Technology and Regulation

On the legal responsibility of artificially intelligent agents: addressing three misconceptions

Antonia Waltermann*

(legal) responsibility, moral agency, legal agency, liability, artificially intelligent agents.

antonia.waltermann@maastrichtuniversity.nl

This paper tackles three misconceptions regarding discussions of the legal responsibility of artificially intelligent entities: (a) that they cannot be held legally responsible for their actions, because they do not have the prerequisite characteristics to be 'real agents'; (b) they should not be held legally responsible for their actions, because they do not have the prerequisite characteristics to be 'real agents'; (c) they should not be held legally responsible for their actions, because to do so would allow other agents to 'hide' behind the AI and thus escape responsibility. (a) is a misconception not only because (positive) law is a social construct, but also because there is no such thing as 'real' agency. The latter is also the reason why (b) is misconceived. The arguments against misconceptions a and b imply that legal responsibility can be constructed in different ways, including those that hold both artificially intelligent and other (human or corporate) agents responsible (c). The paper concludes that there is more flexibility in the construction of responsibility of artificially intelligent entities than is at times assumed.

1. Introduction

The emergence and proliferation of artificially intelligent entities (hereafter referred to also as artificial agents or AI) raises questions of legal liability and responsibility. This is because some artificially intelligent entities do not require human input to perform some action, nor do their actions necessarily follow pre-programmed patterns. Given the developments in machine learning, it seems that (some) artificial agents are acting autonomously and that more artificial agents will be acting more and more autonomously in the future.[1] This leads to an accountability gap in the law.[2] Situations in which harm occurs for which no one is responsible according to current positive law (*lex lata*) but which, it seems, should not have to be borne by the entity suffering it are becoming increasingly likely. How this accountability gap should be closed has been subject to much debate, both politically and academically.[3] In this paper, I will focus on three (interconnected) misconceptions within these debates.[4] Most references will be to tort law, but the ground for legal responsibility, be it tort, contractual, or criminal, does not matter. The three misconceptions are that artificially intelligent entities:

A. *cannot* be held legally responsible for their actions, because they do not have the prerequisite characteristics to be 'real agents' and therefore cannot 'really' act.

B. *should not* be held legally responsible for their actions, because they do not have the prerequisite characteristics to be 'real agents' and therefore cannot 'really' act.

C. *should not* be held legally responsible for their actions, because to do so would allow other (human or corporate) agents to 'hide' behind the AI and escape responsibility that way, while they are the ones who should be held responsible.

The first two misconceptions are connected by the content of the argument put forward ("AI lack the prerequisites to be 'real agents'") but differ in the kind of conclusion that is justified by it, the first conceptual and the second normative. Meanwhile, the second and third misconception are connected by the conclusion of the argument ('AI should not be held legally responsible') but differ with regard to the content of the argument put forward to justify that conclusion.

This paper argues that all three arguments (a-c) are misconceived. The argument to this effect proceeds along the following lines: first, I will briefly outline what I mean by artificially intelligent entities (section 2). Then, I will elaborate on the first misconception (a) that

1 By this, I mean that they act in ways that are not foreseen or predicted and not (easily) foreseeable or predictable. At times, this may go hand in hand with not being (easily) understandable or explainable by programmers/developers. Some more on this in section 2.

2 Cf. Gunther Teubner, 'Digitale Rechtssubjekte? Zum Privatrechtlichen Status Autonomer Softwareagenten' (2018) *Archiv für die civilistische Praxis;* Susanne Beck, 'The Problem of Ascribing Legal Responsibility in the Case of Robotics' (2015) 31 *Ai & Society* 473.

3 Cf. European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)); 'Open Letter to the European Commission Artificial Intelligence and Robotics' http://www.robotics-openletter.eu accessed 26/01/2021; Francisco Andrade and others, 'Contracting Agents: Legal Personality and Representation' (2007) 15 *Artif Intell Law* 357; Joanna J. Bryson, Mihailis E. Diamantis and Thomas D. Grant, 'Of, for, and by the People: The Legal Lacuna of Synthetic Persons' (2017) 25 *Artif Intell Law* 273 for an overview of some political proposals, calls, and concerns.

4 This focus mirrors Ugo Pagallo, 'Apples, oranges, robots: four misunderstandings in today's debate on the legal status of AI systems' (2018) *Phil. Trans. R. Soc.*, although the misconceptions addressed and the arguments used to address them differ.

* Antonia Waltermann is assistant professor at Maastricht University, the Netherlands. I am very grateful for useful comments by Jaap Hage, Daniel On, and Rüta Liepina, as well as by the anonymous reviewers. The usual disclaimers about the final result apply.

Received 9 Apr 2021, Accepted 3 Jul 2021, Published 12 Jul 2021.

legal agency must (conceptually) coincide with 'real' agency (section 3). This is a misconception not only because (positive) law is a social construct, but also because there is no such thing as 'real' agency (section 4). The latter is also the reason why the second argument (b) is misconceived. The argument that there is no 'real' agency will require an excursion into the realm of philosophy and the cognitive sciences,[5] but as I hope to demonstrate, this excursion is highly relevant to the question whether legal responsibility of AI is possible and desirable.

The arguments against misconceptions a and b imply that legal responsibility can be constructed in different ways, including those that hold *both* artificially intelligent and other (human or corporate) agents responsible (section 5), pre-empting the concern that human/corporate agents could 'hide' behind AI responsibility (misconception c). Accordingly, this paper concludes that there is more flexibility in the construction of responsibility of artificially intelligent entities than is at times assumed (section 6). This offers more freedom to law- and policymakers, but also requires openness, creativity, and a clear normative vision of the aims they want to achieve.

Before diving into the argument of the paper, some caveats and clarifications are required.

This paper deals with questions of responsibility and agency, but these terms are used in different contexts with different meanings. In computer science, for example, an agent is an entity that "observes the world through sensors and acts upon an environment using actuators" and "directs its activity toward achieving goals in a rational manner" or, in more technical terms, [a]n agent is a system that receives at time t an observation O_t and outputs an action A_t."[6] Law, meanwhile, knows the concept of an agent in agency law, where a person (the agent) acts as representative of another person (the principal), for example when a lawyer negotiates a contract on behalf of a client. In philosophy of action and in ethical theory, agent again means something else (see section 4). Where this paper uses the term 'agent', this is never in the sense of agency law; instead, the focus is on agents as entities capable of acting (in a sense relevant for responsibility).

When it comes to the terms 'liability' and 'responsibility', a common distinction is between legal liability on the one and moral responsibility on the other hand. Departing from this, I will use '(legal) responsibility' throughout this paper as an umbrella term for all types of liability. Similarly, I will use 'responsible' instead of 'liable'. Even where I omit the prefix 'legal' of 'legal responsibility', I will refer to legal responsibility, as opposed to moral responsibility, unless otherwise stated. In many areas of law (e.g. contract and tort), it would be more accurate to speak of liability than responsibility, but in other areas (e.g. international law), the term responsibility is used. I consider responsibility the more suitable term for the purposes of this paper to indicate a. the proximity to questions of moral responsibility and b. the abstraction from a particular legal field.

The latter relates to a point I want to further emphasise: the argument of this paper is situated at a high level of abstraction: it is not an argument about any particular legal system[7] or area of law. Instead, it is an argument about the relation between law, legal concepts, and concepts and insights from the cognitive sciences broadly construed.

2. Artificially intelligent entities

The European Commission defines artificial intelligence as follows:

> 'Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.
>
> AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications).'[8]

For the purposes of this paper, whether an artificial agent is purely software-based or physically embedded is not relevant; both purely software-based agents such as algorithms used in, for example, insurance risk assessment, and physically embedded ones such as autonomous vehicles or weapons systems can cause harm of the kind that raises questions of (legal) responsibility.

A distinction often made in this connection concerns different levels of autonomy (or independent action) of the artificially intelligent entity: 'from human supervision (level 1), and deterministic autonomy (level 2), to machine-learning (level 3) and multi-agent systems (level 4).' An alternative distinction that focuses on the level of human involvement is between human in the loop, human on the loop (equivalent to level 1) and human out of the loop (ranging from levels 2 to 4). In cases of 'human in the loop', human input is required before an action can be performed. In cases of 'human on the loop', actions can and will be performed without human input, but there is human supervision, and the supervising human can override the artificial agent's decision before the action is performed. An example of this would be a self-driving car with a human 'supervisor' who can redirect the car, or a weapon system that requires authorisation from a human being. In cases of 'human out of the loop', finally, there is no human input or interaction. Here, distinctions can be made between those cases where there is prior human input and the algorithm performs the 'loop' according to deterministic programming (level 2), to those scenarios where the algorithm is capable of learning and adapting its behaviour to what it has learned in ways not anticipated by programmers/designers. One could think of an autonomous car that learns to model its behaviour from other road users, for example. If an autonomous car also communicates with other autonomous cars and adapts its behaviour to information – such as road conditions or the location and length traffic jams – from other autonomous cars, this would be an example of a multi-agent system.[9]

The degree of autonomy is relevant when it comes to the accountability gap in law: current legal instruments, concepts, and arrangements do not seem sufficient for increasingly autonomous artificial agents.

5 I use cognitive sciences here in a very broad sense, including - but not limited to - neuroscience, psychology, and behavioural economics.

6 Woodrow Barfield, 'Towards a law of artificial intelligence' in Woodrow Barfield and Ugo Pagallo (eds), *Research Handbook on the Law of Artificial Intelligence* (Edward Elgar Publishing 2018); Daniel Silver, Satinder Singh, Doina Precup, Richard S. Sutton, 'Reward is enough' (2021) *Artificial Intelligence* 299, 3.

7 Although the author's background is in civil rather than common law, which will be reflected in some of the examples chosen. Nonetheless, the questions raised and argument made (should) hold mutatis mutandis.

8 Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe, Brussels, 25.4.2018 COM(2018) 237 final.

9 Antje von Ungern-Sternberg, 'Artificial Agents and General Principles of Law' (Available at SSRN: https://ssrn.com/abstract=3111881) *German Yearbook of International Law*, 4 f.

This is because the potential solutions that can currently be found in positive law often require a certain level of control and foreseeability by the human or corporate agent producing, owning, or using the artificially intelligent entity or require that the human or corporate agent has acted in a wrongful or culpable way before holding that (corporate or human) agent legally responsible. In cases of contractual breach, for example, an autonomous software agent cannot be held liable according to current German law, given that the software agent lacks the legal capacity to act (rechtliche Handlungsfähigkeit). Consequently, if the (human or corporate) operator of the software agent can demonstrate that they did not themselves violate a contractual obligation, there is no liability, and the other contracting party is left with the damage of the contractual breach caused by the software agent. A similar gap exists with regard to tort liability.[10] More generally, Barfield summarises that 'the use of artificial intelligence begs the question of who is liable if the artificial intelligence controlling smart technology learns and solves problems in ways completely unknown to the human in the system' and '[t]he more autonomous the system, that is, the more the human is removed from the decision-making loops of the system, the more difficult for courts to assign liability to humans when there is a system failure.'[11]

The above gives a broad definition of artificially intelligent entities and outline of the problem, but for the argument of this paper, nothing more specific is required.

3. Misconception a: legal agency must (conceptually) coincide with 'real agency'

The first misconception I tackle in this paper can be summarised as follows: AI cannot be held legally responsible because AI is not an agent.

Coeckelbergh, for example, indicates that

> 'a problem that becomes especially relevant in the case of AI is attribution of responsibility. Since technologies cannot be responsible moral agents and are hence a-responsible, the only way to ensure responsible action is to make humans responsible.'[12]

Dahiyat writes that

> 'Some commentators think that software agents are merely coded information and that we will commit excessive conceptual mistakes if we attribute a legal or moral responsibility to these agents, or if we just assume that they possess whatever else we take to be present when we hold human beings responsible for their actions. This is because, unlike humans who are sensitive, self-determined and moral, software agents lack a number of conditions, which should be fulfilled in order for responsibility to be ascribed.'[13]

Statements such as these indicate, it seems to me, that legal agency must (conceptually) coincide with 'real' agency.[14] In response to this, I will argue that (positive) law as a social construct is (conceptually) independent from any perceived 'real' agency, that is, that law can technically regard entities as legal agents even if they are not 'real' agents. The mere technical possibility, however, does not mean that the law should do so. This is addressed in section 4.

Brozek and Jakubiec identify a spectrum of possible positions regarding the legal responsibility of artificially intelligent entities. The two extremes of this spectrum are 'restrictivism' and 'permissivism'. Restrictivism 'denies the possibility of holding autonomous machines legally responsible on purely metaphysical grounds'[15] while permissivism 'imposes no restrictions on the possible legal constructions'[16]. Restrictivism[17] denies the possibility for holding artificially intelligent entities legally responsible on the grounds that they lack essential qualities necessary for legal (and moral) responsibility.[18] Candidates for these essential qualities are consciousness, intentionality and the capacity for intentional action, (libertarian) free will, autonomy, the capacity for deliberation, alignment between one's reasons for action (in the sense of justificatory reasons, not heuristics or causes) and one's actions, and more. In more legal terminology, AI cannot be held responsible because it lacks both Handlungs- and Schuldfähigkeit, that is, the capacity to act and be culpable.[19]

The restrictivist argument[20] indicates that

(P1) An entity lacking *xyz* characteristics cannot be legally responsible.

(P2) Artificially intelligent entities lack *xyz* characteristics.

(C) Artificially intelligent entities cannot be legally responsible

This presumes that certain entities, possessing certain characteristics, are 'real' agents and 'really' responsible and that the law must conceptually coincide with this extra- or pre-legal reality, that is, that law must accurately map this external[21] reality.

This notion that law (and its concepts) must coincide with extra-legal reality and that it is not (technically) possible for law to do otherwise is clearly a misconception. This is supported by the view that (positive) law is a social construct,[22] which makes it technically possible

10 Teubner (n 2); Gunther Teubner, 'Rights of Non-Humans? Electronic Agents and Animals as New Actors in Politics and Law' (2007) 04 *Max Weber Lecture Series*; Beck (n 2).

11 Woodrow Barfield, (n 6). The chapter offers a number of concrete examples of challenges to the current legal situation.

12 Mark Coeckelbergh, 'Artificial Intelligence: Some Ethical Issues and Regulatory Challenges' (2019) *Technology and Regulation*, 31, cf. Mark Coeckelbergh, 'Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability' (2020) 26 *Science and Engineering Ethics* 2051.

13 Emad Abdel Rahim Dahiyat, 'Law and Software Agents: Are They "Agents" by the Way?' (2020) *Artif Intell Law*, 67.

14 I do not here want to attribute this exact misconception to any particular author. I do, however, want to suggest that it is implicit in the argumentation of many. If I am mistaken about this, all the better.

15 Bartosz Brozek and Marek Jakubiec, 'On the Legal Responsibility of Autonomous Machines' (2017) 25 *Artif Intell Law* 293, 294.

16 Ibid.

17 I use restrictivism and restrivists throughout the following sections and attribute certain views to restrictivists/restrictivism. This should not be taken as a claim that all authors that hold some restrictivist views necessarily hold all the views I here describe. As Brozek and Jakubiec (n 15) point out, this is one extreme on a spectrum of possible views and approaches. An uncharitable interpretation of my approach is that I am constructing and arguing against a strawman, but even if no one were to hold a strictly restrictivist view, it is useful to consider the misconceptions this view rests on. Using the extreme for this purpose serves to highlight the misconceptions.

18 Brozek and Jakubiec (n 15), 294.

19 There is variation in terminology and concepts between different legal fields here; I hope readers will forgive the generalisation.

20 This is essentially what Solum calls the "missing-something" argument applied to legal responsibility, rather than personhood: Lawrence B Solum, 'Legal Personhood for Artificial Intelligences' (1992) *North Carolina Law Review* 70 (4).

21 External to the law, in this case.

22 This sentence does not contain a commitment to a positivist concept of law, as non-positivist law theories account for positive law as a social construct as well. Hage, for example, convincingly argues this point in Jaap Hage, 'The Limited Function of Hermeneutics in Law' in David Duarte, Pedro Moniz Lopes and Jorge Silva Sampaio (eds), *Legal Interpreta-*

to give it any content whatsoever. Brozek and Jakubiec describe it as 'quite possible from [a] purely technical point of view, since the law is a conventional tool of regulating social interactions and as such can accommodate various legislative constructs, including legal responsibility of autonomous artificial agents'.[23] Many others have made the same point in a variety of contexts, not limited to the legal responsibility of artificially intelligent entities.[24] Moreover, differences between different legal systems and cultures as well as across time further support this point: here, one can think of criminal responsibility of animals in the Middle Ages,[25] the legal positions of slaves e.g. in times of the Roman Empire or of the legal position of women in Western societies until quite recently.[26] Lastly, another example is the personhood of anything, 'be it monasteries or corporations, governments or ships in maritime law, rivers in New Zealand or India, down to the entire ecosystem in Ecuador.'[27]

This response to the restrictivist claim that legal concepts must coincide with extra-legal reality leaves open the possibility that there are 'real' agents that can 'really' be responsible and other entities that cannot 'really' be responsible because they lack the essential characteristics for 'real' responsibility. All this response posits is that it is *technically possible* to regard an entity as a legal agent, irrespective of whether it is a 'real' agent or not. *Legal* agency is a legal construct.

This leaves room for a counterargument from the restrictivist perspective: while it may be technically possible for the law to construct legal agency any way it wants, it *should not* do so. Instead, the law should only regard those entities as agents that are 'real' agents, and it should only hold those entities responsible that are 'really' responsible. In other words: law should model its constructs after 'real' agents. Generally, this argument proceeds along the following lines: there are a number of characteristics such as intentionality, autonomy, consciousness, or free will, that are required for 'real' agency and responsibility.[28] Because artificially intelligent entities lack these capacities, they cannot 'really' be responsible agents; instead, human beings can and should be held morally and legally responsible –because they meet these conditions and are 'really' responsible agents.[29]

This is the second misconception I will tackle.

4. Misconception b: legal agency should coincide with 'real agency'

The second misconception, that the law should not attribute responsibility to artificially intelligent entities because these entities are not or cannot be 'real' agents or 'really' responsible rests on the assumption, as pointed out above, that there is such a thing as a 'real' agent or 'real' responsibility.[30]

Intuitively, the idea that there are real agents that are responsible for their actions and that we human beings are such responsible agents makes sense: we distinguish between agents – those entities that make things happen and go through the world seemingly independently of physical laws – and non-agents, things like rocks and puddles or other inanimate objects that behave in predictable ways and are clearly and obviously subject to natural laws.[31] We perceive other human beings as agents whose actions are more accurately and more easily explained by their desires and intentions than by physical laws acting upon them. Not only that, but we also perceive ourselves as agents causally responsible for our actions which are shaped not by physical laws acting upon us, but by our desires and intentions – and we often perceive our actions as something we have willed, something that was the result of our wanting and deciding to do something.[32] Moreover, we are responsible for our intentional and free actions. As Solum already indicated in his seminal paper on legal

tion and Scientific Knowledge (Springer 2019) 5.
Of course, a non-positivist might argue that while it is technically possible for positive law to have any content whatsoever, positive law may well be *wrong*. Depending on the specific non-positivist theory, this may go hand in hand with the claim that the positive law is then not law at all, meaning that it is not, in fact, possible for *law* to have any content whatsoever. While section 4 of this paper does not use non-positivist language, I think it can be taken to address this claim with minor (mental) translations by the non-positivist reader.

23 Brozek and Jakubiec (n 15), 303.

24 For example Hans Kelsen, *General Theory of Law and State* (Harvard University Press 1945), 94 and Bartosz Brozek, 'The Troublesome Person' in Visa Kurki and Tomasz Pietrzykowski (eds), *Legal Personhood: Animals, Artificial Intelligence and the Unborn* (Springer 2017), 8 with regard to natural persons, see also Ngaire Naffine, 'Who Are Law's Persons? From Cheshire Cats to Responsible Subjects' (2003) 66 *The Modern Law Review* 346; Ulfrid Neumann, 'Strafrechtliche Verantwortlichkeit Von Verbänden – Rechtstheoretische Prolegomena' in Klaus Volk, Klaus Lüderssen and Eberhard Kempf (eds), *Unternehmensstrafrecht* (De Gruyter 2012), 16 with regard to corporate criminal responsibility. More generally, cf. Alf Ross, 'Tû-Tû' (1957) 70 *Harvard Law Review* 812.

25 Piers Beirnes, 'The Law Is an Ass: Reading E.P. Evans' the Medieval Prosecution and Capital Punishment of Animals' (1994) 2 *Society and Animals* 27; William Ewald, 'Comparative Jurisprudence (I): What Was It Like to Try a Rat?' (1995) 143 *University of Pennsylvania Law Review* 1889.

26 Married women in the Netherlands, for example, could not perform legal acts without the consent of their husbands until 1957. This example is taken from Robert van den Hoven van Genderen, 'Legal Personhood in the Age of Artificially Intelligent Robots' in Woodrow Barfield and Ugo Pagallo (eds), *Research Handbook on the Law of Artificial Intelligence* (Edward Elgar Publishing 2018).

27 Ugo Pagallo, 'Vital, Sophia, and Co.—the Quest for the Legal Personhood of Robots' (2018) 9 *Information* 230, 9. In my view, arguing analogously from personhood to agency is possible (but not vice-versa) because personhood (generally) presumes agency (but not vice-versa).

28 Dorna Behdadi and Christian Munthe, 'A Normative Approach to Artificial Moral Agency' (2020) 30 *Minds and Machines* 195. While there is debate on whether agency presupposes responsibility and distinctions are made between conditions for (moral) agency and (moral) responsibility, I will not consider these questions here and instead talk about 'responsible agents'. Himma, for example, argues that under the standard view (which I turn to in this section), consciousness is a condition for responsibility, but that 'the very notion of agency itself presupposes consciousness in the sense that only a conscious being can be an agent', Kenneth Einar Himma, 'Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to Be a Moral Agent?' (2009) 11 E*thics and Information Technology* 19, 28 and Coeckelbergh, 'Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability' (n 12) holds (for human beings) that 'agency is normally connected with responsibility. You have an effect on the world and on others, and therefore you are responsible for what you do and for what you decide.'

29 Coeckelbergh, 'Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability' (n 12), 2055.

30 This assumption can be found e.g. in Bryson et al (n 3) with regard to legal personhood. Gunkel outlines how under one view, blaming artificially intelligent entities is 'ontologically incorrect', David J. Gunkel, *The Machine Question: Critical Perspectives on Ai, Robots, and Ethics* (MIT Press 2012) 28. Dahiyat (n 13) holds that 'we will commit excessive conceptual mistakes if we attribute a legal or moral responsibility to these agents'; Coeckelbergh, 'Ethics of artificial intelligence: Some ethical issues and regulatory challenges' (n 12) holds that 'only humans can be responsible agents'.

31 Samir Chopra and Laurence White, *A Legal Theory for Autonomous Artificial Agents* (University of Michigan Press 2011) 11; Joshua Greene and Jonathan Cohen, 'For the Law, Neuroscience Changes Nothing and Everything' (2004) 359 *Philosophical Transactions: Biological Sciences* 1775, 1782.

32 Patrick Haggard and Valerian Chambon, 'Sense of Agency' (2012) 22 *Curr Biol R* 390; Patrick Haggard and Manos Tsakiris, 'The Experience of Agency' (2009) 18 *Current Directions in Psychological Science* 4.

personhood for artificial intelligence, '[o]ur understanding of what it means for a human being to function competently has ties to our views on responsibility'.[33] Fischer and Ravizza describe our ordinary concept of moral responsibility as follows:

> 'An important difference between persons and other creatures is that only persons can be morally responsible for what they do. [...] Whereas both persons and non-persons can be causally responsible for an event, only persons can be morally responsible. [...] [I]n order to be praiseworthy or blameworthy a person must know (or be reasonably expected to know) what he is doing, and he must not be deceived or ignorant about the circumstances and manner in which he is doing it. [...] A second type of excusing condition is force. [...] [A]n agent has the type of freedom necessary to be morally responsible only if he has 'control over his actions,' the act is 'up to him,' he was 'free to do otherwise,' he 'could have acted differently', and so forth.'[34]

They also indicate that 'there seems to be a difference between being *held* responsible and actually *being* responsible.'[35] While it may be possible to *hold* artificially intelligent entities legally responsible, one could say, they *are not* actually responsible – and therefore should not be *held* to be.[36]

The understanding of ourselves as responsible agents I have sketched above takes our (subjective) experience and intuitions as central. As such, it can be termed 'phenomenological'. Phenomenology 'address[es] the meaning things have in our experience, [...] as these things arise and are experienced in our 'life-world'.'[37] This intuitive understanding of ourselves as responsible agents is reflected also in philosophy of action and the notion of moral agency in normative ethics, fields that seek to theorise, systematise, and critically reflect on the intuitions that we have and our social and normative practices.[38] Philosophy of action does so with regard to when an event is an action and when an entity is an agent, normative ethics with regard to when an action is right, wrong, good, bad, permissible, or impermissible or, more generally, with the moral evaluation of actions.[39] The standard understanding of action here holds that human beings are (the only) real agents[40] and that something is an action if it is intentional:[41]

> '[i]ntuitively, an agent is something able to take actions. One way to distinguish agents from other entities is that agents do things, as opposed to have things happen to them; to deny something or someone agency is to deny the capacity to take actions, for the actions of the agent distinguish it from the rest of the world. [...] Related to this notion is the concept of self-directed actions or acting for reasons, for the philosophical sense of 'agency' is linked with the ascription of intentions. To possess agency is to be the originator of action, to be driven by motivations, purposes, desires, and autonomously, freely-chosen decisions.'[42]

According to the standard view, 'moral agents must meet rationality, free will or autonomy, and phenomenal consciousness conditions'.[43] Human beings are 'real' agents because we are capable of acting intentionally, freely, and autonomously, and we are 'really' responsible for our intentional and free actions,[44] that is, because we (seemingly) fulfil these conditions. One aspect of this view is what can be termed (naïve) realism about agents and responsibility: the idea that there are 'real' agents irrespective of (moral or legal) agency-ascriptions and that there is such a thing as 'real' responsibility that is different from being *held* responsible on the basis of moral, social, or legal norms.

33 Solum (n 20).

34 John Martin Fischer and Mark Ravizza, 'Introduction' in John Martin Fischer and Mark Ravizza (eds), *Perspectives on Moral Responsibility* (Cornell University Press 1993) 4. Himma (n 28) identifies this as the standard view: 'for all X, X is a moral agent if and only if X is (1) an agent having the capacities for (2) making free choices, (3) deliberating about what one ought to do, and (4) understanding and applying moral rules correctly in paradigm cases.'

35 Fischer and Ravizza (n 34), 18.

36 This is reflected, for example, in Dahiyat (n 13) and Coeckelbergh, 'Ethics of artificial intelligence: Some ethical issues and regulatory challenges' (n 12). See Behdadi and Munthe (n 28) for an overview of this approach when it comes to moral responsibility.

37 David Woodruff Smith, 'Phenomenology' in Edward N. Zalta, *The Stanford Encyclopedia of Philosophy* (Summer 2018), https://plato.stanford.edu/archives/sum2018/entries/phenomenology, 1.

38 Consider e.g. Fischer and Ravizza (n 34), 7: 'A theory of moral responsibility ought to accommodate these standard excusing conditions in the sense that the ascriptions of responsibility entailed by the theory ought to match our ordinary intuitions about when agents are and are not morally responsible.' John Hyman, *Action, Knowledge, and Will* (Oxford University Press 2015), 32 argues that these fields go (even) further than our intuitive understanding in a kind of 'chauvinism' about action.

39 Julia Driver, *Ethics: The Fundamentals* (Blackwell Publishing 2007), 2.

40 Hyman (n 38), 30. Of course, law holds non-human entities such as corporations responsible. This may be permissible under this view because these composite entities are then, in a sense, 'derived' agents: they derive their agency and responsibility from the fact that one or more human beings have acted. This implies that regarding human beings as legal agents and holding them legally responsible rests on their 'real' agency, while regarding composite entities such as corporations or states as legal agents and holding them legally responsible rests on a legal fiction. Conceiving of corporations and states as such 'derived' agents is, under this view, permissible because they are composed of human beings, the paradigmatic, 'real' agents. For artificially intelligent entities, however, this is not the case. In particular in 'human out of the loop'-scenarios, there is no human agent from whom to derive agency and responsibility. Brozek and Jakubiec (n 15) for example, make this point. Cf. also Jiahong Chen and Paul Burgess, 'The Boundaries of Legal Personhood: How Spontaneous Intelligence Can Problematise Differences between Humans, Artificial Intelligence, Companies and Animals' (2019) 27 *Artif Intell Law* 73 regarding spontaneous intelligence.

41 More specifically, that something is an action if it is intentional under some description or if it is identical to or derived from an intentional action. Markus Schlosser, 'Agency' in Edward N. Zalta, *The Stanford Encyclopedia of Philosophy* (Fall 2015), https://plato.stanford.edu/archives/fall2015/entries/agency, para 2. What this means is that if you unknowingly alert a burglar by intentionally turning on the light, alerting the burglar is an action of yours because it is either the same action as turning on the light under a different description (after all, you alerted the burglar by turning on the light) or it is derived from your intentional action of turning on the light. For the purpose of this paper, not much rides on whether an event is an action if it is intentional under some description or identical to or derived from an intentional action; what matters is that intentional action is the fundamental conception of action on this view. Not all philosophers of action take this view. Hyman (n 38), for example, argues that intentionality is not decisive.

42 Chopra and White (n 31), 11 f.

43 Behdadi and Munthe (n 28), 197.

44 This is a broad outline that does not leave room for nuanced differentiation between different theories. For a more elaborate overview on agency, see e.g. Schlosser (n 41) or Matt King and Peter Carruthers, 'Responsibility and Consciousness' in Derk Pereboom and D.K. Nelkin (eds), *Oxford Handbook on Moral Responsibility* (Oxford University Press, forthcoming). An overview of different views related to actions and responsibility can be found in Joseph Keim Campbell, Michael O'Rourke and Harry S. Silverstein, *Action, Ethics, and Responsibility* (Bradford Books 2010) and Fischer and Ravizza (n 34). The standard view of (moral) agency is often contrasted to the functionalist view, under which 'agency requires only particular behaviors and reactions which advocates of the standard view would view as mere indicators of the capacities stressed by the standard view.' Behdadi and Munthe (n 28), 197. I focus here on the standard view, as that is the view underlying the misconception I am addressing.

Given this, the second misconception can be rephrased as follows: *'really' responsible agents exist and law should only ascribe agency and responsibility to those entities that are 'real' agents.* Is it likely, however, that we are 'real' agents' and 'really' responsible in the way our intuitions and the standard view indicate? And are our intuitions and the phenomenological view sufficient basis for making choices about the (legal) ascription of agency and responsibility?

I argue that they are not. My argument rests on insights from the cognitive sciences broadly construed that suggest that the phenomenological view is misguided, particularly as concerns (naïve) realism about agency and responsibility. This implies that our intuitions about ourselves and the criteria for responsible agency are not as strong a justification for choices about the legal ascription of agency and responsibility as we assume. In the following, I briefly touch on a number of different arguments that challenge the distinct 'realness' of human agency.[45]

There is increasing evidence that there are two systems for human decision-making, including moral and legal decision-making: one that is unconscious, fast, and instinctive or automatic, the other conscious, slower, and controlled.[46]

> 'Dual-process theories of thinking and reasoning quite literally propose the presence of two minds in one brain. The stream of consciousness that broadly corresponds to System 2 thinking is massively supplemented by a whole set of autonomous subsystems in System 1 that post only their final products into consciousness and compete directly for control of our inferences, decisions and actions.'[47]

That we sometimes make 'gut decisions' and sometimes carefully consider our choices may not seem particularly radical or challenging to the (phenomenological) view we have of ourselves as agents. What is challenging is the degree to which we make choices unconsciously and to which biases and heuristics play a role in those choices we think we have made rationally and without any other factors at play, according to dual-process theory and the evidence substantiating it. Implicit biases such as racism or sexism have a large impact on our judgments and behaviour, as in this 2005 study:

> 'Subjects were asked to rate the suitability of two candidates for police chief, one male and one female, where one candidate was presented as 'streetwise' but lacking in formal education while the other one had the opposite profile. Despite the fact that Uhlmann and Cohen varied the sex of the candidates across conditions – so that some subjects got a male streetwise candidate and a female well-educated candidate while other subjects got the reverse – subjects considered the male candidate significantly better qualified in *both* conditions. [...] Rather than being conscious of the sexist attitude, the agent is conscious of a confabulated criterion which itself seems plausible – i.e. the importance of being streetwise or highly educated.'[48]

Beyond that, situational factors shape our behaviour in ways we are not aware of, such as a scramble-sentence test including words relating to rudeness makes subjects considerably more likely to interrupt a conversation (67%) than the control group (38%) or those subjects whose scramble-sentence test included words related to politeness (16%); the presence of a briefcase (as opposed to a backpack) triggering more competitive behaviour; or the time since the last food break having significant impact on how judges ruled in decisions relating to prison parole.[49]

Neuroscientific studies have corroborated the dual-process theory and found neurobiological correlates.[50] These insights challenge the presupposition that we are generally rational and that all, most, or even many of our actions are intentional. Further evidence that our intuitions about our own actions and their causes are far less reliable than they seem to us comes from insights related to confabulation. Carruthers indicates that '[t]here is extensive and long-standing evidence from cognitive and social psychology that people will (falsely) confabulate attributions of judgments and decisions to themselves in a wide range of circumstances.'[51] This evidence indicates that we are 'inaccurate in reporting the *causes* of [our] judgments or behavio[u]r' and decisions. For instance, subjects of an experiment instructed to move a finger and to freely decide which finger upon hearing a noise reported that they had decided to move the finger that they moved – but the actual cause of the digit moving was focal magnetic stimulation of areas of the relevant motor cortex areas. These subjects believe that they have acted on the basis of an intentionally made choice, that is, that they are the ('real') agent, but this is a confabulation.[52] Our intuitions about our actions being intentional are not reliable. Specifically with regard to our sense of agency (defined as the experience of controlling one's own actions and thereby events in the world), Haggard and Chambon write that this experience of agency can be tricked and is sometimes illusory.[53]

The assumption that our intention is causally relevant for our actions, that is, that our intentional choices cause, direct, and guide our actions, is further called into question by insights from and following from the Libet experiments. In these experiments, it was found that a 'readiness potential' for action in the brain preceded not only the voluntary movement, but also awareness of the conscious intention to move.[54] Some of these results have been interpreted in such a way that consciousness plays less or even no causal role when it comes to our actions. This is also the conclusion of the social psychologist Daniel Wegner who holds that

> 'each human mind has an abbreviated view of itself, a self-portrait

45 These arguments will necessarily brief and behind each of them is a discussion that cannot be reproduced here in full. My aim here is not to give an exhaustive account of the insights, debates, and nuances; to do so would go far beyond the scope of this paper. The arguments mainly refer to empirical, rather than philosophical insights, although I agree with authors such as Caruso that 'philosophical arguments on their own are sufficient for showing that people are never morally responsible for their actions in the basic desert sense' (Gregg Caruso, 'If Consciousness Is Necessary for Moral Responsibility, Then People Are Less Responsible Than We Think' (2015) 22 *Journal of Consciousness Studies*, 54). I will not reiterate these arguments here.

46 Cf. Daniel Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux 2011); Joshua David Greene, Moral Tribes: Emotion, Reason, and the Gap between Us and Them (Penguin Press 2013); Jonathan St B. T. Evans, 'In Two Minds: Dual-Process Accounts of Reasoning' (2003) 7 *Trends in Cognitive Sciences* 454.

47 Evans (n 46), 458.

48 Caruso (n 45), 52.

49 Caruso (n 45), 54.

50 Evans (n 46), 455.

51 Peter Carruthers, 'How We Know Our Own Minds: The Relationship between Mindreading and Metacognition' (2009) 32 *Behavioral and Brain Sciences* 121, 130.

52 Ibid, 131, reviewing, inter alia, Nisbett and Wilson (1977), Brasil-Neto et al. (1992) and Wegner & Wheatley (1999).

53 Haggard and Chambon (n 32).

54 Benjamin Libet and others, 'Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential): The Unconscious Initiation of a Freely Voluntary Act' (1983) 106 *Brain* 623.

> that captures how it *thinks* it operates, and that therefore has been remarkably influential. The mind's self-portrait has as a central feature the idea that thoughts cause actions, and that the self is thus an origin of the body's actions. This self-portrait is reached through a process of inference of *apparent mental causation*, and it gives rise to the experience that we are consciously willing what we do. Evidence from several sources suggests that this self-portrait may often be a humble and misleading caricature of the mind's operation—but one that underlies the feeling of authorship and the acceptance of responsibility for action.'[55]

These interpretations are debated, particularly when making the strong claim that consciousness plays no causal role whatsoever; nonetheless, they offer further support for the thesis that we are far less intentional and conscious agents than we think and that while we have a feeling of authorship and responsibility, such feelings do not offer privileged information about causal responsibility. Another element of the phenomenological view as outlined above is that unlike rocks, stones, or even more complex ordinary matter such as bees or mice, we have the power to freely bring about one event or some alternative event, that is, the power to do otherwise.[56] This understanding of freedom to choose between events is in conflict with causal determinism and quantum indeterminacy, thereby further calling into question the phenomenological view.[57]

While none of these arguments and insights by themselves prove that the phenomenological view and its notions of 'real' agency and 'real' responsibility are mistaken, they demonstrate that the presuppositions of this view and the intuitions that support it are neither as plausible nor as solid as our (unexamined) intuitions may make them appear. Given this, our intuition that some entities (namely human beings) are 'real' agents which can be 'really' responsible does not provide a good argument against ascribing legal agency and responsibility to other entities (that is, AI) by itself: the insight that our intuitions and our understanding of ourselves – of what causes our actions and decisions – are often based on mistaken confabulations calls into question the phenomenological view and thereby also the normative implications that should attach to it. If we are often wrong about our understanding of ourselves and others as responsible agents, if there are good reasons to doubt the accuracy of our intuitions, why should we attach normative consequences solely to the belief that we are 'real' agents and other entities are not?

To be clear, I am not making an argument that we should disregard our intuitions entirely. I am making the argument that it does not suffice to say 'law should not attribute agency and responsibility to AI *because AI are not 'real' agents or 'really' responsible*'. Instead, it seems to me that a normative argument that does not rely solely on our – likely mistaken – intuitions and reference to the phenomenological view is required.[58] What could such an argument look like? One example can be found in Brozek and Jakubiec who argue that while it is possible for law to attribute agency and responsibility to AI, it should not do so because this would take law too far from the life-world and therefore, any such rules would remain 'law in book' rather than 'law in action'.[59] This is an argument from legal efficacy and our intuitions and phenomenological view. Whether it is the case that any such rules would be inefficacious is an as of yet unanswered empirical question.[60] This argument demonstrates, however, that to call into question the phenomenological view's presuppositions does not necessarily mean negating or disregarding the fact that people do have the intuitions that feature in the phenomenological view. Instead, the demand for an argument that goes beyond the phenomenological view indicates a different place for these intuitions in the argument: they are empirical information that needs to be embedded in a normative argument, instead of indicators of absolute, external truth.

Another example of a normative argument of the kind I have in mind as necessary in the debate whether law should attribute agency and responsibility to AI is the following:

> '[A]scribing responsibility to software agents might hide the real source of the problem, mask the human creator of the harm, and might also be used as an excuse for some people to evade their responsibility and behave recklessly.'[61]

This argument, found more frequently in the literature,[62] can be rephrased as follows: law should not attribute agency and responsibility to artificially intelligent entities because to do so would allow other (human or corporate) entities to escape responsibility in cases in which they (the human or corporate entities) should be held responsible.

This brings us to the third misconception I want to address in this paper.

5. Misconception c: hiding behind AI responsibility

There are (at least) two possible ways to demonstrate that it is a misconception to believe that holding AI responsible would necessarily allow other agents to escape responsibility: this can be demonstrated by looking at (conceptual) possibility and by looking at current legal practice.

The first approach to the second misconception relates back to the point made in section 3. of this paper: (positive) law is socially con-

55 Daniel M. Wegner, 'The Mind's Self-Portrait' (2003) 1001 *Annals of the New York Academy of Sciences* 212. Wegner holds further that '[e]xperiences of conscious will thus arise from processes whereby the mind interprets itself – not from processes whereby mind creates action. Conscious will, in this view, is an indication that we think we have caused an action, not a revelation of the causal sequence by which the action was produced.' Summary taken from Daniel M. Wegner, 'Frequently Asked Questions About Conscious Will' (2004) 27 *Behavioral and Brain Sciences* 679; see also Daniel M. Wegner, 'The Mind's Best Trick: How We Experience Conscious Will' (2003) 7 *Trends in Cognitive Sciences* 65; Daniel M. Wegner, *The Illusion of Conscious Will* (MIT 2002).

56 Fischer and Ravizza (n 34), 8.

57 Fischer and Ravizza (n 34) offer an overview of this incompatibility as well as the different positions that have been taken in the debate, mainly libertarianism and compatibilism. See also "Jaap Hage and Antonia Waltermann, 'Responsibility, Liability, and Retribution' in Bartosz Brozek, Jaap Hage and Nicole Vincent (eds.), Law and Mind: A Survey of Law and the Cognitive Sciences (Cambridge University Press 2021).

58 The call for a normative approach when it comes to the (in this case moral) responsibility of artificially intelligent entities can be found also in Behdadi and Munthe (n 28). The arguments leading to the conclusion of their article and mine strike me as compatible and can be read in conjunction.

59 Brozek and Jakubiec (n 15), 293.

60 My intuition on this question is a different one than that of Brozek and Jakubiec: I believe such rules could very well be(come) efficacious, in part because it seems to me that we take the intentional stance quickly, in part because law influences our life-world. Cf. S. Marchesi and others, 'Do We Adopt the Intentional Stance toward Humanoid Robots?' (2019) 10 *Front Psychol* 450.

61 Dahiyat (n 13), 69.

62 Bryson et al (n 3) consider it the main case of potential abuse and (rightly) point out that lawmakers must provide solutions for this. See also Gunkel (n 30).

structed. Its rules are created (be it by legislators such as parliaments, or by judges), which means that we (read: our law creators) can set up the system in such a way that it works for us,[63] as well as change it if it has adverse effects or does not lead to the desired results.[64]

Accordingly, it is – technically, in theory – possible to attribute agency and responsibility to more than one entity. Whether this is desirable and for what reasons it is or is not desirable cannot be addressed in this paper but understanding the ontological nature of agency and responsibility (both within and outside of the law) as a social construct allows us to understand the degree of control that we (or in this case: our lawmakers) have over the situation.

In how far is it necessary to adapt existing laws and legal concepts to do so?

When it comes to tort liability, the law already knows circumstances in which more than one entity is regarded as the tortfeasor. Landes and Posner distinguish between 'simultaneous' and 'successive' joint tort: the first covering those cases where 'the victim suffers a single or indivisible injury as a result of the tortious activity of two or more parties',[65] and the second covering those cases where 'one tortfeasor aggravates an injury inflicted by the other, as where a driver negligently hits a pedestrian and a physician negligently treats, thereby aggravating, the pedestrian's injury'.[66] In the Principles of European Tort Law,[67] Title V outlines rules for multiple tortfeasors, either under solidary or under several liability:[68]

Art 9:101 Solidary and several liability: relation between victim and multiple tortfeasors

1) Liability is solidary where the whole or a distinct part of the damage suffered by the victim is attributable to two or more persons. Liability is solidary where:

   a) a person knowingly participates in or instigates or encourages wrongdoing by others which causes damage to the victim; or

   b) one person's independent behaviour or activity causes damage to the victim and the same damage is also attributable to another person.

   c) a person is responsible for damage caused by an auxiliary in circumstances where the auxiliary is also liable.

2) Where persons are subject to solidary liability, the victim may claim full compensation from any one or more of them, provided that the victim may not recover more than the full amount of the damage suffered by him.

3) Damage is the same damage for the purposes of paragraph (1)(b) above when there is no reasonable basis for attributing only part of it to each of a number of persons liable to the victim. For this purpose it is for the person asserting that the damage is not the same to show that it is not. Where there is such a basis, liability is several, that is to say, each person is liable to the victim only for the part of the damage attributable to him.

These already existing conceptual tools could, it seems to me, be employed to prevent a situation in which corporate or human agents escape liability, although outlining the specific form this should take goes beyond the scope of this paper.[69] When it comes to criminal liability, it is similarly true that more than one person can be liable as principal, with notions such as joint perpetration, perpetration-by-proxy, instigation, and aiding further delineating situations of multiple agents.[70] However, in criminal law, matters are made more complicated by the fact that some legal systems construe the act requirement for criminal liability more stringently and at times less explicitly normatively than when it comes to tort or other liability, such as Germany regarding corporate criminal liability, for example.[71] This is a subject for another paper and cannot here be addressed. Equally, tort liability and criminal liability are not the only liability regimes that one could and should consider when it comes to responsibility of artificially intelligent entities.[72] For present purposes, however, it suffices to say that there are means, both when it comes to *lex lata* and *lex ferenda*, to ensure that attributing legal responsibility to artificially intelligent agents does not allow other agents, human or corporate, to escape responsibility.

This demonstrates that it is not *necessarily* true that AI responsibility would preclude the responsibility of other agents. Whether AI should be held responsible and the most suitable means of implementing such responsibility in practice if it is found to be desirable are important matters for both academic and political discussion, but not the aim of this paper. In this paper, I only seek to address a limited number of misconceptions, not give all-things-considered recommendations or conclusions.

6. Conclusion

This paper has addressed three misconceptions regarding the legal agency and responsibility of artificially intelligent entities: first, that law cannot attribute agency and responsibility to such entities because they are not 'real' agents or 'really' responsible; second, that it should not do so for the same reason; third, that if the law were to attribute agency and responsibility to such entities, it would allow other (human or corporate) agents to escape responsibility, while

63 It is more complicated than that, of course: what rules will have what impact is at times very difficult to predict. Moreover, parliaments are not single entities but composed of different individuals belonging to different political parties, which may pursue different aims. And so on. Nonetheless, the general point stands.

64 Beck (n 2) offers different possibilities, including some discussion of advantages and disadvantages.

65 William M. Landes and Richard A. Posner, 'Joint and Multiple Tortfeasors: An Economic Analysis' (1980) 9 *The Journal of Legal Studies* 517, 518. This can be further divided into 'joint care' and 'alternative care' cases, that is, cases in which both parties have to take care to avoid the damage occurring, and cases in which it would be sufficient if only one party had taken care.

66 Ibid, 518.

67 While the Principles of European Tort Law are non-binding guidelines, they try to merge different traditional approaches with a modern perspective on how the law of torts should develop in the future and as such provide a good exemplification of what concepts of tort law exist and may be implemented in the future in Europe.

68 For a comparative law overview of multiple tortfeasor liability in Europe, W. V. H. Rogers and W. H. van Boom, *Unification of Tort Law: Multiple Tortfeasors* (Kluwer Law International 2004).

69 Cf. Lewis A Kornhauser and Richard L Revesz, 'Sharing Damages among Multiple Tortfeasors' (1989) 98 *The Yale Law Journal* for a law and economics approach to different liability regimes and their potential effects in situations involving multiple tortfeasors.

70 Cf. Laura Peters, *Acting Together in Crime* (Eleven International Publishing 2018).

71 The German view is that corporations can neither act nor be culpable and that they lack the capacity for both. Therefore, Germany does not know corporate criminal liability. Instead, an administrative (quasi-criminal) approach is used. David Roef (2019) 'Corporate Criminal Liability' in Johannes Keiler and David Roef (eds) *Comparative Concepts of Criminal Law* (Intersentia 2019).

72 The possible contractual liability of artificial agents should not be disregarded, for example; the possible legal responsibility of autonomous weapons in humanitarian law situates questions of agency- and responsibility-ascription (also) in the international legal sphere.

they should be held responsible.

Given that (positive) law is a social construct, it is clearly technically possible for law to attribute agency and responsibility to artificially intelligent entities. Legal historical and comparative legal research shows that this has been done; legal theory demonstrates why it can be done. However, the mere technical possibility does not mean it should be done. The second misconception argues that agency and responsibility should be attributed to 'real' responsible agents, presupposing that there are such 'real' and 'really' responsible agents. This presupposition, I have argued, fits with the phenomenological view of the world and our place in it, as well as the standard view on agency and responsibility: we (human beings) are the paradigmatical responsible agents because we possess consciousness, intentionality, and rationality. However, insights from the cognitive sciences demonstrate that the presuppositions of this view and the intuitions that support it are neither as plausible nor as solid as we may assume. Given this, I have raised the question why we should attach normative consequences to the belief that we are 'real' agents and other entities are not in itself? The view that our intuitions about 'real' agency are not in themselves sufficient basis for refusing to attribute agency and responsibility to artificially intelligent entities does not necessitate disregarding these intuitions; they can inform normative arguments and be embedded in them.

A normative argument against attributing legal agency and responsibility to artificially intelligent entities is that it would allow other agents (human or corporate) to hide behind the artificially intelligent entities and escape responsibility that way, while they should be held responsible. However, understanding that (legal) agency and responsibility are constructed also means that who is regarded as an agent in law and held responsible can be changed in such a way as to produce the desired consequences. This includes the possibility to hold both artificially intelligent agents *and* human and/or corporate agents responsible *at the same time*. Investigating whether this should be done and if so, what form this should take goes beyond the scope of this paper, but there is no technical or conceptual impossibility to do so.

Artificially intelligent entities pose a challenge for policy- and lawmakers due to the accountability gap they create. This paper has addressed three misconceptions in debates about one possible means to close the accountability gap, namely, to regard artificially intelligent entities as agents responsible for their own acts. As such, the explicit scope of this paper has been relatively narrow. Nonetheless, I think that implicitly, this paper also demonstrates another challenge that artificially intelligent entities pose (for policy- and lawmakers, scholars, citizens, and so on): by investigating how (legal) concepts do (or do not) apply to artificially intelligent entities, we have to address our assumptions about ourselves and our place in the world, especially where these are not as accurate as we have long thought. This requires intellectual humility[73] but at the same time, understanding the ontological nature of (legal) agency and responsibility, both that of artificially intelligent entities and ourselves, as a social construct allows us to understand the degree of control that we (or in this case: our lawmakers) have over the situation. It shows us the freedom we have to shape and create practices of agency and responsibility that suit our (normative) goals. Thus, there is more flexibility in the construction of responsibility of artificially intelligent entities than one might assume, which offers freedom to law- and policymakers, but also requires openness and creativity as well as a clear, normative vision of the aims we and they want to achieve.

73 Cf. Kathryn Schaffer and Gabriela Barreto Lemos, 'Obliterating Thingness: An Introduction to the "What" and the "So What" of Quantum Physics' Foundations of Science' (2019) *Foundations of Science*.