Why Autonomous AI cannot be an agent in Criminal Law

Mark Dsouza

I recently participated in an excellent conference on Artificial Intelligence and Criminal Law in Bergamo, Italy, and in my talk, I defended the relatively unsexy claim that the emergence of artificial intelligence does not necessitate any radical changes to the foundations of substantive criminal law. For the most part, I was attempting to dispel the worry that the increased use of AI tools might make it harder to prosecute the human tool user, because of the complexity of the AI ecosystem (which might include several interacting AI tools), the potential unpredictability of an AI tool's chosen conduct tokens, and the inscrutability of an AI tool's internal logic. But I noted with surprise the fact that a few of the other participants at the conference suggested that they thought that the criminal law should evolve to be applied to autonomously choosing AI entities, in the sense that the entity could be an addressee of the criminal law, and potentially, a criminal defendant. There was, of course, some push back to that idea from sections of the audience, but it seems clear that this idea is (unlike my presentation) sexy. And so, it is likely to continue to be seductive in academic circles.

I want therefore, in this short post, to set out why I think that nothing that is recognisably a system of criminal law can ever treat an AI entity as an agent to which criminal laws are addressed and that might be made a defendant in a criminal trial. But before I get around to doing that, a note on how I report on the ideas of others in this post. The ideas I am reporting on are not all published, and some were conveyed to me in conversations around presentations rather than in presentations itself. For the purposes of this blog, I am relying only on my sketchy (and borderline illegible) notes in a small notebook that I used to jot down the claims that struck me as most interesting. I'm afraid that these notes do not extend to clearly identifying who said what, and so I cannot provide references in this piece when I refer to the ideas of others. Besides, I doubt that my rough and ready reconstruction of their sophisticated ideas will do them justice, and in any event, my main reason for rejecting these arguments relates to a claim about the nature of our criminal law that I situate at a deeper, more conceptual level.

Some arguments for according AI criminal agency

Let me start with a slapdash survey of some of the arguments I encountered in support of the idea of according criminal law agency to an AI entity. As I say, I do not claim authorship of any of these ideas – I am merely attempting to reconstruct some of the very thought-provoking arguments that I heard from a notebook of semi-coherent scribbles.

1 We should accord criminal law agency to AI entities because they have, or will soon have, intelligence comparable to that of humans, and so will behave as if they were agents. AI entities are capable of being given instructions, and of working out how to comply with them. But, just like humans, they are sometimes unpredictable, and do not simply obey. Some are capable of autonomously choosing their conduct, and in so choosing, they are capable of exercising their capacities to inflict harm on people, or their direct or indirect interests. In other words, they can perform the actus reus of an offence. This already qualifies them to commit offences of strict liability. Additionally, they can satisfy the mens rea of (at least some) criminal offences, insofar as they can be shown to have either been unaware of information they ought to have known (which is the essence of objective fault), or to have chosen their actions despite having information that contraindicates that conduct (which is the essence of subjective fault). When certain harms occur at the hands of a human agent with the necessary mens rea, criminal consequences typically follow. There is no reason, the argument goes, that the same should not be true when an AI entity brings about those harms with the necessary mens rea. Besides, AI entities can be 'punished'—they can be deactivated. In sum, it is contended that there is enough to suggest that it would be appropriate and effective to apply the criminal law even when the harm is brought about by an autonomously choosing AI entity.

2 The criminal law already extends criminal agency beyond humans. Common law jurisdictions hold corporations liable, and until about the 18th century, various legal systems have even held animals criminally liable. So why baulk at doing the same for AI entities?

3 Accepting that autonomously choosing AI entities can be criminal law agents will close worrying potential liability gaps. There are two separate worries here that this solution is thought capable of addressing.

3.a The first worry is that since autonomous AI entities can choose for themselves, it may not be possible to hold the humans (or corporations) responsible for the AI entity criminally liable in their own rights. These (already recognised) criminal law agents could, it is feared, plausibly deny mens rea for any offence that the AI entity autonomously chooses to commit. It was suggested that making the AI entity liable in its own right as a principal would make it easier to hold the responsible human or corporation behind the AI entity liable as an accessory to the AI entity's offence.

3.b The second worry is that even if the first worry can be addressed, the complexity of the AI ecosystem, with several extremely sophisticated AI entities interacting autonomously and not-entirely-predictably with each other, may result in liability gaps where there would ordinarily have been criminal liability for human agents. The problem is that the inner workings of each autonomously choosing AI entity could be so inscrutable that we are unable to say for certain which went wrong when harm occurs, and so it may not be possible to identify the appropriate responsible human to hold liable. It was suggested that making AI entities potentially criminally liable in their own right would allow us to reduce the stakes of a (potentially wrongful) conviction. So, if we are unable to identify which of a set of interacting AI systems made a rogue choice that caused a criminal harm, then given the lower stakes of an error, we could employ presumptions of guilt with reverse burdens of proof to identify which AI system(s) should be deactivated.

Some quick responses

No doubt I have not set out all the arguments that one might make in support of according criminal law agency to AI entities. Nor can I say with confidence that my descriptions have done justice to even those arguments that I have described. So, while I will offer tentative responses to the enumerated arguments, I realise that even if they convince, my responses cannot support the conclusion that we should not accord criminal law agency to AI entities. Therefore, I will also offer a separate, and more conceptual, argument of my own against according criminal law agency to even autonomously choosing AI entities. But first, some quick responses to the arguments I mention above. The numbering I use below corresponds to the numbering I used when setting out the argument to which I am responding.

1 A criminal offence requires more than the actus reus and mens rea. At a fundamental level, it requires qualifying agency. This is what is denied by pleas of insanity, infancy and the like. The fact that some entity is capable of performing the actus reus of an offence with its mens rea cannot compensate for the absence of qualifying agency. That is why it is no response to a plea of insanity or infancy that the defendant did perform the actus reus with the mens rea.

Moreover, the fact that an AI entity can be deactivated does not itself mean that it can be punished. That would be to conflate *consequences* with *punishment*. Punishment includes a communicative dimension, and what it communicated is moral disapprobation of the agent punished. That seems to be entirely missing when deactivating a rogue AI entity. A malfunctioning toaster can be unplugged, but that hardly amounts to punishing it.

2 While doing so is relatively standard in the common law world, the merits of treating corporations as criminal law agents remains contested in large parts of the world. It is therefore not as solid an example from which to argue by way of analogy as one might assume. But even granting, for the sake of argument, the appropriateness of treating corporations as criminal law agents, there is a clear sense in which corporations are distinguishable from AI entities. If we were to pierce the corporate veil we would (eventually) find a human being pulling the strings. We are reluctant to pierce the corporate veil for reasons to do with the economic value of maintaining the separateness of the corporate person, but the fact that somewhere behind the veil is a human controller gives us reason to want to hold the corporation criminally liable. To fail to do so would allow natural persons to immunise themselves against the criminal consequences of their actions by donning the shroud of the corporate form. That argument simply does not apply in relation to autonomously choosing AI entities. To the extent that an AI entity is subject to some human control, it is a mere tool, and so it cannot immunise the human against criminal liability anyway. (And note that not a lot of control is required to continue to hold the human controller criminally liable—a misfiring or not entirely predictable tool is still a tool.[1]) But if the AI entity is entirely uncontrolled by any human, then it is not analogous to a corporation, and so the argument by analogy to the corporation fails. As for the fact that common law jurisdictions used, a few centuries ago, to put animals on trial, well, previous follies offer no argument for new ones.

3 I am not sure that the advent of autonomous AI entities will create concerning liability gaps, but even if they do, I doubt that the proposed solution is apposite, or even effective.

3.a Regarding the first concern, as I previously mentioned, I think there's good reason to think that even if an AI entity makes autonomous choices, it can be treated, in law, as a mere tool being deployed by a human (or, as the case may be, corporate) controller. When a person (D) uses something as a tool, she exercises control over it and thereby treats it as an extension of herself in respect of that usage.[2] Therefore, conduct performed through a

---

1 M Dsouza, 'Don't panic: Artificial intelligence and Criminal Law 101' in D Baker & P Robinson (Eds.), *Artificial Intelligence and the law: Cybercrime and Criminal Liability* (Routledge, 2020) pp. 247-264.

2 JK Feibleman, 'The Philosophy of Tools' (1967) 45(3) Special Forces 329, 330. See also Dsouza, 'Don't Panic' (n1).

tool is conduct performed by D herself. When D *trains* her dog to steal sausages from the local butcher, *D* appropriates the sausages and is potentially guilty of theft. But if the dog were to steal the sausages of its own accord, then even if the owner knew, but did not care, that it was greedy and not well-trained, we would not say that D herself had appropriated the sausages. D could, of course, be liable for other offences with different actus reus stipulations; consider for instance the offence of being the owner of a dog that causes injury while dangerously out of control under s.3(1) of the Dangerous Dogs Act 1991. But when that offence applies, D she is not convicted of causing the injury herself, just as D cannot be convicted of herself stealing the sausages from our earlier example. On the other hand, if D *trained* the dog to injure someone, D could certainly be convicted of an offence involving *D* causing the injury.[3] Along similar lines, if D *deliberately* uses an AI entity as her tool, the AI entity's conduct can be attributed to D. And notice that D can intend to use the AIT as a tool even if the AIT retains some measure of autonomy over if, when, and how it does the specific conduct. An unpredictable, or not entirely predictable tool, is still a tool. If D were to throw a fox into V's chicken coop in order to disrupt V's poultry farming business, D would have caused the damage, even though in principle, it would be up to the fox to (autonomously) choose *whether* to attack the chickens, and if so, in what order, and for how long. To that extent, worries of a liability gap seem overblown. Of course, if D did not deliberately use the AI entity as a tool to perform some potentially criminal action, the conduct of the AI entity could not be attributed to D. But then again, in this circumstance, it is not clear that the AI entity's conduct *should be* attributed to D, or should attract *criminal* liability in its own right. At most, we could hold D liable for being the owner of an AI entity that was not subjected to enough control to prevent it from causing criminal harm. But that form of criminal liability, along the lines of the criminal liability under s.3 of the Dangerous Dogs Act 1991, can be enacted entirely without treating the AI entity as a qualifying agent in the criminal law.

Nor is it clear to me that there will be cases in which we will be unable to prove D's mens rea as a principal for an offence, but will have no trouble proving D's mens rea as an accessory. Take the English criminal law rule on accessorial liability as an example. To hold D liable as an accomplice to some principal P's offence, we must show that D intended by their own conduct to assist or encourage P's criminal conduct with (at least)[4] the knowledge that in performing that conduct, P would commit a full criminal offence. A

---

[3] *Murgatroyd v Chief Constable of West Yorkshire* [2000] All ER (D) 1742.

[4] There is some disagreement here, but it is not pertinent to the central claims in this post

person with that level of mens rea could easily be convicted as a principal for the same offence if P were not a qualifying criminal law agent, but were instead a tool. According criminal law agency to an AI entity then, seems to be a solution in search of a problem.

3.b Once again, my instinct is that the worry about how difficult it would be to identify the faulty AI entity in a network of interacting AI entities is overblown. This seems to be no greater problem than identifying which part of the car is malfunctioning when all I can say, as a would-be driver, is that it isn't working. But I will readily concede that I am not an expert on just how intricate AI entities and their networks can be, so let us accept, for the moment, that we might struggle to identify the rogue AI entity in a network of such entities. Even then, the proposed solution is inapt. Reversing the burden of proof will hardly make it any easier to identify the source of the problem. Even if the manufacturers of AI entities have better access to proprietary information about the coding of the AI entity, they may *also* struggle to ascertain whether it was their own AI entity that went rogue. And besides, if the only way for them to avoid criminal liability in the form of the deactivation of one AI entity was to publicly disclose their trade secrets in a criminal trial, many would rather absorb the liability so as not to surrender their competitive advantage. It isn't clear that presenting pushing AI manufacturers to have to make this choice is a desirable course of action—it might push less financially secure innovators out of the market. But even if this were considered an acceptable risk, we would still need good reasons to shift the consequences of a gap in our practical epistemic capacities (i.e. our ability to find out what went wrong) onto potentially several manufacturers, users, and servicers of AI entities. In more quotidian criminal law contexts, when the culprit is one of two or more people, and we don't know which, the criminal law accepts that since there is reasonable doubt as to the identity of the actual perpetrator, we cannot convict any of the suspects.[5] It is settled, as a matter of policy, that the consequences of our inability to find out who committed the wrong should not be shifted from where they naturally fall (i.e. on the person or persons victimised) onto the pool of plausible suspects. There is no reason to think that when the pool of suspects includes one or more AI entities (and the consequences of imposing liability are likely to be felt by the legal persons who own, control, or use the AI entity), this settled policy decision must be disturbed.

[5] *R v Banfield* [2013] EWCA Crim 1394.

An independent positive (but negative) argument

Let me turn now to my positive argument for why any recognisable system of criminal law cannot treat AI entities, no matter how autonomous, as qualifying agents. As promised, this argument operates at a deeper, more conceptual level than the arguments alluded to above.

Most theories of the criminal law take it that the criminal law is morally distinctive—that it has an important, and even characteristic, connection to some underlying system of morality. Opinions may diverge as to whether this morality is critical morality—referring to *objective* truths that can be discovered by perfecting our reasoning, or positive morality—referring to widely *accepted* truths. But either way, these moral theories, like all *relevant* moral theories, treat 'the moral good' as being, in some way, logically contingent on the addressees of the moral theory being humans. For instance, Aristotle explicitly treats 'the good' as being particularised to humans as participants in the ethical system he describes—Book 1 of his Nicomachean Ethics is even titled *The Human Good*. Hobbes too, in *Leviathan*, says that nothing is absolutely good or evil; whatever is the object *of a person's* desire is good, and whatever is the object *of a person's* aversion is evil. And in *The Concept of Law*, even the famously positivist HLA Hart reserves a place within his conception of the institution of law for a certain minimum content of natural law characterised by statements 'the truth of which is contingent on human beings and the world they live in retaining the salient characteristics which they have.'[6]

These axiomatic value statements of the moral good show that our morality, and therefore any system of criminal law based on our morality, is situated within the domain of a community of human agents. Human agents are part of the background to the rules for moral and criminal law guidance. And even to the extent that we are comfortable with allowing for corporate criminal law agents, it remains true that if one strips away the corporate veil, one will eventually find one or more humans pulling the corporate strings.

AI entities, like non-human animals, can never be the right *kind* of agent to be a member of that moral community, and so are not the right *kind* of agent to be addressed by the criminal law. They can be the *objects* of the criminal law, but they can't be its *subjects*. To be clear, a system that purported to address AI entities (or non-human animals) as agents is not inconceivable; it is simply not recognisably a (proper) system of criminal law, since it will no longer be morally distinctive in the sense I have described. This would be true even if the non-human entity being considered for the status of 'agent' superseded human abilities in one or all respects. If a society of alien beings, more advanced in all respects than us, decided to share Earth with us, it would make no more sense for us to apply our

---

[6] For more detail on these arguments, see M Dsouza, *Rationale-Based Defences in Criminal Law* (Hart, 2017) Ch.3.1.

system of criminal law to them, than it would for us to apply it to ants—or to AI entities. They are not part of the relevant moral in-group, and any system of norms that governed our interactions with them would not be the (right kind of) moral system to spawn a morally distinctive system of criminal law.

Of course, not everyone thinks that the criminal law is a morally-distinctive system of norms. But although I cannot argue for it here, I suspect that even theorists who see the criminal law as merely another tool of public law will agree that the criminal law is a subset of a system of norms that governs the manner in which humans (or entities like corporations that are, at base, entirely controlled by humans) should conduct themselves. Hence, no system of norms that sought to govern how non-human entities should conduct themselves would be recognisable as a system of criminal laws.

But maybe ChatGPT disagrees…