhaps they really do gain greater security if it is publicly known that the last person who wronged them was penalised. But essential to all of this is that the entity that committed the crime is dissuaded from doing so again. This is also why tort settlements against companies can be outrageously high. When an elderly woman was awarded an enormous settlement after receiving third degree burns from McDonald's coffee, it was not because the woman needed the money, but rather because a smaller settlement would not have persuaded McDonald's to reduce the temperature at which it kept its coffee[1]. Similarly, the penalties the EU proposes against tech giants who violate EU privacy or competitiveness laws are set not for redress as much as for dissuasion.

I mentioned that humans have co-evolved with our intuitions about justice. Think about it: why is it punishment to put someone in jail, or label them a felon, or take away their home, or to fine a person (including a corporation) for an enormous amount of money? It is because humans have an enormous systemic aversion to isolation and losing power. We share this with other social species—even a guppy will die of stress if it is isolated from its society [cf. 3]. Again, just as with uniqueness, if AI were to also display this aversion, it is a consequence of design decisions taken. In fact, there are fantastic amounts of extant AI and none of it minds at all that it is entirely treated as a tool, subordinate to human will, turned off, traded back in to Apple for the new iPhone, etc. Humans have so much trouble understanding how an intelligent entity could not feel betrayed by such action that they refuse to recognise vastly superhuman intelligence as intelligence. Can you do arithmetic as well as your phone? or spell as well? Even if they do recognise it, then they make up a new term for intelligence that *would* mind, like 'conscious' or 'general'. Unfortunately, these terms already have other mean-

[1]Liebeck v. McDonald's Restaurants, P.T.S., Inc., No. D-202 CV-93-02419, 1995 WL 360309 (Bernalillo County, N.M. Dist. Ct. August 18, 1994)

ings entirely irrelevant though sometimes coincidental to the real matter at hand here.

What matters is that none of the costs that courts can impose on persons will matter to an AI system in the way the matter to a human. While we can easily write a program that says "Don't put me in jail!" the fully systemic aversion to the loss of social status and years of one's short life that a human has cannot easily be programmed into a digital artefact. Even if we could program it, what right would we have to make something that will be bought and sold capable of suffering? But generally speaking, well-designed systems are modular, and systemic stress and aversion is therefore not something that they can experience. We could add a module to a robot that consists of a timer and a bomb, and the timer is initiated whenever the robot is alone, and the bomb goes off if the timer has been running for five minutes. This would be far more destructive to the robot than ten minutes of loneliness is to a human, but it would not necessarily be any kind of motivation for that robot. For example again of a smart phone, if you added that module to your smart phone, what other component of that phone would know or care? The GPS navigator? The alarm clock? The address book? This just isn't the way we build artefacts to work.

Law has been invented to hold humans accountable, thus only humans can be held accountable with it. As I mentioned when I was describing shell compnaies, even the extension of legal personality to corporations only works to the extent that real humans who have real control over those corporations suffer if the corporation is to do wrong. Similarly, if you build an AI system and allow it to operate autonomously, it is essential that the person who chooses to allow the system to operate autonomously is the one who will go to jail, be fined, etc. if the AI system transgresses the law. There is no way to make the AI system itself accountable.

Having said that, it is quite easy to make the people (human or corporate) who use AI accountable, more so than within ordinary human organisations. What we can do is require that the way that any intelligent system is built—and if it has machine learning, is trained—is fully documented, and that that documentation is encrypted and secured. Further, many of the operations of the system—its decisions, and what it perceived when it made those decisions which determined those outcomes—can be recorded, a process that is called logging. This can make the system accountable in the sense that you can do accounting with the AI system, just like you can use books to make a company accountable for its finances. But the true executive of that company

is the one that has to be held responsible with the evidence gathered from these methods, whether the conventional books of accounting, or the digital logs of AI.

In our article, for the sake of argument, we admit that some people might possibly find that there are rewarding aspects to building unique, suffering AI that really would benefit from legal personhood. But what we argue is that the probable costs of social harm from corporations and individuals evading their responsibilities by offloading them to AI far, far outweigh any benefit that would come to society by creation of such a vulnerable and needy form of AI. I mean, think about it. Why would we want to motivate corporations to fully automate part of their business process (that is, get rid of any human employees) by allowing them to cap their legal and tax liabilities at the costs or establishing their new artificial legal personality? The European Parliament (EP) asked the European Commission (EC) to consider this possibility; fortunately it didn't take the EC long to consider and dismiss it. Probably part of the motivation of the EP was European Car Manufacturers lobbying because they are worried about competing with Apple and Google in the driverless market, because those tech giants have more money than they can legally spend, so are fully willing to take on all liability for their driverless cars. The injustice of this vast economic inequality does need to be addressed, but not by exposing European Union citizens to bazillions of new shell companies on wheels.

[1] J. J. Bryson, M. E. Diamantis, T. D. Grant, Of, for, and by the people: the legal lacuna of synthetic persons, Artificial Intelligence and Law 25 (2017) 273–291.

[2] S. M. Solaiman, Legal personality of robots, corporations, idols and chimpanzees: A quest for legitimacy, Artificial Intelligence and Law (2016) 1–25.

[3] F. B. M. de Waal, Good Natured: The origins of right and wrong in humans and other animals, Harvard University Press, Cambridge, MA, 1996.