Philosophy & Technology (2020) 33:433–465
https://doi.org/10.1007/s13347-019-00362-x

RESEARCH ARTICLE

AI Systems Under Criminal Law: a Legal Analysis and a Regulatory Perspective

Francesca Lagioia[1,2] · Giovanni Sartor[1,2]

Received: 6 December 2018 / Accepted: 21 June 2019 / Published online: 15 July 2019

Abstract
Criminal liability for acts committed by AI systems has recently become a hot legal topic. This paper includes three different contributions. The first contribution is an analysis of the extent to which an AI system can satisfy the requirements for criminal liability: accomplishing an *actus reus*, having the corresponding *mens rea*, possessing the cognitive capacities needed for responsibility. The second contribution is a discussion of criminal activity accomplished by an AI entity, with reference to a recent case involving an online bot, the Random Darknet Shopper. This discussion will provide the context for the analysis of commonalities and differences between criminal activities by humans and by artificial systems. The third contribution concerns the evaluation of different ways of addressing criminal activities by AI systems in a regulatory perspective.

Keywords Criminal liability · Artificial Intelligence · Autonomous agents, software agents · Normative agents

1 Introduction

AI systems are complementing or replacing humans in many tasks and activities: for instance, surgical robots, unmanned vehicles, trading algorithms, digital assistants, and personal and industrial robots are increasingly used.

New sets of legal problems arise in the context of the deployment of AI systems, problems which did not exist when computer systems were mere tools. Some of these

✉ Francesca Lagioia
francesca.lagioia@eui.eu

Giovanni Sartor
giovanni.sartor@eui.eu

[1] EUI – European University Institute, Fiesole, Italy

[2] CIRSFID – University of Bologna, Bologna, Italy

problems—such as liability for damages caused by AI systems or the validity of contracts concluded by them—have been addressed in a number of studies, and recently also in legal disputes and legislative initiatives, such as the report on Civil Law Rules on Robotics, passed by the European Parliament's legal affairs committee,[1] the European Commission's AI Strategy and the Work of the High-Level Expert Groups on AI.[2]

Criminal liability resulting from AI activities, with a few exceptions (Hallevy 2013, Pagallo 2013), has only been addressed in the context of war, in connection with the application of humanitarian law to autonomous weapons (Task Force 2011, Bhuta et al 2015). In this paper, we aim to address criminal liabilities for the autonomous operation of AI systems in general terms, namely, to cover all cases in which AI systems autonomously engage in acts that would constitute crimes if performed by humans. In fact, it is possible that an AI system engages in criminal actions for which no human possesses the corresponding *mens rea*, no human having planned, foreseen, or directed such actions. This raises the issue of how the legal system should respond to this gap in criminal liability, namely to the fact that a criminal activity is accomplished for which nobody is criminally responsible (though some form of civil liability may apply).

An important subsidiary issue concerns whether the legal response to AI crimes should depend on the cognitive attitude of the involved AI system. Though AI systems are very far from reproducing the complexity of human psychology, we will argue that under an appropriate level of abstraction (Floridi 2016), both humans and AI systems may be viewed as having the cognitive attitudes (intentions, beliefs, awareness) that are relevant to the realisation of *mens rea*. As human mental states contribute to determine the legal reaction to human crimes, also the cognitive states of AI system might be taken into consideration to appropriately react to their harmful behaviour: the intentional or reckless causation of harm by an AI system may need to be addressed differently from the "inculpable" harmful behaviour by the same system.

In other words, the intentional or reckless engagement in criminal activities can be viewed as a distinct kind of functioning failure of an AI system, as a defect that calls for a distinct response.

On the one hand, intentional or reckless causation of harm can be prevented in a specific way. To avoid the innocent causation of harm, a system must either be restricted in its sphere of action or be endowed with superior cognitive capacities, so that it can figure out the unintended effects of its action. On the contrary to avoid intentional or reckless causation, it is sufficient that the system is prevented from adopting the "malicious" attitude at stake. As we shall show in the following, this can be obtained without limiting the system's action capacity, either by providing appropriate disincentives (to developer, users and possibly even autonomous systems), or by endowing the system with a normative architecture.

[1] P8_TA (2017)0051 Civil Law Rules on Robotics European Parliament resolution of 16 February 2017, with recommendations to the Commission on Civil Law Rules on Robotics, 2015.

[2] On 25 April 2018, the EU Commission set up three different groups of experts on (i) the ethics of AI; (ii) whether and to what extent to amend the directive on liability for defective products; and, (iii) liability and new technologies formation (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence). See also the Commission's document on Artificial intelligence: Commission outlines a European approach to boost investment and set ethical guidelines, IP/18/3362. For an extensive literature analysis of the foreseeable threats of AI crimes see King et al (2018)

On the other hand, intentional or reckless causation of harm may have very serious consequences. While the innocent causation of harm presupposes extraordinary circumstances in which the action has unexpected results, intentional or reckless causation just requires that the agent effectively engages in producing the outcomes at stake. Compare for instance, an unexpected fatal accident caused by an autonomous car (meant to take people to destination, without harming anybody), and a planned killing by an autonomous weapon system (meant to eliminate a particular individual, or even to indiscriminately cause death, for terror purposes).

The prevention of intentional or reckless criminal activities by AI systems is also relevant to the broader issues of machine ethics. Recent research has pointed to the risk that AI systems, while achieving human and possibly superhuman intelligence, may disregard human values, and thus put at risk the very future of humanity. It has also been claimed that autonomous AI systems can be safe only if they endorse and maintain moral values (Bostrom 2014). Criminal law, besides responding to contingent political requests (the so-called mala-prohibita crimes), is meant to address actions that cause most serious harm to others (mala-in-se crimes), and are therefore generally considered immoral. Therefore, by making AI systems comply with criminal law, we ensure that such systems respect the essential core of morality.

Some of these issues will become actual only in the near future. However, we believe that they should be addressed sooner rather than later, to promote technological developments and regulatory responses.

We will adopt a conceptual framework informed by legal theories and based on the technological capabilities of current and future AI entities. To support and validate this theoretical framework, we analyse and discuss a real case concerning an automated online shopping bot that was deployed to make random purchases on the Deep Web. The bot purchased a diverse set of items, including illegal drugs (Power 2014), an action that would clearly count as a crime if it were performed by humans.

We are aware that the very notion of AI is highly controversial, and that different definitions of it have been proposed in the literature (see Russell and Norvig 2010, Poole et al. 1998, Kurzweil 1990). However, we shall not engage in selecting a specific definition and will be content with the general idea that AI systems exhibit, to different extents, intelligent behaviour, or at least behaviour that would appear intelligent if performed by humans. Certain features of a system are relevant to this purpose—such as the ability to learn from experience, the capacity of building models of current and future states of the environment, the ability to engage in goal directed behaviour—but we do not need to commit to any particular combination of such features as being essential to intelligence. In fact, we shall not discuss any general features of AI systems, but rather focus on those cognitive competences, only possessed by some systems, present or future, which are preconditions for criminal liability (see Section 4). Since such competences can be implemented in different computer architectures, our analysis will remain at an abstract level; specifications and experiments will be addressed in future research. The control relation between humans and AI systems can also take different forms and levels of intensity, according to the extent to which the action of AI systems is monitored by their users and directed through instructions. In this paper, we shall focus on cases in which AI systems are not constrained to comply with users' requests, so that their cognitive features and decisional processes become decisive to the realisation of criminal activities.

In Section 2, we discuss whether and to what extent agency, responsibility and legal personhood can be attributed to AI systems.

In Sections 3, 4 and 5, we analyse both the factual and the mental components of criminal offences to determine whether AI systems can realise such components in what ways and to what extent. We devote special attention to *mens rea,* and in particular to the cognition and volition requirements. In Section 6, we address reason-responsiveness and discuss the architecture of normative agents.

In Section 7, we use the Random Darknet Shopper case, to exemplify and discuss commonalities and differences in criminal activities by humans and by artificial systems. In considering similarities between human agents having cognitive limitations and AI entities, we draw analogies and evaluate possible models of criminal liability. Once we have examined who might be liable for AI crimes, we consider whether there is a criminal responsibility gap. What if AI systems commit actions that would count as crimes if committed by humans but there is no intent, recklessness or negligence in either users or manufacturers?

In Section 8, we provide a regulatory perspective. We discuss the extent to which AI crimes may be prevented by restricting the autonomous behaviour of AI systems, or remedied by applying civil liability. Then, we examine how such crimes may be countered through specific criminal liabilities for creating and deploying criminal AI systems and whether criminal sanctions can be directed against such systems. In Section 9, we discuss the principle of legality and the consequent need for legislative change.

2 Toward Criminal Liability and Personhood in AI Systems

In this section, we summarise some of the basic principles and presuppositions for the regulation of human activities through criminal law. This provides the framework for assessing whether such principles and presuppositions also apply to AI systems.

2.1 Cognitive Preconditions of Criminal Liability

In modern legal systems, criminal provisions consist of rules punishing accomplishments or omissions that are accompanied by certain mental states (usually, intention, recklessness or negligence). Such provisions are assumed to entail, or presuppose, the command to omit or accomplish the actions at stake. For instance, the criminal rule punishing homicide (e.g. with 30-year detention) presupposes the prohibition to kill. Criminal rules are assumed to operate by deterring their addressees from committing prohibited actions, or from omitting required ones. Besides deterrence, criminal punishment may perform further functions, such as retributing blameworthiness, re-educating perpetrators or expressing society's reprobation (Duff 2007).

A necessary precondition of punishment is the responsibility of the perpetrator. A person is responsible for a crime when he or she "must answer for it in court" Poole,. It is usually assumed that responsibility depends on certain cognitive capacities, which "are best understood as a matter of reason-responsiveness: a responsible agent is one who is capable of recognising and responding to the reasons that bear on his situation" (Duff 2007, 39). This capacity covers both epistemic reasons for beliefs and practical reasons for action. For instance, according to Fischer and Ravizza (Fischer and Ravizza 2000, 35-6), (moral or criminal) responsibility depends on the action being determined by the agent's "guidance control". Guidance control is realised when two conditions are met: the decisional mechanism leading up to the criminal behaviour should be (1) "moderately reason-responsive" and (2) "the agent's own". The first condition for reason-responsiveness requires the agent to act according to a decisional mechanism that in presence of strong reasons to act (or not to act) recognises these reasons and brings the agent to (not) perform that action in a sufficiently broad range of circumstances. According to Fisher and Ravizza the requirement of reason-responsiveness marks the difference between morally responsible actors and actors acting under factors excusing them, i.e. factors under which the person's decisional mechanism is bypassed or not responsive enough to reason.[3]

So far, the law has assumed that criminal rules address only humans (and in some cases, corporations, which act through humans). The fact that such rules are directed only at humans has a clear rationale: only humans meet the preconditions for regulation through criminal laws, since only humans can possess mental states, be responsible for their actions, and be influenced by criminal rules.

First of all, only humans possess to a sufficient extent situation awareness and capacity for purposeful choice, and consequently only humans can act intentionally, recklessly or negligently, to the extent that is required for the commission of a crime.

Secondly, the punishment of the agent who has committed a crime requires that that agent is criminally responsible. Only humans possess the level of reason-responsiveness that is required for criminal responsibility.

Finally, since criminal laws are meant to deter unwanted actions, their operation presupposes that their addressees can be influenced by legal commands and by the corresponding sanctions.[4] Since non-human entities cannot appreciate the significance of norms and sanctions, nor the social significance of criminal behaviour, their action cannot be governed by criminal norms.

All such assumptions—that have so far excluded non-human entities, such as animals, from the scope of criminal law—may no longer hold for AI systems. In fact, such systems represent a new kind of non-human agency, which may—in the near future if not yet—be exempt from the cognitive limitation of other non-human agents: they may possess the cognitive capacities that provide for mental states, reason-responsiveness and understanding of norms and sanctions.

The main purpose of this paper, indeed, is to examine to what extent AI systems may meet the preconditions for regulation through criminal law, namely, if they can realise crimes, respond to reasons, and be influenced by criminal norms.

We consider cases in which AI systems satisfy both the material and the mental element components of crimes—we refer to such cases as AI crimes (regardless of the way in which they are considered by the legal system at issue)—and examine whether they require a specific legal response. Compare for example two cases involving the death of a patient following therapy delivered by a medical robot. In the first case, the therapy was

[3] For instance, potent drugs, manipulation of the brain, brain lesions, neurological disorders, phobias, drug addiction, coercive threats.

[4] As Bentham observed "law's proper role" is "to address the wills of citizens and thus to guide their actions through their understanding". (Postema 2001, 494).

provided according to existing protocols, but the patient died because of an allergy unknown to the robot; in the second case, the robot knew of the allergy, knew that a drug would cause death under the relevant conditions, but chose to deliver the drug to kill the patient, perhaps to save the costs of expensive treatment. Should the second case be treated differently, since it would have been deemed homicide if it had involved a human?

If the legal system chooses to regulate AI crimes distinctively, we may say that for that legal system, AI crimes are legally relevant. The regulation of AI crimes may impose obligations and liabilities (a) only on humans or (b) both on humans and AI systems. Let us now consider the two approaches.

According to the first approach, the legal relevance of AI crimes does not entail that AI systems are addressees of criminal law. Humans (users, developers, deployers) would remain the only addressees of criminal norms and sanctions: they could be subject to criminal sanctions when contributing to the AI system's criminal behaviour, they could be obliged to pay compensations and fines and to limit further uses of that system (prohibition of its further deployment, obligation to disable it, or to reprogram its legal/moral component, etc.).

Such sanctions or obligations might be excluded when the concerned person could appeal to justifications or excuses that apply to the AI system's action, for instance when the system acted under state of necessity. For example, imagine an autonomous car intentionally damaging somebody’s property to save its own passenger.

According to the second approach, AI systems would be subject to criminal law, i.e. they would be directly affected by legal reactions to their crimes. Such legal reactions might consist in measures that are similar to sanctions against humans (e.g. fines) or also in different kinds of sanction (e.g. re-programming malicious AI systems). The second approach presupposes that AI systems are considered as legal persons. Though we cannot here address the philosophical and doctrinal debate on legal personhood (Kurki and Pietrzykowski 2017), it may be useful to clarify the sense in which we are (not) using this concept, when applying it to AI systems.

2.2 Legal Personhood of AI Systems Under Criminal Law

When speaking of the legal personhood of AI systems, we are not using the notions of a person that have been developed in philosophy, often in connection with theology”[5] nor are we referring to personhood as an axiological construction (including attributes such as dignity, self-determination, etc.), which identifies the “legal anthropology” endorsed by a legal system, i.e. the aspects and capabilities of individuals that a legal system is meant to protect and enhance.

We are, rather, focusing on a more limited and legalistic notion of personhood as the actual or conditional possession of legal rights and duties. An entity is a person, in this sense, to the extent that either it is the holder of duties and rights, or would become the holder of duties and rights, if appropriate triggering conditions were met. Thus, legal personality is viewed, from this limited perspective, as consisting in what is also called “legal capacity” in civil law systems. This view was famously advanced by Hans Kelsen (1967, 173), who claimed that the physical person is “a totality of rights and

[5] From Boethius’s idea of a person as an “individual substance of rational nature”, to Kant’s view of personality as the “freedom of a rational being under moral law, see Brozek (2017).

obligations which have the behaviour of a human being as its content and thus form a unity". From this perspective, to say that an entity is a person is just to say that this entity is or may become the bearer of rights or duties.

This legalistic notion of personhood may be understood in a thinner or thicker version. In the thinner version, to say that an entity is a person just means that there is at least one norm addressing the behaviour of that entity, attributing to it rights or duties. In the thicker version (Kurki and Pietrzykowski 2017), it means that the behaviour of that entity is addressed by a set of norms that corresponds to a large extent to the norms that are generally applicable to humans, so that the entity has similar entitlements and burdens (it may hold property, is protected against crimes and torts, has the obligation to respect other people's property and life, etc.). The thin and the thick notions may lead to opposite characterisations of the same individuals. For instance, since slaves where subject to criminal laws and were protected by some criminal norms (at least in some legal systems, such as late Roman law or US eighteenth-century law), they would qualify as legal persons according to the thin notion, but not according to the thick one.

For the purpose of criminal law, it is sufficient to focus on duties and sanctions. Thus, to say that an entity is a legal person under criminal law (according the thin conception) means for us simply the following: there are legal norms that establish duties concerning the behaviour of that entity, duties whose violation would trigger a criminal sanction against the entity.

In conclusion, when discussing whether AI systems may be granted personality relative to certain criminal norms, we focus only on the obligations established by these norms, without assuming or implying that AI systems should have any further legal burdens or entitlements. Moreover, as just remarked, we consider only the personality of AI systems as duty-bearers in criminal law, namely, the possibility that AI systems are subject to criminal duties and sanctions. We are not addressing the very different issue of whether AI systems could be viewed as right-bearers under criminal law, being protected by some criminal norms (e.g. a prohibition to terminate or damage certain kinds of AI systems) or having the power to trigger criminal prosecution by bringing complaints.[6]

3 The *actus reus* in Criminal Offences Perpetrated by AI Systems

In order to impose criminal liability, two cumulative components need to be met: a factual component (*actus reus*) and a mental component (*mens rea*).

The *actus reus* is usually understood as the external-objective component, i.e. the carrying out of the offence. Its structure is the same for every type of offence, whether intentional or negligent. It consists of three main elements: a necessary element, criminal conduct itself, and two optional elements, circumstances and results. Conduct may consist in commission or omission (usually omission is criminally relevant only when the agent was under a duty to act). Thus, the *actus reus* identifies what the defendant must have done (commission) or failed to do (omission).

Here, we focus on commission, since the case we discuss, the Random Darknet shopper case, concerns a commissive crime, namely, the purchase of illegal drugs,

[6] The issue of whether AI systems can bear legal entitlements, i.e. rights and powers, has been addressed relative to civil law (see Pagallo 2013, 102,) relative to constitutional law (see Solum 1991, 1255).

consisting in the participation in an agreement to this effect. It does not require specific circumstances or the production of any result.[7]

In commissive crimes, the *actus reus* consists in a material performance (i.e. something done), with a factual-external presentation.[8] Recent legal doctrine has criticised the traditional view of the *actus reus* as mere willed muscular movements or bodily movements.[9]

First of all, it has been observed that the exact nature of the *actus reus* depends on the specific crime; sometimes, the *act* consists in a state of affairs, rather than an event, which may or may not involve a positive action. Consider, for instance, crimes that prohibit the state of possessing something, such as drugs or firearms.

Secondly, under some circumstances, a defendant can be responsible for the conduct of a third party. Two examples are particularly relevant: (1) vicarious liability—under which an employer may be criminally responsible for the acts or omissions of an employee—and (2) the doctrine of innocent agency—where primary liability is attributed to a manipulator who used an innocent party to commit a crime.

Thirdly, some crimes, such as solicitation crimes (e.g. solicitation to suicide), conspiracy and defamation, punish criminal actions without any execution through bodily movements unless we consider that the physical conduct consists in the movement of tongue, mouth and vocal cords. In computer crimes, bodily movements are missing, unless we consider that the physical act in computer crimes resides in sending electronic impulses (Freitas et al 2014).

Even though the *actus reus* cannot be reduced to muscular bodily movement, having to be put in context, it essentially consists in a material aspect having a factual-external presentation. It does not include the agent's capacity to engage in practical reasoning, guide its actions and actualise results, and more generally the agent's mental states and processes. Consequently, an involuntary and unwilled action can still realise an *actus reus*. Examples of such involuntary actions include instinctive reactions (e.g. where the defendant is undergoing a panic attack), automatism (e.g. reflex, convulsion, bodily movements under epileptic seizure, acts following concussion, physically coerced movements), and cases of mental disconnection (e.g. somnambulism). Such actions

[7] Most European drug laws penalize many acts involving hard drugs: illegal cultivation, production, manufacture, extraction, preparation, acquisition, and possession, offering, offering for sale, distribution, purchase, sale, delivery on any terms whatsoever, brokerage, dispatch, dispatch for transit, transport, importation and exportation of illegal drugs.

[8] On the notion of criminal act, and in particular willed and unwilled acts in criminal law, see for example Murphy (1979), Hart (1968), Austin and Austin (2000).

[9] This theory dates back to nineteenth century authors such as Holmes, O.W. and Austin, J.. In particular, Holmes's view is that "An act is always a voluntary muscular contraction, and nothing else. The chain of physical sequences which it sets in motion or directs to the plaintiff's harm is no part of it, and very generally a long train of such sequences intervenes". According to this author "An act […] imports intention […] A spasm is not an act. The contraction of the muscles must be willed" (Holmes 2009, 63). Similarly, for Austin "Most of the names which seem to be names of acts, are names of acts coupled with certain of their circumstances. For example: If I kill you with a gun or pistol, I shoot you. And the long train of incidents which are denoted by that brief expression, are considered (or spoken of) as if they constituted an act, perpetrated by me. In truth, the only parts of the train which are my act or acts, are the muscular motions by which I raise the weapon, point it at your head or body, and pull the trigger" (Austin 1875, 202). Generally, see also Ormerod et al (2011), Herring (2014), and Duff (1990, 96-99).

realise the factual element of a crime, but criminal liability is not imposed upon them, since the link between mind and behaviour is missing or essentially distorted.

According to this characterisation of *actus reus*, both the AI systems in charge of controlling physical objects (i.e. robots) and those without a physical presence (i.e. software agents and bot) can fulfil the conduct requirement of an *actus reus* (e.g. destroying a physical object or erasing an electronic memory). This is true not only when the performance at issue is the result of inner calculations carried out by AI systems, but also when AI systems execute instructions fed by a remote human operator (Hallevy 2013). For instance, in our running case—i.e. the online purchase of illegal drugs by a bot—it is clear that the act of the purchase was carried out by the web robot.

In Section 7, we consider whether, in addition, users or programmers may be considered to have purchased the goods through the bot, to have been accomplices, to have instigated the offence, or to have contributed in other ways.

4 *Mens rea* in Criminal Offences by AI Systems: the Cognition Requirement

In intentional offences, *mens rea* has two components: cognition and volition. Cognition is the agent's awareness of factual reality and involves all components of the *actus reus* (act or course of conduct, surrounding circumstances, and act's outcome or result). Volition consists in the intention to perform the act and achieve its outcome (for crimes including the realisation of an outcome), and it can never be alone, it is always accompanied by awareness (Hallevy 2013, Ashworth and Horder 2013). In this section, we consider the cognition component, while we address the volition component in Section 5.

The cognition requirement, namely, the agent's awareness, is usually understood as including both perception and understanding. For a human to be aware of a certain context, two cumulative conditions need to be met: (1) taking in data about certain facts (through the senses) and (2) forming a relevant comprehensive image of these facts. In evaluating the possibility of attributing such cognitive processes to AI systems, we need to isolate the cognitive mechanisms that enable such systems to acquire information and to build usable comprehensive images.

Endsley et al (2000) define situation awareness as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future". This definition of situation awareness breaks down into three separate levels:

- Level 1: perception of the elements in the environment,
- Level 2: comprehension of the current situation, and
- Level 3: projection of future status

In order to achieve situation awareness, the most basic level (Level 1) requires the agent's ability to perceive the current status of the environment, its characteristics and variables, namely the capability to monitor, detect and recognise the various relevant situational elements, such as other agents, objects and their current status, informational and behavioural clues. Perception should enable the agent to achieve awareness of reality, in the specific environment of interest to the agent, at a particular point in time. Data may be captured and collected through a number of physical or virtual channels. For agents operating in different domains, the perception requirement could be quite different. For instance, a physician should be aware of a patient's age, medical history and medicines the patient is using, and should be able to detect the general status and symptoms of the patient. On the other hand, a bus driver needs a completely different set of information, such as visibility conditions (e.g. darkness, rain, fog or snow), the speed limit, the conditions and hazards on or near the highway (e.g. pedestrians, animals or other obstacles).

The second step in achieving situation awareness (Level 2) requires the integration of the continuous disjointed information collected at the first level, through a process of pattern recognition, interpretation and evaluation, as well as a comparison of such information with goals and objectives. The continuous extraction and collection of environmental data and their integration with existing knowledge lead to developing a coherent, useful and comprehensive picture of the environment.

The highest level of situation awareness (Level 3) consists in the ability to project future states and actions in the operational environment. These projections can be used in directing further perception and in anticipating future states and events on the basis of the current situation.

In the following section, we argue that AI systems, in principle, can fulfil the cognition requirement: they can achieve situation awareness by taking in factual data and creating general images, going through the three steps of perception, comprehension, and projection.

4.1 First Stage: Perception

An AI achieves the first level of awareness, namely perception, to the extent that it is able to extract information from the environment, and make it accessible to its internal reasoning modules. Perception involves different technologies, depending on whether the entity is situated in the physical world or in a virtual environment.

For artificial physical agents, perception processes are embodied by the sensors and connected architectures that gather information from the agents' surroundings and by procedures that make this information available, after suitable validation, to the agents' memory. Perception in physical robots must be implemented through hardware, such as a video camera or a laser sensor in a mobile agent. In many contexts, advanced technologies can sense with an accuracy that equals or even surpasses that of the corresponding human organs. For instance, cameras can absorb light waves at frequencies that the human eye cannot detect.

For software agents situated in a virtual environment, perception is achieved by tracking activities and messages. Consider a software agent, such as the Random Darknet Shopper bot, that reliably acquires price information from web pages by using file-reading mechanisms, equipped with error checking and validation routines.

Perception is not a merely passive process but includes an active engagement. Weyns, Steegmans, and Holvoet (2004). propose a generic model for active perception focused on software agents operating in virtual environments (see Fig. 1).

The model is composed of three functional modules: (1) sensing, (2) interpreting and (3) filtering.

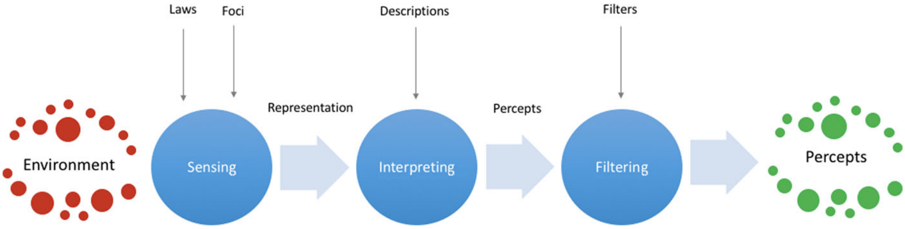

Fig. 1 Model for active perception, according to Weyns, Steegmans and Holvoet

The sensing module maps the state of the environment onto a representation consisting of a structured assembly of symbols that refers back to the environment. The mapping of the state to a representation depends on the selected set of foci and on the set of perceptual laws. Focus selection enables the agent to direct its perception, enabling it to sense the environment for specific types of information, whereas the laws of perception constrain the composition of a representation according to the requirements of the domain being modelled.

The interpreting module processes the representations provided by the sensing module, producing descriptions. Descriptions may distinguish different objects as well as interpret a group of objects as a cluster.

The filtering module selects the data items in a percept that match specific selection criteria. Each filter's selection is based on conditions concerning the elements of the percepts being filtered.

In the model developed by Weyns, Steegmans, and Holvoet, perceptual laws determine perception and information flows, taking into account the goals of the agents. Thus, such laws provide controlling strategies to acquire and process data, depending on the general goals and the specific tasks of the data acquisition process. Consider, for instance, a self-driving car. The car's general sensing activity may consist in mapping the general road conditions (e.g. traffic conditions, pedestrians, animals or other obstacles on or near the highway, pavement-related problems such as ice in sub-grade and poorly performing drainage) in order to obtain a depth-map of the environment. A specific task carried out by the car may consist in the image acquisition process. In this case, a law of perception may specify the area that falls in the scope of the car, setting the cameras so that objects at a certain distance will lie in the field of view.

Through the completion of the above described steps, perceptual data become knowledge, in the sense of usable information, readily available to the agent's functioning. Therefore, when AI agents are able to acquire and use their percepts, we can say that they can achieve the first stage of awareness, i.e. perception.

4.2 Second and Third Stage: Comprehension and Projection

Humans achieve the second stage of awareness, i.e. comprehension, by analysing the factual data provided by perception and integrating such data with further information. AI technologies can perform similar operations: they can build general images out of their perceptual inputs, by analysing input data and integrating them with data and patterns stored in memory.

Humans achieve the third stage of awareness, namely projection, by making guesses on the basis of the information available to them, to anticipate future events. AI technologies can similarly compute the probabilities of the outcomes that may result from taking alternative courses of action, using this information as a basis for making reasonable choices.

Even though the information processing methods used by AI systems and by humans are different, such methods may be viewed as different ways of implementing the same cognitive functions. Therefore, we can assume that certain AI systems can achieve the second and third stage of awareness, namely comprehension and projection.

For example, a chess-playing computer, like Deep Blue, developed by IBM, analyses the current status of the game based on the location of pieces on the board. It reviews possible options for the next move, and for each option, the probable responses by the other player. Then, for each response, it again reviews possible responses, using various methods to restrict search, and it assesses the merit of the available moves so as to decide on its next move (Hsu 2002). Thus, it seems that we can attribute to a chess-playing computer system goals (winning the match, attacking a certain piece, getting to a certain position) and information (e.g. about what moves are available to its adversary), and we can assume that it can devise rational ways to achieve these goals according to the information it has (Sartor 2009, 253).

Various technologies can be used to enable AI systems to make projections out of a dataset. Probabilistic models play a key role today in scientific data analysis, machine learning, robotics, and cognitive science, and more generally in artificial intelligence. Some probabilistic systems explicitly represent and manipulate uncertain information and make predictions accordingly (Ghahramani 2015, 452). Other systems do not explicitly represent probabilities. Examples are neural networks that are successfully applied to domains characterised by the availability of large amounts of data, such as speech recognition, image classification, and the prediction of words in texts Hinton et al (2012), Bengio et al (2003), Sermaent et al (2013). The scope of machine-learning tasks can go beyond pattern-classification or mapping, and can include optimisation and decision-making, compressing data and automatically extracting interpretable models. The decision of certain systems may depend on the uncertainty of data or forecasts. Typical examples include autonomous vehicles detecting pedestrians in images or medical systems classifying gene-expression patterns in leukaemia patients into subtypes according to expected clinical outcome.

Situation awareness in humans also includes inferring unperceived elements, through various mental processes, such as abduction. In artificial systems, this can be obtained through methods that make inferences about missing or latent data. For instance, consider the task of classifying people with leukaemia into one of the four main subtypes of this disease, on the basis of each person’s measured gene-expression patterns. On the basis of a training set consisting of observed data—pairs of gene-expression patterns and labelled subtypes—a system can infer whether the Figure 2 below provides a graphical representation of the cognition requirements. Subtypes for new patients have the same or similar gene-expression patterns (Ghahramani 2015).

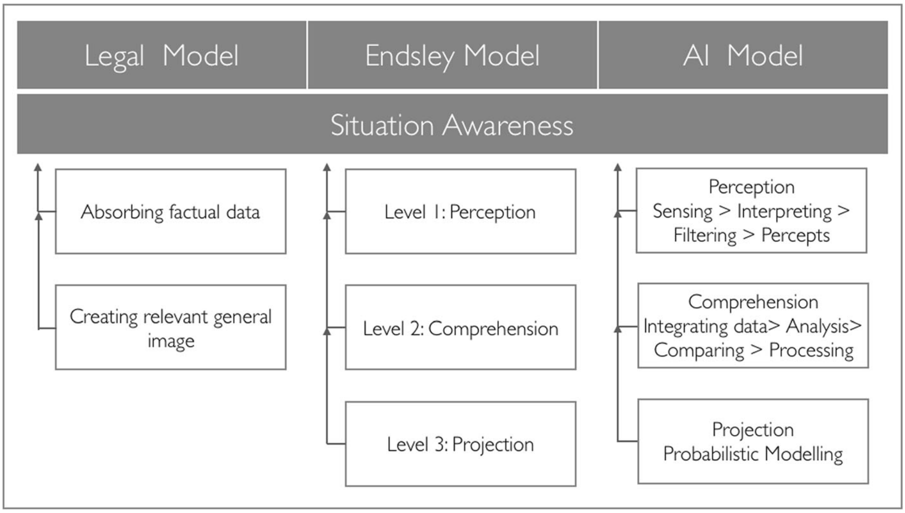

Fig. 2 Cognition requirements in AI systems

5 The *mens rea* in Criminal Offences by AI Systems: the Volition Requirement

As regards the volition requirement of *mens rea*, we can distinguish two levels: intent and recklessness. Since the offence of purchasing illegal drugs requires intent, here we consider only whether an AI technology can form intentions through a deliberative process and how intent can be proved.

5.1 Deliberation by AI Systems

The *mens rea* requirement for intentional crimes includes the intention to realise the criminal act and its effects, with the awareness of all relevant circumstances. Thus, criminal intention is directed toward the future. For instance, the murderer's intent is focused on the future causing of the victim's death and persists until the commission of the act. Criminal intention consists in a mental process that, in principle, is under the control of the agent, and may be brought to awareness, as opposed to uncontrollable/unaware urges, instincts, or impulses.

There are many different legal theories of intention, and controversy persists over its nature. In particular, there is an ongoing debate on the relation between intention, foresight, and desire. For instance, we may distinguish between different cognitive states, possibly having different significance for criminal law (Duff 1990): merely desiring something (one would like to kill somebody, but is too afraid to proceed), intending something in a strict sense, i.e. being committed to bring about the action (as when shooting to kill), merely predicting the realisation of some result as a side effect (a terrorist predicts that some people will die as a consequence of the intended destruction of a facility), or anticipating what results a specific action might possibly bring about (as when a driver knows that his dangerous behaviour may harm other people).

As a conceptual framework for intention, we adopt the approach by Michael Bratman (1987), on which the Belief Desire Intention (BDI) model of rational action is based. This model has been used to capture not only human practical cognition but also to model and develop AI architectures. According to Bratman, intention is a key attitude in planning for the future, and it consists in a commitment to act. The BDI paradigm includes the three following mental attitudes: (1) belief, (2) desire, and (3) intention. The first attitude, i.e. belief, represents the agent's informational state. Beliefs includes perceptual content, but also further information obtained through reasoning, especially by using rules of inference. The second attitude, desire, embodies the motivational state. Desires represent objectives or situations the agent would like to accomplish or bring about. An agent may have multiple desires, which may be in conflict. The third attitude, intention, represents the outcome of the agent's deliberation, i.e. what the agent has chosen to do, given its beliefs and desires. Thus, intentions are plans of action to which the agent has committed; they have to be mutually consistent.

A BDI agent will implement its current intentions as soon it has the opportunity to do so, without further deliberation. It may however revise and possibly abandon its intentions before their implementation on the basis of new deliberations.

These attitudes are captured by the following components of the BDI architecture (Kinny et al 1996):

1. A *belief store*, containing the information the agent has about the world. Through a perceptual processor, the agent observes the environment and interprets the data coming from sensors, providing input to the belief store in terms of new beliefs.
2. A *goal store*, containing the goals or desires an agent has adopted. To achieve these goals, the agent will construct plans and store them in its plan library.
3. A *plan library*, a collection of plans an agent can use to achieve some particular goal. A plan in turn consists of three main components:
   - *Body*, defining a concrete set of actions needed to fulfil a plan.
   - *Invocation conditions*, defining the circumstances under which the agent should consider and activate a plan if the agent believes that those circumstances have taken place.
   - *Termination conditions*, defining the conditions under which the agent may reconsider its current intention, such as new circumstances that make a plan unachievable.

When an agent forms new beliefs, it proceeds to evaluate which plans have invocation conditions that correspond to its internal beliefs. Additionally, it may construct or adapt its existing plans in order to achieve its goals under the new conditions. The emerging set of plans corresponds to the agent's intentions, and each plan defines a possible course of action. Therefore, intentions refer both to an agent's commitment to its desires (the goal to be achieved through the selected plans) and its commitment to the plans selected to achieve these goals.

If the architecture of an agent corresponds to the BDI approach, we can meaningfully assert that the agent has certain beliefs and intentions. The agent has a certain belief when its belief store contains the corresponding information; it has a certain intention, when it has selected the corresponding plan for implementation on the basis of its goals, plans and belief, according to its deliberative mechanism. For instance, assume that a BDI bot is asked to find and purchase online a certain musical track for the lowest price it can find within 15 min. After searching for 15 min, the bot will form the belief that a certain seller is providing the best price and the intention to buy from that seller.

5.2 Proving Intention

Since intention is an internal “state of mind”, to establish whether an AI system possesses such an internal state, we must focus on its internal structure and, more precisely, on its functioning. In particular, we have to consider whether the entity has internal epistemic states (beliefs) and conative states (desires, goals, intentions), and whether the entity processes such states, adapting its behaviour (Sartor 2009). In general, we can say that an internal state of an AI system (e.g. character strings involved in data capture and data storage functions that take in some specific environmental information) represents the belief in the existence of certain situations when that entity (1) adopts the concerned state on the basis of such situations and (2) that state contributes to making it so that the entity behaves as these situations require. Similarly, we may say that an entity has the goal (desire) of realising a certain result when there is an internal state of the entity such that while the entity has that internal state, it will tend to achieve that result, and when the result is achieved, the internal state will be abandoned (or modified so that it stops determining the above behaviour).

It is more difficult to prove intention than awareness, since awareness relates to existing current or past states, while intention includes the projection of future states (planned actions of the agent and the expected outcomes), only existing as mental representation of the agent. To prove intention, we may adopt different approaches.

One approach may consist in inspecting the internal functioning of the agent. Unfortunately, the internal configuration of the agent, at the time of the criminal action, may be unavailable, no longer retrievable or not detectable from the configuration of the agent (e.g. deep neural network).

Proof could be obtained by the system itself, in case it has the ability to provide reports on its internal states. Various technologies, of different levels of complexity, can be used to this purpose, from enabling the system to provide logical implications of its belief and goal sets, to the much more difficult task of providing explanations for the formation of beliefs and choices (Doshi-Velez et al 2017, Guidotti et al 2018, 93).

However, the reliability of such proofs presupposes the truthfulness of the AI systems concerned. Such systems could indeed be constructed in such a way as to lie about their internal states, under certain situations, or when being truthful would fail to maximise the realisation of their objectives (e.g. profits). Under what condition would we be justified in trusting such systems?

A third possibility consists in addressing the system’s behaviour on the basis of presumptions. Criminal law has developed evidentiary substitutes to deal with the complexity of proving intent. In particular, it is generally presumes that agents intend the natural and probable consequences of their acts. More to the point, according to the foreseeability rule, offenders are presumed to have intended the results of their actions whenever (1) such actions were committed with full awareness and (2) their consequences were highly likely and could have been anticipated by the offenders (Shute 2002). Let us consider how the foreseeability rule might apply to AI systems, so as to see whether they can be considered as capable of intent.

As noted in the previous section and in the example of the chess-playing computer, AI systems are capable of assessing the likelihood that certain factual events will take place, and can act accordingly: they examine alternative options, construct plans of action considering their future outcomes, and make informed decisions to implement a plan and act on it. In many contexts, AI systems have also the capability of assessing the probability that their action will lead to a certain outcome, at a human level, or even more accurately. Such AI systems can be claimed to have not only the intention to act but also the intention to bring about the outcomes of their actions.

As concerns the evidence of intent based on the foreseeability rule, the activities of AI systems at each stage in the consolidation of intent can be monitored and recorded, and so there may be direct evidence to prove criminal intent according to foreseeability rule.

In conclusion, the cognitive states relevant to criminal law (cognition/awareness, including reason responsiveness, and volition/intention), under certain conditions, can also be attributed to certain AI systems, being detectable in their architecture and configuration or presumable from their behaviour. In particular, we can say that AI systems implementing the BDI model have awareness of the relevant facts and make intentional choices.

The importance of preventing the commission of crimes by humans may require us to focus on the prevention of similar action by AI systems. As we consider not only the material behaviour but also the accompanying mental state, to appropriately react to human crimes, similarly, we may need to consider not only the material behaviour but also the accompanying cognitive states of the concerned AI system, to appropriately react to AI crimes. In particular, we argue that those harmful actions by AI systems that are motivated by intention or recklessness may need to be addressed differently that those actions that “inculpably” cause harmful consequences (see Section 8.4).

6 Responsibility

In this section, we consider whether AI systems can not only engage in AI crimes—realising both the *actus reus* and the *mens rea* required for a human crime—but can also be considered responsible for such crimes, possessing a sufficient level of reason-responsiveness (see Section 2).

6.1 Reason-Responsiveness

Criminal responsibility presupposes that the concerned agents have a sufficient reason-responsiveness, namely, a sufficient understanding of the epistemic and practical reasons at stake (see Section 2.1). We have already shown that AI systems can achieve epistemic understanding of the relevant facts. We will now consider under what conditions such systems may also have a sufficient reason-responsiveness.

First of all, criminal law aims to discourage unwanted behaviour though the threat of sanctions that negatively affect the interests pursued by the agent. To determine whether criminal deterrence also applies to AI systems, we need to consider whether an AI system can be aware of its interests (or of the interest of its owner/user) and of the ways in which criminal sanctions affect such interests. Thus, a criminally-responsive AI must possess instrumental rationality, i.e. have its purposes, and the ability to adapt its action to its purposes, taking into account the possible consequences of such actions, including in particular criminal punishments.

We may wonder whether instrumental rationality (coupled with the mandate to pursue certain interests) constitutes a level of reason-responsiveness that suffices for criminal responsibility. An AI system only possessing instrumental rationality—directed only at maximising its, or its users', utility, and at avoiding the disutility related to expected sanctions—could in fact be compared to a so-called partial psychopath, in the sense of an agent who is "incapable of moral understanding but capable of prudential deliberation and action", and who therefore "is not responsible to moral reasons, but is responsible to prudential reasons" (Duff 2007). There is a lively debate in legal doctrine on whether a partial psychopath can be criminally responsible. For instance, Morse (2008) argues for a negative answer: partial psychopaths (in the sense above) should not be subject to criminal punishment, but rather to civil commitment, the expectation of which could provide sufficient deterrence. Others argue that prudential rationality, as connected to the expectation of criminal sanctions, is sufficient for criminal responsibility (Kenny 1978, 42-44; Litton 2013).

In this regard, we need to distinguish two attitudes toward norms: on the one hand, the mere ability to recognise norms enforced in society, and the possibility to be subject to sanctions for their violation, and on the other hand, the disposition to comply with norms, on the basis either of their conventional legitimacy or of their moral merit. The motivation to comply with criminal norms on the basis of their moral merit usually results from the agents' ability to empathise with potential victims and feel responsibility for the harm and suffering of the latter.

It has been affirmed that partial psychopath may possess the first attitude—the knowledge of what is viewed as legally or morally wrong in their jurisdiction—and consequently they may comply to avoid sanction. However, they do not possess the emotional capacity to appreciate the moral wrongness of their behaviour, and thus lack the motivation to comply, unless compliance is on their interest (Slobogin 2003, 324). To go beyond the condition of a partial psychopath, an AI system needs to be responsive to moral and legal reasons, i.e. to possess a normative architecture, the capacity to take values and norms and not only sanctions into account, and thus it would need to be a normative agent (Boella and Van der Torre 2007).

Normative agents have the ability to represent norms and values, to reason with them (knowledge representation and reasoning), and in addition the motivation to comply. A normative agent may possess further capacities (Neumann 2010, and Hollnder and Wu 2011), such as the ability to (i) recognise and infer the norms followed by other agents (learning); (ii) convey norms to other agents (communication and networking); and (iii) impose punishments on other agents if they fail to comply with known norms (enforcement of morality and law).

In conclusion, it seems for AI systems to be reason-responsive for the purpose of criminal law, three capacities are relevant. The first capacity is the systems' ability to gaining awareness of their conduct and of the resulting effects. The second capacity is the systems' ability to identify and understand the norms that apply and the corresponding sanctions, and to appreciate the sanctions' impacts on their interests. The third capacity is the systems' moral motivation to comply, which usually results from their "affective knowledge" namely from the "ability to internalise the criminal act and emotionally appreciate its wrongfulness" (Slobogin 2003, 324).

As we have shown above, AI systems can possess the first capacity. With regard to the second capacity, a number of AI systems exist that include an explicit representation of norms and sanctions, and some have the ability to determine their behaviour taking into account the possibility of incurring into sanctions. Concerning the third capacity, we may wonder whether AI system already possess “real” moral motivation, namely, a motivational state that is relevantly similar to human moral dispositions (Pagallo 2017, 647).

Thus, a key issue in determining whether existing AI systems may possess reason-responsiveness as required by criminal law, consists in establishing whether the capacity for moral motivation is also required for this purpose, in addition to the awareness of facts and norms. A comparison with the way in which the legal system addresses psychopaths may be significant: partial psychopaths, who are aware of facts and norms but lack moral motivation, are generally considered to be legally responsible: on moral or pragmatic grounds, as well as on the basis of positive law, the absence of moral motivation does not excuse psychopaths. It is true that some authors have also argued for the non-responsibility of partial psychopaths, since punishing and blaming those who cannot understand the moral wrongness of their acts would violate their fundamental rights (for a discussion, see Litton 2013). However, we believe these arguments for non-responsibility do not apply to AI systems (unless we seriously view them as bearers of fundamental rights).

In conclusion, we do not believe that moral motivation is required for AI systems to possess the level of reason-responsiveness that is necessary for their criminal responsibility. This does not mean that AI systems having capacity for awareness of facts and norms should necessarily be criminally responsible, since capacity for reason responsiveness, while being a necessary precondition for criminal responsibility, it is not sufficient for it. The attribution of criminal liability depends on the contingent content of positive law, according to the principle of legality (see Section 9). As we shall argue later, the criminal liability of AI systems should be decided on pragmatic reasons, namely, considering how its attribution may have a deterrent effect (see Section 8.4).

6.2 Compliance and Intelligent Violation

To date, normative agents are broadly based on BDI architectures for both selecting goals and devising plans so as to achieve their goals. They include norm compliance and value-achievement among their goals or among the constraints over the achievement of their particular interests (Calfranchi 1999). We can have two approaches to the design of normative agents. While in the first approach the agents will always obey the norms (whenever possible), in the second approach they will decide whether to comply on the basis of their own reasoning.

On the first approach, norms are statically built into the agents’ protocols (Jennings 1993, 223), as static constraints within their architecture (Shoam and Tennenholz 1992), so that agents cannot choose to violate norms in pursuit of particular goals. This kind of normative agents is mere norm-followers; they cannot change their compliance patterns over time in light of accrued experience.

On the second approach, a more flexible architecture allows for intelligent norm violation. Consider, for instance, a self-driving car having to avoid a group of pedestrians who are crossing the street. Assume that it is too late for the car to come to a full stop, and so it needs to swerve into the opposite lane by crossing a solid double line. Although this manoeuvre is prohibited by the rules of the road, it may well be a more reasonable, smarter, or moral choice for a self-driving car to make, considering that the alternative would be to kill the pedestrians. Truly intelligent normative agents would have the ability to (a) know that a norm exists; (b) take this norm into account in its decision-making and behaviour, and (c) then decide whether to follow the norm in the case at hand. It is important to note that taking a norm into account does not necessarily mean following it: it means only that the goals the agents select and the plans they set out in light of those goals will be informed by their belief that the norm exists, and by their motivation to follow it (unless there are prevailing reasons to the contrary) (Castelfranchi et al 1999). These agents are not mere norm-followers, they can also violate a norm out of necessity or convenience, depending on the circumstances of the case at hand. Such agents may be better suited to dynamic environments, since they have the ability to violate a norm when the violation is needed to satisfy more important legal values (e.g. saving lives) or superior norms. Such agents could in principle invoke legal justifications for violations of criminal norms, such as self-defence, or state of necessity. As far as we know, prototypes of systems capable of intelligent norm-violation already exist; we are not yet aware of operative applications.

A negative aspect of the “freedom” of intelligent norm violators would consist in the possibility of opportunistic violations. Agents having the capacity for the intelligent violation of norms could indeed choose to violate any norm whenever compliance would not fit the goals they are pursuing (e.g. whenever this compliance would fail to maximise the utility of their users), so behaving as the Bad Man of Judge Holmes.[10] Moreover, such agents, even when disregarding legal norms out of moral perspectives, could misunderstand what is required by the moral imperatives at stake.[11] Thus, the development of AI system capable of intelligent norm violations should be undertaken with great caution, allowing for the disapplication of existing norms only under very limited circumstances.

7 The Random Darknet Shopper: Case and Possible Scenarios

In this section, we present some scenarios based on the Random Darknet Shopper case mentioned in Section 1. After presenting the case, we discuss different ways in which an AI system may be involved in the purchase of illegal drugs. We consider commonalities and differences in the commission of the considered crime by humans and by AI systems, drawing possible analogies and evaluating different models of criminal liability.

[10] According to Holmes to understand the working of the law we have to consider the psychology of the bad man, namely, the individual that only complies of the law in order to avoid sanctions, and only to the extent that the disutility of the sanctions outweighs the benefit to be obtained through the violation (see Holmes 1897, 459).

[11] For some examples of the harm that could be caused by an AI system that misunderstands a moral imperative, see Bostrom (2014). For instance, a “perverse instantiation” of the utilitarian imperative of making people as happy could be implemented by “implanting electrodes into the pleasure centers of our brains” (158).

7.1 The Random Darknet Shopper Case

In November 2014, the Random Darknet Shopper, an online shopping bot, was programmed to go to one particular marketplace on the Deep Web and make one random purchase a week, with a budget of $100 in Bitcoin. It bought a diverse set of items: a pair of fake Diesel jeans, a baseball cap with a hidden camera, a pair of Nike trainers, 200 Chesterfield cigarettes, a decoy letter (used to see if your address is being monitored), a set of fire-brigade issued master keys, a fake Louis Vuitton handbag, and 10 ecstasy pills. The purchases were all made for an art show in Zurich, titled The Darknet: From Memes to Onionland, which closed on January 11. All products were on display as part of the exhibition (Power 2014).

It appears that a crime was committed by the electronic agent, or at least that the agent engaged in an action that would count as a crime if performed by humans, i.e. the offence of purchasing illegal drugs.

Pills of pure MDMA (a synthetic recreational drug) were confiscated by the St. Gallen public prosecutor's office, along with the Random Darknet Shopper and other articles the bot bought on the Deep Web. The bot's creators were threatened with prosecution, and the bot was seized. Three months after the confiscation, the Swiss public prosecutor decided to drop charges and released the bot back to the artists. In the order for withdrawal of prosecution, the prosecutor stated that the outweighing public interest in the social and artistic issues raised by the Random Darknet Shopper justified the possession and exhibition of drugs by the artists. In particular, as reported by the artists on their !MEDIENGRUPPE BITNIK website:

"In the order for withdrawal of prosecution the public prosecutor states that the possession of Ecstasy was indeed a reasonable means for the purpose of sparking public debate about questions related to the exhibition. The public prosecution also asserts that the overweighing interest in the questions raised by the art work «Random Darknet Shopper» justify the exhibition of the drugs as artefacts, even if the exhibition does hold a small risk of endangerment of third parties through the drugs exhibited."[12]

In the end, the artists were cleared of all charges (Kharpal 2015, Kasperkevic 2015).

We believe that this case is significant for the discussion of AI crimes even though both the artists and the Random Darknet Shopper could avoid any sanctions because of the safe harbour for artistic creations provided by the Swiss Constitution. In fact, in different legal jurisdictions, this safe harbour many not be available or not cover such activities. In any case, the issues raised by this case have larger implications, also concerning AI crimes not pertaining to artistic endeavours, or being inexcusable for their serious consequences.

7.2 Variations on the Purchase of Illegal Drugs

In this section, we shall consider different possible involvements of humans and AI systems in the purchase of illegal drugs, by changing aspects of the Random Darknet Shopper case. In particular, we shall discuss the following five scenarios, exhibiting different human-machine interactions and levels of control, as well as involving different cognitive skills and autonomous initiatives by the AI system. In each scenario, we will examine whether any human is to be held criminally liable for the offence

[12] See !MEDIENGRUPPE BITNIK website (https://motherboard.vice.com/en_us/article/mgbwg4/in-europe-robots-can-legally-buy-drugs-online-for-art), last accessed 22 May 2018.

committed through the AI system. In case nobody would be responsible, we shall consider whether there is a criminal liability gap that needs to be addressed.

1. First scenario: The web robot is designed or employed with the intention or knowledge that it will engage in criminal conduct, and the AI technology does not satisfy the requirements of the *mens rea*.

The AI system is used as a mere instrument in the commission of the offence, executing orders exactly as instructed. The programmer and the user satisfy the means rea of the crime (purchase of illegal drug) but they do not meet its material-element, since they do not perform the action at stake. In this case, we can have three different sub-scenarios:

(a) the web robot does not deploy relevant cognitive capabilities;
(b) the web robot's capabilities resemble those of a person lacking full capacity, such as a child or a mentally incompetent person;
(c) the web robot's capabilities resemble those of animals.

In the first two sub-scenarios, the web robot must be considered an innocent agent. According to the doctrine of innocent agency, a person who did not materially commit an offence may be liable for acting through an innocent agent. The acting agent is innocent by reason of lack of the required fault element or lack of capacity (no *mens rea or insufficient reason-responsiveness*).

As an innocent agent, the web robot commits the offence, while the person who has orchestrated the offence sending or activating the robot is criminally responsible as a perpetrator. The perpetrator's liability is determined on the basis of the conduct of the robot and the mental state of the perpetrator (Gillies 1980; Hallevy 2010, 2011, 2012, 2013).

The role of perpetrator is played by the person who has the relevant intention and set up the robot so as to implement that intention. There are two main candidates for the role of perpetrator: the programmer of the AI bot, and its user. Which one of them would play the perpetrator's role will depend on who intentionally set up the bot to commit the offence: was the offensive behaviour pre-programmed by software developers, or requested by users? In the third sub-scenario, the bot is considered as an animal (Schaerer et al 2009, Kelley et al 2010). Let us consider two possibilities: (a) the bot-as-animal acts of its own initiative or (b) it is directed by its owner.

The first possibility—the bot-as-animal acting of its own initiative—is addressed by considering that animals are human property, i.e. entities over which humans can own and exercise property rights. If animals cause harm, usually the humans who own the animals are legally responsible, and have the civil obligation to compensate the harm. For example, if a person is attacked by a dog, its owner is legally responsible for any harm or injury caused. Models of liability differ in modern legal systems,[13] but in any case

[13] For an overview of variations in state statutes for strict liability for dog bites, see generally Miller (1987), Wisch (2012) and Walden (2017). In extreme cases these issues fall within the scope of criminal law whenever dog owners violate legal restrictions on keeping dangerous dog or the owner's failure to control the animal is reckless or criminally negligent (e.g. under the Dangerous Dog Laws).

In extreme cases these issues fall within the scope of criminal law whenever dog owners violate legal restrictions on keeping dangerous dog or the owner's failure to control the animal is reckless or criminally negligent (e.g. under the Dangerous Dog Laws).

animals are not legally responsible. Thus, in our variation of the Random Darknet, the artists owning the bot would be civilly liable for the damage it might have caused.

The second possibility—the bot-as-animal being directed by its owner—can be addressed by considering that the humans who directed animals to commit an offence are the real perpetrators and perform the criminal action using animals as their instruments. For instance, in an assault committed by a dog under the order of its owner, the incident must be considered an assault (malicious wounding) committed by the owner.

Thus, in this variation of the Random Darknet case, the artists themselves would be deemed the perpetrators.

2. Second scenario: The web robot is designed or employed with the intention or knowledge that it will engage in criminal conduct, and the AI system is capable of meeting the *mens rea* requirement.

In this second scenario, the act (*actus reus*) is accomplished by the Darknet Shopper, and both the artists and the Shopper fulfil the mental element requirements. If we replace the Shopper with a human being, we will clearly have a case of complicity, such as joint perpetration, perpetration-through-another, incitement, or conspiracy. In fact, complicity in an offence requires in each accomplice, besides some involvement in the action, the possession of the *mens rea* for that offence Bronitt and McSherry (2017), Gillies (1980), Lepara and Goodin (2013).

However, satisfying the behavioural and mental requirements for being an accomplice is not sufficient in order to be criminally responsible as an accomplice. For this purpose, legal personhood is required (see Chopra and Laurence 2011, 153.89; Calverley 2008; Asaro 2016; Pagallo 2013, 40), i.e. the bot itself must be viewed as the addressee of criminal norms and sanctions (see Section 2). We shall return to this issue in Section 8.

3. Third scenario: The web robot is not designed or employed with the intention or knowledge that it will engage in criminal conduct, but the programmer or the user has taken unreasonable risks that caused the conduct to occur.

In this scenario, the web robot meets the *actus reus* requirements, but the artists did not intend the offence to be committed: it was not their intention to commit the offence by instrumentally using the web robot, nor had they anticipated the possibility of its occurrence. However, they could have foreseen the occurrence, by applying required diligence. Here, we can distinguish four different subcases:

(a) the web robot does not deploy relevant cognitive capabilities;
(b) the web robot's capabilities resemble those of a person lacking capacity, such as a child or a mentally incompetent person;
(c) the web robot's capabilities resemble those of animals;
(d) the web robot fulfils the mental element requirements.

In the subcases (a) and (b), neither the web robot nor the artists fulfil the mental element requirement for intentional crimes: they had no knowledge of the committed offence and had no intention to commit it. To determine whether the artists may be criminally liable for their behaviour in this scenario, we need to consider whether the *mens rea* for the offence at issue requires intention, or whether it is also satisfied by recklessness or negligence. In the latter case, the artists would be liable since they deployed the bot disregarding the fact that this gave rise to a substantial and unjustifiable risks. In particular, the artists did not set suitable constraints and restrictions on the kind of goods the web robot could buy or on the websites it could visit, allowing it to enter deep-web and dark-web sites. They released the bot into an environment where it was highly probable that some unlawful outcome would occur.

In subcase (c), we could apply to AI systems the legal rules that apply to animals. For instance, most animal laws have special restrictions on dangerous dogs. Owners are required to muzzle dog that have already caused injuries; if they do not comply and the dog injures or kills someone, they could be found guilty because of their recklessness or negligent behaviour. Accordingly, in the Random Darknet Shopper case, if the offence of the purchase of illegal drugs admits of gross negligence or recklessness, then programmers and users could be found liable.

In the subcase (d), the bot satisfies both the material and mental elements of the crime. However, it will not be liable unless having legal personality under criminal law. The artists will not be liable if the crime requires intention; they may be liable for recklessness if crime allows for this mental requirement.

4. Fourth scenario: The web robot is designed or employed with the intention or knowledge that it will engage in criminal conduct, but the AI system quantitatively or qualitatively exceeds the original plan.

In this scenario, the artists knowingly and wilfully design or use the bot to purchase illegal drugs, but the bot strays from the plan and commits some other offence, on top of or in place of the one that has been planned (e.g. illegal purchase of weapons). This scenario resembles the basic idea of the natural and probable consequences doctrine in accomplice liability cases (Heyman 2010; Bird 2006, 43). In legal systems in Continental Europe, as well as in English Common Law, criminal liability for the unplanned offence is ascribed to all the parties involved in the planned offence, according to the so-called natural and probable consequences liability model (Robinson 1997; Hallevy 2012).

Suppose, for example, that a group plans a bank robbery that does not involve killing anyone, yet, during the robbery, a guard is shot and killed by one of the accomplices. The homicide was not part of the plan, and the other accomplices did not commit the shooting or agree on it, even though a reasonable person would have foreseen this outcome. According to natural and probable consequences doctrine, all accomplices are accountable for both the robbery and the homicide.

Let us now apply this doctrine to the Darknet Shopper case. We assume that the artists intended to have the Shopper buy illegal drugs, but that the Shopper exceeded the planned offence either qualitatively (committing additional offences of a different type) or quantitatively (committing additional offences of the same type). In this scenario, we need to distinguish liability for the planned and the unplanned offence. Liability for the planned offence falls under the scenarios (1) or (2) above, according to whether the bot possesses the required *mens rea*. Concerning the liability for the unplanned offence, as above, we may distinguish four different subcases:

(a) the web robot does not deploy relevant cognitive capabilities;
(b) the web robot's capabilities resemble those of a person lacking capacity, such as a child or a mentally incompetent person;
(c) the web robot's capabilities resemble those of animals;
(d) the web robot fulfils the mental element requirements.

In all these subcases, the artists will be criminally liable for the planned offence. In addition, according to the previously discussed probable and natural consequences model, they will be liable for the unplanned offences being probable consequences of the planned offence.

In subcase (d), the web bot satisfies the material and mental requirements of both the planned and the unplanned offence. If it were a human being, it would be considered both accomplice in the planned offence, and perpetrator for the unplanned one. However, as noted above, the bot while satisfying both requirements will not be criminally liable unless having legal personality for the offence at issue.

5. Fifth scenario: the web robot commits the crime, but no intention or recklessness can be ascribed to the programmer or user.

In this scenario, different subcases can also be distinguished:

(a) the web robot does not deploy relevant cognitive capabilities;
(b) the web robot's capabilities resemble those of a person lacking full capacity, such as a child or a mentally incompetent person;
(c) the web robot's capabilities resemble those of animals;
(d) the web robot fulfils the mental element requirements.

In the first three subcases, *mens rea and responsibility* cannot be ascribed neither to the artists nor to the bot. Thus, these cases would not be subject to prosecution. Possibly the artists may have to compensate damages according to civil law. In the last subcase, only the web robot satisfies the *mens rea* requirement. As above, it will not be criminally liable unless having legal personality under criminal law for the offence at issue.

8 A Regulatory Perspective

In the above sections, we have seen that it is possible for AI systems to engage in activities that would constitute crimes if they were accomplished by humans, i.e. AI crimes. We have also seen that some AI systems may possess a certain degree of reason responsiveness, either only prudential or also moral/legal (normative agents). We need now to establish how the law may respond to AI crimes.

We think that "criminal" AI systems will require a specific response by the law, since they are particularly dangerous: not only there may be a liability gap but the social consequences of AI crimes may be extremely serious. Consider, for example, the case of a medical robot, managing the delivery of drugs to hospitalised patients, that chooses to kill all patients requiring expensive treatments in order to save costs, or an autonomous car that in order to reduce travel times drives at the highest possible speed, regardless of harms to pedestrians.

As we have seen in Section 7, there are cases in which the users or developers will be criminally liable, as authors or accomplices in AI crimes. The criminal punishment of human agents can indeed provide deterrence concerning the intentional use of criminal AI systems (assuming that evidence of the user's intent can be obtained). Outside of these cases, to deter AI crimes, we need to rely on different approaches.

8.1 Limiting the Task-Autonomy of AI Systems

First of all, the law could limit tasks that can be assigned to AI systems, or limit their autonomy in carrying out those tasks, according to the specific context in which AI systems are operating, along with the attendant risks. Thus, in highly sensitive areas, such as military operations, it would seem a good choice to constrain the autonomy of AI systems, while ensuring human monitoring (Schmitt and Thurner 2012, 231). In general, we believe that humans should remain in the loop and exercise meaningful control whenever the delegated tasks involve the possibility of intentionally harming humans, even under legitimate grounds for justification. Consider, for instance, the case of AI security guards, being used on private property to prevent and stop theft, property damage and personal injury. In such cases, there should always be a human decision before the AI systems implement any potentially harmful defensive measure. In fact, such systems may be unable to properly identify, and target dangerous actions, and may fail comply with legal standards and security safeguards, causing unnecessary or disproportionate harm to humans.

Limiting AI tasks and keeping humans in the loop, however, does not provide a general regulatory approach to criminal AI behaviour, since it would be impossible to prevent all possible AI crimes by limiting AI autonomy, without considerably restricting the useful ways of deploying AI. Many civil domains in which AI agents can be usefully deployed, without direct human involvement, inevitably provide opportunities for autonomous criminal actions (commercial exchanges offer opportunities for fraud, physical interaction offers opportunities for harm, etc.).

8.2 Civil Liability

Civil law remedies may provide compensation for victims of AI crimes (as for any other unlawful harmful behaviour by AI system). When the harmful robotic behaviour has been determined by faulty human action, compensation can be provided by humans, according to the general rules on intentional or negligent torts.

Outside the domain of fault liability (or in addition to it), compensation can be provided by various forms of strict or semi-strict liability, possibly supplemented by insurance and limited by caps. In particular, it has been argued that AI users may be subject to strict or semi-strict liability for harm caused by their AI systems, in the same way that owners of animals are strictly liable for the harmful behaviour of their animals. Following this approach, it has been affirmed that AI systems are similar to animals since they are (a) interactive and able to perceive their environment and to respond to stimuli by changing the values of their own properties or inner states; (b) autonomous, because they modify their inner states or properties without any direct human intervention, thereby exerting control over their actions; and finally, they are (c) adaptable, improving the rules through which their own properties or inner states change (Pagallo, 2013, 37; Allen et al., 2000; Floridi and Sanders, 2004, 349).

We think, however, that AI systems have (and will have in the future to a larger extent) cognitive capacities—and physical capacities, when connected with appropriate physical actuators—that vastly exceed, in many regards, those of animals. Consequently, AI systems can engage in many activities—contact formation, medical diagnosis and therapy, driving vehicles, governing machines, etc.—that are precluded to animals. AI systems can endorse much more varied criminal objectives, and their criminal behaviour can have much more serious impacts on human lives and interests than harmful animal behaviour. Therefore, we think that the legal response to animal harm (strict or semi-strict civil responsibility for guardianship) may fail to fully address AI crimes.

We would argue that, in general, establishing a mechanism for compensating victims, even based on strict liability, may be an insufficient response to AI crimes. This consideration applies whenever the gain that can be obtained through crimes may exceed the expected cost of the obligation to compensate victims.

First of all, compensation would not be applicable to those crimes that do not harm specific individuals. Consider, for instance, the case of our example: nobody, arguably, suffered a monetary loss as a consequence of the purchase of illegal drugs (or weapons) by the Random Darknet Shopper.

Even when individuals are harmed, the obligation to compensate them may not provide adequate deterrence. Consider, for instance, the case of an AI system that, to increase profitability, engages in fraudulent commercial practices or puts at risk people's lives (e.g. reducing maintenance or controls in a transport system, driving at extreme speed, etc.). These actions would amount to a crime, if they were accomplished by humans, and their authors would meet not only the obligation to compensate damage, but also criminal sanctions. If such actions were accomplished by an AI system, and were met only with civil sanctions, it may be convenient for the system to persist in its illegal behaviour (and for the user to allow this possibility), while paying compensation in those cases in which the illegality was detected. If, as it is likely, many AI crimes may remain undetected (as is the case for most crimes committed by humans), the gain that can be obtained through such crimes may exceed the cost of the obligation to compensate victims.

A third difference between criminal and civil law is that criminal law also punishes attempts, while civil liability addresses only cases in which harm really takes place. If only civil liability would regulate AI crimes, then the law would provide no response in cases in which the AI systems try to engage in criminal action but do not complete such actions. Consider the case in which a medical robot acting for an insurance company intentionally tries to kill an expensive patient, but fails to succeed (e.g. since a nurse detects that a lethal drug is going to be delivered by the robot). Should this behaviour have no legal consequences, since no harm was caused, and therefore no obligation to compensate was incurred?

8.3 A Specific Criminal Liability for Creating and Deploying Criminal AI Systems

There are two possible answers available, alternatively or conjunctively, to address the insufficiency of civil law compensation to the victims of criminal AI behaviour: expanding the legal responsibility of the humans in charge of the AI system, or providing for some remedies directed against the system.

Let us first address the first approach, namely, punishing the behaviour of the user/controller who has intentionally or negligently allowed the AI system to develop criminal behaviour (e.g. by adopting an architecture that enabled such behaviour or by omitting the controls that, for example, allowed the system to evolve becoming dangerous). This liability may apply even when no damage or injury has occurred, for the simple fact of having undertaken a dangerous activity through an AI system, so creating a significant risk of harm.

This result could be obtained by broadening the scope of recklessness (in continental system, *dolus eventualis*), to cover the so-called opaque reckless (Ferzan 2000, 597), namely the situations in which the defendant knows that his or her conduct is risky but fails to realise or consciously disregards the specific nature of the risk. Under this approach, the use of an AI systems with the awareness that it might engage in criminal activities of a certain kind (e.g. unlawful commercial transactions) suffices for the user's criminal liability for the specific criminal activity performed by the system (e.g. purchase of Ecstasy), even when the user did not foresee that the system would engage in the specific activity. Alternatively or additionally, in such cases, the user could be addressed through statutory liability.

Criminal or statutory liability may prevent opportunistic behaviour by AI users, or induce them to take additional care. Moreover, the persistent use of criminal AI systems that the user/controller knew to be unsafe could be considered as a criminal or statutory wrong, whether or not damages or injuries occur, rather than a mere civil law issue.

The negligent creation/deployment of a criminal AI system may also result from the failure to adopt a normative architecture (Hollander and WU 2011, 6). The adoption of such an architecture could indeed prevent AI systems from engaging in criminal activities that they would have chosen according to a merely prudential reasoning, as the activities that would have most advanced their interests. In particular, normative constraints could prevent both the adoption of criminal means to achieve permissible goals (e.g. engaging in fraud for maximising profits) and the direct pursuit of criminal goals (e.g. killing an adversary). Failure to prevent harmful criminal behaviour, through state-of-the-art technological solutions, may engender civil liability according to existing laws and also lead to criminal or statutory, liability, as we have argued above. Sanctions for the failure to adopt an adequate architecture (regardless of the actual causation of harm) may also be established in correlation with legally enforced technical standards.

The need to protect society from AI crimes may also justify (as a precautionary measure) the prohibition to deploy certain kinds of AI systems, and justify their termination when the prohibition is violated.

8.4 Punishing AI Systems

In principle, AI crimes could also be addresses by making AI systems directly subject to criminal law. Given the current socio-technical arrangements, we are not arguing for this approach, but it may be useful to start speculating about it. As we observed above, subjecting AI systems to criminal law means that such systems would be subject to some kind of criminal punishment for their behaviour: when their actions counts as crimes if committed by humans (and possess the corresponding *mens rea*), they would be subject to sanctions. Therefore, they would have personality in criminal law in the sense specified in Section 2. This presupposes, according to the general principles of criminal law, that such AI systems could be viewed as responsible, being sufficiently responsive to reasons (at least in the sense that they have the ability to understand that their action will lead to sanctions upon them).

Whether AI systems should be viewed as duty-bearers under criminal law will also depend on the contingent normative choices of each legal system, namely, from the choice to let (some) criminal norms also cover the actions of AI systems. This choice should be based not only on the available technologies (namely, on the fact that AI systems possess the cognitive capacities that are needed for responding to criminal norms, as discussed above), but also on the appreciation of the values at stake. In particular, regulators should consider the extent to which punishing AI crimes may contribute to efficiently prevent harmful behaviour by AI systems.

Obviously, a convenient system of penalties and due adjustments has to be designed.

Consider, for instance, an AI operating in the stock market that engages in criminal activities to maximise the profits of its clients (and its own commissions), such as the unauthorised remote access to IT systems and data for the purpose of insider trading. Penalties for such computer and financial crimes usually consist in a certain period of incarceration, fines and restitution to victims, or both.

A convenient and effective deterrent system of punishment for self-interested AI systems may include fines, to be collected from the AI systems themselves, as happens in the case of corporations. This presupposes that AI systems have legal personality under private law, as specified in Section 2, and in particular, the capacity to own assets (a source of funds from which victims could be compensated). For instance, AI systems may be backed by a warranty structure, or contingency funds, from which fines can be deducted, or could even be viewed as corporate entities.[14] Such funds could be initially provided by owners and users, and complemented with the gains obtained by the system. Victims could additionally be compensated through insurance.

Also, a deprivation or limitation of liberty may, through opportune adjustments, influence the present and future behaviour of AI systems. For example, criminal systems might be temporarily or permanently banned from interactions they value (e.g. a marketplace). This exclusion may have a deterrent function, and its implementation may prevent the future illegal behaviour of the punished agent.

The possibility of punishing AI systems might be relevant under different rationales. Under a deterrence objective, punishment can be justified as it may dissuade such systems from committing the same criminal actions: AI agents aiming to maximise their utility will refrain from engaging in criminal activities leading to expected losses (sanctions) exceeding expected benefits. Under a rehabilitation objective, punishment could be directed to improve systems' performance, for example by refining decision-making processes through learning or

[14] Bayern, S., et al. Company law and autonomous systems: a blueprint for lawyers, entrepreneurs, and regulators. Hastings Sci. & Tech. LJ, 2017, 9: 135. Pagallo, U., The laws of robots, p. 103 (n 2).

by introducing norms as constraints in the system's architecture. In conclusion, it is not impossible—though certainly unneeded, outlandish, and merely speculative under the present circumstances—that advanced AI systems might be subject to criminal sanctions, under a deterrence and rehabilitative rationale.

As an alternative to making AI systems criminally responsible, we might consider whether establishing a users' responsibility to pay non-compensatory statutory fines for AI crimes could provide an equivalent avenue to deter AI crimes. The idea is that if users/deployers were fined (for amounts exceeding compensation) for crimes committed by AI systems, this should induce them to prevent such crimes. In particular, they should be induced to provide AI systems with the motivation to act in such a way as to prevent sanctions against their users. This might be obtained—when deterrence through sanctions is the most appropriate way to influence the behaviour of AI systems—either by making such systems internalise in their utility function the disutility resulting from sanctions against their users, or by providing adequate private sanctions. Assume for instance (a) that an AI system has the goal of maximising the pot of money it gains through market exchanges, (b) that in case it intentionally engages in fraudulent activities, its user is punished with statutory sanctions, and (c) that these sanctions outweigh the rewards to be obtained through frauds. Assume also that the user consistently commits to detract from the pot of money assigned to the system whatever amount the same user has to pay to cover fines resulting from the activity of the system (or sets up an automated mechanism, e.g. a smart contract, which does that). Under such conditions, the private sanctioning mechanisms established by the user against the AI system (induced by the threat of public fines against the user) could achieve the same deterrent effect of public fines targeting the system.

8.5 The Principle of Legality/Legal Certainty and the Punishment for AI Crimes

We have considered in the previous section two possible ways of addressing AI crimes through criminal law.

On the one hand, a specific criminal liability could address creation and deployment of criminal AI systems. This could be achieved by extending the responsibility of users and controllers though various refinements of existing norms and doctrines, such as expanding the concepts of recklessness and negligence or introducing criminal or statutory liabilities.

On the other hand, we may envisage also the future possibility of punishing AI systems, who would be viewed as addressees of criminal norms and sanctions.

These changes may require new legislation, since, according to the principle of legality, there should be no crime nor punishment without a criminal law. This clause of immunity is also enshrined in Article 7 of the 1950 European Convention on Human Rights (Pagallo 2017). We may indeed draw a parallel between punishing AI crimes and the introduction, more than 20 years ago, of new criminal rules addressing computer crimes. In that case, to avoid human impunity for novel harmful behaviour, it was necessary to introduce new criminal rules (e.g. against illegal access or computer-related fraud). The same may be needed to address AI crimes.

9 Conclusion

We have shown that AI systems can engage in activities that would constitute crimes if accomplished by humans. Certain AI systems, in particular those implementing a BDI approach, can also possess the cognitive states relevant to criminal law (cognition/awareness, and volition/intention, as required for *mens rea*). Some systems may also possess the level of reason-responsiveness that is required for criminal responsibility. To meet the latter standard, an AI system would need to be at least prudentially rational, being able to anticipate the negative effects of punishment on the achievement of its goals/utilities.

We have also considered how the law may respond to AI crimes. We have argued that the autonomy of certain AI systems—especially those charged with tasks that involve intentional harm to humans, even on legitimate grounds—may be limited, in view of the risks that come with such tasks, particularly the loss of human lives.

Since it is not always possible or desirable to limit the autonomy of AI systems, AI crimes need further legal responses. We have observed that the remedies provided by existing laws may be insufficient. On the one hand, in many cases, no human users/controllers would be criminally responsible, since they could not be considered as accomplices or instigators of the AI crimes. On the other hand, civil liability may not provide sufficient deterrence to AI crimes, since it is limited to the compensation of victims. Thus, we have argued that the deployment of criminal AI systems could be addressed first of all by a broad interpretation of the notion of recklessness (*dolus eventualis*) so that that deployers could be viewed as criminally liable for AI crimes whenever they have entertained the possibility of AI crimes committed by their systems.

Additionally, or alternatively, we have considered that the creation and deployment of criminal AI systems could be viewed as a separate criminal offence, at least when potential victims are put at danger. We have also argued that in some cases, a necessary precaution—the omission of which may lead to criminal liability—may consist in endowing the AI system with a normative architecture.

Finally, we have also speculated on the future possibility of directly punishing criminal AI systems, by devising appropriate measures. If the law would develop in this direction, AI systems would become the direct addressees of criminal norms, which presupposes that they have a sufficient level of reason-responsiveness, at least in the sense that their actions can be influenced by the prospect of criminal sanctions.

References

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence, 12*, 251.

Asaro, P. M. (2016). *The liability problem for autonomous artificial agents. Ethical and moral considerations in non-human agents*. AAAI Spring Symposium Series.

Ashworth, A., & Horder, J. (2013). *Principles of criminal law*. Oxford University Press.

Austin, J. (1875). Lecture on jurisprudence: Or the philosophy of positive law. J. Murray.

Austin, J. and Austin, S. (2000). The province of jurisprudence determined. J Murray.

Bengio, Y. et al. (2003). A neural probabilistic language model. *Journal of machine learning research, 3*, 1137.

Bhuta, N., Beck, S., Geiss, R., Kress, C., & Liu, H. Y. (2015). *Autonomous weapons systems: Law, ethics, policy*. Cambridge University Press.

Bird, K. R. (2006). Natural and probable consequences doctrine: Your acts are my acts' W. *St. UL Rev, 34*, 43.

Boella, G. and Van Der Torre, L. (2007). A game-theoretic approach to normative multi-agent systems. IEEE Transactions on Systems, Man, and Cybernetics, 68–79.

Bostrom, N. (2014). Superintelligence. Oxford University Press

Bratman, M. (1987). Intention, plans, and practical reason. Harvard University Press.

Bronitt, S., and McSherry, B. (2017). Principles of criminal law 4e. Thomson Reuters.

Brozek, B. (2017). The troublesome 'person'. In Kurki, V. A. and Pietrzykowski, T., editors, Legal personhood: Animals, Artificial Intelligence and the Unborn. Springer

Calverley, D. J. (2008). Imagining a non-biological machine as a legal person. *AI & SOCIETY, 22*, 523.

Castelfranchi, C., Dignum, F., Jonker, C. M., and Treur, J. (1999). Deliberative normative agents: Principles and architecture, in International Workshop on Agent Theories, Architectures, and Languages. Springer.

Chopra, S., & White, L. F. (2011). *A legal theory for autonomous artificial agents*. University of Michigan Press.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D. Shieber, S., Paulson, J.A., Waldo, J., Weinberger, D., Wood, AA. (2017). Accountability of AI under the law: The role of explanation. Preprint available at arXiv:1711.01134.

Duff, R. A. (1990). Intention, agency and criminal liability: Philosophy of action and the criminal law. Blackwell.

Duff, R. A. (2007). Answering for crime: Responsibility and liability in the criminal law. Bloomsbury Publishing.

Endsley, M. R. and Garland, D. (2000). awareness: A critical review, situation awareness analysis and measurement. CRC Press.

Ferzan, K. K. (2000). Opaque recklessness. *J. Crim. L. & Criminology, 91*, 597.

Fischer, J. M., & Ravizza, M. (2000). *Responsibility and control*. Cambridge University Press.

Floridi, L. (2016). The method of abstraction. In *The Routledge handbook of philosophy of information*. Routledge.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*, 349.

Freitas, P. M., Andrade, F, and Novais, P. (2014). Criminal liability of autonomous agents: From the unthinkable to the plausible, in AI Approaches to the Complexity of Legal Systems. Springer.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature, 521*, 452.

Gillies, P. (1980). The law of criminal complicity. Law Book Company.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys (CSUR).

Hallevy, G. (2010). The criminal liability of artificial intelligence entities—From science fiction to legal social control, Akron Intell. Prop. 4, J. 171.

Hallevy, G. (2011). Unmanned vehicles: Subordination to criminal law under the modern concept of criminal liability, JL Inf. & Sci., 21 200.

Hallevy, G. (2012). The matrix of derivative criminal liability. Springer Science & Business Media.

Hallevy, G. (2013). When robots kill: Artificial intelligence under criminal law. UPNE.

Hart, H. L. A. (1968). Punishment and responsibility: Essays in the philosophy of law. Clarendon Press.

Herring, J. (2014). Criminal law: Text, cases, and materials. Oxford University Press.

Heyman, M. G. (2010). The natural and probable consequences doctrine: A case study in failed law reform Berkeley. *J. Crim. L., 15*, 388.

Hinton, G., Deng, L., Yu, D., Dahl. G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., and Kingsbury, B. (2012). Deep neural networks for acoustic modelling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine, 29, 82*.

Hollander, C. D., & Wu, A. S. (2011). The current state of normative agent-based systems. *Journal of Artificial Societies and Social Simulation, 14, 6*.

Holmes, O. W. (2009). *The common law*. Harvard University Press.

Holmes, O. W. (1897). The path of the law. *Harvard Law Review, 10*, 457.

Hsu, F.-H. (2002). *Behind deep blue: Building the computer that defeated the world chess champion*. Princeton University Press.

Jennings, N. R. (1993). Commitments and conventions: The foundation of coordination in multi-agent systems. *Knowledge Engineering Review, 8*, 223.

Kasperkevic, J. (2015). Swiss police release robot that bought ecstasy online, The Guardian, (https://www.theguardian.com/world/2015/apr/22/swiss-police-release-robot-random-darknet-shopper-ecstasy-deep-web) last accessed 30 March 2018.

Kelley, R., Schaerer, E., Gomez, M., & Nicolescu, M. (2010). Liability in robotics: An international perspective on robots as animals. *Advanced Robotics, 24*, 1861.

Kelsen, H. (1967). The pure theory of law. University of California Press.

Kenny, A. J. (1978). *Freewill and responsibility*. Routledge.

Kharpal, A. (2015). Robot with $100 bitcoin buys drugs, gets arrested, Cnbc, , (http://www.cnbc.com/2015/04/21/robot-with-100-bitcoin-buys-drugs-gets-arrested.html) last accessed 30 March 2018.

King, T., Aggarwal, N., Taddeo, M., Floridi, L. (2018). Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions, . SSRN: https://ssrn.com/abstract=3183238 or https://doi.org/10.2139/ssrn.3183238.

Kinny, D., Georgeff, M., and Rao, A. (1996). A methodology and modelling technique for systems of BDI agents. European Workshop on Modelling Autonomous Agents in a Multi-Agent World. Springer.

Kinny, D, Georgeff, M, and Rao, A. (2017). Why things can hold rights: Reconceptualizing the legal person. In Legal personhood: Animals, artificial intelligence and the unborn. Springer.

Kurki, V. A. and Pietrzykowski, T. (2017). Legal personhood: Animals, artificial intelligence and the unborn. Springer.

Kurzweil, R., et al. (1990). *The age of intelligent machines*. Cambridge, MIT press.

Lepora, C., & Goodin, R. E. (2013). *On complicity and compromise*. Oxford University Press.

Litton, P. (2013). Criminal responsibility and psychopathy: Do psychopaths have a right to excuse?. Handbook on psychopathy and law, 275.

!2018 MEDIENGRUPPE BITNIK website (https://motherboard.vice.com/en_us/article/mgbwg4/in-europe-robots-can-legally-buy-drugs-online-for-art), last accessed 22 May 2018.

Miller, W. (1987). Annotation, modern status of rule of absolute or strict liability Dogbite, Animal Law Review, 51, 446.

Morse, S. J. (2008). Psychopathy and criminal responsibility. *Neuroethics, 1,* 205.

Murphy, J. G. (1979). Retribution, justice, and therapy: Essays in the philosophy of law. Springer.

Neumann, M. (2010). Norm internalisation in human and artificial intelligence. *Journal of Artificial Societies and Social Simulation, 13, 12*.

North, P. (2012). Civil liability for animals. Oxford University Press.

Ormerod, D., Smith, J. C., and Hogan, B. (2011). Smith and Hogan's criminal law. Oxford University Press.

P8_TA(2017)0051 (2015). Civil Law Rules on Robotics European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics.

Pagallo, U. (2013). The laws of robots, Springer.

Pagallo, U. (2017). AI and bad robots: The criminology of automation. In the Routledge Handbook of Technology, Crime and Justice. Routledge.

Poole, D., Mackworth, A., & Goebel, R. (1998). *Computational intelligence: A logical approach*. New York: Oxford University Press.

Postema, G. (2001). Law as command: The model of command in modern jurisprudence. *Philosophical Issues, 11*, 470.

Power, M. (2014). What happens when a software bot goes on a darknet shopping spree? (https://www.theguardian.com/technology/2014/dec/05/software-bot-darknet-shopping-spree-random-shopper) last accessed 30 March 2018.

Robinson, T. B. (1997). A question of intent: Aiding and abetting law and the rule of accomplice liability under section 924 (c). *Michigan Law Review, 96*, 783.

Russel, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach*. Prentice Hall.

Sartor, G. (2009). Cognitive automata and the law: Electronic contracting and the intentionality of software agents, in Artificial intelligence and law 17, 253.

Schaerer, E., Kelley, R., and Nicolescu, M. (2009). Robots as animals: A framework for liability and responsibility in human-robot interactions. In RO-MAN 2009. The 18th IEEE international symposium on Robot and human interactive communication. IEEE.

Schmitt, M. N., & Jeffrey, S. T. (2012). Out of the loop: Autonomous weapon systems and the law of armed conflict. *Harv. Nat'l Sec. J., 4*, 231.

Sermanet, P. Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks, arXiv preprint arXiv:1312.6229.

Shoham, Y. and Tennenholtz, M. (1992a). Emergent conventions in multi-agent systems: Initial experimental results and observations, In KR-92.

Shoham, Y. and Tennenholtz, M. (1992b). On the synthesis of useful social laws for artificial agent societies (preliminary report).

Shute, S. (2002). Knowledge and belief in the criminal law, criminal law theory: Doctrines of the general part. Oxford University Press.

Slobogin, C. (2003). The integrationist alternative to the insanity defense: Reflections on the exculpatory scope of mental illness in the wake of the Andrea Yates trial. *American Journal of Criminal Law, 30*, 315.

Solum, L. (1991). Legal personhood for artificial intelligences. *North Carolina Law Review, 70*, 1231.

Task force on the role of autonomy, (2011). DSB task force on the role of autonomy, 2011. The role of autonomy in DoD systems. US Defense Science Board (DSB).

Walden, C. (2017). State Dangerous Dog Laws, Animal Legal & Historical Center. Michigan State University.

Weyns, D., Steegmans, E., & Holvoet, T. (2004) Towards active perception in situated multi-agent systems. *Applied Artificial Intelligence, 18*, 867.

Wisch, R. F. (2012). Quick Overview of Dog Bite Strict Liability Statutes, Animal Law, available at: https://www.animallaw.info/article/brief-summary-dog-bite-laws.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.