Metacognition, Accountability and Legal Personhood of AI

Beatriz A. Ribeiro, Helder Coelho, Ana Elisabete Ferreira, and João Branquinho

Abstract One of the puzzles yet to be solved regarding Artificial Intelligence (AI) is whether or not robots can be considered accountable and have, eventually, legal personhood. With inputs from Philosophy, Psychology, Computation and Law, the paper proposes an interdisciplinary approach to the question of legal personhood in AI. In this paper, we examine, firstly, the concepts of Object (a mere tool) and Agent, in order to understand in which category AI may belong to. Secondly, we analyze how Metacognition, broadly defined as the cognition about cognition, which results in mental processes that control an entity's thoughts and behavior, can be applied to law as a minimum requirement for accountability. For instance, we shall see that both children and people with mental diseases, besides being two categories of subjects that have a very restricted legal capacity, also show some limitations when it comes to Metacognition. In other words, we argue that the main difference between a non-responsible and a responsible Agent depends on the metacognitive processes that can be carried out by the entity. Ultimately, we discuss how to transpose this idea to AI, debating the possible terms of legal personhood of AI.

B. A. Ribeiro (✉)
Vieira de Almeida & Associados, Communications & Digital, Lisbon, Portugal

H. Coelho
University of Lisbon, LASIGE, Computation Department, Lisbon, Portugal
e-mail: hmcoelho@fc.ul.pt

A. E. Ferreira
University of Coimbra, Department of Law, Coimbra, Portugal

J. Branquinho
University of Lisbon, Philosophy Department, Lisbon, Portugal
e-mail: jbranquinho@campus.ul.pt

H. Sousa Antunes et al. (eds.), *Multidisciplinary Perspectives on Artificial Intelligence and the Law*, Law, Governance and Technology Series 58,
https://doi.org/10.1007/978-3-031-41264-6_9

1 Introduction

One of the puzzles yet to be solved regarding Artificial Intelligence (AI) is whether or not robots can be considered accountable and have, eventually, legal personhood. With inputs from Philosophy, Psychology, Computation and Law, the paper proposes an interdisciplinary approach to the question of legal personhood in AI. In this paper, we examine, firstly, the concepts of Object (a mere tool, not subject to legal personhood) and Agent, in order to understand in which category AI may belong to.

Secondly, as the concept of Agent presents many difficulties, namely because it seems to have a different meaning according to each of the above mentioned domains of knowledge, a common denominator was identified, which it was found to be the voluntary act. If there is a voluntary act, we must, then, conclude that we have an Agent before us. Accordingly, and as long as AI acts voluntarily, it makes sense to argue that complex robots (in the sense of strong AI) are Agents, thus not mere tools.

Thirdly, since children, animals and people with mental illnesses act voluntarily but are still not held accountable (either have no legal personhood or limited exercise of such personhood), the paper investigates what is missing in these cases, in order to draw a line between accountable and non-accountable agents.

At last, we analyze how Metacognition, a concept borrowed from Psychology, which is broadly defined as the cognition about cognition, resulting in mental processes that control an entity's thoughts and behavior, can be applied to law as a minimum requirement for accountability and eventually legal personhood. For instance, we shall see that both children and people with mental diseases, besides being two categories of subjects that have a very restricted legal capacity, also show some limitations when it comes to Metacognition. In other words, we argue that the main difference between a responsible and non-responsible Agent depends on the metacognitive processes that can be carried out by the entity. Ultimately, we discuss how to transpose this idea to AI, debating the possible terms of legal personhood of AI.

There's no doubt that the Law depends, to a certain extent, on the description and classification of the problem (Birks 1997) we have before us. In other words, when confronted with a given situation, we are forced to list its essential features and see if those features match the legal norm. If it does, we have found ourselves a legal solution for the problem; if not, we must keep searching for a match.

When it comes to legal personhood, there are some basic requisites which, in absence, rule out any chance of even considering ascribing it to a certain entity. For instance, no one thinks about describing a deceased person as a legal person, though some rights might be extendable after death (such as right to honor). Legal personhood regarding human beings implies being alive, as this *status* begins when we are born. Whenever something doesn't quite fit the categories that we, humans, created, for instance if we're somewhat alive and not yet born (the unborn child), it becomes unclear for us what must be done regarding that entity.

In this sense, it has been argued (Boulangé and Jaggie 2014) that the first step in order to build a legal framework, in the case of any sort of robots, is to determine its *status*, meaning define its concept and boundaries and then confront it with the available legal options. In this regard, Pagallo (2013) developed extensive work on understanding the main traits of each type of robot that is planned in the near future, in his book *The Laws of Robots: Crimes, Contracts, and Torts.*

Globally, the author divides the possibilities into three categories: (1) Legal Person, (2) Proper Agent and (3) Source of Damage. What it means, in practice, is that we must check whether a given robot shares sufficient attributes with human beings, therefore leading us to grant it Legal Personhood (Hypothesis 1). If it has much more similarities, meaning more features in common, with the concept of tool, thus being considered a mere object, then the answer is to treat it as such (Hypothesis 3). What can also happen is the robot not being completely alike to any of those categories and yet share a fair number of attributes with each one. We have, then, a Proper Agent (Hypothesis 2), whatever legal terms we might want to apply to it.

Accordingly, in a preliminary stage, it is relevant to understand what it means to be an Agent. If an entity is an Agent, it is, therefore, not a thing, because the logic law of non-contradiction doesn't allow this to happen. Given the fact that one thing opposes to the other (and they do, since they show different and opposite properties) the sentence *The robot A is an Agent* and the sentence *The robot A is a thing* cannot, ever, be true at the same time. For instance, an Agent, as we shall see, acts voluntarily, while a thing doesn't act at all. It seems obvious that one entity cannot act voluntarily and don't act at all at the same time.

By understanding what an agent is and arguing that a robot is an Agent, we exclude, automatically, the idea that it can be a thing. In a second phase we'll look into what it means to be a legally responsible Agent.

On the other hand, and endorsing the idea stated by Asaro (2007), the mere comprehension of the concept of Agent might as well help us to draw the boundaries of legal personhood, since the first concept walks hand-in-hand with the latter. In other words, Agency might conceal important clues in this domain.

Predictably, understanding the concept of Agent and list its main features is nearly impossible. Every single area of knowledge uses the notion of Agent, and yet, consensus has not been found. To name a few, Psychology, Philosophy, Law, Computation, Economy and Neuroscience, each *stole* the concept of Agent and filled it out with the attributes that most suited the domain. In this regard, Shardlow (1990) has a very interesting thesis where he reached, precisely, to this conclusion, even though the author investigated mainly three areas: Philosophy, Psychology and Computation. Confronted with this fact, we would be forced to argue that the concept of Agent is a dead end. Nevertheless, there may be something that can be done about this dead end.

There's this method in programming and computation, that programmers use when they must describe a complex problem: they draw the base-case. The base case is, simply put, the description of the simplest possible case in the complex situation. In a second stage, then, comes the building and writing in code of complex cases

and respective exceptions. What is, then, our base case in matters of Agency? What is the one thing or, rather, the only feature that, regardless of the area we look into, is always there?

As it shall be argued, is it the voluntary act. However, it will be also shown that this is not enough, since children, animals and people with mental disabilities do act voluntarily but are not considered legally responsible.

The following step was to determine what was missing in these cases, with resource to the domain of Psychology, which was found to be certain types of metacognitive processes, related to the ability of feeling guilt and the capacity of planning complex behavior.

In this sense, besides the capacity of acting voluntarily, any responsible entity has to show a specific kind of metacognitive processes. Only then accountability is an option. For a comprehensive understanding of the paper, the next page provides a visual outline of its structure.

2 What Is the Common Denominator in Agency?

Intuitively, each one of us has an idea of what it means to be an Agent. It's an entity, whatever kind, capable of acting and execute actions, opposing to others entities that merely tolerate or accept events that happen to them.

In order to find a consensual definition, however, we must increase the level of abstraction. In this abstract sense, and for this purpose, an Agent is an entity which acts continuous and autonomously in time, in a dynamic environment, where other processes exist, and other Agents are present (Coelho 2008).

In Philosophy, two of the most prominent theory are the *Standard Conception* and the *Standard Theory.* Both argue that and Agent is a being which is capable of intentional action.[1] The difference between these two theories has to do with whether or not the intentionality of the action includes unwanted actions.

For instance, let's imagine Asimov wishes to reach for his glass of water, in the middle of the night, and turns on the light in order to do so. We would assume that the latter was desired by him, and intentional, since he had, before actually acting, the thought about turning on the light in order to get the glass of water. However, if there was a burglar on the outside of his house and he was not aware of this fact, he might as well let the burglar know he was home, even though it was not what he intended to do.

Even though he wanted to turn on the light, Asimov's thought was definitely not about alerting the burglar and yet he did it. This is what an unwanted action is. The *Standard Conception* argues that intentional action includes both turning the

[1] Intentional action not in the sense of having the intention to do something but instead in the sense described by Anscombe (1957) and Davidson (1963), which relates to acting for a reason (a mental state of believing that the specific action is the best to achieve a certain goal).

light for the glass and turning the light and warn the burglar; on the other hand, the *Standard Theory* holds that only the first is an intentional action. Despite not agreeing about the meaning of intentional action, both theories believe an Agent is an entity capable of intentional action. Thus, according to these perspectives, an entity is an Agent if it can act voluntarily, since the act depends on the belief that the specific action in question is the best to achieve a certain goal.

Naturally, and especially not in Philosophy, this is not the sole theory at the center of the debate. Other theory was described by Dennett (1987). This author argues that we have an Agent before us if we can predict his behavior, accurately, by means of its mental states. Accordingly, Allen and Bekoff (1997) used this idea, arguing that it could be applied to non-human Agents.

More recently, Barandiaran et al. (2009) focused on extremely simple entities, such as bacteria. In the author's opinion, the fact that these kind entities can't be included in the category of Agent, given the before mentioned Philosophical theories, doesn't mean they shouldn't be regarded as Agents. In this sense, Barandiaran outlined three main requisites for what he calls *minimum Agency.* Besides individuality (which is the clear distinction between the Agent and its environment) and normativity (meaning the existence of goals and rules that the Agent uses to guide its action) he also argues that interactional asymmetry is crucial. This last precondition for Agency concerns the ability to exchange energy and matter with the environment. In other words, the Agent must be able to collect the necessary energy to act and being a passive entity in the environment is not enough.

So far, in Philosophy, it seems that the voluntary act is a relevant requisite to ascribe Agency. As we shall see later on, this is not the only domain of knowledge where this ability is a precondition.

In fact, that's precisely what happens in Computation. While Minsky (1967) saw the artificial Agent as a Finite State Machine (FSM), a description often seen as reductive, other authors such as Russell et al. (1995, p. 33) see the Agent as an entity that analyses the surrounding environment and acts according to the input of that same environment.

Another very praised view is the one described by Wooldridge and Jennings (2009) which defines the Agent as the entity that presents properties such as autonomy, social skills, reactivity to the environment and proactivity (ability to initiate action). According to the authors, an entity that shows these cumulative attributes has what they call *weak Agency*. Conversely, if we're looking for a *strong* Agency, the Agent must show some degree of cognitive processes, including beliefs, desires and intentions (Taylor 1966, p. 98; Shoham 1993).

It is not possible to simply look into every single area of knowledge in order to discover what it means to be an Agent in each one. There's, still, one more to go and is an especially complex domain: Law.

In Law, an Agent is typically considered the author of an illicit action (for instance a crime), which he did by means of a voluntary act. The biggest issue in this matter is that in order to be considered an Agent, in the sense used by Law, there's the implicit idea that the Agent has legal personhood. Since we're trying to

do the opposite, meaning we're trying to get to accountability and legal personhood through the notion of Agent, this isn't particularly helpful.

What we can do, instead, since the concept of legal person can be considered as the basic unit of law, in order to act in legal relationships (Derham 1958), is examine what makes the difference when it comes to giving legal personhood to an entity. In other words, it's important to investigate the reasons behind the lawmaker's decision to grant or not this legal status to an entity.

The first reason to give legal personhood is, obvious and naturally, because the entity is a (born and yet not deceased) human being (Solaiman 2017). Artificially, we also consider companies to have legal personhood, with theories justifications that go back to Savigny and that by no means this paper intends to discuss.

In this sense, there are two main theories, regarding the matter, in analytic jurisprudence: the will theory and the interest theory (Kramer et al. 1998). Most of the nineteenth-century German legal academics who wrote on this topic based their theories on the Kantian ideas of freedom and autonomy as the central concepts. Human beings possess, according to this theory, innate moral freedom, which grounds their capacity to hold rights and thus their legal personhood. Yet, the minority view, advocated an interest-based understanding of rights. Modern analytic theories of rights are usually classifiable as either one of these theories. However, hardly any of the theories can be said to have 'won' the debate (Kurki 2019).

Additionally, these theories are still not enough in order to draw the line between responsible entities and non-responsible ones. Anglo-Saxon Judges reflected extensively upon the concept of Agent, long before it became a foregone conclusion to us. Salmond (1913), argued that in order to be a juridical person, one must show the capacity of being a part in juridical relations. In another direction, Dewey (1926) described how we do not think about conceding legal personhood to things, since their behavior would be exactly the same, whether you ascribe or not legal duties to it. In the author's words, we grant legal personhood to either entities whose behavior can be modulated by the legal norm or to entities through which we wish to regulate human's behavior, this being the reason why ships were once given legal personhood.

More recently, Dario and Palmerini (2012), based on the before mentioned theories of Legal Personhood related the concept of legal personhood to the idea of duty and the thought of being able to act in order to enforce that same duty.

Today, and in general, several authors (for instance, Mathew Kramer and Joel Feinberg, this last author regarding animals) have supported a specific conception of legal personhood: the one that argues that any entity who is capable of carrying legal rights should be granted legal personhood (Kurki 2016).

This vision has been somewhat applauded, constituting, inclusively, the main grounds for a case in December 2014, in the NY Supreme Court about a chimpanzee named Tommy. Tommy's representation asked for the extension of the concept of legal person, in order to be able to request *habeas corpu*s later on. The representative argued, precisely, that animals can carry at least one legal right, and that this was enough to get a specific type of legal personhood, in accordance with the rights

proclaimed (Kurki 2016). Pietrzykowski (2017) described a similar case in a Court of Argentina, about an Orangutan in a Buenos Aires' Zoo.

There's intense literature when it comes to this debate. Other relevant views include Rationality as the main criteria for legal personhood (Morse 2000) and Intentionality[2] (Calverley 2008; Chopra and White 2011).

It's important to state that we cannot, ever, disconnect the Law from the reality where it operates. Law is permeable to reality and culture (Ferreira and Pereira 2017) and this is a crucial relation if we want to avoid an obsolete and useless legislation. This is why all these different theories in Law are so important in this research.

It is also relevant to point out that it seems that regardless of the view supported, there's always this idea of being able to act (in the sense that if one is capable of carry a legal right or obligation one must be capable of acting accordingly) hovering over all the mentioned theories. The same occurs in Computations and Philosophy, though wearing different vests. In conclusion, it appears that different words are used to name the same thing.

As described before, each area of knowledge took the concept of Agent to itself and designed it in its image and likeness. Despite this fact, however different the definitions of Agent might be, the condition of having the power to act, voluntarily, is always present.

3 What Is a Voluntary Act?

Markby, in Elements of Law—Principles of Jurisprudence (1889) defined voluntary act as the body movement that follows the will. Coincidentally, on another domain of knowledge—in Classic Philosophy—Davidson used this exact same description, 84 years later, when writing his theory of Agency. The same was argued again and again throughout the twentieth and twenty-first centuries—in Law, though with different words—namely by Cook (1917) and Yaffe (2012).

In Psychology, James et al. (1890) described the voluntary act as the opposite of involuntary act, in the sense that the latter occurs without foresight. In recent Philosophy, the similar was argued by Olsaretti (1998), who supports the idea that we have a voluntary act if we have not an involuntary act. The action will not be voluntary, in the author's thesis, if there is no other acceptable option, according to some objective criteria (though the author doesn't exactly explain what is this objective criteria). For Olsaretti, an unacceptable option is the one that causes specific damage to the Agent or when a moral rule is imperative to the point that makes all other options unacceptable. She also states that the voluntary act is deeply related to the motivations of the Agent, in the sense that it depends, inevitably, on the beliefs the Agent has about his options. If the Agent is mistaken about his options,

[2] In the sense previously described in Philosophy.

he might have a good option but be unaware of its existence. Thus, an act can be involuntary for misinformation.

That's precisely what Aristóteles (2004) argued, in Nicomachean Ethics: that the only two reasons that would make an act involuntary would be ignorance or major external forces.

In conclusion, an act seems to be voluntary when there's a bodily movement, guided by will, as long as it is not undermined by ignorance or an external force.

The following question is: does AI act voluntarily? AI might have a previously defined (by humans) structure of their beliefs, desires and intentions, but after that initial definition, more complex (or stronger) AI is able to act upon the environment autonomously, and possibly according to the goal they set for themselves. We have come to the point when AI is so advanced that in some cases not even creators know exactly why the robot did what it did. In normal conditions, the robot is well informed about his choices, as it is capable of collect the essential information in order to create a model of the world. Also in normal circumstances, they will not be coerced to do anything, though they might be.

So, do robots belong in the category of Agent? It appears that in the cases of strong or complex AI robots (the so-called robust AI) seem to have the minimum requisite to be considered as one: they act voluntarily.

It's important to disclaim that by referring to complex AI, namely, machines that use cognitive processes or machine learning, we are not describing objects that clearly act as tools and that are perceived and intended to act as such, like smart air conditioners which adjust according to the temperature or lights change intensity according to the hour of the day.

As mentioned before, if complex AI belongs in the category of Agents, it cannot be considered merely a tool. What matters now is to learn how much responsibility they can take, if any at all.

4 What Makes an Agent a Legally Responsible One?

The next step is trying to understand what makes the Law ascribe responsibility or not to an individual.

According to the previous definition of Agent, it seems obvious that children and animals are also Agents. However, we don't consider them as legally responsible Agents. In other words, simply being an Agent and acting voluntarily isn't enough for the Law. In this sense, where should we draw the line between responsible agents and non-responsible ones?

There is one very relevant legal concept that might help us in this query, which is the notion of imputability. However, the sense that we want to grasp here is the lack of imputability, which relates to a specific category of people to whom, either because they are under aged or suffering from a mental illness, we cannot ascribe legal responsibility to, even though they have legal personhood. Though there are many reasons and theories on why Law does not deem these individuals as

accountable, one of the major reasons concerns is the fact that these subjects do not present the capacity of feeling guilt (Pizarro de Almeida, 2000 p. 21).

In the legal sense, guilt is understood as the capacity that the subject has of acting in a responsible way, meaning he is able to understand what an illicit behavior is and therefore opt by not performing that behavior. In this sense, the subject must be capable of reflecting upon a certain conduct and assert a positive or negative value to that same conduct.

In other words, we can only ascribe any legal responsibility when we assume that the Agent has the minimum requirements, from a physical and psychological point of view, in order to respond positively to normative rules. In the presence of this set of minimum requirements then we have an imputable Agent (Muñoz Conde and Arán 1996).

Other than helping in the judgement in criminal cases, the guilt also relates to a negative valuation that the society develops towards the Agent's behavior. There's no point, at all, in addressing a negative valuation of conduct towards an Agent that is not capable of understanding that judgement. It simply will not be effective. In these cases, the cognition of the Agent might be so compromised that even though he can act voluntarily, according to some desires or goals, he cannot reflect upon those (primary) mental states that originated the behavior.

In Philosophy, as well as in Cognitive Psychology, these mental states about other primary mental states, goes by the name of Metacognition.

5 Metacognition: Shaping Legal Responsibility

It seems fair to say that we are allowed transpose concepts from one domain of knowledge to another. Most of the foundations of Modern Law came from authors such as Kelsen, Hart and Austin, all of them also philosophers, who set the grounds for Philosophy of Law. On the other hand, we cannot legislate about the world around us without fostering concepts of the mundane. For instance, we wouldn't be able to legislate Medicine if we were not capable to grasp the concepts of that specific area of knowledge. Moreover, some authors such as Morse (2003) argue that Law itself uses models of actions that derive from Folk Psychology.[3] In other words, it is legitimate for us to use concepts long used in other areas of knowledge, is this case, the notion of Metacognition, which is a relatively old concept in Philosophy and Cognitive Psychology.

In general, Metacognition is the cognition about cognition (Fleming et al. 2012), being useful in order to control and/or monitor behavior and mental processing (Nelson and Narens 1990).

[3] Folk Psychology is traditionally used to denote our everyday (intuitive) understanding, or rationalizing, intentional actions in mentalistic terms (Hutto and Ravenscroft 2021).

Frankfurt (1971), a philosopher, argued that the main difference between human beings and other types of Agents is rooted in the structure of the will, in the sense that only human beings reflect upon their own motivations, which results in second order mental states. For instance, let's imagine Wall-e has to study for an exam. In order to succeed in this exam, Wall-e must, beforehand, list the study methods he knows, analyze his own strong characteristics and his weaker ones, so he can choose the best study method for him, considering the specific subject he has to study. Learning is, in itself, a cognitive process. By reflecting on this cognitive process (choosing a study method), he is using this second order mental states or, as many authors describe, secondary cognition. To Frankfurt, the difference between human beings and other Agents, which also act voluntarily, is Metacognition. We can, then, argue that there is a distinction, between Agents who act voluntarily but do not show Metacognitive Processes, and Agents who act voluntarily and do present this capacity.

Agents who act voluntarily and present metacognitive processes can do so in several ways, as this type of cognition has many shapes and forms, and not all will be described in this paper. However, as we shall see, to hold an entity accountable, at least two kinds of metacognitive processes are required: strategic and monitoring processes. Both will be explained henceforth by this order.

As Cox (2005) stated, any intelligent Agent, when confronted with a choice (any choice, therefore including the choice to practice an illicit act or not), he must decide three things: (1) which action, given the possible ones, is the most adequate in the present situation, (2) if the choice he is making is sufficiently informed or if more information is required and (3) if something has gone wrong, understand why it happened. This is a critical auto-reflexive type of thought, which translates the analysis that an individual makes in terms of the quality of the options presented in decision-making. In turn, this process is undoubtedly linked to Metacognition.

Accordingly, one of the most essential components of Metacognition described by literature is knowledge of cognition (Lai 2011). This implies awareness of our own capacities and limitations, including internal and external factors that may affect or reduce our cognitive performance (Flavell 1979). This component is extremely relevant when it comes to defining strategies in action, since it is the reason we chose one strategy to the detriment of other strategy (as it happens in the above mentioned example of the study methods).

What is important to point out is that any person who wants to commit act illegal act has, necessarily, the strategic analysis that was described in the previous paragraph. A mentally ill person can act wrongfully but his intention was to act merely and not to act illegally. On the other hand, someone who plans an illegal act, thinks about the final goal, reflects on his own capacities and limitations and other external factors that might affect his performance, defines a strategy, all things considered in the light of the possibility of being caught.

Supporting this idea, it might also be useful to look into the theory of planned behavior, from the area of Psychology (Ajzen 1991). Summarily, the author argues that the Agent's intention is modulated, mainly, by three things: (1) individual attitudes regarding the behavior at hand, (2) individual pressure concerning the

specific conduct and (3) behavior control. Simplifying, what we have is a certain behavior, linked to an intention which in turn is modulated by these three factors. There hardly can be any doubts about the existence of strategic metacognitive processes on planned behavior, including illicit planned behavior.

On the other hand, Metacognition is said to have three levels of consciousness in any storyline. The first one concerns the story or the behavior itself. The second one relates to the thoughts that the Agent has towards that occurrence. The third and last one is about the reflexive work about the thoughts of the second level (Cox 2005).

Translating the theory to a practical example, let's imagine we have a subject, HAL, shopping at the local store. Someone tries to steal something, and the police is called to the store. The thief is caught and taken into custody. HAL watched closely everything that happened. This is the first level of consciousness, the occurrence, story or behavior (in this case, someone stealing in the shop). HAL then kept on with his life, meditating about the event, its legal value, and the punishment he saw being applied to the thief. We have, then, a level two of consciousness. Finally, as a healthy human being, HAL is also capable of having second order thoughts about that first reflection. For instance, he might initially have thought that the punishment was not fair but then feel ashamed by his own thought. Or realize he didn't think stealing was wrong and then feeling scared that he might act in a similar way.

What we have at hand is a judgement made about other judgements, with the purpose of monitoring behavior. As explained through the example above, being able to feel guilt, can also be considered to have this purpose.

As previously described, one of the reasons why law does not account people with mental disabilities is, precisely, the inability to feel guilt, which implies a kind of complex agency. This complex agency implies the capacity of understanding what an illicit behavior is and opting by not performing that behavior, which in turn implies metacognitive processes, in this case, not in the sense of strategic analysis (needed when planning and illicit behavior) but rather in the sense of monitoring behavior (which concerns the process of reflecting upon behavior and decide whether or not commit the crime).

In conclusion, among the several forms of metacognitive processes that an individual may have, to perform and understand an illicit behavior, an individual will need, at least, two types of metacognitive processes: strategic and monitoring. This is the core of accountability.

Without knowing, Law has been using this concept of Metacognition across time. Animals are not directly responsible, nor children are, having instead someone who is responsible for them. In the first case, animals are able to understand that the occurrence getting a biscuit happened because they rolled over when asked to. However, they cannot, in general, have complex and second order thoughts about the best way or method to do it, which leaves us only with a second level of consciousness and hardly any metacognitive processes. Accordingly, animals are not held accountable for their acts, nor are granted legal personhood.

Children's situation is clearly different, as they show some type of Metacognition, and it gets more complex while growing up. There are many studies in this regard, for instance the ones described by Georghiades (2004), in *From the general*

to the situated: three decades of Metacognition, which shows precisely this. They are, inclusively very early in their lives, able to learn (and learning implies a certain kind of Metacognition). They do not present, however, strategic Metacognitive processes, which, as described in the previous paragraphs, is the specific kind we're looking for when discussing legal accountability. We're talking about a formally stated operational thought (Piaget 1976), which rarely is attributed to children (Brown and DeLoache 1978). Studies also show that strategic Metacognition starts developing around 14 years old, even though it might not be completely developed until later on (Schraw and Moshman 1990). Although children do have legal personhood, truth is, by chance or not, the law only ascribes criminal responsibility to underage individuals when they turn 16 years old, believing that at this age they are sufficiently developed to understand the consequences of their actions.

This type of strategic Metacognition is also missing in the case of some mental illnesses (Saxe and Offen 2010), though the consequences in consciousness might change from disease to disease and from person to another person (David et al. 2012).

In programming and computation, Metacognition relates to what the system knows about its own cognition and also about cognition in general. As Crowder et al. (2011) describe it, in AI this concept is intertwined with introspection, in the sense that allows the machine to form beliefs about its own internal states, instead of simply analyze the environment where it moves.

Traditionally, in computation, metacognitive processes are used for specific problem solving, such as algorithm selection from the efficiency point of view (Cox 2005).

In this sense, Crowder & Friess argue that there are at least three types of Metacognition in this domain of knowledge:

(a) Metacognitive knowledge, which relates to what the system knows about itself, as a cognitive processor (Kosko 1986);
(b) Metacognitive regulation, regarding the control of cognition and learning, which may include the knowledge the system has about what it knows and does not know (LaBar and Cabeza 2006);
(c) Metacognitive experience, which concerns past experiences that somehow relate to the present mission of the system (Crowder et al. 2011), allowing the system to create expectations or predictions about what may happen, given those experiences that took place before that moment of analysis.

In this sense, its seems fair to acknowledge that AI can has some degree of metacognitive processes. However, it does not match the type of Metacognition necessary in order to consider an entity as accountable. In fact, none of these processes translate in strategic or monitoring metacognitive processes. Hence, AI should not, at least for now, be held accountable for its behavior, the same way kids, animals and people with mental illnesses are not.

6 Accountability and Legal Personhood

Up to this moment, we linked Agency, to voluntary act, the latter as a minimum requirement for the first, and accountability to metacognition. There is still one round left, regarding the connection between accountability and legal personhood.

We are fully aware that legal responsibility and legal personhood are not the same concept, an often-made mistake regarding AI, either by scholars or official entities, as Pagallo (2018) pointed out in his research. In fact, they're different concepts and might also mean different legal consequences. But they must be intrinsically intertwined.

In this paper, we described how animals, children and people with mental illnesses were not to be considered accountable from the legal standpoint. In this sense, it was also highlighted that, even though children and people with mental disabilities do have legal personhood in most jurisdictions, they do so within a limited scope and a restricted exercise of their rights. We also pointed out how animals do not have legal personhood, at all, in most jurisdictions, although some extensions of this instrument were granted in specific cases.

In fact, it appears that legal personhood in its full sense exists to the extent that the entity is capable of exercising its rights. As we have seen, there are entities (e.g. children and people with mental disabilities) that while being granted legal personality, do not present legal capacity or have their legal capacity restricted, and therefore are not considered legally responsible. In other words, the scope of their legal personhood is limited.

On the other hand, any entity who is considered to have some sort of accountability, e.g. people in general, have both personality and capacity. Their legal personhood is at its fullest.

This means, in principle, that even though legal personhood can be granted either way, if we don't have accountability, we hardly can have legal capacity. In other words, accountability fills the capacity of the entity, thus determining the actual content and size of the legal personhood.

This is consistent with the idea described by Visa A. J. Kurki of what constitutes an active legal personhood, opposing to a passive legal personhood, being a concept that "requires that one can perform acts-in-the-law (being endowed with legal competences) and be held legally responsible (onerous legal personhood)". In his research intitled "A Theory of Legal Personhood", the author states that the key elements of active legal personhood are centred on legal responsibility and legal competences.

In fact, one cannot be interested in the idea of a "shallow legal personhood". Take the example of the robot Sophia, the humanoid robot built by Hanson Robotics, which "jokingly" stated AI would destroy humans in the near future. Sophia was granted citizenship by Saudi Arabia, in 2017. Besides all the hype and attention this circumstance has received, from a legal stance, this citizenship is hollow, in the sense that there is no actual point in granting such status. In reality, the word "jokingly" must be used with caution since the robot Sophia as no idea what a joke,

in practice, means, let alone the meaning of a legal duty. Sophia may have been granted citizenship but has no means to exercise its rights as a citizen.

The same logic should be applicable to legal personhood in the case of AI. If no legal consequences can be drawn from it, similarly to the citizenship of the robot Sophia, there is no actual benefit in granting it. Moreover, we should only do it, when we recognize the utility of this step, as it occurred in the case of corporations. Legal persons, gained its fictional legal personhood, when humans started to understand the importance of attributing legal obligations to companies. In other words, when humans started to recognize the utility in it.

Still, we could argue that both children and people with mental illnesses lack either or both the competence and the accountability elements of legal personhood, and still it is granted to them (although, as Kurki puts it, it is a kind of passive legal personhood), meaning that there would be no reason to avoid doing the same in the case of AI.

However, there are specific reasons for such thing to happen. As Savigny and many other authors stated, the original concept of legal person is typically a match with the concept of human being, based on the presumption that human beings possess legal capacity (Kurki 2019). In this sense, to both children and people with mental disabilities, legal personhood is attributed by the mere fact that they're both categories of born human beings, a criteria that, surely, cannot be applied to AI. This circumstance tells us that we must look for a different criteria in this case. In this paper, it is argued that this criteria should be the possibility of playing an active role in legal personhood, through competence but, in special, legal responsibility.

Additionally, to children, legal personhood is typically attributed according to the Hegelian understanding that there is a potential of rationality and freedom and that children start to accumulate the capabilities required of a duty-bearer at some point (Kurki 2019).

In conclusion, without metacognition, there can hardly be any legal responsibility. On the other hand, without accountability, there is no reason why AI should have legal personhood, because without this element, there are no useful legal consequences to be drawn from it. Such legal consequences may only exist the day we find AI to be accountable. Otherwise, legal personhood in AI will mean nothing more than an empty shell.

7 Conclusions

This paper sought to draw a line between accountable and non-accountable AI, using several areas of knowledge, such as Philosophy, Psychology, Computation and Law.

In this sense, the paper argues that the problem of whether or not to ascribe legal personhood to AI can be solved through the notion of metacognition, a concept that, without knowing, Law has been using all along to decide upon this matter.

To achieve this purpose, we started by examining the meaning of Agent, in order to assess whether or not AI should be considered as such. As the concept presented

many difficulties, a common denominator was needed, which it was found to be the voluntary act. If there is a voluntary act, we must, then, conclude that we have an Agent before us. Accordingly, and as long as AI acts voluntarily, it makes sense to argue that complex robots are Agents, thus not mere tools.

However, as stated before, this does not necessarily mean that an AI must be held accountable just because it fits the category of Agent. Animals, people with mental illnesses and children are intuitively considered Agents and yet not held accountable.

Hence, the other argument that was made is that in order to ascribe responsibility to an Agent, that entity must show, at least, strategic and monitoring metacognitive processes. These elements take part in the ability of being accountable, which in turn composes, along with the concept of legal competence, the notion of an active Legal Personhood.

Considering the above conclusions, two other ideas must follow. If the entity does show Metacognitive processes, then we might consider grant the said entity with legal personhood. On the other hand, if it doesn't show this capacity, then we need an autonomous and, if necessary, new, applicable law, as we have in the case of children, animals and mental illnesses.

When it comes to the state of AI, today, it seems that it does not yet stands in a sufficiently complex level in terms of metacognitive processes in order to being held accountable for their actions, notwithstanding showing simple metacognitive processes.[4]

References

Ajzen I (1991) The theory of planned behavior. Organ Behav Hum Decis Process 50:179–211

Allen C, Bekoff M (1997) Species of mind: the philosophy and biology of cognitive ethology. MIT Press, Cambridge

Anscombe GEM (1957) Intention. Harvard University Press, Cambridge

Aristóteles (2004) Ética a Nicómaco (trans: Caeiro DAC). Quetzal, Lisboa

Asaro PM (2007) Robots and responsibility from a legal perspective. In: Proceedings of the IEEE, pp 20–24. http://peterasaro.org/writing/ASARO%20Legal%20Perspective.pdf. Accessed May 2019

Barandiaran XE, Di Paolo E, Rohde M (2009) Defining agency: individuality, normativity, asymmetry, and spatio-temporality in action. Adapt Behav 17(5):367–386. https://doi.org/10.1177/1059712309343819

Birks P (1997) Definition and division: a mediation on institutes. 3.13. In: Birks P (ed) The classification of obligations. Clarendon Press, Oxford, pp 1–21

[4] *See* generally, on the imitation of humans by Robots, in this book M C Patrao Neves and A B Almeida—Before and Beyond Artificial Intelligence: Opportunities and Challenges; and M N Duffourc and D S Giovanniello—The Autonomous AI Physician: Medical Ethics and Legal Liability.

Boulangé A, Jaggie C (2014) Ethique, responsabilité et statut juridique du robot compagnon: revue et perspectives. Cognition, Affects et Interaction. https://www.researchgate.net/publication/278625871_Cognition_Affects_et_Interaction. Accessed May 2019
Brown AL, DeLoache JS (1978) Skills, plans, and self-regulation. In: Siegler RS (ed) Children'sthinking: what develops? Lawrence Erlbaum Associates, Inc., Hillsdale, pp 3–35
Calverley DJ (2008) Imagining a non-biological machine as a legal person. AI Soc 22:523–537
Chopra S, White L (2011) A legal theory for autonomous artificial agents. University of Michigan Press, Ann Arbor
Coelho H (2008) Teoria da agência: Arquitectura e cenografia. Edição do Autor, Lisbon
Cook WW (1917) Act, intention, and motive in the criminal law. Yale Law J 26:645–663
Cox MT (2005) Metacognition in computation: a selected research review. Artif Intell 169:104–141
Crowder J, Friess S, Ncc M (2011) Metacognition and meta memory concepts for AI systems. In: Proceedings on the international conference on artificial intelligence (ICAI), Athens
Dario P, Palmerini E (2012) Robot companions as case-scenario for assessing the "subjectivity" of autonomous agents. Some philosophical and legal remarks. In: First workshop on rights and duties of autonomous agents, pp 24–31
David AS, Bedford N, Wiffen B, Gilleen J (2012) Failures of metacognition and lack of insight in neuropsychiatric disorders. Philos Trans R Soc Lond Ser B Biol Sci 367:1379–1390
Davidson D (1963) Actions, reasons, and causes. J Philos 60:685–700
de Almeida P (2000) Modelos de inimputabilidade: Da teoria à prática. Almedina, Janeiro de
Dennett DC (1987) The intentional stance. MIT Press, Cambridge
Derham DP (1958) Theories of legal personality. In: Webb L (ed) Legal personality and political pluralism. Melbourne University Press, Melbourne, pp 1–19
Dewey J (1926) The historic background of corporate legal personality. Yale Law J 35:655–673
Ferreira AE, Pereira D (2017) Partilhar o mundo com robôs autónomos: A responsabilidade civil extra- contratual por danos, Introdução ao problema, Cuestiones de Interés Jurídico. IDIBE, Alicante
Flavell JH (1979) Metacognition and cognitive monitoring: a new area of cognitive–developmental inquiry. Am Psychol 34:906–911
Fleming SM, Dolan RJ, Frith CD (2012) Metacognition: computation, biology and function. Philos Trans R Soc Lond Ser B Biol Sci 367:1280–1286
Frankfurt H (1971) Freedom of the will and the concept of a person. J Philos 68:5–20
Georghiades P (2004) From the general to the situated: Three decades of metacognition. Int J Sci Educ 26:365–383
Hutto D, Ravenscroft I (2021) Folk psychology as a theory. In: Zalta EN (ed) The Stanford encyclopedia of philosophy
James W, Drummond R, Henry Holt and Company (1890) The principles of psychology. Henry Holt and Company, New York
Kosko B (1986) Fuzzy cognitive maps. Int J Man Mach Stud 24:65–75
Kramer MH, Simmonds NE, Hillel S (1998) A debate over rights: philosophical enquiries. Oxford University Press, Oxford
Kurki VAJ (2016) Revisiting legal personhood. Paper for Spanish-Finnish Seminar in Legal Theory. PhD Candidate, University of Cambridge
Kurki V (2019) A theory of legal personhood. Oxford University Press, Helsinki Legal Studies Research Paper No. 58
LaBar KS, Cabeza R (2006) Cognitive neuroscience of emotional memory. Nat Rev Neurosci 7:54–64
Lai ER (2011) Metacognition: A literature review. Pearson Research Report. Pearson Education, Upper Saddle River
Markby W (1889) Elements of law, considered with reference to Principles of general jurisprudence. Clarendon Press, Oxford
Minsky M (1967) Computation: Finite and infinite machines. Prentice-Hall, Englewood Cliffs

Morse SJ (2000) Rationality and responsibility. Faculty Scholarship at Penn Law, p 524. https://scholarship.law.upenn.edu/faculty_scholarship/524. Accessed 29 Sept 2021
Morse SJ (2003) Diminished rationality, diminished responsibility. Ohio State J Crim Law 1:289–308
Muñoz Conde F, Arán MG (1996) Derecho penal: Parte general. Tirant Lo Blanch, Valencia
Nelson TO, Narens L (1990) Metamemory: a theoretical framework and new findings. In: Bower GH (ed) Psychology of learning and motivation. Academic, San Diego, pp 125–173
Olsaretti S (1998) Freedom, force and choice: Against the rights-based definition of voluntariness. J Polit Philos 6:53–78
Pagallo U (2013) The laws of robots - crimes, contracts, and torts. Springer, Dordrecht
Pagallo U (2018) Vital, Sophia, and Co.—the quest for the legal personhood of robots. Information 9:230
Piaget J (1976) The grasp of consciousness. Harvard University Press, Cambridge
Pietrzykowski T (2017) The idea of non-personal subjects of law. In: Kurki VAJ, Pietrzykowski T (eds) Legal personhood: animals, artificial intelligence and the unborn. Springer International Publishing, Cham, pp 49–67
Russell SJ, Norvig P, Davis E (1995) Artificial intelligence: a modern approach. Prentice Hall, Upper Saddle River
Salmond JW (1913) Jurisprudence. Stevens and Haynes, London
Saxe R, Offen S (2010) Seeing ourselves: what vision can teach us about metacognition. In: Dimaggio G, Lysaker PH (eds) Metacognition and severe adult mental disorders. Routledge, Hove, pp 13–30
Schraw G, Moshman D (1990) Metacognitive theories. Educ Psychol Rev 7:351–371
Shardlow N (1990) Action and agency in cognitive science. Master's thesis, University of Manchester
Shoham Y (1993) Agent-oriented programming. Artif Intell 60:51–92
Solaiman SM (2017) Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy. Artif Intell Law 25:155–179
Taylor R (1966) Action and purpose. Prentice-Hall, Englewood Cliffs
Wooldridge M, Jennings NR (2009) Intelligent agents: theory and practice. Knowl Eng Rev 10:115–152
Yaffe G (2012) The voluntary act requirement. In: Andrei M (ed) The Routledge companion to philosophy of law. Routledge, New York, p 174