Don’t Panic: Artificial Intelligence and Criminal Law 101

MARK DSOUZA

The advent of artificial intelligence technology (AIT) challenges established notions about how we do things in almost every domain of our lives. It creates possibilities and efficiencies, but also raises worries: are we ready to cope with the new challenges and dangers that AIT might pose? As AIT adapts, we are being pushed to adapt with it. It isn’t surprising therefore that concerns have arisen about whether our laws are fit for purpose in the brave new world of AIT. The challenges posed to our laws by AIT are almost impossible to overstate. In this essay though, I will argue that at least in relation to the substantive law of core criminal offences – the kind that form the foundations of substantive criminal law teaching at universities – these challenges *are* sometimes overstated. I think there is no immediate need to panic; the criminal law already has the resources to cope with criminal activity involving AIT for at least the medium term. But the process of explaining why this is so gives us a chance to reflect once more on the resources that the criminal law has at its disposal, and reconsidering these in this new context brings out nuances that are sometimes overlooked.

Even restricted to core criminal offences, my claim sounds very strong. It suggests that AIT does not change much in relation to our ordinary practices of identifying courts with appropriate jurisdiction, appropriate defendants, the offence’s actus reus, mens rea, and defences. I think that a plausible defence of this strong claim is possible, but in this essay, I will undertake a rather more limited task. I restrict myself to considering only cases in which defendants might be accused of committing a crime within the territorial and procedural jurisdiction of a court, in circumstances in which the putatively criminal event was mediated by an autonomous choice made by AIT. The sorts of autonomous choices

I have in mind are those made by weak or narrow AIT[1] – systems that are intelligent when solving a specific problem, but would not pass general intelligence tests such as the Turing Test.[2] AIT of this sort makes decisions or choices in pursuit of its functions 'for itself' i.e. not under the express directions of a human controller, but based on its own evaluation of its environment, and with some foresight of the consequences of its choices.[3] However, it does not set, or reflect on the value of, the functions that it performs. It is not capable of being 'moved' by the reasons presented to it. It is not, in other words, itself an agent that qualifies for criminal liability. Narrow AIT can be contrasted with 'general AIT', which refers to intelligent machines characterised by their ability to replicate or surpass a broad range of human intellectual capacities, and characterised even more by the fact that they do not exist now, and will probably not exist in the determinate future.

Narrow AIT already exists and is commonplace. Chess-playing robots, autopilot technology, and self-driving cars (or even just self-parking cars and cars with cruise control): these are all science fact; not science fiction. My interest is in how we ought to evaluate cases in which this sort of AIT chooses 'for itself', i.e. not under the express directions of a human controller, but based on its own evaluation of the circumstances and the appropriate response thereto, and as a result brings about what looks like a crime.

For ease of analysis, in this essay, I will refer to a generic defendant as D. But I also have names for more specific defendants. The person who programs the AIT will be Penelope the programmer (P); the user who gives direct instructions to the AIT will be Ursula the user (U). In some examples I will

---

[1] See B. Goertzel, C. Pennachin (eds.), *Artificial General Intelligence* (Springer, 2007) 1; S.D. Baum, B. Goertzel, and T.G. Goertzel, "How long until human-level AI? Results from an expert assessment" Technological Forecasting & Social Change 78 (2011) 185–195 at 185

[2] S. Legg and M. Hutter, "Universal Intelligence: A Definition of Machine Intelligence" (2007) 17(4) Minds and Machines 391

[3] P. Stone et.al., "Artificial Intelligence and Life in 2030" (2016) Stanford University Research Report <http://ai100.stanford.edu/2016-report> accessed 24 November 2019, 12-13

also talk about Humera the hacker (H), and Olivia the owner (O). And should the analysis call for it, Vrinda the victim (V) will also make an appearance. The AIT in this story will usually be a self-driving car. But the ultimate question will always be the same – when the AIT chooses to act in a manner that suggests that an offence might have been committed, who, if anyone, is liable?

I. The defendant

The first thing to note is that given the scoping assumptions of this essay, the self-driving car, qua AIT, will not be itself criminally liable, since it lacks qualifying agency. For the same reason, it will not make anyone else liable under the innocent agency doctrine[4] either. Much like the gun or knife used by a murderer, the AIT is simply not an agent, innocent or otherwise.[5] Similarly, the limited doctrine of vicarious criminal liability is also inapplicable in the cases with which I am concerned. That doctrine makes one agent vicariously liable for the crimes of another,[6] but the AIT is not an agent. Accordingly, it can neither be held vicariously liable for the crimes of an agent, nor make an agent vicariously liable for its 'own' wrongdoing. The real defendant therefore is the agent (usually, a human) behind the AIT. This also means that the analysis of defences that call into question the agent's responsible agency, such as insanity, infancy, diminished responsibility, and (on some accounts of the defence) loss of control, remains unaffected to that extent. The insertion of AIT into the story makes no difference to the human defendant's agency.

II. The actus reus

---

[4] A.P. Simester et.al., *Simester and Sullivan's Criminal Law* (7th edn, Hart 2019) 227-30; D. Ormerod and K. Laird, *Smith, Hogan and Ormerod's Criminal Law* (15th edn, OUP 2018) 182-33; For a discussion of the underlying theoretical basis of the innocent agency doctrine see M. Dsouza, "A Philosophically Enriched Exegesis of Criminal Accessorial Liability" (2019) 8(1) UCL Journal of Law and Jurisprudence 1, 17-19

[5] Whatever views one might have about gun-control, in substantive criminal law at least, the NRA slogan, "Guns don't kill people; people kill people", is surely right.

[6] Simester & Sullivan (n4) 283; Smith, Hogan (n4) 266

Next, we need to consider whether the criminal law has the resources to help us identify cases in which the AIT's seemingly criminal activity (if any) should be attributed to a human defendant. I will address offences with three different types of actus reus stipulations separately since they raise different issues in the attribution analysis. These are, specific conduct offences, specific consequence offences, and state of affairs offences.

1. Specific conduct offences

Specific conduct offences are offences for the commission of which, a specific type of conduct is required. Examples include dangerous driving under s.2 of the Road Traffic Act 1988, careless and inconsiderate driving under s.3 of the Road Traffic Act 1988, and theft under s.1 of the Theft Act 1968 (which requires an 'appropriation'). These offences can sometimes *also* require a specific consequence to ensue – for instance, the offences of causing death by dangerous driving under s.1 of the Road Traffic Act 1988, and causing death by careless or inconsiderate driving under s.2B of the Road Traffic Act 1988, require both, a specific type of conduct ('driving in a dangerous manner', or 'driving in a careless or inconsiderate manner', respectively), and a specific consequence (the death of a human being).

In all cases in which an offence requires the performance of specific conduct, the human defendant must perform that conduct directly, or through an innocent agent,[7] or through a tool.[8] In the cases in which we are interested, the AIT performs the conduct, so we can rule out the first possibility – *ex hypothesi* the defendant has not done the thing directly. We can also rule out the second possibility – as mentioned above, innocent agency is inapplicable, since the AIT is not an agent. But what about the third possibility? Is the AIT a tool?

When 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.[9] Therefore, the thing done through the tool is conduct performed by D herself. So, 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

---

[7] Some, like S.H. Kadish, "Complicity, Cause and Blame: A Study in the Interpretation of Doctrine" (1985) 73 Cal L Rev 323, 373–77; and Simester & Sullivan (n4) 228–29, argue that certain types of conduct (what they call 'nonproxyable conduct') cannot be performed through an innocent agent. Accordingly, they would exonerate the defendant in cases requiring those types of conduct, if she did not perform the conduct herself. This is not my view. I have argued elsewhere that there is no reason to think that a human cannot perform specified conduct through the innocent agency of another. See Dsouza (n4) 17-19. Whatever view one takes on that controversy, no one seriously contests that at least *some* forms of conduct can be performed through the innocent agency of another. Hence, it is possible to steal through the innocent agency of a child who is told to take from the shop, or of an adult that one misleads as to the ownership of property. The innocent agent in these cases would be performing the actus reus conduct element of theft, *viz.* appropriation.

[8] *R v Clarence* (1888) 22 QBD 23; *R v Martin* (1881) 8 QBD 54; *R v Lewis* [1970] Crim LR 647

[9] J.K. Feibleman, "The Philosophy of Tools" (1967) 45(3) Special Forces 329, 330. The significance of the exercise of control is also hinted at in S. MacDonald, *Text, Cases and Materials on Criminal Law* (2nd edn, Pearson 2018) 48. One might analogise the exercise of control in this manner to a deliberate treating of something as a tool. Simester & Sullivan (n4) 227 emphasises the requirement in innocent agency cases, that the defendant intend to treat the intervening agent as an innocent agent. I have explained elsewhere that this intent to treat as an innocent agent is effectively an intention to treat an agent as a mere tool in the defendant's plans, and that the innocent agency doctrine works by creating a legal fiction that mirrors the defendant's intention, such that in law, the concerned legal agent is treated as a mere tool and drops out of the responsibility analysis. See Dsouza (n4) 17-19. Presumably, the same intent to treat something as a tool is also required in cases in which the putative tool is not a legal agent.

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 has appropriated the sausages (though D could be liable for other offences with different actus reus stipulations). And while the owner of a dog that causes injury while dangerously out of control may be convicted of 'being the owner' of a dog that did so under s.3(1) of the Dangerous Dogs Act 1991 she is not convicted of causing the injury herself. The applicable offence is a state of affairs offence.[10] 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.[11]

Similarly, if D *deliberately* uses the AIT as her tool, the AIT's conduct can be attributed to D. Note 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.

In sum therefore, the actus reus of specific conduct element offences can be performed through an AIT, but only in the rare cases in which D deliberately uses the AIT as a tool to perform the specific conduct.

2. Specific consequence offences

Some offences require that the defendant's conduct result in a specific consequence. For some of these offences – including murder and many forms of manslaughter, any conduct will do. For others, such as causing death by dangerous driving, and causing death by careless or inconsiderate driving under ss1 and 2B of the Road Traffic Act 1988 respectively, some specific conduct must have been performed by the defendant. However, *all* specific consequence offences require *some*

[10] I address these below in §II.3

[11] *Murgatroyd v Chief Constable of West Yorkshire* [2000] All ER (D) 1742. See also Smith, Hogan (n4) 664; Simester & Sullivan (n4) 457

'qualifying conduct', though the conduct need not necessarily have been performed by the defendant (or even by a qualifying agent). So for instance, in the state of affairs 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, the qualifying conduct is performed by the dog, though the defendant faces the criminal liability based on the existence of a state of affairs – her owning the dog.

Where the qualifying conduct must be performed by the defendant, she can do so through the instrumentality of an autonomously choosing AIT by deploying the AIT as a tool. This will be rare, given that an autonomously choosing AIT makes for an unwieldy tool.

In all specific consequence offences, if the qualifying conduct requirement is met, we must then check whether the consequence can also be attributed to that conduct. So for instance, imagine that the specific consequence that our offence proscribes is the death of Vrinda the victim. Where the offence requires that the qualifying conduct be performed by a human defendant, our prospective defendants include Penelope the programmer, Ursula the user, and Humera the hacker (or more than one of them). They might perform conduct like programming, using, or hacking into and reprogramming the AIT respectively. We need to ascertain when this conduct can be linked to the ultimate death of V.

The criminal law uses the rules of causation to link a principal's conduct to consequence.[12] Under these rules, the qualifying conduct can be linked to the proscribed consequence if it was a factual cause of the consequence,[13] and it was legally significant,[14] and salient to it.[15] The most important part of our analysis will relate to the application of the novus actus interveniens rules.

[12] Simester & Sullivan (n4) 90; See also Dsouza (n4) 4

[13] *R v Dalloway* (1847) 2 Cox CC 273, *R v White* [1910] 2 KB 124

[14] *R v White* [1910] 2 KB 124; *R v Cato* [1976] 1 WLR 110

[15] *R v Wallace (Berlinah)* [2018] 2 Cr App R 22

Doctrinally, an intervention breaks the chain of causation between an agent and a consequence if it was an unforeseeable natural (or at least non-agential) event,[16] or an independent intervention by the victim or some third party.[17] The AIT is not the victim, and it is not a third party, since it is not a qualifying agent, but

a. U or H could break the chain of causation between P's conduct and the consequence; or

b. H could break the chain of causation between P's or U's conduct and the consequence; or

c. U could break the chain of causation between H's or P's conduct and the consequence.

Each of these possibilities can be analysed under existing rules of criminal causation. But could the AIT's own choice break the chain of causation? The existing rules of criminal causation are perfectly well-equipped to answer this question as well. If the AIT's choice, though not dictated, or even foreseen, was foreseeable, then the answer is 'no'.[18] So in these 'foreseeable' cases, it is possible to bring about the actus reus through the AIT. In 'unforeseeable' cases, there is no particular reason to think that the human behind the AIT ought to be held criminally responsible for the consequence. But if we did want to hold her criminally responsible, state of affairs liability flowing from ownership of the AIT would remain possible even with AIT inserted into the story.

Where the offence requires that the qualifying conduct be performed by an autonomous non-agent like a dog, or what is of interest to us, an AIT, much the same causation analysis applies, with two qualifications. Firstly, the human defendant must obviously be replaced by the AIT as the performer of the conduct in the analysis. Secondly, and less straightforwardly, we need some

[16] *Environment Agency v Empress Car Co. (Abertillery) Ltd* [1999] 2 AC 22, 34-36; Simester & Sullivan (n4) 97

[17] *R v Kennedy* [2007] UKHL 38; *R v Jordan* (1956) 40 Cr App R 152. See also A.P. Simester, "Causation in (Criminal) Law" (2017) LQR 416

[18] This rule applies both to non-agential interventions and agential interventions. In respect of the former, see (n16). In respect of the latter, see *R v Roberts* (1972) 56 Cr App R 95; *Wallace* (n15)

explanation of the link between the AIT and the human defendant (in this case, Olivia the owner), and how it can support a blaming judgment. I address this second issue next.

3. State of affairs offences

Some offences have a third type of actus reus stipulation. These are the growing set of offences that do not require conduct, but rather a state of affairs, such as 'being in possession of something'. Examples include 'Possession of a controlled drug' under s.5 of the Misuse of Drugs Act 1971 and 'Going equipped for stealing' under s.25 of the Theft Act 1968.[19] So, how are these offences affected by the insertion of AIT into the mix?

It is conceivable that in some such offences, the possession of the AIT itself might be what is criminalised. For instance, D commits the s.25, Theft Act 1968 offence, if she is outside her home and has with her any article for use in the course of or in connection with burglary or theft. It is entirely possible that the article for use in the course of or connection with a burglary or theft may be an AIT. But this is an uninteresting case, insofar as it is not one in which the AIT actually does anything.

What if, instead, an AIT makes (unpredicted and not deliberately risked) choices that bring it into possession of contraband? In cases involving these 'out-of-control' acquisitions by the AIT, does D, the human behind the AIT, perform the actus reus of a possession offence? In 2014, a bot called Random Darknet Shopper, set up to make random purchases off the darknet for an art exhibition in Switzerland, went rogue and started buying ecstasy and fake designer handbags.[20] The police did not press charges against the art gallery, possibly because this was treated as a work of art, and therefore receives special protection under the Swiss constitution. However, that is the sort of case I have in

[19] This offence, despite its name, does not require the defendant to be 'going' anywhere. It requires only that she not be at her place of abode.

[20] M. Power, 'What happens when a software bot goes on a darknet shopping spree?' *The Guardian* (5 Dec 2014) <https://www.theguardian.com/technology/2014/dec/05/software-bot-darknet-shopping-spree-random-shopper> accessed 24 November 2019

mind here. There is some controversy as to whether one can unknowingly be in possession of something. In *Warner v Metropolitan Police Commissioner*[21], Lord Guest thought that "[i]f someone surreptitiously puts something into my pocket, I am not in possession of it until I know it is there", but some commentators express doubts about this proposition.[22] But to the extent that this controversy is live, it arises even outside of cases involving AIT. We can therefore bracket this concern off. What is more certain is that in cases involving the deliberate creation or underwriting of a state of affairs (such as cases of voluntary possession), D treats the AIT as a tool, and so 'performs' the actus reus of possession through it.

III. The mens rea

1. Preliminaries

Even after we show that one or more of P, U, H, or O have committed the actus reus associated with an offence (say) involving causing the death of V (despite the supervening autonomous choice made by the AIT), we are faced with the problem of establishing that the concerned human defendant had the mens rea to commit the prima facie offence. To some extent this will depend on the mens rea standard stipulated for the offence concerned.

But we might wonder whether offences that require subjective fault can *ever* be proved against a human defendant when the AIT autonomously chooses its actions. In fact, they can. For instance, autonomous non-agential intervening choice is compatible with having an intention as to the resulting consequences. The story of Androcles and the lion illustrates this. When Androcles was cast into the pit to be devoured by a lion, there was certainly an intention to kill him, and this intention existed despite the fact that the lion could in principle (as it did

[21] [1969] 2 AC 256. The House of Lords held that one could only possess something if one knew that it, or something not completely different from it, was in one's possession.

[22] See for instance, Smith, Hogan (n4) 162-64; Simester & Sullivan (n4) 179-82

in the story) decide not to eat Androcles.[23] There is no reason to think that the same is not true for lower subjective mens rea states like knowledge or recklessness.

2. Intention

So when can we say that P, U, H, or O intend, for instance, to cause the death of V through a supervening AIT? The answer, I propose, requires us to refer to the standard tests of intention. Before we get to the question of oblique intention, we should ask whether D desire the death of V either as an end in itself,[24] or as a necessary means to some other end.[25] We should also bear in mind when asking these questions that a conditional intention to cause the death of V (or any other specified consequence) is, for legal purposes, the same as an unconditional intention to cause the death of V (or other specified consequence).[26] The answers to these questions depend on the facts of the case before us, but one thing to note especially in relation to the second question is that a conditional *desire* is not enough to prove either unconditional intention, or conditional intention.

[23] There is some suggestion that this might also possible through autonomous agential intervening choice. Perhaps we could say that an evil supervillain who orders her assassin to kill V intends to cause the death of V intends to kill V. This view was alluded to by the SC in *R v Jogee* [2016] UKSC 8 [para 90].The SC stated that in many cases – especially of concerted physical attack – "there may often be no practical distinction to draw between an intention by D2 to assist D1 to act with the intention of causing grievous bodily harm at least and D2 having the intention himself that such harm be caused. In such cases it may be simpler, and will generally be perfectly safe, to direct the jury (as suggested in Wesley Smith and Reid) that the Crown must prove that D2 intended that the victim should suffer grievous bodily harm at least." I have my doubts about this proposition, but I will not explore them here, since an AIT is not an autonomous agent.

[24] *DPP v Smith* [1961] AC 290, 327; *Hyam v DPP* [1975] AC 55, 79

[25] *Hyam* (n24) 74

[26] *Attorney-General's Reference (Nos 1 and 2 of 1979)* [1980] QB 180; Simester & Sullivan (n4) 152

Imagine that P the programmer contemplates the possibility that the AIT she is programming might encounter a situation in which it must choose between killing one and five people.[27] She thinks that in such a situation, the AIT should choose to kill the one, and she programs the AIT accordingly. This programming choice does not *ipso facto* mean that P the programmer intends, or even conditionally intends, to kill (or cause the death of) eventual victim V at the time of making that programming decision. At the time when the programming decision is made, nobody needs saving, and it is possible (and in fact exceedingly likely) that P hopes that this trolley-like situation will not arise. If so, then P does not desire the death of V any more than she desires her burglar alarm to go off when she sets it before leaving her house. Since P would not be disappointed if the AIT never encountered a trolley-like situation and had to choose to kill a person, P does not intend, even conditionally, to kill the person.

If P had actually (and not conditionally or contingently) *desired* that the AIT encounter a trolley like situation, then that would of course be instructive. Consider D's intention to steal from a bag, if there be something of value in it: if an external condition is met (there is something valuable in the bag), then D plans to bring about something – a permanent deprivation of someone else's property (by appropriating it, in this case, for herself). It seems very likely that D also *desires* that the external condition be met. This desire is not contingent or somewhere in the future – it exists in the present. And the fact that D has this present desire that the condition be met also suggests that D has a present (albeit conditional) desire to bring about the permanent deprivation of the other's property. After all, if there were nothing valuable in the bag, D would be disappointed, because she would be unable to steal anything valuable. This desire to do something indicates a direct intention to do it. On these facts, it makes sense

[27] This is a variation on the classic trolley problem. See P. Foot, 'The Problem of Abortion and the Doctrine of Double Effect' (1967) 5 Oxford Review 1, 3. See also J.J. Thompson, 'The Trolley Problem', in *Essays on Moral Theory* (Harvard, 1986) 94–116

to say that D (presently) intends to steal anything valuable that may (in the hypothetical future) be in the bag.

In other words, a *contingent* desire *that something happen* – in this case, that the AIT choose to kill the one rather than the five if ever faced with the trolley-like situation – is not instructive in the same way as a present desire that something happen, since it gives us no indication of *present* desires as to the performance of the potentially criminal conduct.

Nor can we say that P has an oblique intent[28] to cause anyone's death. Oblique intention will suffice for almost every offence for which intention is required,[29] but one has merely to state the test for oblique intention to see that it is inapplicable on the facts above. It is not virtually certain that a trolley-like problem will ever arise – in fact, it is very unlikely. So even the first condition for oblique intent is not met. The fact that the AIT would make a foreseeable choice should it ever be faced with a trolley-like situation, does not suggest that the programmer realises that it is virtually certain that the AIT *will* kill V (or anyone else).

On the other hand, if the AIT was virtually certain to cause the death of a person, and P realised that, then applying the standard test for oblique intention, the jury can find that she intends that death when she deploys the AIT anyway. Therefore, although it will be a rare case in which P, U, H, or O will intend to kill through the medium of an autonomously choosing AIT, there appears to be no need to come up with new tests for intention to deal with the advent of narrow AIT.

3. Knowledge/belief:

[28] *R v Woollin* [1999] AC 82

[29] But not every such offence – for instance, the offence under s.44 of the Serious Crime Act 2007 seems to insist on direct intent.

Some offences are so defined that they can be committed when the defendant acts knowing or believing that certain circumstances exist, or that certain consequences will ensue. In either case, the addition of an autonomous AIT makes little difference to the analysis of the human defendant's knowledge or belief. We simply need to ask whether D performed her conduct (be it programming, using, hacking into, or, on a stretched interpretation of the word 'conduct', knowingly owning the AIT) with the knowledge or belief required for the offence. The knowledge or beliefs of the AIT (assuming that AIT actually forms beliefs in the same sense as humans form beliefs)[30] are irrelevant, since it is not the defendant. This mens rea state cannot be satisfied by showing that the human defendant acted despite the fact that the AIT she programmed, used, owned, or hacked into, believed or knew some facts – the culpability in offences requiring knowledge or belief traces to the defendant's advertent choice to act *despite* knowing or believing certain facts.[31] The defendant's choice to act is simply not culpable in the same way when she herself did not know, or believe, the relevant facts, even if the AIT 'knew' or 'believed' them.

4. Recklessness and Negligence:

Other mens rea standards such as recklessness and negligence could also be met by P, U, H, or O. If that is the allegation, and recklessness or negligence satisfies the mens rea for the prima facie offence, the analysis is simpler. A person is reckless as to circumstances or consequences if she was subjectively aware of the risk of these circumstances existing, or consequences ensuing from her conduct, and unreasonably took that risk.[32] A person is negligent if she unreasonably took a risk, and ought to have known of the existence of the

[30] See D. Dennett, *Intuition Pumps and Other Tools for Thinking* (Penguin 2014) 91-97 for some scepticism as to this.

[31] M. Dsouza, 'Corporate Agents in Criminal Law – an Argument for Comprehensive Identification' (September 2, 2019) 1-32, available at SSRN: https://ssrn.com/abstract=3446666, 24

[32] *R v G & R* [2004] 1 AC 1034

risk.[33] For both these mens rea states though, the common issue is the reasonableness of the risk-taking. The key question here is not, "How did/will the AIT respond to the situation?" It will respond to the situation in accordance with its programming or the instructions given by humans. Unlike humans, AIT will not choose as a moral agent. It will also not choose arbitrarily (except perhaps if it was programmed or instructed to choose arbitrarily, in which case too it is choosing as programmed or instructed). Even to the extent that AIT is empowered to teach itself the best ways to achieve its goals, the AIT is told its goals, and is programmed to be able to learn. Therefore, we focus instead on what the defendant did – program or instruct the AIT. If the defendant's programming or instruction of the AIT was reasonable, then the defendant lacks the mens rea for offences predicated on negligence or recklessness. So for instance, a defendant who was properly sensitive to V's interests while programming or instructing the AIT was not reckless or negligent as to harming V, because she acted reasonably.

What it was reasonable to program or instruct the AIT to do depends, to some extent, on the epistemic capabilities of the AIT, since these will affect what one may sensibly ask of the AIT. It is theoretically possible for the AIT's sensors to be of such poor quality that it would be unreasonable to deploy the AIT at all, since there would be no reasonable instruction that one could give such an AIT in respect of a foreseeable situation that it might encounter. Assuming that the AIT's sensors meet industry specifications, and that the industry specifications are a good proxy for a reasonable quality and quantity of sensors, that seems unlikely, and so we can set this possibility aside for now.[34] The reasonableness of the defendant's programming

[33] G. Williams "Carelessness, Indifference and Recklessness: Two Replies" (1962) 25 MLR 49, 57; Smith, Hogan (n4) 113; Simester & Sullivan (n4) 166

[34] Of course, if we did want to expand this analysis to include this question, we could. It might, for instance, become relevant in the context of new technology that does not yet have industry standards, or for backyard inventors or tinkerers who fine tune their AITs to meet what they consider higher performance specifications. If so, we can also introduce Malena the manufacturer into the picture. But in order to keep this essay to manageable proportions, I will not attempt to do so here.

or instructions then depends on our evaluation of the quality of the guidance given to the AIT. This in turn, depends on what the AIT's sensors can sense, how fast it can process that information, and how fast it can process and act on its guidance as to how to respond.

Humans give the AIT this guidance. They tell the AIT what to do if faced with a choice between life and property; or between two or more lives; or between different pieces of property; etc. So long as this guidance conforms to prescribed industry standards (if any, and subject to a qualification to follow), or in the absence of those, to enough of the jury's standards of reasonableness to make a conviction impossible, the defendant does not act negligently or recklessly as the case may be.

In sum, the key issue is not what outcomes occur in variations of the trolley question – it is, what advance guidance is reasonable to give. That depends on the information and capacities we have when giving the guidance.

Decisions on the reasonableness of guidance might perhaps be facilitated by a centralised AIT ethics checklist – something like Asimov's three rules.[35] But that would be a limited solution – someone who figured out a loophole and deliberately set out to exploit it to cause harm ought not to be able to escape liability by pointing to formal compliance. In fact, the intent to exploit a loophole should itself inculpate, and this would be possible if compliance with our AIT ethics checklist was treated as evidence of reasonableness, rather than being constitutive of it. It would be sensible to treat our ethics checklist, should we have one, as a set of rules of thumb

[35] I. Asimov, "Runaround" in *I, Robot* (Doubleday, 1950) 40. Note however, that Asimov's rules leave far too many things unclear. For one thing, although the first, and most important, rule is that "A robot may not injure a human being or, through inaction, allow a human being to come to harm", how should a robot respond in a situation in which it will injure (or allow injury to) a human no matter what it decides to do? For a brief discussion of the flaws in Asimov's rules, see J. Tasioulas, "First Steps Towards an Ethics of Robots and Artificial Intelligence" [2019] *Journal of Practical Ethics* <http://www.jpe.ox.ac.uk/papers/first-steps-towards-an-ethics-of-robots-and-artificial-intelligence/>, accessed 24 November 2019

– a work constantly in progress. It should constantly be open to evolving, perhaps with the state periodically underwriting the current statement of ethical principles, which may then be re-evaluated based on academic and industry critiques, and in the event that we discover problems, through trial and error. But to be clear, developing such a set of rules is not the task I set myself in this essay.

IV. Consent

One question that cuts across the actus reus/mens rea divide is whether the putative victim's consent, where it is legally valid, will affect our defendant's liability, even when the defendant (or as the case may be, the AIT) could not have been aware of, and responsive to, that consent when performing the qualifying conduct. In doctrinal criminal law, where consent matters, the granting of consent is best conceived of as a factor that negates the actus reus of an offence,[36] whereas a belief (or sometimes, a reasonable belief) in consent may negate the mens rea of an offence.[37]

To the extent that consent implicates the actus reus of the offence, the defendant's reliance on or uptake of the consent is not necessary for the negation of the actus reus. Hence the putative victim's consent, if valid, means that the AIT does not occasion even a prima facie offence. And insofar as the mens rea of the offence is concerned, given that the conduct performed by the defendant will usually be remote from the circumstances in which V refuses or grants consent, the defendant would rarely be able to credibly claim that she believed that V had consented.[38]

---

[36] M. Dsouza, "Undermining Prima Facie Consent in the Criminal Law" (2014) 33 Law and Philosophy 489, 494-97

[37] ibid at 497-98. A belief in consent may also occasionally support a claim to a rationale-based defence. However, for reasons that will become apparent, I do not address the role of consent in defences separately.

[38] The same applies when the defendant relies on a belief in the victim's consent in claiming a rationale-based defence. For this reason, I do not address the analysis in relation to such claims separately.

But again, none of this is a departure from standard principles. The key questions remain what they always have: "Did the putative victim grant legally valid consent?" and "Did the defendant believe (or reasonably believe, as the case may be) that the putative victim had consented?".

V. Contemporaneity

Even so, there remains another problem. A prima facie offence is only committed when the actus reus and the mens rea exist contemporaneously. However, it appears that P, U, H, and O will in most cases have formed their respective mens rea states in advance of the AIT choosing to bring about the actus reus. Again though, this problem is hardly a new one for the criminal law, and the criminal law already has the resources to deal with it. The continuing act principle allows the criminal law to stretch the conduct involved in the actus reus beyond the point at which D was actively doing something, up until the mens rea was formed,[39] or to link it to some previous conduct performed at a time when there was mens rea such that the entire chain of conduct is seen as a single continuing act. Again, the advent of AIT doesn't require us to reinvent the criminal law.

VI. Rationale-based defences

However, the set of concerns most frequently voiced are to do with instances in which the AIT is faced with a difficult choice about which criminalised outcome to bring about. This often manifests in asking how AIT should deal with Trolley Problems and who, if anybody, should bear the responsibility for the choices that AITs make in such cases. Trolley Problems are usually puzzles relating to the sorts of defences that are known as 'rationale-based defences'. These defences include claims of self-defence, prevention of crime, arrest, duress, and necessity. In making any of these claims, the defendant asserts that she deliberately did what amounted to a prima facie offence, but did so for good reasons. They are not generally available in respect of prima facie offences that were constituted by the defendant's

[39] *Fagan v Metropolitan Police Commissioner* [1969] 1 QB 439; *R v Church* [1966] 1 QB 59; *R v Le Brun* [1991] 3 WLR 653

negligent conduct – in those offences, the defendant makes no deliberate choice to commit the prima facie offence, and so it is meaningless to talk about her motivations for offending. I have, in previous work, set out a theoretical framework within which we should understand rationale-based defences.[40]

The invocation of Trolley Problems in relation to AITs is usually misleading – the trolley problem is a problem involving a claim to a defence when an *agent* chooses to commit a prima facie offence in response to a perceived present threat. But in AIT cases, a non-agent is making this choice in response to the perceived present threat. The human agent's choice is usually made well in advance of any perceived present threat, when P, H or U interact with the AIT. At that stage they plan generalised threat-management strategy rather than responding to specific extant threats. They cannot therefore seek exculpation for their actions by claiming a rationale-based defence – instead, they may rely on the reasonableness of their actions to deny mens rea. I have already outlined how that argument would proceed.

Still, it can, on rare occasions, be appropriate to consider whether a rationale-based defence is available, and so a brief examination of such claims is useful. Any claim of a rationale-based defence involves two stages. First, the defendant must form the belief that facts exist that necessitate the use of defensive force (and decide to use the defensive force for those reasons). Depending on the defence in question, this belief may have to satisfy an objective test of reasonableness as well. Once the defendant crosses the first stage (and only then)[41], do we reach the second stage, at which the defendant considers her response. Only at this stage, will she consider the normative appropriateness of her defensive option(s). It is just about within the realm of possibility that U or H might be responding to a present threat. For instance, one can just about imagine U or H seeing a present (though slow-burn) threat and thereupon instructing or reprogramming their AIT in the hope that it

[40] M. Dsouza, *Rationale-Based Defences in Criminal Law* (Hart 2017)

[41] *R v Field* [1972] Crim LR 435 – no need to retreat because no present threat; *R v Cockburn* [2008] 2 Cr App R 4 – no 'pre-threat' defensive measure). M. Dsouza, "Retreat, Submission, and the Private Use of Force" (2015) OJLS 1, 23-24

will (autonomously, but foreseeably) react to the threat in a particular way, which involves committing a prima facie offence. On this peculiar set of facts, the defensive claim potentially available would be rationale-based. But the analysis required to adjudicate this claim would be exactly the same the one we would undertake in non-AIT infused claims to a rationale-based defence. We would ask: "Did the defendant think there was a present threat (D's conclusion might sometimes need to also be objectively reasonable)?", and "Was the defendant's chosen response appropriate (this question must be answered by reference to objective standards)?" Again, there is no need to reinvent the wheel here.

VII. Application

Let us now apply the analytical approach sketched above in relation to liability as a principal to our prospective defendants, starting with Penelope the Programmer. Let us say that she programmed the AIT – a self-driving car – such that down the line, it chose to do something seemingly criminal. Doing that thing could itself be an offence, or the consequence of doing that thing – V's death might make it an offence. If the offence requires the performance of some conduct (say dangerous driving), we should ask whether P deliberately used the car as her tool to perform this conduct. If so, then the analysis may continue, but if not, then P does not commit the actus reus of the offence. If the offence also requires a consequence – say the death of V – then we apply the usual rules of causation, and consider whether either a human user or hacker, or the car's own autonomous choice broke the chain of causation. This last question depends on whether the car's autonomous choice was foreseeable. If it was, then the chain of causation is not broken, and the inquiry may continue.

We next consider mens rea. If the mens rea required includes intention as to one of the actus reus elements, we apply the ordinary tests for intention in the criminal law to ask whether P had the requisite intention at the time she programmed the car. If the mens rea required includes knowledge or belief as to something, then we ask whether P knew or believed that thing. If the mens rea required includes recklessness or negligence as to a circumstance or consequence, we consider what P

subjectively knew of or foresaw the concerned circumstance or consequence, or what she ought to have known or foreseen it as the case may be, and consider the reasonableness of P's programming choices in view of that epistemic profile. If at the end of this analysis P is still prima facie liable for the offence, we might consider whether she has any rationale-based defences available to her. The reasonableness analyses necessary to make this determination would also proceed along the same lines as the one made in respect of assessing recklessness and negligence.

Almost the entirety of this analysis can also be applied to Ursula the User, who instructs the car such that, down the line, it chose to do something seemingly criminal. The only difference is that the programmer's actions are unlikely to be potential breaks in the chain of causation.

Finally consider Humera the Hacker. Most of the analysis above will apply to her as well, with one added complication. H's conduct would presumably amount to at least two different sets of (prima facie) offences; those relating to her very act of hacking the AIT, and those relating to the criminalised outcome brought about by the car as a consequence of that hacking. Although our focus here should be on the latter set of offences, it is clear that our findings in relation to these will be influenced by our findings in relation to the former set. If there is no basis for thinking that it was right, or at least acceptable behaviour for H to have hacked the car, then it seems unlikely that there would be a basis for thinking that reprogramming the car to act as she did was reasonable. The converse is equally true as well. If we thought that it was reasonable for H to have reprogrammed the car as she did, we would usually think that it was also right or societally normatively acceptable for H to have hacked the car in the first place. But these things could, in principle, come apart. We could imagine that H had installed a backdoor in the programming of the car for some malicious fun months ago, but now, seeing an impending disaster, she uses the backdoor to reprogram the car and try to save the day, albeit at some possible cost to some innocent bystander – V. In such a case, we would analyse H's actions in relation to each prima facie offence separately.

VIII. Secondary Liability

In addition to liability as a principal, the criminal law also imposes secondary liability on agents, either as accessories to others' crimes, or by way of inchoate liability for crimes that were not necessarily[42] ever completed. I cannot enter into a detailed discussion of whether the advent of AIT affects the analysis for these forms of liability here, but I briefly set out my reasons for thinking that it does not.

1. Accessorial liability:

D is an accessory to an offence committed by another if she aids, abets, counsels or procures the commission of that offence.[43] But how much does the traditional analysis of this form of liability change when D is the programmer, owner, user, or someone who hacked into and reprograms an AIT that then autonomously does something that aids, abets, counsels, or procures an offence committed by a human principal? The answer is, not much.

As far as the actus reus for being an accessory is concerned, D can theoretically aid, abet, counsel or procure the criminal offence of another through the instrumentality of an autonomously choosing AIT. But this will be exceedingly rare, since D could only do so by using the autonomously choosing AIT as a tool. This seems unlikely. Since the AIT will choose its conduct autonomously, it would be a relatively unreliable tool. But should such a case arise, it has already been demonstrated that doctrinal criminal law has the analytical tools to address it.

The mens rea for accessorial liability is notoriously complicated – we must show that D intended to perform the conduct that she did perform, intended thereby to aid, abet, counsel, or procure the principal's conduct,[44] and knew that in performing the conduct assisted or encouraged, the

[42] Note that most inchoate offences can technically be charged even if a completed offence was committed. See s.6(4) of the Criminal Law Act 1967 in the context of criminal attempts; s.56 Serious Crime Act 2007 in the context of the offences of encouraging or assisting crime.

[43] s.8 Accessories and Abettors Act 1861

[44] *R v Bryce* [2004] EWCA Crim 1231; *R v Derek William Bentley (Deceased)* [2001] 1 Cr App R 21

principal would commit an offence.[45] Even so, it is composed of various subjective fault elements that have already been analysed above. In none of those cases, did the insertion of AIT into the story stretch the analytical resources already at the disposal of doctrinal criminal law, and there is no reason to think that the combination of these factors would do so either.

2. Inchoate offences:

There are several different inchoate offences, but once again, they are all composed of actus reus and mens rea elements that have previously been analysed. Thus inchoate offences will typically require the defendant to perform some conduct,[46] but will typically not require any consequence to follow. Once again, D can perform this conduct through the instrumentality of the autonomously choosing AIT by using it as a tool, but once again, for obvious reasons, D's chosen tool is unlikely to be an AIT that chooses its conduct for itself. The mens rea stipulations for different inchoate offences differ, and there are too many of these to analyse in detail in this piece. Suffice it to say that each of these are constructed out of subjective fault states that, as previously established, can be applied also to cases involving AIT. Therefore, as is true for the plethora of offences that have mens rea stipulations that include various permutations and combinations of these fault states, there is no reason to believe that they will create special difficulties in combination.

Conclusion

This piece has involved a necessarily brief survey of the main topics considered during undergraduate substantive criminal law modules, with a view to examining whether the introduction of narrow AIT

[45] *R v Jogee* [2016] UKSC 8; *Johnson v Youden* [1950] 1 KB 544

[46] Doing something more than merely preparatory to the commission of the offence for attempts liability under s.1 Criminal Attempts Act 1981; making an agreement (containing specified terms) with another person for conspiracy liability under s.1 Criminal Law Act 1977; doing something capable of encouraging or assisting an offence for inchoate liability under Part 2 of the Serious Crime Act 2007, etc.

necessitates a rethinking of the fundamentals of criminal law. While the re-examination of these foundational matters in a new context is always valuable, and often sheds new light on less prominent features of the criminal law, my main conclusion is that the substantive law of core criminal offences has the resources to deal with cases involving narrow AIT. And given that general AIT is not yet appearing on the horizon anytime soon, for the present at least, the emergence of AIT gives us no reason to panic.