Criminal Liability of Autonomous Agents: from the unthinkable to the plausible

Pedro Miguel Freitas[1], Francisco Andrade[1] and Paulo Novais[2]

[1] Law School, Universidade do Minho, Braga, Portugal
pfreitas@direito.uminho.pt
fandrade@direito.uminho.pt
[2] Informatics Department/CCTC, Universidade do Minho, Braga, Portugal
pjon@di.uminho.pt

Abstract. The evolution of information technologies have brought us to a point where we are confronted with the existence of agents - computational entities - which are able to act autonomously with little or no human intervention. And their behavior can damage individual or collective interests that are protected by criminal law.

Based on the analysis of different models of criminal responsibility of legal persons - which constituted an interesting advance in the criminal law in relation to what was hitherto traditionally accepted -, we will appraise whether the necessary legal elements to have direct criminal liability of artificial entities are present.

Keywords: Criminal liability • Software Agents • Autonomous agents • Objects • Legal Persons

1 Software Agents and Objects – an Introduction

An “agent” is a computational entity (software/hardware) that “being located in a defined environment acts upon it by autonomous actions, having a defined goal to accomplish” (Wooldridge and Jennings 1994). Thus being, the “agent” performs tasks on behalf of a user in a predefined computational environment, with little or no human intervention at all (Wong and Sycara 1999, 1). The agent is capable of analysing its environment and the problem data and of deciding accordingly, in an independent way (Durfee and Rosenschein 1994; Wooldridge 2009).

An important distinction must be considered between "agents" and "objects" (Wooldridge and Jennings 1994). Obviously, the degree of autonomy is much bigger in "agents" (Wooldridge and Jennings 1994). But also the definition of communication mechanisms and used language must be considered (Wooldridge and Jennings 1994). An object is an entity capable of storing an inner state, of using a set of methods acting upon that inner state and of communicating through messages (Durfee and Rosenschein 1994; Wooldridge 2009); the object has autonomy in the sense that it controls its own state but, contrarily to agents, is not capable of controlling its own behaviour (Durfee and Rosenschein 1994; Wooldridge 2009). Decision control centres are different in objects and in agents. And it can be said that objects have a static behaviour while "agents" have a dynamic behaviour (Brito and Neves 2000; Durfee and Rosenschein 1994; Wooldridge 2009). Another difference arises out of the definition of the dialog mechanism, more complex in "agents" than in objects (Jennings 1999; Wooldridge 2009). And it may be said that while both "agents" and objects have an identity, a state and a behaviour of its own, actually "agents" may be described in terms of a set of characteristics integrating knowledge, beliefs, desires, intentions, aims and even obligations (Fasli 2007; Georgeff *et al* 1999).

An "agent" is thus a program capable of acting in a flexible way, on behalf of its owner, user or client[1], in order to reach defined goals. So, it must present a set of properties or characteristics (Wooldridge 2002) such as autonomy (capacity of taking decisions on which actions to undertake without having to be constantly inquiring the user), reactivity (Weitzenboeck 2001, 4) (capacity of properly responding to prevailing circumstances in dynamic and unpredictable environments), proactivity (Weitzenboeck 2001, 4) (capacity of acting in anticipation of future goals), communication, cooperation and sociability (Weitzenboeck 2001, 4) and adaptive behaviour. This said, it must be stated that "agents" are not limited to data interchange (such as EDI – Electronic Data Interchange) but are capable of communicating in complex conversational environments and of assuming different roles, as well as adapting to diverse situations (Fasli 2007, 59).

Autonomy is one of the most relevant features of "software agents", implying the possibility of acting and performing tasks without any human intervention. A "software agent" is independent and acts autonomously, having control both of its inner state and of its behaviour, being capable of clearly understanding the goals of its mission and of defining a strategy in order to reach the defined goals. Of course, the levels of autonomy may greatly vary (Russel and Norvig 1995, 35) and although the "software agent" may decide autonomously, without any human intervention, the user may have more or less capacity of controlling the parameters influencing the behaviour of the agent (Chavez and Maes 1996, 8). But the greater autonomy of new generations of "software agents", capable not only of acting within pre-established parameters but also of having initiative and deciding, by themselves, what, when and how to do, upon favourable conditions (in the perspective of the "software agent"!) may force us to distinguish the situations in which the user will still have some control upon the

[1] Obviously, we are not considering a legal framework (which does not exist) for the actuation of software agents.

strategy to be followed by the "software agent" or at least upon the main parameters of decision (Chavez and Maes 1996, 8) from the cases when this control will be totally lost and only the trust (or lack of trust) of the user in the "software agent" capabilities remains. And we may even have to face the possibility of the "software agent", reasoning upon the available data, overcoming what the user may reasonably have foreseen (Dowling 2000, 3).

Software agents are not considered as persons. Yet, they have this capacity of autonomous acting and their acting may well modify the legal position of legal persons. Furthermore, it may be considered that software agents have something more or less equivalent to a "will" or at least what may be called "intentional states" (Sartor 2003, 23-51; Sartor 2009, 253-290).

The intentionality of software agents brings along the issue of the legal consideration of the acts of software agents (Andrade *et al* 2007). For the moment being, software agents are not considered as legal persons and the most plausible solution for the consideration of their legal acts is the one suggested by Giovanni Sartor of having commercial corporations specially created for the use of software agents (2002). Thus being, liability for the acts of software agents would impend on commercial corporations (Wettig 2003; Wettig and Zehendner 2004). This will force us, when concerning criminal responsibilities, to analyse the issue of the consideration of criminal liability of the corporations.

2 Criminal Liability of Legal Persons

In fact, one of the most troublesome questions that criminal law is currently facing is the criminal liability of corporations or legal persons. Should criminal penalties be solely imposed upon an individual or should also a legal person be subjected to those penalties and, being so, in what way would that occur?

It is a rather old question but still widely disputed in the context of criminal policy and criminal law, for which, to use an example that is closer to us, only in 2007, with several amendments to the Criminal Code, the Portuguese legislator gave a pragmatic and definitive answer.

Despite that, the replacement of the old principle of Roman law *societas delinquere non potest* has been gradually accepted in countries of Anglo-Saxon legal tradition, such as the United States, which is believed to be one of the first countries to do it, or the UK and in many countries of different legal traditions.

Such a solution – the criminal liability of legal persons – appears to be not only essential for a timely and adequate response of criminal law to an increasingly complex human society, but it also meets the requirements imposed by various international bodies such as the European Union, the Council of Europe and the United Nations, which require States to adopt the necessary legislative measures in order to sanction legal persons for acts that constitute certain offenses.

However, despite this demand for the accountability of legal persons, there is no consensus on the actual manner this should be done. Many different models of responsibility of legal persons exist, and they range from mere tort liability to criminal liability.

The specific reason why there are doubts whether criminal law, a body of law which, as we all know, should only be applied as last resort, when all legal remedies are insufficient, should apply, has to do with two fundamental concepts in criminal law theory: agency/conduct (the common law *actus reus* or the german *Handlung*) and blameworthiness (*mens rea* or Schuld[2]). These are the core challenges necessary to overcome in order to legitimate criminal liability of legal entities.

On the one hand, some authors, including most German authors, defend that the notion of action in the criminal law framework demonstrates that legal entities are not able to act for themselves. Only the natural or physical persons may carry out behaviours that are criminally relevant. And as such, criminal responsibility cannot fall on a legal entity, but rather on the individual (Correia 2010, 234; Martín 1996, 43-45).

On the other hand, many say that it is impossible to morally and ethically judge legal entities for not acting lawfully (despite having the opportunity to abide by the law), due to the fact that blameworthiness for an unlawful action demands the existence of an agent that has free and conscious will and chooses to break the law in an hypothesis where he/she could and should have acted differently (Castro e Sousa 1985, 114) - in this sense only individuals possess "personal qualities necessary to be censured for not acting differently"(Castro e Sousa 1985, 114). Therefore, the conclusion should be obvious: regarding the lack of ontological unfitness to be blamed, legal entities cannot be held criminally accountable (Dias 2012, 196). Accordingly, only the individuals that have committed the relevant criminal acts on behalf of legal entities or in their interest can suffer criminal sanctions, and not the legal entities themselves.

However, one must not neglect that we live in a rapidly evolving society characterized by the discourse of the global risk society (Beck 1992), which entails a profound paradigm shift in our cultural, economic, sociological and technological dimensions as a community, and brings paramount changes to the way criminality materializes. There is an increasing criminality that involves a greater complexity and organization, frequently having corporations, societies and associations as key actors. Thus, it seems accurate the idea expressed by Figueiredo Dias: if we chose to only prosecute and punish physical or biological persons acting on behalf of legal entities, completely waiving criminal accountability of the latter, that would mean that (given the degree of complexity, not only of the committed crimes, especially those against the economy, but also of these legal entities' organizational structures) it would be im-

[2] Despite not being totally equivalent, the English legal term *mens rea* and the german legal term Schuld are, for simplicity reasons, treated as functional equivalents for the purpose of this paper. It should be noted however that depending on the perspective that one assumes on the concept of crime, more especifically, whether it is a classic approach (of such authors as Liszt, Beling and Berner), a neoclassical approach (Mezger), a teleological theory (Welzel) or a functional-teleological and rational system (Schünemann and Roxin), the translation of *mens rea* to german can be *subjektiver Tatbestand* and/or *Schuld*.

possible to specifically determine the individuals that should be held responsible. And so there would be absolute impunity (Dias 1983, 51).

Thus, assuming that there is a need for a real and autonomous criminal responsibility of legal persons, how can we overcome these dogmatic obstacles upheld by the traditional thinking of criminal law?

The answers vary. In the Portuguese legal order, Figueiredo Dias rejects the arguments of inability of agency and blameworthiness of legal entities and considers adequate the implementation of a so-called *analogic model* (2012, 198). According to this Author, "the individuals can be replaced, as criminally responsible, both objectively and subjectively, ethical and social hubs, by their collective work and materialization, such as legal entities, associations, groupings or corporations, in which free beings express themselves" (2012, 198).

Faria Costa, on the other hand, although not recognizing in his first writings the criminal liability of legal entities (1981, 43), later admits the plausibility of their punishment in light of the theory he coined as material rationality of opposite places (1992, 537). The legitimacy of this type of criminal liability is based on a material analogy between the behaviour of natural persons and legal persons: if under 16, natural persons, although having capability to act, are exempt from criminal responsibility, as stated by article 19 of the Portuguese Criminal Code, then it would not be totally unreasonable to accept the punishment of legal persons despite not being physically and anthropologically capable of acting. According to Faria Costa, the criminal justice system, through "axioms developed by criminal dogmatics" (1992, 551), constructs "a space of normativity whose essential feature is represented by the absence of a particular characteristic" (1992, 552). This space of normativity can "enlighten and justify, in terms of material rationality, its opposite place" (1992, 553). And so, if with the infancy defence "we have the curtailment of ontological segments of action, here, inversely, there is an extension of a communicational act, criminally relevant; if with infants we limit and remove blameworthiness, here, inversely, the notion blame is reconstructed and the legal person becomes a true centre of imputation" (1992, 553).

These theoretical solutions, of greater expressiveness in Portugal, are obviously not exclusive. Examining comparative law, particularly civil law (continental law) countries, we observe that this topic has been debated to exhaustion and to the same exhaustion answers have been offered which aim to support and implement the notion of blameworthiness of legal entities (Dias 2012, 299): imputation model – *Zurechnungsmodell* – according to which guilt and action of the responsible corporation boards are imputed to the legal person; model of the culpability of the organization – *Modell des Organisationsverschuldens* (Tiedemann 1988) – that recognizes the existence of a specific and autonomous blameworthiness of the legal entity, which derives from the idea that the legal entity provides a favourable environment for the practice of certain crimes; model of prevention – *Präventionsmodel* (Schünemann, 1994) – which acknowledges the possibility of sanctioning legal entities with *security measures*; and, finally, the model of analogue blameworthiness (Heine 1995, 241), where an analogue imputation of blameworthiness is shed on the legal entity, having as a criteria of criminal imputation an appraisal of the way business was carried out (*Betriebsführungsschuld*).

From the point of view of Common Law (Ormerod 2008, 247), we reach the conclusion that opinions are mainly divided between the doctrine of identification and the vicarious liability (or agency doctrine)[3]. The first theory sets up an overlap between the conduct and blameworthiness of individuals in positions of leadership and the conduct and blameworthiness of the legal entity. In other words, those individuals represent the "body" and "mind" of the legal entity and therefore the acts carried out by them must be regarded as being done by the legal entity itself. Meanwhile, the advocates of the second theory stress the liability of the legal entity for the conduct of its agents, meaning that the legal entity is charged with criminal responsibility for the actions of agents such as directors, supervisors, etc.

3 Models of Criminal Liability of Autonomous Agents: an Appraisal

Having outlined the main points of interest on criminal liability of legal entities, a few questions arise: Are there such substantial differences between legal entities and autonomous agents that justify the exemption from criminal responsibility of the latter? Is it plausible to conceive criminal responsibility of autonomous agents or AI entities?

Traditionally AI entities are considered not to have legal personhood. They are said to be mere objects[4]. And this is may be the *punctum crucis* of this question. Throughout the different branches of law (civil law, administrative law, etc.), and in particular criminal law, there is one key distinction that is commonly made between subject (or agent) and object. According to George P. Fletcher, "[a] subject is someone who acts, and an object is someone or something that is acted upon" (1998, 43). Although simple in its wording, it encompasses complex issues, which we face namely when addressing "software agents" or "artificial entities".

The criminal liability of artificial entities has been a rather unknown territory for legal scholars[5]. There are a few exceptions however.

---

[3] There are other approaches to corporate liability, such as the aggregation theory (also termed as collective knowledge doctrine), the culpable corporate culture, the reactive corporate fault.

[4] Here we use the common notion of objects and not "objects" in the sense we referred in the beginning of this paper.

[5] We can find an interesting account on this topic on LEGAL-IST Consortium's Report on Legal Issues of Software Agents, Doc. No. D14, Rev. No. 2, 29 March 2006, which for liability purposes drafts a fruitful analogy between software agents (owner of certain cognitive capabilities and mental states) and trained dogs, coined by the authors as the dog model. This model starts by assuming that both software agent and trained dog are programmed to autonomously pursue assigned tasks and goals. Depending on the direction and level of training/programming, the dog's and agent's cannot be completely foreseen in advance,

Gabriel Hallevy has proposed three models of the criminal liability of artificial intelligence entities: Perpetration-via-Another Liability Model; Natural-Probable-Consequence Liability Model; Direct Liability Model (2010a, 2010b, 2010c, 2012). They present a sound foundation on which this topic could be further developed and, as such, we should therefore understand the main characteristics of each model.

The first model considers that the AI does not possess any human attribute and so denies the possibility of having the AI as a perpetrator of an offense. It is seen as akin to mentally limited persons, such as a child, a person who is mentally incompetent or one who lacks a criminal state of mind. The AI entity is an innocent agent that is a mere instrument used by the real perpetrator, who architects the offense and constitutes the real mastermind behind it. As such, the person behind the AI is to be held accountable for the conduct (*actus reus*) of the AI, albeit the subjective or internal element (*mens rea*) is determined by the perpetrator-via-another's mental state.

The perpetrator-via-another can either be the programmer or the user: the programmer, when he designs an AI entity with the purpose of committing criminal offenses, or the user (end-user), that, albeit not designing the AI entity, is in control of it, and uses it to commit offenses.

It should be noted that this model assumes that the AI is completely dependent on either the programmer or the user. It is not self-ruling or self-determining, but solely an instrument (equivalent to a hammer or even a dog used for illicit purposes) for which no specific mental state is required, *e.g.* a programmer creates an AI entity to destroy computer data.

Accordingly, this model would not be implemented in hypotheses where the AI entity decides to commit an offense based on its own accumulated experience or knowledge; commits an offense despite not being programmed to do so; acts as a semi-innocent agent[6].

The second model – coined Natural-Probable-Consequence Liability Model – presupposes that the programmer or user of the AI entity, despite not programming or using it for the purpose of committing a certain crime, might be held accountable for the crime committed by the AI entity, if the offense is a natural and probable consequence of the AI's conduct. Even though the programmer or user was not aware that the offense was committed until it had already been committed, did not plan to commit any offense and did not take part in the commission of the offense, if there is evi-

---

which in turn can lead to unwanted results or even illicit. Disregarding the possibility of holding the AI (Artificial Intelligence) entity directly liable for its actions, there could be criminal responsibility of the developer or user (or in dog's case, the trainer or the owner) for their negligence, imprudence or unskillfulness – this is in essence what is described in the Natural-Probable-Consequence Liability Model. In the hypothesis of wilful misconduct by the trainer/owner/developer/user, criminal liability would rest upon the subject to whom the fact can be led back.

[6] Hallevy, Gabriel: The Matrix of Derivative Criminal Liability, p. 38. Springer, Heidelberg (2012), describes a semi-innocent agent as "a negligent party that is not fully aware of the factual situation while any other reasonable person could have been aware of it under the same circumstances".

dence that they could and should foresee the potential commission of offenses, then they might be prosecuted for the offense.

So, this model does not require the criminal intention of the programmers or the users, as the first model does, but only their negligence, which is criminally relevant due to the fact that a diligent and reasonable programmer and user should be able to foresee the offense and prevent it from happening[7], *e.g.* a programmer sets up an AI entity to protect a computer system and the latter decides, as part of its mission, to seriously hinder a computer system which it considers a potential threat.

Finally, the Direct Liability Model – the third and last model – aims at providing a theoretical framework for a functional equivalence between AI entities and humans for criminal liability purposes. For this reason, this model deserves greater attention in our analysis, as it constitutes the main focus of our paper. Gabriel Hallevy's reasoning stems from the idea that criminal liability implicates solely the fulfilment of two different requirements: *actus reus* (external element) and *mens rea* (internal element) and if AI entities were able to fulfil them both then criminal accountability would follow.

We have no doubt that if such liability of AI entities were to exist, it should not replace the programmer or user's liability. Both could co-exist, if all the legal requirements were fulfilled, meaning that the criminal liability of AI entities would not exclude the individual responsibility of programmer or users nor would it depend on the criminal accountability of those – similar to what is commonly done when punishing legal entities, where criminal punishment of the individuals behind the legal entity does not constitute a requirement to have the criminal punishment of legal entities themselves. But the problem remains: do AI entities fulfil all necessary requirements to trigger criminal liability?

On one side, regarding the *actus reus* requirement, it is insufficient to propose its fulfilment only when AI entities control a mechanical or other mechanism to move its moving parts (*e.g.* robots). In our view, this argument should clearly be regarded as unbearably limited. If we were to establish the criminal liability of AI entities, why should those be solely responsible when it is proved that they controlled mechanical instruments or others of the same sort? It seems to be nothing more than an unjustifiable overlap between AI entities and robots. The former, as we know, is not the same as the latter. One example that clearly shows that this confusion between terms can lead to unjust results has to do with computer offenses. Let us imagine that the AI

[7] Under this model, Gabriel Hallevy devises two situations that bring different outcomes. The first situation is when programmers or users did not want to commit any offense but negligently programmed or used the AI entity and an offense occurred. In this hypothesis, programmer and user should be held accountable for an offence, as long as there is a negligent offense stated by criminal law for that type of cases. The second situation deals with accomplice liability cases, namely when programmers or users programmed or used the AI to commit one offense, but the latter committed another, in addition or instead of the planned one. The author proposes the punishment of the programmer or the user as if they acted with knowledge and intent. Alongside the criminal liability of the programmer or the user, the AI entity, provided that did not act as an innocent agent, could be directly held liability for its actions.

entity, merely software, intentionally decides to target a computer system with a denial-of-service attack (DoS attack). Shouldn't the AI entity be held criminally responsible here as well?

To perceive the fulfilment of the *actus reus* requirement as having willed muscular movement (in this case, mechanical) or bodily movement is to ignore that there are crimes without *actus reus* or *acts* in a traditional sense – *e.g.* computer crimes. Unless we consider that the physical act in computer crimes resides in electronic impulses – which seems to be a far-fetched and unnecessary argument –, to suggest that *actus reus* equals the traditional definition of act is inadmissible. As Figueiredo Dias (2012, 240) and David Ormerod (2008, 51) remind us, it is misleading or even strange to say that, for example, in the crime of defamation the relevant act corresponds to the movement of one's tongue, mouth and vocal chords. For these reasons, the traditional view of acts as willed voluntary movements is seen, in recent years, as outdated (Herring 2012, 106)[8].

More importantly we should emphasize the fact that in order to occur the criminal liability of an agent, the conduct proscribed by a certain crime must be done voluntarily. What this actually means it is something yet to achieve consensus, as concepts as consciousness, will, voluntariness and control are often bungled and lost between arguments of philosophy, psychology and neurology, leading the judiciary and legal scholars alike to prefer stating the cases where there is not a voluntary act (Hamilton 2011; Saunders 1987-1988, 447). In these cases, as Jonathan Herring affirms, "an involuntary action is one for which not only is the defendant not responsible, it is not even properly described as *his* act" (2012, 105). So, the voluntariness requirement serves the purpose of excluding from criminal liability those acts that are mere automatisms (Ormerod 2008, 55; Dias 2012, 305) or done unconsciously. This fact makes clear that AI entities should only be made criminally accountable if they voluntarily acted, which means that must be an act done with will, volition or control. Accordingly, we cannot say that an AI entity voluntarily acted if the presence of one of these internal elements, depending on what particular theory one follows on the characterization of the "voluntarily" concept, is not found in a certain situation. While these elements describe a certain internal state of the agent, they should not however be confused with *mens rea* (Saunders 1987-1988, 443-445). There can be volition without *mens rea*, but the contrary is not true[9]. Thus, before turning to a closer insight on *mens rea*, it becomes necessary to call volition (or will or control) into question. While we may find easy to note that volition and human acts generally appear hand in hand, and so in the acts of legal entities, to plunge into the same conclusion as to AI entities' acts would arguably be precipitated.

Additionally, criminal courts and legal scholars demand the existence of a human action, which means that this voluntary act, whatever it may be, must be carried out

[8] When it comes to punish an absence of behaviour (omission) it must be proved that there was a duty to act and the agent failed to perform such a duty.

[9] Saunders (1997-1988) gives the example of the athlete who, during an athletic competition, throws a javelin, after being sure that no person was in his path, but a bystander is hit by the javelin and dies. Despite not having *mens rea* in causing the death of the bystander, there is a voluntary act which consists in throwing the javelin.

by humans and not inanimate objects or animals. This, for us, shows that voluntariness being expressed as a requirement is deeply tangled with demanding human agency. But, as we stated previously, human agency is no longer an absolute and unsurpassable criteria: legal entities are now criminally liable for certain offenses – which could open the path for having criminal responsibility of AI entities.

Finally, recognizing *mens rea* of AI entities can pose a difficult challenge to overcome. There is first a matter of determining the specific level of development of a particular AI entity. Not all AI entities bear the same capabilities, *e.g.*, cognitive skills and abilities, and this should be reflected on whether *mens rea* can be attributed to an AI entity. Secondly, a certain state of mind, which differs from one crime to another, must be attributed to the accused. Some Authors remind us that the only mental requirements needed to impose criminal liability are knowledge, intent, negligence, among others, and peremptorily affirm that knowledge and specific intent can be attributed to AI entities when these have sensory receptors of factual data, which in turn is analysed by the AI entity (Hallevy 2010, 188). Even if AI have sensors which provide them with data that could be processed internally, can we say that the AI entity understands or comprehends what is being processed? This would lead us to the highly controversial "Chinese Room Argument" of John Searle, which is the subject of a never-ending debate with inconclusive results.

Additionally there is the problem that predicates on determining blameworthiness of AI entities. *Mens rea* can be referred to in its general sense or in its special sense (Kadish and Schulhofer 2001, 203; Herring 2012, 134; Wilson 2008, 116). To demand the presence of a certain mental state in the agent, which is described by the offense, is to demand *mens rea* in its special sense. But this is not sufficient. Criminal law must ensure that there is only punishment when the agent is at fault (Jefferson 2008, 86). So we must pose the question: can there be any blameworthiness in AI entities' actions that enables their legal punishment?

Criminal conviction encompasses a censure (Herring 2011, 67) of the agent for acting in a certain fashion. And this relates to the element of guilt/blame/*Schuld* that has to be present. Guilt or *Schuld* is seen, by some Authors (*e.g.* Kaufmann), as censuring someone for acting unlawfully when he could have acted differently; or for acting unlawfully as a result of not promoting a law abiding character or personality (e.g. Mezger). But blameworthiness supposes a free being – with conscious and free will (Dias 2012, 279) – that has a choice in determining his essence. Although criminal law was used, until late eighteenth century, to punish animals for crimes such as homicide and theft (Wilson 2008, 117), it seems now that invoking criminal law for these cases is, in light of the reasons behind criminal punishment – either retribution, deterrence, rehabilitation or restoration, rather useless and unjust[10]. But as far as science goes, animals lack this ability to become cognizant and influence the "self", at least at

[10] There are however recent studies that challenge the traditional deterministic view of animal behaviour - Brembs, Björn: Towards a scientific concept of free will as a biological trait: spontaneous actions and decision-making in invertebrates. Proc. R. Soc. B, vol. 278, no. 1707, 930-939 (2011). And those who proclaim the idea that animals share with humans the possession of neurological substrates that generate consciousness, see The Cambridge Declaration on Consciousness, July 7, 2012.

the same level humans do (Morris 2006). On the other hand, remembering what was stated above on criminal liability of legal entities, there is a theory that could well be called into action: the analogic model (Dias 2012, 298). Individuals, or biological people, are free beings that, for criminal purposes, can and should be replaced by their work - as ethical and social cores that too are "products of freedom" or "materialization of free beings" (Dias 2012, 298). Provided that AI entities have self-awareness, self-consciousness, free and conscious will, ability to apprehend the (un)lawfulness of their behaviour and means to guide themselves by law, the minimum requirements to call forth their blameworthiness and, hence their criminal responsibility, are present, since they too - AI entities - could embody social and ethical cores, as they are human creations, either directly or indirectly. As a result, in this hypothesis, we reach the dogmatic, juridical and technological apparatus to enable AI entities as active legal actors in criminal justice.

4 Conclusion

The criminal liability of legal person persons has constituted an innovative breakthrough in criminal law and the models used to support such an advance can provide us with invaluable clues to unveil a plausible dogmatic framework for the criminal responsibility of artificial entities. But more importantly, it demonstrates a certain degree of flexibility shown by criminal law when criminal policy demands so.

A flexibility that can be used provided that certain dogmatic premises are met, to justify the punishment of AI entities. The question then will not be anymore whether "can we do it?" but "should we?", "why?" and "how"?

Relying on previous studies put forwarded by Reynolds and Ishikawa, Ugo Pagallo considers three examples of criminal robots (Pagallo 2011, 311-313): Picciotto Roboto[11]; Robot Kleptomaniac[12] and Robot Falsifier[13], and then points out that today's state-of-the-art in technology is not capable of producing a "Robot Kleptomaniac". It may be so. Legal personality and criminal accountability of AI entities may be no-

[11] The Picciotto Robot hypothesis deals with a robot security guard, deprived of free will or moral sense, which is used by a gang to carry out criminal enterprises. Reynolds and Ishikawa conclude: "As such, it seems that the robot is just an instrument just as factory which produces illegal products might be. The robot in this case should not be arrested, but perhaps impounded and auctioned." Reynolds, C., & Ishikawa, M.: Robotic thugs. Ethicomp Proceedings, Global e-SCM Research Center & Meiji University, pp. 487–492 (2007).

[12] The Robot Kleptomaniac has free will and self-chosen goals and, when confronted with a fixed supply of energy that is running low, chooses to rob batteries from a local convenience store.

[13] The Robot Falsifier example creates awareness for the fact that the Legal Tender project claimed that viewers could remotely operate a robotic system to physically alter purportedly authentic money.

where soon. But, living in an ever-evolving world as we do, means that the notion of fully autonomous AI entities or robots is not totally unthinkable, either in battlefields or in our civil life. This argument surely gives grounds to further legal and technical investigation on this topic.

Acknowledgments

This work is part-funded by CROWDSOURCING project (Reference: DER2012-39492-C02-01).

References

Andrade, Francisco, Novais, Paulo, Machado, José, Neves, José: Contracting Agents: legal personality and representation. Artificial Intelligence and Law. 15, 357-373 (2007)

Beck, Ulrich: Risk society: towards a new modernity. Sage Publications, London (1992)

Brembs, Björn: Towards a scientific concept of free will as a biological trait: spontaneous actions and decision-making in invertebrates. Proc. R. Soc. B, vol. 278, no. 1707, 930-939 (2011)

Brito, Luís & Neves, José: A execução paralela em sistemas multiagente: comunicação, distribuição, coordenação e coligação. Vol. I, p. 4-5. Universidade do Minho, Braga, Portugal (2000)

Castro e Sousa, João: As pessoas colectivas em face do direito criminal e do chamado direito de mera ordenação social. Coimbra Editora, Portugal (1985)

Chavez, A. & Maes, P.: Kasbah: an agent marketplace for buying and selling good. AAAI Technical Report. 8-12 (1996)

Correia, Eduardo: Direito Criminal, I. Almedina, Portugal (2010)

Costa, Faria: Aspectos fundamentais da problemática da responsabilidade objectiva no direito penal português, Coimbra Editora, Portugal (1981)

Costa, Faria: Responsabilidade jurídico-penal da empresa e dos seus órgãos (ou uma reflexão sobre a alteridade nas pessoas colectivas, à luz do direito penal). Revista Portuguesa de Ciência Criminal, Ano 2, no. 4, 537-559 (1992)

Dias, Figueiredo: Direito Penal, Parte Geral, vol. I, 2nd ed. Coimbra Editora, Portugal (2012)

Dias, Figueiredo: Pressupostos da Punição e Causas que Excluem a Ilicitude e a Culpa. In: Centro de Estudos Judiciários (org.): Jornadas de Direito Criminal, I, 41-83, Lisbon (1983)

Dowling, C.: Intelligent agents: some ethical issues and dilemmas. In: Proceedings of 2nd Australian Institute of Computer Ethics Conference (AICE2000), http://crpit.com/confpapers/CRPITV1Dowling.pdf, 1-5 (2001)

Durfee, Edmund H. & Rosenschein, Jeffrey: Distributed Problem Solving and Multi-Agent Systems: Comparisons and Examples. Proceedings of the International on Distributed Artificial Intelligence, 1-10. Seattle, USA (1994)

Fasli, Maria: Agent Technology for e-commerce. John Wiley and Sons Ltd, England (2007)

Fletcher, George P.: Basic Concepts of Criminal Law, p. 43. Oxford University Press, Oxford (1998)

Georgeff, Michael, Pell, Barney, Pollack, Martha, Tambe, Milind and Wooldridge, Michael: Belief-Desire-Intention Model of Agency in Intelligent Agents V: Agents Theories, Architectures, and Languages, LNCS Volume 1555, 1-10 (1999)

Hallevy, Gabriel: I, Robot - I, Criminal – When Science Fiction Becomes Reality: Legal Liability of AI Robots Committing Criminal Offenses. Syracuse Sci. & Tech. L. Rep., vol. 22, article 1, 1-37 (2010a)

Hallevy, Gabriel: The Criminal Liability of Artificial Intelligence Entities – from Science Fiction to Legal Social Control. Akron Intellectual Property Journal, vol. 4, issue 2, 171-201 (2010b)

Hallevy, Gabriel: The Matrix of Derivative Criminal Liability. Springer, Heidelberg (2012)

Hallevy, Gabriel: Unmanned Vehicles – Subordination to Criminal Law under the Modern Concept of Criminal Liability. J. of Law, Info. & Sci., vol. 21, n. 2, 200-211 (2012)

Hallevy, Gabriel: Virtual Criminal Responsibility. Original Law Review, vol. 6, no. 1, 6-27 (2010c)

Hamilton, Melissa: Reinvigorating Actus Reus: The Case for Involuntary Actions by Veterans with Post-Traumatic Stress Disorder. Berkeley Journal of Criminal Law, vol. 16, iss. 2, art. 2, 346-390 (2011)

Heine, Gunter. Die strafrechtliche Verantwortung von Unternehmen. Nomos, Germany (1995)

Herring, Jonathan: Criminal Law: Text, Cases, and Materials, 5 ed. Oxford University Press, Oxford (2012)

Jefferson, Michael: Criminal Law, 8th ed. Longman, England (2008)

Jennings, Nicholas R.: Agent-Oriented Software Engineering. MAAMAW. 1-24 (1999)

Kadish, Sanford H., Schulhofer, Stephen J.: Criminal Law and Its Processes: Cases and Materials, 7th ed. Aspen Publishers, Inc., New York, USA (2001)

Martín, Gracia: La responsabilidad penal de la propias personas jurídicas. In Puig, Mir & Peña, Luzón (eds.): Responsabilidad penal de las empresas y sus órganos y responsabilidad por el producto, 35-74. J. M. Bosch, Editor, Barcelona, Spain (1996)

Morin, Alain: Levels of consciousness and self-awareness: A comparison and integration of various neurocognitive views. Consciousness and Cognition, vol. 15, no. 2, 358–371 (2006)

Ormerod, David: Smith and Hogan, Criminal Law, 12.ª ed., Oxford University Press, Oxford (2008)

Pagallo, Ugo: Robots of Just War: A Legal Perspective. Philos. Technol. 24, 307-323 (2011)

Reynolds, C., & Ishikawa, M.: Robotic thugs. Ethicomp Proceedings, Global e-SCM Research Center & Meiji University, 487–492 (2007)

Russel, S.J., Norvig, P.: Artificial Intelligence: a modern approach. Prentice Hall, USA (1995)

Sartor, Giovanni: Agents in Cyberlaw. Proceedings of the workshop on the law of electronic agents – LEA 2002 (2002)

Sartor, Giovanni: Cognitive automata and the law: electronic contracting and the intentionality of software agents. Artificial Intelligence and Law. 17, 253-290 (2009)

Sartor, Giovanni: L'intenzionalitá dei sistemi informatici e il diritto. Rivista Trimestrale di Diritto e Procedura Civile, Anno LVII. 23-51 (2003)

Saunders, Kevin W.: Voluntary Acts and the Criminal Law: Justifying Culpability Based on the Existence of Volition. U. Pitt. L. Rev., vol. 49, 443-476 (1987-1988)

Schünemann, Bernd: Die Strafbarkeit der juristischen Person aus deutscher und europäischer Sicht. In Schünemann, Bernd & Gonzalez, Carlos Suarez (eds.): Bausteine des europäischen Wirtschaftsstrafrechts. Madrid-Symposium für Klaus Tiedemann, 265-295. Heymanns, Berlin, Germany (1994)

Tiedemann, Klaus: Die "Bebußung" von Unternehmen nach dem 2. Gesetz zur Bekämpfung der Wirtschaftskriminalität. Neue Juristische Wochenschrift, 1169-1232 (1988)

Ugo Pagallo: The Laws of Robots. Springer, Heidelberg (2013)

Wettig, Steffen & Zehendner, Eberhard: A legal analysis of human and electronic agents. Artificial Intelligence and Law. 12, n. 1-2, 111-135 (2004)

Wettig, Steffen & Zehendner, Eberhard: The electronic agent: a legal personality under German law?. Proceedings of 2nd workshop on law an electronic agents – LEA 2003, p. 57-112 (2003)

Weitzenboeck, Emily: Electronic agents and the formation of contracts. International Journal of Law and Information Technology, vol. 9, no. 3, 204-234 (2001)

Wilson, William: Criminal Law, 3th ed. Longman, England (2008)

Wong, H. Chi & Sycara, Katia: Adding Security and Trust to Multi-Agent Systems. Proceedings of Autonomous Agents'99 - Workshop on Deception, Fraud and Trust in Agent Societies. 1-13 (1999)

Wooldridge, Michael: An Introduction to MultiAgent Systems, 2. ed. Wiley, UK (2009)

Wooldridge, Michael & Jennings, Nicholas R.: Agent Theories, Architectures, and Languages: A Survey. ECAI Workshop on Agent Theories, Architectures, and Languages. 1-39 (1994)