Artificial intelligence, criminal acts, criminal responsibility

Manolis Melissaris*
PhD in the Philosophy of Law, author

Harm or risk of harm is frequently caused in the course of the operation of artificial intelligence (AI) machines.[1] In 2015, a stationary robot at a Volkswagen factory grabbed and crushed against a metal plate the worker setting it up, resulting in the worker's death.[2] Facial recognition software regularly leads to wrongful, and biased, arrests.[3] ChatGPT falsely identified a law professor as having committed sexual harassment.[4] An algorithm, dubbed Random Darknet Shopper, created but not controlled by two artists, "bought" illegal drugs online.[5] Image-generative AI "creates" child abuse images on an alarmingly large scale.[6]

Generally speaking, AI machines display capacities, such as reasoning, learning, remembering and planning, normally associated exclusively with humans. In fact, it is on the basis of their resemblance of their activity to that of human actors that they are customarily juxtaposed to other, non-intelligent, machines.

First generation AI machines could process a very narrow amount of data and perform rather limited tasks. They could complete a vast number of calculations very fast. Programmers fed into the machines data *and* very simple and clear rules on a massive scale. That gave the machines the capacity to know at each junction what they should "do" so as to carry out an operation so adeptly as often to outperform humans, even defeat grand masters at chess.

* I am indebted to Dr Apostolos Georgiadis for enlightening me on artificial neural networks.

1. Readers that way inclined can follow a regularly updated report of AI-related incidents here: https://incidentdatabase.ai/.
2. https://www.theguardian.com/world/2015/jul/02/robot-kills-worker-at-volkswagen-plant-in-germany
3. https://theconversation.com/ai-technologies-like-police-facial-recognition-discriminate-against-people-of-colour-143227
4. https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/
5. https://www.bbc.com/future/article/20150721-my-robot-bought-illegal-drugs
6. https://www.theguardian.com/technology/2023/oct/25/ai-created-child-sexual-abuse-images-threaten-overwhelm-internet

Eventually, it was felt that this model was too limited, not least because it required a vast amount of programming in order to draft exhaustively the rules that the machines were meant to follow.

Enter neural network technology.

Recent, advanced robots operate with interconnected artificial neural networks (henceforth ANNs), which are mathematical functions, modelled on the human brain. The technology had been theoretically known for a while but only became possible to develop in the past twenty years with the increase in computational power and the decrease of running costs. Programmers no longer provide the rules. They only provide the data and the machine processes those through ANNs. Advanced AI displays the reactive skills of its predecessors but also has the ability to make sense of, even imperfect, information, and develop further practical, problem-solving abilities based on the knowledge that it accumulates in the process.

Crucially, and in contrast to earlier machines, although we know what kind of task the machines will perform, we have no idea *how* they do it, at least not yet. They learn by example and through experience, not by rule-following. This means that they are autonomous and, very importantly, opaque and unpredictable even to their own programmers. They are, as the expression goes, "black boxes".

It is those machines that present an interesting challenge in terms of legal responsibility generally and criminal responsibility in particular and it is those machines that I have in mind in this paper.

Where lies the challenge?

The intuition is, no doubt, strong that the injurious or dangerous incidents such as the ones outlined in the introduction, are more than mere accidents. Many express this in terms of an analogy: if a human killed or defamed or purchased illegal substances online, they would be criminally liable.[7] This, however, begs the question, in that it assumes that a wrong *has* been committed and that, accordingly, the conditions of responsibility hold, although it is unclear how exactly this may be so. That, however, is precisely what the enquiry is about.

The intuition does not need such overstating. An accident is something sufficiently distant, though not necessarily entirely unconnected, to one's actions. They can be caused, say, by

7. Himmelreich, for instance, draws the analogy in relation to autonomous war machines. J. Himmelreich, "Responsibility for killer robots", *Ethical Theory and Moral Practice*, 22, 2019, 731–747.

natural phenomena or the inevitable wear and tear of objects beyond the proximate control of anyone in particular. In the case of AI machines, though, the fact that the robots simulate human behaviour, suffices for us to sense strongly that the harm does not come about randomly, that it would not have happened had it not been for *some* action determined by *some* kind of decision-making. This makes it perfectly reasonable for us to ask the responsibility question, in order to determine whether a wrong has been committed.

No sooner have we done so, however, that we stumble upon obstacles. As we saw, their programmers or owners have relinquished control over ANN-operated machines. They feed data into the robot but how these are processed and the results of the process are determined entirely by the robot's "decisions" and "actions". This intervention has enough of an impact on the outcome to appear impermissible to hold the programmer criminally responsible.

This leaves us with only one alternative, to blame the machine. Most people would consider this a conceit. Brożek and Jakubiec, for example, consider AI machines not even *prima facie* eligible as bearers of legal responsibility. They argue that the law adopts a folk psychological understanding of agency. The legal agent is one who can respond to and relate to her environment in the relevant manner, which includes relations of duty and responsibility. Machines are, at least for the time being, incapable of this and should therefore be discounted.[8] Brożek and Jakubiec are broadly right but the argument needs fine-graining and substantiating with greater focus. As will become evident later, this is what I will try to do in this paper.

But, even if they passed this threshold eligibility test, most seem to emphasise that robots would fail subsequent ones, as they do not display, or very convincingly hide, characteristics that our institutions of responsibility require, mainly the practical reasoning ability that allows one to form judgement. Therefore, "punishing" robots would be as confused as it would be risky, because it could undermine the rule of law and the moral foundations of our criminal law institutions.

Many consider it a frustrating impasse that, on the one hand, machines "do" things, which trigger in us reactions that we would normally reserve for blameworthy behaviour[9] and, on the other hand, because of the nature of robots and the fundamental principles of our institutions of responsibility, no one can actually be held responsible, let alone accountable. This "responsi-

8. B. Brożek and M. Jakubiec, "On the Legal Responsibility of Autonomous Machines", *Artificial Intelligence and Law*, 25, 2017, 293–304.

9. This echoes Strawson on responsibility. See P.F. Strawson, "Freedom and Resentment", in *Proceedings of the British Academy*, 48, 1962, 1–25.

bility gap", as dubbed by Andreas Matthias already in 2004, is only going to widen further the more the technology advances and continue to undermine our basic practices of blaming and punishing.[10] Matthias's proposed solution was to curb the development and spread of Artificial Intelligence. Almost twenty years on, his call has so far obviously fallen to deaf ears.[11]

Some share Matthias's concern and have accepted and developed further the idea of a responsibility gap.[12] Others are unconvinced that there is a gap at all.[13]

Finally, there are those who do not give up on the conceptual possibility of ascribing responsibility to AI machines. Focusing exclusively on criminal responsibility, and assuming that the problem relates to *mens rea*, Abbott and Sarch suggest three possible ways of breaking the impasse. [14] The first is to impute to machines the mental and volitional states of their creators or initial users. The second option is a kind of strict liability. The third, and more interesting for our purposes, is to impose direct criminal liability to the extent that "the AI is programmed to be able to take account of the interests of humans and consider legal requirements, but ends up behaving in a way that is inconsistent with taking proper account of these legally recognized interests and reasons", because then the robot will have displayed the disregard for others that is the subject of reproach by criminal law".

Similarly, Gabriel Hallevy believes that AI machines are cognitively equipped to form the knowledge required to intend for the purposes of criminal law. Since their acts also meet all the relevant requirements, so Hallevy argues, then there are no conceptual obstacles to establishing criminal responsibility.[15]

10. A. Matthias, "The responsibility gap: Ascribing responsibility for the actions of learning automata", *Ethics and Information Technology*, 6(3), 2004, 175–183.

11. While this paper was being written, a summit of political, tech, and business leaders took place to discuss the risks posed by the development and expansion of the use of AI, and ways to regulate it (london.theaisummit.com/)

12. See F. Santoni de Sio and G. Mecacci, "Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address Them", *Philosophy and Technology*, 34, 2021, 1057–1084; J. Danaher, "Robots, law and the retribution gap", *Ethics and Information Technology*, 18(4), 2016, 299–309; R. Sparrow, "Killer robots", *Journal of Applied Philosophy*, 24(1), 2007, 62–77.

13. See S. Köhler, N. Roughley, H. Sauer, "Technologically blurred accountability", in C. Ulbert, P. Finkenbusch, E. Sondermann, T. Diebel (eds.), *Moral agency and the politics of responsibility*, Routledge 2018, 51–68; D.R. Tigard, "There Is No Techno-Responsibility Gap", *Philosophy & Technology*, 34, 2021, 589–607; F. Hindricks and H. Veluwenkamp, "The risks of autonomous machines: from responsibility gaps to control gaps", *Synthese*, 201, 2023, at 21.

14. R. Abbott and A. Sarch, "Punishing Artificial Intelligence: Legal Fiction or Science Fiction", *UC Davis Law Review* 53(1), 2019, 323-384.

15. G. Hallevy, "The Criminal Liability of Artificial Intelligence Entities - from Science Fiction to Legal Social Control", *Akron Intellectual Property Journal*, 4:2, 2010, 171-201.

Robot acts

We tend to think of AI machines as driving, writing, diagnosing, predicting, recognising, in other words as acting in much the same way as we do. Is it really exactly so?

The roughly one hundred billion neurons in our body, located mainly in the brain and the spine but also in peripheral ganglia, are cells that receive sensory input from the external world, communicate with each other and send motor commands to our muscles. The stimuli are received by the dendrites, so named because they resemble tree branches. A signal is then passed on to the axon, which runs like a cable through the neuron. The axon transmits an electrical charge to the synapse, the point of connection between neurons (each neuron has approximately one thousand synapses, which adds up to a staggering total of one hundred trillion). Subsequently, it is briefly converted into a chemical neurotransmitter and then back into an electrical signal to the dendrite of another neuron. The messages are finally carried to muscles, the muscle fibres contract and our bodies move.

ANNs work in much the same way only, this time, the signals are in human-made code and the neurons are mathematical functions. To put it simply, by hitting our keyboard we send voltages into the computer, i.e. binary 01 signals, which constitute the data input. This triggers the algorithm, the necessary computations are made and passed on from neuron to neuron in the hidden layers, with some toing and froing, until the machine is satisfied that it has the correct answer or course of action at the output layer.

For the machine to produce substantial results, it needs to be trained. Unlike older technology machines, which required detailed rules drafted by and known at all times to the programmer, new generation robots learn by example and experience. For example, ChatGPT is fed data one tiny morsel at a time so that when it is next presented with the same sequence, it will be able to work out what follows. Consider that trillions of texts, pretty much the entirety of the content of the internet, have been inserted into Chat GPT and you will see how it manages to successfully, at least on the surface, compose complete texts in response to prompts or questions.

So, on one level, one might, plausibly enough, argue that there is no difference between robot and human actions. When computers see, say, a cat and recognise it as such, they break down, with unthinkable speed, incredible accuracy and, as we have already mentioned, near perfect opacity, the image of a cat into pixels, run the parts by the data collection that they have accumulated and conclude that it is indeed a cat. When *we* see a cat, the light hits the retina, it is then converted into an electrical signal, which is transmitted to the brain for neurons to take over and

for us eventually (barring any malfunctions, all this happens in nanoseconds, which makes it feel instantaneous) to recognise the object as a cat.

Be that as it may, we are interested in something more specific, namely whether AI machines can be held to be acting *in a criminal law relevant manner*, whether their acts satisfy the *actus reus* requirement.

Actus reus and the consciousness assumption

As I mentioned earlier, some assume that AI machines not only act but that they also act in a guilty manner.[16] The *actus reus* requirement being fulfilled, the only outstanding and pertinent question is whether computers can form the requisite *mens rea*. I find this view rather hasty and ultimately mistaken, because it underappreciates the nature of *actus reus*.

Humans perceive themselves as possessing what has come to be called consciousness. There is little consensus among scientists and philosophers as to what consciousness *is*, let alone if it has any real counterpart at all. For all we know, we might all be plugged into a computer and what we think of as a unique faculty that differentiates us from other beings is an illusion or a trick played on us. Nevertheless, this does not change our *perception* of the layers of our existence. Taking the first-person perspective, we experience and represent the world as it appears to be and we place ourselves inside it in a relationship of interdependence – changes in the world affect us and vice versa. We do so with some coherence across time and space, remembering ourselves in the past and picturing ourselves in the future. At the same time, we also distance ourselves from it, thus being able to develop a reflexive attitude towards it.

This consciousness assumption is built into our institutions, the law being chief amongst them. Crucially, it underpins conceptions of legal responsibility and is reflected in the act requirement in criminal law, in two ways.

i. The first is rather obvious but surprisingly absent from the discourse on AI and criminal responsibility. When criminal law requires that its subjects act (or not act, when under a duty so to do) for them to be criminally liable, it requires them to do so consciously, with self-awareness, and by exercising volitional control over their movements. In other words, it expects them

16. More often than not, unquestioningly so. For example: "*It is relatively simple to attribute an actus reus to an AI system. If a system takes an action that results in a criminal act, or fails to take an action when there is a duty to act, then the actus reus of an offence has occurred*". J.K.C. Kingston, "Artificial Intelligence and Legal Liability", in *Research and Development in Intelligent Systems XXXIII: Incorporating Applications and Innovations in Intelligent Systems XXIV*, Springer 2016, 269-279.

to intend to perform the movements that bring about the requisite result. The *actus reus*, the objective aspect of the offence, is not devoid of all mental elements, although this residual subjectivity remains normatively neutral and should carefully be distinguished from the intention to commit the act *qua* offence, which relates to a moral attitude that the defendant must have developed.

ii. Apart from the intention to move one's body, criminal law offences typically also require knowledge of or a belief regarding the circumstances of the commission of an offence, the difference being that knowledge is justified belief whereas sheer belief is purely subjective. I would suggest, however, that both these requirements are placed within the framework of a wider-ranging faculty attributed by criminal law to its conscious subjects.

Here's an example. One of the ways of committing the offence of fraud in England and Wales is by false representation. According to sections 1 and 2 of the Fraud Act 2006, the *actus reus* requirements of the offence are: making a representation, in relation to fact ("I am the heir to a world's richest person"), the law (e.g. "you owe me X amount of money"), or the state of mind of the person making the representation or a third person (e.g. "your boss orders you to give me X amount of money"), made either orally, in writing or through a machine (so as to cover fraudulent use of bank cards and the like). Note that it is not required that the representation have any result. Neither does its addressee need fall for the deception nor does the person making it need make any gain or cause loss (though the intention to bring about one or the other is a *mens rea* requirement of the offence).

For one to commit fraud by false representation, one must indeed voluntarily perform some bodily movements, for example uttering or typing the words that describe a state of affairs. This in itself, however, is not enough. The utterance refers to something outside its physical manifestation, and draws its meaning from that remote level. The existence, in one way or another, of that level is therefore built into the act of representation. This is not a philosophical point regarding the nature and function of language. The law does not deal in the subtleties of whether meaning is socially constructed or inherent in the world. It incorporates a common, folk view, which most people recognise and accept as plausible, of how we relate to the world and ascribe meaning to it.

Although section 2 of the Fraud Act does not state it explicitly, it does not suffice that the defendant make an utterance that amounts to a representation. Fraud is a deception offence,

even though it is not required that anyone actually be deceived,[17] and it takes two to dance that particular tango, an addressor and an addressee. This presupposes that the statement that amounts to a representation be *communicable* to another party, which, in turn, includes, among other things, that the representation be intelligible and made in an effective way. The presuppositions at play are further specified in relation to the *modus operandi*. For instance, making a fraudulent representation by computer presupposes other, remote parties able to gain access to the representation with the appropriate technology and so forth. Similarly, the requirement that the representation be untrue or misleading depends on the possible existence of a different description of facts as well as a forum or criteria for judging which version of the facts is true or accurate. And for D to know or suspect that the representation is or might be false, D must have access to these criteria of judgement.

The criminal law of fraud expects the defendant to be able to grasp these background presuppositions of making a false representation. This includes the ability to know or believe specific circumstances but, when we cast the net wider, this in turn depends on the broader ability to form a picture of the world as a reality and as a possibility. Although the ability to have a sense of the wider picture is in itself neither a matter of knowledge, not least because much of it consists of future projections, nor a matter of belief, not least because it might include competing and irreconcilable alternatives, which one cannot simultaneously believe, it is a precondition for knowing or believing the specifics surrounding the commission of the offence. It is what we might call a matter of *imagination*, a horizon of awareness of the state of the world in its generally accepted social meaning.[18, 19]

I am only hurriedly sketching something complex and possibly contested in its details. I am, however, confident that the basic idea coheres with a widespread understanding of the law and that it is, therefore, uncontroversial. If this is so, it will suffice to argue the following. Just as the possibility of exercising voluntary control over our movements is an upshot of the consciousness assumption on which the law relies, so is the faculty of imagination, of the ability to place our acts within a wider context, of which we form a mental image. It follows that one who lacks this consciousness-related faculty of imagination, is incapable of acting as a subject of criminal law in the same way that one lacking the capacity to will one's movements cannot commit a criminal offence.

17. So much so that soon after it was enacted, it was unfavourably described as criminalising lying. See e.g. D. Ormerod, "The Fraud Act 2006 – criminalising lying?", *Criminal Law Review*, 2007, 193.

18. There seems to be an increasing interest in imagination in law. See, for example, M. del Mar, *Artefacts of Legal Inquiry: The Value of Imagination in Adjudication*, Bloomsbury 2020, which focuses on imagination in adjudication.

19. I would suggest that the faculty of imagination is at play even in offences in which it appears that all that is required is voluntary control over one's movements but this argument will have to be made in a different context.

This is not to say that the criminal law subject must exercise the capacity of imagination, which the law assumes it has, *correctly* or *accurately* for her act to count as one. Consider the doctrine of factual impossibility. In the law of England and Wales, and doubtlessly in other jurisdictions too, being mistaken as to the factual circumstances of the act is not a defence; for the purposes of ascribing criminal responsibility, facts are taken the way the defendant believed them to be. To return to our running example, if, say, D makes the fraudulent representation by sending telepathic signals or by speaking in a language that D has entirely made up and which no one else speaks, D would still be liable for having attempted fraud. But the important point is that it is a precondition of liability that D have the conscious cognitive ability to imagine her act within a context, which includes its consequences and others' reactions to it.

One might argue that the mental state that accompanies acting according to the law can also be explained microscopically in terms of the operation of neural networks, and that expressions such as "intention" are only shorthand for the operations of our nerves and muscles. This might be so. But it would also be a view that does not cohere with criminal law's conception of the acting subject. Law assumes that conscious intention and imagination and acting *qua* physical movement are two distinct layers, that the criminal law agent can genuinely take a removed standpoint in relation to her existence, which of course includes her own movements. This allows it to perform two crucial functions.

First, the consciousness assumption, which translates into the voluntariness and imagination requirements, *attributes the harmful act to the actor in a responsibility-specific way*. Consider automata. Say A's arm is raised up in the air and hits B. In the first scenario, his arm moves inadvertently because of a muscular spasm. If we think of the intention to move as the product of neural interactions, then nothing stops us from saying that A voluntarily hits B. In everyday parlance, we may lazily even do so, although we would not go on to blame A for doing so. This would be inaccurate in criminal law terms. It is not that A is not blameworthy, because she had no control over her acts; she *didn't act at all*, because she did not will her muscle movements. A moving as an automaton is not tantamount to A acting as a conscious subject.

It also makes it possible to attribute remote harms to the agent. Presenting a risk, for example, requires the ability to foresee the possible results of one's actions, which can only be a feature of a subject that has a mental grasp of the wider picture of the world. This is not to take sides as to the fairness of holding people responsible for remote harms or risk-creation. It is simply to say that the consciousness assumption is a necessary precondition for even beginning to consider attributing criminal responsibility for remote harms or risk-creation.

Second, the consciousness assumption and its specifications help to *individuate the actor*. It is *the defendant*, and not someone else, who is held to have performed an act, because and only to the extent that she had control over the relevant physical movements *and* she willed them. This is a more fine-grained manifestation of the individuating function of consciousness *qua* subjectivity, which in criminal law is completed with the remaining requirements pertaining to the *mens rea*. Criminal law does not take its subjects as processes but as centrally controlled bundles of overlapping and corresponding physical and mental characteristics, and it is on this precondition that it apportions responsibility. That subject is taken to be identical across time

AI and AR

As far as we know and as the technology currently stands, machines, however well trained, respond to electric signals triggering mathematical functions. What follows is a complex, back and forth interaction of artificial neurons, resulting in, say, another object moving, texts forecasting future events or diagnosing a medical condition, and so forth. They are so impressively adept and, at least mostly, successful in carrying out their tasks that we tend to believe that they act just as we do. Rather, we consider their acts to be as we take our acts to be, namely as being motivated and controlled by some centre that is removed from our purely physical existence, what we call consciousness. By extension, we also accept that AI machines act in a criminally responsible way, an impression reflected in our linguistic practices about advanced machines generally and the criminal law related ones in particular. We tend to say that the robot at the Volkswagen "killed" the worker or that ChatGPT "defamed" the law professor, assuming that these actions belong to the machine just as they would to a human being.

Is that more than an impression? Jaap Hage offers a more nuanced argument in support of the possibility of holding AI machines as proper acting subjects of the criminal law.[20] Given that mental states cannot cause one to act, we have to accept that the volitional element in the act requirement is only attributed by law to its subjects in reflecting our tendency to make the same attribution to others. Not being able to be in one's mind, we assume that what they do, or what they fail to do, is by choice.

So far, so good. I have been assuming the same in this paper. Hage, however, takes it one step further, to argue that, in light of the above, there is no obstacle to extending this attribution to non-human agents.

20. Jaap Hage, "Theoretical foundations for the responsibility of autonomous agents", *Artificial Intelligence and Law*, 25, 2017, 255-271.

Because attribution is mind-dependent, agency and responsibility may theoretically be attributed to anything, and on any grounds. It is possible to consider events as the acts of animals or of gods, or as the acts of organizations, and we may hold animals, gods and organizations responsible and liable for these 'acts'. This however is only from a historical perspective done by analogy to the attribution of agency to human beings. Ontologically speaking, there is no difference between attribution of agency to humans and to other agents.[21]

The attribution of mental states being a psychological matter, it *must be the case* that there is a sufficiently widespread psychological disposition rather than being institutionally possible to extend the scope of the law. This is so without a doubt in relation to our attitudes to other humans. It might even be scientifically verifiable – we could run an opinion poll that confirms that most of us are on the same page – but the point is that it does not need verification. It is what sustains our everyday interactions, either direct or mediated by institutions. Can the same really be said about our attitudes towards non-human agents? Sure enough, we tend to anthropomorphise animals, for example, and to speak of them as acting in ways similar to us but we always stop short of attributing to them the full set of faculties that we consider ourselves to possess. The same is reflected in our institutional practices; cases in which non-human agents were held to be human subjects are few and far between, and invariably short-lived.[22]

If the first obstacle to attributing volitional elements to machines is that we don't *in fact* do it, the second is that we don't even have *prima facie* indications that it might be appropriate to do so. If anything, we have reasons to think exactly the opposite.

As far as we know, the activation of an artificial neuron and its communication with other neurons are automatic reactions. They may be patterned and unpredictable, but this is a far cry from thinking of them as subject to the machine's will, as motivated by a mental state distinct from their activity. Their training might allow for various courses of action but their choices are still determined by the data. Even the mistakes that they make, and we know that they do err on occasion, amount to glitches in placing the bits of a vast jigsaw puzzle of information rather than opting for one choice over another, as we understand ourselves to be doing. One might counterargue that we too cannot genuinely will to do wrong but the objection would misfire. The issue here is not whether the will is subject to deontological constraints but whether machines have the basic faculty of developing a distanced attitude in relation to their "actions".

21. Hage, n. 21, at 261.

22. Granting rights to non-human agents is another matter. It does not make them subjects of the law but rather its ward.

Nor do machines have the ability to picture their environment as criminal law expects of its subjects. Their understanding of it is piecemeal, they have to break everything down to miniscule constituent parts and when they reconstitute them as output, they do not have a holistic cognition of it as we would in its social meaning.

Imagine that an analyst for a mortgage lender, let's call him Adam, is tasked with predicting which existing borrowers are likely to be late with repayments or default altogether. Adam inputs into a computer all the information deemed relevant, such as age, address, occupational history and status, credit history and so forth, as well as general historical records of default. Let us suppose that, due to some technical glitch, borrower Betty's social insurance contributions were not being registered and, as a result, she appeared to be unemployed for a year. Taking that into account, the computer concludes that Betty is likely to default within the next quarter. Adam then reads the results and emails them to his line manager.

To start with, the computer does not know that it is making a representation, that its calculations reflect a state of affairs outside it, let alone that there is a real person called Betty to whom those calculations refer and whose future will be affected by them. Nor does it know that the calculations are communicated to anyone. Nothing exists outside its ANNs. The computer grasps its environment, only when it internalises it as data. Adam, on the other hand, is in a position, or so the criminal law assumes, to place his statement in a context that exists outside him, to attach meaning to it. For the same reason, the computer is incapable of misleading anyone, because, as far as it is concerned, there is no one to mislead. Once again, not so for Adam. The machine's representation can also never be false. Everything fed into it, even the incorrect information regarding Betty's social insurance, it cannot but treat as true, although it might attach a different value to different sets of information, for it lacks the ability to contrast it, indeed, to even consider contrasting it, to any alternative.

If all this is right, then there are never any circumstances under which computers fulfil the mental state aspect of the *actus reus*.

There are two possible upshots of the machine's imagination deficit.

First, it could be regarded as compromising the machine's criminal capacity. One is excluded from criminal responsibility, when one cannot, or is assumed not to be able to, as children standardly are across most legal orders, distinguish between right or wrong. The same principle properly extends, at least philosophically if not institutionally, to those who lack the capacity to understand the consequences of their actions and, *a minore ad maius* those incapable of placing

their actions in a context *at all*. But even if we do not accept this extension, because it does not have sufficient institutional support, complete inability to make sense of the factual context would at the very least afford them the excuse of insanity. If we take this tack, we accept that machine activity may fulfil the *actus reus* of offences but, given their nature, it is practically impossible to hold machines criminally responsible.

Second, I would suggest, possibly somewhat more ambitiously, that the machine's imagination deficit cancels out the very character of its activity as criminal.

Whatever tack one takes on law's normativity, whether one believes it to be independent from social attitudes or considers it a matter of widespread social acceptance, one must accept that intelligibility is, at the very least, a condition of its existence. Law must take its subjects and the world as they are. For the law to be intelligible, it must obviously be expressed in a comprehensible language. Second, the meaning of legal concepts might be system-specific and determined by legal practice but must nevertheless tap into and cohere with the folk meaning of words. Third, and most importantly for our present purpose, the law must be such as to capture the world in an imaginable way.

To reiterate the earlier point, this presupposes the capacity of imagination on the part of the law's subjects. To emphasise, it is the general capacity that is crucial, not the ability to exercise it correctly. Therefore, even those with temporary or permanently impaired ability to foresee the consequences of their actions or attach a generally accepted meaning to them, still bear the faculty.

Let me return to Hage. Recall that he argues that, it being institutionally possible to attribute to machines intentional states, the real question then is whether it is desirable. In the absence of any evidence that machines actually do take from their actions the distance that conscious subjects regard themselves as taking, this attribution would be a legal fiction, we would be treating machines *as if* they possess the requisite capacities. Why, one might ask, can we not accept that provided that the law is intelligible to a sufficient number of subjects, it maintains its normativity not only over those but also over anyone who does not meet the preconditions of intelligibility?[23]

23. This obviously raises the fairness-related question of whether an action-guiding order may fail to guide the actions of its subjects; but this is not of immediate concern to us in this context.

Because it would be contradictory. When the law claims to bind A, it claims to do so in a possible world, in which all its subjects bear the exact same characteristics as A. Now imagine a possible world, in which all subjects bear the characteristics of an AI machine, including the imagination deficit. In that world, the law would be totally unintelligible *in fact*, therefore its normativity would be fatally undermined. A legal system that holds AI machines as its proper subjects would, at the same time, be unable to have AI machines as its proper subjects. It follows that the law may not bind AI machines as subjects. Therefore, their "acts" can never fulfil the *actus reus* requirements.

The responsibility gap

Let us remind ourselves why many are concerned about the techno-responsibility gap. AI machines can cause harm. When they do so, they themselves cannot be held criminally responsible, because they lack the capacity to form *mens rea*. Nor are the programmers criminally responsible, to the extent that they have relinquished control over the machines. This leaves us frustrated, because there is no one to whom we can direct our reactive sentiments to what appears wrongful and not just an accident.

What impact is there on the responsibility gap, if it is indeed the case that machines cannot act in a criminal law relevant way? At first sight, the gap seems to widen. The appearance of wrongfulness and the feeling that the harm is accidental do not go away, and our blame reactions persist. That there is no action or actor at all makes the frustration even more nagging, because we are left with the sense that amending our institutional order so as to right such wrongs feels like even more of an uphill struggle.

This, I would suggest, is a false impression. What ostensibly gives rise to the responsibility in the first place is the intervening "act" of the machine, which exculpates the programmers or owners, at least for the harm caused at the end of the line (this obviously does not preclude liability for other criminal offences committed independently at an earlier stage). I have been arguing, however, that machines do not act in a criminal law relevant way, therefore making it impossible to attribute criminal acts to them or individuating them as actors. If this is right, then the buffer between the acts of the programmer and the end harm is removed.

Two options are opened up. The first is to accept that machines do act but always, without exception, as automata or without the capacity to appreciate the nature and meaning of their acts. The law would then have to treat them as innocent agents. Should the relevant conditions be met, liability would roll back to programmers or users, who would be held as having committed

the offence through an innocent agent.[24] This is a possibility, but at the same time we would be left with the rather awkward situation of having a whole class of subjects of criminal law which, however, can *never* meet the requirements of criminal liability.

Legal responsibility would again roll back to programmers or users, if we accept that machine activity does not qualify as *actus reus* at all, only this time there would be nothing at all separating them from the harm or risk of harm caused.

The conditions will, of course, vary. The fact that machines remain opaque "black boxes" still makes a difference and will determine crucial aspects of liability such as the remoteness of the harm from the programmer's or user's acts, the gravity of the risk, and so forth. But difficulties in fairly apportioning criminal responsibility there might be, a techno-responsibility gap there isn't.

24. This is one of the models of liability put forward by Hallevy. The difference is that, because he assumes that advanced machines do actually act in a criminal law relevant way, he reserves this option for less advanced computers.