Automative Agency and Responsibility

Timur Cengiz Uçan[0000−0001−7620−6601]

[1] Bordeaux Montaigne University, 33607 Pessac, France
[2] UMR *Sciences, Philosophie, Humanités*, 33170 Gradignan, France
timur.ucan@u-bordeaux-montaigne.fr
https://timurcengizucan.net

Abstract. Increase in uses of variously designed automated artificial agents renders pressing the resolution of conceptual and practical problems related to involvements of responsibilities and correlative juridical liabilities. Can an automated artificial agent be held responsible and eventually juridically liable for the realization of an action? In some cases, to ascribe juridical personhood to an automated artificial agent might seem required to render some actions judicable. But in some cases, to refuse juridical personhood to such agents might seem required to account for the responsibilities and juridical liabilities of their conceivers, producers, and users. Floridi summarized different approaches to address this problem (2023, 113-135). The first consists in ascribing juridical liability to such agents (Hallevy, 2011, 12-17). The second, in considering such agents as means of actions (Hallevy, 2011, 4; Pagallo, 2017, 21). The third involves integrating the importance of relevant knowledge to assess actions realized by such agents (MacAllister, 38-40). And the fourth involves evaluating the sufficiency of risk assessments of their developers to account for the judicability of actions achievable by such agents (Hallevy, 2011, 7-12). In this article, I argue that relative congruency of these approaches should be further integrated to address the tension generated by apparently mutually incompatible constraints relevant to relatable yet different contexts. For this purpose, I argue that Turing's account of the involvements of thought ascriptions to machines (1950) and Wittgenstein's criticism of 'private language' (Uçan, 2024; Wittgenstein, 2009) should be integrated to think the derivativeness of action ascriptions to automated artificial agents.

Keywords: Agency · Automation · Private Language · Responsibility.

1 Introduction

Accounting for the actions of artificial agents, among which those constituted by or equipped with artificial intelligence systems, involves addressing a tension with respect to our conceptions of the relations of responsibility and agency. If one treats an artificial agent as a person in a juridical sense, one removes responsibility of its action/s from the one/s of its creator/s or user/s. On such approach, whatever is the truly imputed considered action, the artificial agent is sole responsible for its realization. An action of an artificial agent can then be juridically evaluated, but not a correlative action of its creator/s or user/s. On the contrary, if one does not treat an artificial agent as a person in the juridical sense, one removes responsibility of its action from the artificial agent. On such approach, whichever is the truly imputed considered action, the creator/s or user/s is or are the sole responsible/s for its realization. But then, as there is no juridically assessable action of the artificial agent, the extent to which the action of its creator/s, or user/s can be judged remains unclear. And we accordingly could be lead to the deceptive would-be conclusion according to which neither an action of the artificial agent nor an action of its creator/s, or user/s can be judged at all. Contemporary approaches addressed this tension by arguing in favor of either direct or indirect ascribability of actions, and correlatively, in favor of either the direct or the indirect responsibility of artificial agents. On such approaches, whether to attribute or to defer juridical responsibility of an action to an artificial agent involves an initial mediation by an opposition between the mediate and the immediate, the direct and the indirect. However, the presupposition of such an initial and unavoidable opposition tends to reiterate the problems it was meant to solve. Surely, some artificial agents or instrumental complexes can start realizing actions which undeniably have causal consequences. And directly ascribing responsibility to such agents responds to and satisfies to an extent the need for identifying the initiator of such an action. But thusly effectuated, such ascription generates the problem of its own *truthfulness*, as the intentional character of the initiation of such action, legitimately expectable to account for its judicability, cannot fail to remain *merely arbitrary*. By contrast, the actions of such artificial agents can be also be considered as means for non-artificial initiators. And indirectly ascribing responsibility to such agents responds to and satisfies the need for identifying the *intentional* initiator of such actions. But thusly effectuated, such ascription generates the problem of its own *sufficiency*, as the causal responsibility involved by the initiated action, legitimately expectable to account for its judicability, cannot fail to remain *indecisively* questionable. A finer-grained account of the mutual involvements of responsibilities of artificial and non-artificial agents is thus required. To achieve this objective, I first propose a study of contemporary approaches to responsibility-determination in the case of actions involving automated artificial agents. I then argue that model-based approaches rely upon a reductive and privatist conception of mind as occluded consciousness, that can and should be criticized both to avoid oversimplifying responsibility attributions in complex cases involving collective responsibilities, and also to preserve the open-endedness which is required for almost unpredictable further complexities which shall result from interlaced technological and social developments.

2 The liabilities of artificial agents

2.1 The opposition between direct and indirect conceptions of the liabilities of artificial agents

The increasingly rapid conceptions, developments, productions and deployments of artificial agents equipped with artificial intelligence systems, which *somehow* can take decisions or act independently from other agents, generate many unprecedented contexts, occasions and events. Reflexion to assess and judge intentional and unintentional individual or collective actions and consequences that involve artificial agents thusly became necessary. As much as possible, no legal vacuum could be left about such possibilities of action to which are internally related human lives, ecosystems, and which are at the core of major interests of private enterprises. Peculiarly, the question whether an artificial agent is liable, that is, *legally accountable* for eventually relevantly and correctly ascribable actions, became of central importance. One central tension and correlative opposition can be discerned among responses which can be provided to the question whether an artificial agent is liable. This tension is between conceptions according to which an artificial agent is directly liable and the conceptions according to which an artificial agent is indirectly liable.

Indirect conceptions of the liability of artificial agent respond to deep and central requirements and aspects of ordinary knowledge of human agency in relation to tools or machines. Tools or machines are means for actions realized by human agents. Artificial agents are to this extent relatively indistinct from tools or machines. For, artificial agents also should be considered as means used by human agents to realize actions. Correlatively, an artificial agent can be held *accountable* for one's actions, and can be considered as a *legitimate source* of moral actions (Floridi and Sanders, 2004) of which the considered artificial agent can be causally *responsible*. But the artificial agent cannot be considered as *morally responsible* for the realization of a *moral action* (Floridi, 2023, 209). It is the human agent whose legal responsibility is somehow involved by the artificial agent who is liable, that is, legally responsible for the action of the artificial agent. In this sense, the artificial agent is not directly, but *indirectly* liable through the direct liability of its creator/s, designer/s, producer/s or user/s.

> This liability model does not attribute any mental capability, or any human mental capability, to the AIUV.[3] According to this model, there is no legal difference between an AIUV and a screwdriver or an animal. When a burglar uses a screwdriver in order to open up a window, he uses the screwdriver instrumentally, and the screwdriver is not criminally responsible. The screwdriver's "action" is, in fact, the burglar's. This is the same legal situation when using an animal instrumentally. An assault committed by a dog by order of its master is, in fact, an assault committed by the master. (Hallevy, 2012, 6).

[3] Artificial Intelligence Unmanned Vehicles are vehicles equipped with artificial intelligence systems. Some of these vehicles, as for example "autonomous cars", can carry their drivers, while others, as some "unmanned underwater vehicles" or "UUVs" are not designed to carry anyone.

Interestingly enough, on this approach, there is no such thing as the action of the artificial agent. For, inasmuch as the realization of the action of the artificial agent is dependent upon the realization of an action *by* a (human) agent, the action of the artificial agent can be explained out as an action or sub-action of the (human) agent. Hallevy's powerful and deep remark concerns a deep aspect of intentional action, to which we shall come back. At this stage, let us just remark that the distinction between means and ends is sharp, that means could not act, and that agents use means to achieve their ends.

However, the very powerful and apparent intuitiveness of the indirect conception of the liability of artificial agents does also constitute its major weakness. For then, nothing excludes that the ascription of an action of an artificial agent to a human agent strictly implies the reduction of every action-achievement to actions of human agents. That is to say, on this approach, strictly speaking, only humans can achieve (intentional) actions, and no non-human agent (as non-human animals) could achieve any action. Let us then underline that the establishment of these distinctions is relative and necessarily compatible with the ends of courts of law, which are those of establishing and judging some actions realized by persons as human agents. To this extent, at least some of the versions of the indirect conceptions of the liability of artificial agents are not strictly incompatible with the direct conceptions of the liability of artificial agents.

By contrast, direct conceptions of the liability of artificial agent respond to a falsely naïve aspect of the action of artificial agents. We somehow need to be able to correctly ascribe actions to artificial agents if we are to judge their liability. Even if the achievement of actions by artificial agents is dependent upon the achievement of actions by (human) agents, we need to be able to distinguish among actions, those which are realized by (human) agents, from those which are realized by artificial agents. And we need to be able to correctly attribute any such action to an artificial agent. For example, not only that we need to be able to think that some AIUVs can bring you to a desired place, but also, identify the given AIUV which has brought you to a given place. Or, to provide the example of a case which concerns the establishment of the juridical liability of an artificial agent, we need to be able to judge that an artificial agent harmed or killed a (human) agent (Hallevy, 2012, 8). Suppose for example that the decision taken by the artificial agent clashed with the decision taken by its (human) user, and that the action realized by the artificial agent consequently to this decision lead directly or indirectly to the death of a (human) agent. Even if only to accuse, charge or blame one or several (human) agent(s) (as for example, in the imagined case, the private enterprise which produced the AIUV), the correct attribution of an action to an artificial agent is required. That is to say, even if only as a transitional step towards the establishment of the liability of a (human) agent, the liability of an artificial agent remains required. And remark that this is an unproblematic aspect of our ordinary practices. Even if we never contributed to the conception or the production of such system, we can remark that the operating system of a computer can update itself, that the fridge can keep food and beverages cold, that some robots can deliver pizzas as others clean the room

and others cut the grass of the yard. Each of these ordinary remarks involve that we unproblematically ascribe actions to artificial agents, and not just any one to any one, but exactly those for which these artificial agents have been conceived, produced, and used, to those artificial agents.

That is exactly the reason for which, as remarked by Hallevy, it is so important to correctly assess the liability of artificial agents. For, cases for which the assessment of legal responsibilities are done do not restrict to cases of intentional uses of artificial agents by (human) agents against other human agents, or artificial agents of other human agents. On the contrary, harm can result from actions of artificial agents which neither result from the intentional action of any human agent, nor result from an action that belongs to the range of actions for which these artificial agents were conceived, produced and used. The case in which an AIUV kills its driver due to a clash between its decision and the one of its driver needs to be analyzable as a *consequence* of an action of an artificial agent, rather than as the *aim* of its action.[4] It is the artificial agent whose legal responsibility is somehow involved by the (human) agent that is liable, legally responsible, for its own action. That is to say, it would be misleading to consider that the artificial agent is indirectly liable through the direct liability of its creator, producer, or user. The artificial agent is *directly* liable, in the sense that its own direct liability is involved by one's contribution to the realization of an action.

We so far considered an opposition and correlative tension central to conceptions of the liability, the legal and juridical responsibility of artificial agents. The opposition is between the direct or immediate and the indirect or mediate conception of the responsibility of artificial agents. And such opposition is relatively intelligible inasmuch as *merely* rejecting any incompatibility between the two sorts of conceptions would just amount to miss that requirements of epistemic standards fluctuate across contexts and through practices. Misleading would be to confuse the direct responsibility of a user of an artificial agent (e.g. a military drone equipped with weapons and an artificial intelligence system) in the killing of another person with the indirect responsibility of the used artificial agent. And equally misleading would be to confuse the direct responsibility of such an artificial agent (previously *imagined, conceived*) with the indirect responsibility of the creator or producer of the artificial agent which self-used itself.

2.2 Fluctuations of epistemic demands concerning liability-assessments standards across contexts

The consideration of the opposition between the direct or immediate and indirect or mediate conceptions of the liabilities of artificial agents thusly involved distinguishing several ranges of cases to which these conceptions (or "models") apply (Floridi 2023; Hallevy 2012; Pagallo 2017). Consideration of further contexts and

[4] For, consideration of the existence of artificial agents conceived, produced and used *in order to harm, kill or destroy* agents (artificial or not), suffices to obtain relevant contrast (on this see McAllister, 2017).

variations of epistemic *standards*[5] of assessment across such contexts is nevertheless required to provide a more accurate account of the liabilities of artificial agents. Indeed, the need and the relevance of liability-assessments standards is relatively contextual. Such standards are appealed to and applied in diverse contexts according to diverse needs relative to planned, ongoing or achieved actions. On the basis of the earlier brought out tension, we can further distinguish two interrelated yet distinct ranges of cases, imagined or not. In the first – imagined – range of cases, epistemic standards of liability assessments are to be raised due to high risks of criminal deviation related to the *intentional* use of artificial agents in strictly regulated and controlled practices, as that of the interrogation of persons in the context of investigations (McAllister, 2017). In the second, epistemic standards of liability assessments are to be raised due to the eventually disastrous *consequences* of the un/intentionally *neglectful* uses of artificial agents by their users (Hallevy, 2012, 7-12). Consideration of these two ranges of cases suffices to remark that the epistemic standards of liability-assessments which apply to artificial agents independently from (human) agents, and those which apply to (human) agents dependently on the possible actions of artificial agents, can be raised in eventually complementary, but also, independent ways which can radically transform relations, institutions, and societies.

The first range of cases draws our attention to the risk of the conceivable deterioration of our concept of liability and of the impartial application of juridical procedures as a result of an unprecautious utilization of artificial agents to apply procedures usually applied by (human) agents. Criminal ab/uses (as torture by robo-interrogators, to use McAllister's example) could result from the unprecautious uses of artificial agents to interrogate persons. To this extent, the autonomous actions of competent human agents in the preservation of the impartiality of basic treatment-conditions of (human) agents are of primary importance. One could have expected that the consideration of a transfer of competencies from human to artificial agents would result in the increase of the liabilities of artificial agents. But in fact such consideration reinforces the human responsibilities of human agents, and thereby constitute a further reinforcement of the argument in favor of the indirect conception of the liabilities of artificial agents.

For, as remarked by Floridi (2023, 133-135) the "perpetration-by-another" model and the "command responsibility" model are compatible. And in fact the relations of these 'models' are tighter than a mere compatibility relation. Inasmuch as liabilities are concerned, we studied that the use of an artificial agent by a person can legally be treated as the (instrumental) use of a means by that person to achieve an end, such that, as a result, the (human) person but not the artificial agent is correctly assessed to be responsible. And obviously, some of the ends that can conceivably be achieved by means of the use of a means are legally forbidden. Thus, some conceivable uses of artificial agents, peculiarly those which can be made of artificial means in attempts to achieve forbidden ends are also forbidden. Such considerations obviously also concern the

[5] By contrast with *models*.

liabilities of officers who can give orders by means of which they can prescribe, impose, force some uses to achieve ends. In such cases, both the orders and the uses can be considered as instrumental means for the achievement of such ends. And as earlier mentioned, neither every end, nor every means to every end are legally authorized. Thus, a person who is responsible for the commandment of a group of person can be expected to competently predict the consequences of the applications of one's own commandments, give appropriate orders, and take responsibilities for given orders. And an instance of the international criminal court would, if required, precisely take into consideration the production of orders whose applications resulted into war crimes to judge the eventual war crimes of their author(s). Thusly approached, the article 28 of the Rome statute (2021) could be read as implicitly containing the legal means for an appropriate legislation of the new ranges of cases which have or are going to emerge with the conception, the development, and the production of artificial agents equipped with artificial intelligence systems. However, as argued by McAllister:

> (...) an Additional Protocol should unequivocally state that the use of robots in interrogative spaces by state actors may foreseeably result in human rights violations of which human indirect perpetrators will be held liable. (2017, 2571)

Such protocol could be added to the Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment (1984) to preventively forbid uses of artificial agents to indirectly achieve human rights violations which could not be legally achieved otherwise. Indeed, independently from the philosophical and conceptual problem, it is the eventuality of the practical bypass of international regulations which constitutes the first technical problem to which a practical resolution needs to be provided. However, if the contexts considered so far involve extreme violence and destruction, less extreme, yet distinctively destructive and potentially much more widespread and common misuses and consequences of actions of artificial agents are conceivable. The correlative elevations of epistemic standards of assessments of liabilities and contexts are those that the creators, developers, designers, and producers of artificial agents, and peculiarly those equipped with artificial intelligence systems, are demanded to be responsive to. (Human) Agents involved in the conception, the development, the production, and the deployment of artificial agents are liable of eventual consequences of their uses in ways which are different from those of mere or simple users of artificial agents.

Producers and users are in asymmetric relations with respect to artificial agents. It is relevantly expectable from the creators or producers of an artificial agent to have foreseen ways in which the artificial agent can dysfunction, and correlative ways for emergency situations to be satisfactorily brought to end. But the same does not apply to users in most cases. And precisely for this reason, negligence concerns first creators, developers and producers rather than users. The gravity of the risks involved by the use of an artificial agent is in many cases determinative of negligence-ascription distributions. Negligence by the user of the knowledge of the functionings of, say, the autopilot system of an airplane is likely to have disastrous consequences. Negligence by a user of

the knowledge of the functionings of an evolved heating system equipped with a thermostat and an artificial intelligence system is much less likely to have similar consequences. Nevertheless, even in cases in which increased attention and knowledge is required from the user, the liabilities of the creator, designer, producer and the user of an artificial agent could not be equivalent. In cases in which relatively autonomous decisions can be taken by an artificial agent, the liabilities of the creator or producer, and those of the user further diverge.

The elaboration of a more pervasive account of the considered fluctuations of epistemic demands thusly involves the consideration of a central range of cases. Concerned cases are the ones of (unwanted) consequences resulting from the achievement of an action by artificial agents and which were not intended by its creators or users.[6] In such cases, there is no chain along which liability transmits, as the one between, an officer and a soldier. However, the creators or the producers of the device possibly can be held accountable for the consequences which resulted from the action of the artificial agent (Hallevy, 2012, 7-12). Negligence of the consideration of a pertinently conceivable dysfunction of an artificial agent during its conception and production can thus result in the attribution to the creator or the producer of the responsibility of its action/s. Remark that although what is called the "natural-probable-consequence liability model" does not concern actions of artificial agents *intended* by non-artificial agents, such approach potentially concerns every action or consequence of an action of artificial agents which was not *intended*. That is to say, suppose that creators or users intended to use an artificial agent to achieve a legally forbidden aim. And further suppose that for unpredicted motives, the artificial agent failed to achieve the legally forbidden aim, but, achieved a different legally forbidden aim. In such case, the artificial agent did not achieve an action the creator or user wanted the artificial agent to achieve. And, the artificial agent did achieve an action the creator or user did not want the artificial agent to achieve. But although the action achieved by the artificial agent is different from the forbidden action initially planned by the (human) agent, the human agent is responsible for the unintentional, yet achieved, realization of a different forbidden action by the artificial agent.

The relevance of liability-assessments standards is to this extent determinately contextual. The occurrence of grave events motivate the elevation of liability assessments standards. And the fluctuations of the demands for applying such standards of evaluations are mainly relative to preventive concerns, whether legal preventive concerns to forbid practices with unwanted consequences, technical preventive concerns to predetermine ways for technical problems to be solved without unwanted consequences, and juridical concerns about the determination of the liabilities of involved persons after the occurrence of an unwanted event.

[6] As in the earlier considered example of the killing of a driver by the AIUV as a result of a clash of decisions between the driver and the artificial agent.

2.3 Philosophical and methodological difficulties related to reductionism and functionalism

We considered that the opposition between the direct or immediate, and, indirect or immediate conception of the liabilities of the artificial agent is an opposition between relatively incompatible ways of solving the issue of the contextuality of liability-ascriptions. Both (sorts of) conceptions solve the problem of the judicability of actions achieved *by* or *by means* of artificial agents, although in opposite ways. However, the position of this opposition of conceptions, and the correlative debates can and should be criticized for epistemological motives. Indeed, a shared supposition of the debated and opposing conceptions is that artificial agents – machines, could *lack* or *be provided* with a mind or consciousness. Correlatively, such agents would relevantly be characterizable, through their *mindlessness* or their *mindedness*, as *capable* or *incapable* of intentionally achieving actions achieved by *humans*.

Such presupposition is common both to direct and indirect conceptions of the liabilities of artificial agents. For example and notably, when Floridi introduces the direct conception of the liability of artificial agents he presents Dennett's approach as the idea that "for practical purposes we may treat any agent as if it possesses mental states" (2023, 133). That is to say, independently from the eventual answer to the question whether an artificial agent 'possesses' mental states, we non-artificial agents, could treat artificial agents *as if* such agents do 'possess' mental states. On this approach, to be minded involves to 'possess' states of mind. Floridi rightly underlines the problem that is in fact involved by the reductionist move involved by Dennett's approach. Surely for practical purposes we might consider artificial agents *as* minded agents. However, is this so clear that we are not going to thereby neglect and miss aspects of artificial agents, of our agency, of our relations with artificial agents, and of events which involve the interrelated actions of artificial and human agents? Can we really, for some practical purpose, negate without confusion any distinction between artificial agents and minded agents? Further, the supposition that the eventual mindedness of artificial agents could be thusly established generates the obverse problem that the mindedness of non-artificial agents could be eventually similarly revoked, an eventually even more problematic consideration. Finally, problematic aspects of the supposition that minded agents 'possess' mental state remain uninterrogated and unenvisaged.

To this extent, both the reductionism that is involved by the conception according to which artificial agents are minded and capable of intentional actions, and its negation, which grants its relevance, are problematic and generate opposites difficulties to the resolution of the problem of the liabilities of artificial agents. Reductionism with respect to mind (as *reducible* to its operations) provides a practical resolution of the problem of the liabilities of artificial agents, but leaves the conceptual problem unresolved. By contrast, the negation of reductionism provide satisfactory elements of resolution of the conceptual problem, but leaves the practical problem unresolved. A problem cannot but raise in relation to the supposition that artificial agents *can* or *cannot* be morally responsible.

Asymmetries between agents and possibilities of action are presented as impossibilities at the occasion of the comparison of humans and artificial agents (Uçan 2024 108; 115). And these pseudo-impossibilities are then further thematized and presented as incompatibilities between models. On these approaches, models quasi-incorporate the impossibilities derived from the comparison of artificial agents with non-artificial agents. Mind or consciousness is then conceived as element/s which supplement/s a body, and artificial agents as *mindless* agents, that is, agents *deprived* of *mind*, *soul*, *consciousness*, the element/s which would supplement a body or bodies. A common aspect of these conceptions or 'models' of mind is their privatism. For then mind should be conceivable as internal occlusion of consciousness. However, and for motives yet to be considered, privatism could not have turned out coherent.

3 The responsibilities of automated artificial agents

3.1 Turing and the disjunctive entrapment

Rapidly providing definite answers to arising juridical and practical problems related to the use of automated artificial agents might thus have turned out more difficult than initially assumed. Problems such as the one of the ascribability of a legal personhood to artificial intelligence systems attest of such difficulty. Interestingly enough, professionals of juridical instances are not necessarily reluctant even to the idea that artificial intelligence systems could be granted rights (On this see Forrest, 2024).[7] However, at this stage, two remarks are to be made. The first is the elucidation of the asymmetry of the concepts of liability and responsibility. While liability can be understood as legal responsibility, responsibility is not similarly derivatively explainable, independently from the very possibility of its own enactment.[8] This point matters to remark both that liabilities and responsibilities are intertwined in the understandings of our practices, and that providing a better account of the responsibilities and a better account of the liabilities of (automated) artificial agents are tasks whose achievements mutually contribute to each other.[9]

The second remark that can provide satisfactory elements of resolution of the mentioned difficulty is one that is historical, epistemological and philosophical. For, it is prior to the material construction of artificial intelligence systems that fundamental reflexions about artificial intelligence were achieved, notably by Turing. "Computing machinery and intelligence" (1950) contains and addresses fundamental problems related to the concept of artificial intelligence. Turing there explicitly posed the question "Can machines think?". Independently from readings which have been made of this passage and which ascribe to Turing the idea that the question "Can machines think?" *cannot* be replied to, I underline that Turing explicitly proposed a replacement strategy to pose and address the question. Remarkably enough, the question presents formally and exactly elements of a dilemma that can be assumed to be unsolvable. To make sure, suppose that the question "can machines think?" is a closed question, i.e a question that can be truly answered only by means of one and only one answer that is either positive or negative. Then the correct reply to the question is either

[7] An idea critically discussed by Putnam (1975a; 1975b).

[8] By contrast, in most cases, acts are juridically judicable even if committed unintentionally.

[9] Indeed, focusing only on the legal aspect of responsibility-involvements of (automated) artificial agents in events, risks to make us miss the ways in which uses of automated artificial agents are likely to reshape and contribute to ordinary routines, practices, understandings, and forms of life. In other words, focusing on eventual misuses of automated artificial agents and on the legal resolutions of legal problems arising from such actual or potential misuses, might also lead us to miss the ways in which automated artificial agents can enable the practical resolution of practical problems. Medical robots equipped with artificial intelligence systems remarkably provide us with such ways.

that machines think, or that machines do not think, but *cannot be* that machines think and do not think. In one sense, the replacement strategy provides a replacement to this pseudo-dilemma. For the truth of the direct ascription of the capacity to think to machines would imply that this capacity which traditionally has been thought to be distinctive of humans (by contrast with non-human animals) is in fact not distinctive of humans at all. While the truth of the direct rejection of the capacity to think to machines would imply that humans could be pretending to determinate a capacity that conceivably could not be humanly determinable at all. Would the positive and the negative answers be the only possibly relevant answers to the question, then the question becomes simplistic, demiurgical and untechnical, but also unphilosophical and dogmatical. Turing rightly and effectively contested the narrow, simplistic, untechnical and dogmatical understanding and response to the question. Nevertheless, Turing also radically misconceived solipsism by equating solipsism with the thesis according to which *the only way to know that – the fact that – someone thinks is to be that (particular or individual) person and feel oneself thinking* (on this, see Uçan, 2024). Briefly, the problem with such conception of solipsism is not, strictly speaking, that it is false, but that the problematic of solipsism cannot be (over)reduced to this only aspect without major loss. The important point for our present purpose is that Turing was entirely right with respect to the problematicity of the disjunctive entrapment that the question "Can machines think?" was meant to unavoidably imply. If the question of the correctness of the ascription of thoughts to machines is meant to be answerable to by means of the identification of the mindedness or mindlessness of machines, then there is a sense in which the question both cannot lack and cannot be provided a (true and satisfactory) answer. To this extent the problem of the apparent unavoidability of the dilemma both needs to be posed and addressed.

3.2 The relevance of Wittgenstein's criticism of private language to the thought of automative agency

One central aspect of Wittgenstein's criticism of so-called 'private language' in *Philosophical Investigations* is the dissolution of the delusory unavoidability presented by the dilemma implicitly criticized by Turing. Thereby, we shall see that Wittgenstein's criticism of 'private language' also concerns the problematic of the responsibility of (automated) artificial agents. To bring out and explicit this aspect, let us recall that (internal) 'possession' of mind, soul, or consciousness, and correlatively of language – conceived to be more or less equivalent with the disposition to articulated language, has been traditionally conceived and presented as what whose 'possession' would distinguish humans from non-human animals. Without oversimplifying, Descartes sometimes defends this conception (1996, 278). (Non-human) animals do not speak, thus do not have souls as humans do. Nevertheless, inasmuch as soulfulness would be rendered sensible through enactments of linguistic dispositions, the possibility that soulfulness could be inexistent during the absences of linguistic enactments would not be absurd. Correlatively it could be the case that linguistic enactments could be

merely delusory appearances of soulfulness or mindedness. Only 'metaphysical' assurance and correlative infallible grounds of infallible certainty could render possible the exclusion of such doubt. But at this stage of the reasoning, differences in modalities and availabilities of thoughts, questions, and true answers might be delusorily taken as *indicative* of the truth of two traditional assumptions. First, the past exercises of a capacity to ('internally') *access* to thoughts is involved by the intelligibility of our thoughts of and about differences. Second, the intelligibility of accessed access-differences would be tied to an oppositive conception of the distinction between direct or mediate, and, indirect or immediate. In other words, although in one sense there is no one else to ask to know one's own thoughts, there is no one else than an other to ask to learn one's – that person's – own thoughts. Correspondingly, the superfluity of an activity realized and mediated by someone else to know one's own mind (the so-called 'first-person' perspective) would be indicative of the *directness* of the relation of the minded agent to one's own thought. And the necessity of an activity realized and mediated by someone else to know one's thought (the so-called 'third-person' perspective) would be indicative of the *indirectness* of the relation of the minded agent to the thoughts of the other minded agents. In metaphorical words, others' thoughts would always be screened to us, forever undirectly available and necessarily mediated to us rather than directly available and necessarily unmediatedly as our owns, or, the opposite.

Reflexions about the contribution of the environment and of others would suffice to elucidate that the oppositive conception of the distinction between the direct and the indirect consists in an oversimplification of the diverse ways in which our thoughts are eventually available, or can become such, to ourselves. For, both environments and others rendered *possible* activities which we realized prior to our realizations of the existence of differences in the ways in which we know that thoughts are made available to ourselves by ourselves. The development and the learning of necessarily linguistically mediated differences of the modalities of the availability of thoughts necessarily involve the existence and the fact of the past contributions of others and environments.

By contrast, a 'private language' could and would constitute the metaphorical sealing stone of the delusory and alienatory confinement that would result from the overimposition of the simplificatory account of our relations to our thoughts, to each other and to ourselves upon our ordinary linguistic uses. Wittgenstein elucidated that 'private language' should be criticized, that its very conceivability and intelligibility should be radically contested (Wittgenstein, 2009, §243).[10] Let's radically elucidate: it is not the case that each ordinary set of ordinary linguistic practices is more complex than those involved in and by the oversimplified account. But, and that is required and sufficient to philosophically pose the problem, as notably Wittgenstein (and also Sartre) did: it is not the case that each ordinary set of ordinary linguistic practices is less complex than the oversimplified account. To this extent it is false that the oversimplified account necessarily

[10] On Wittgenstein's elucidation of the intrinsic open-endedness and the eventually tragic ungroundedness of language (1979, 168-190) see Cavell (1979, 168-190).

involves a philosophical result by comparison with the non-philosophical conceptual equipment. The traditional conception of philosophy and of epistemology which is tied to the oversimplified account is misleading and should be criticized. Nothing *in* language(s), *of* language(s), could imply the sort of delusory confinement that could and would allegedly be unavoidably implied by the existence of an instance of 'private language'.

Then the criticism of 'private language' is both relevant to dis/solve the disjunctive entrapment problem addressed by Turing with the replacement strategy, and to address the problematic of the responsibility of automated artificial agents. The presupposition of the unavoidability of the oppositive conception of the direct or immediate and the indirect or mediate in fact turns to be necessarily misleading in its applications to our linguistically mediated relations with others. At best improbable is that such presupposition will not similarily turn out when applied to our relations with automated artificial agents.

For, according to the criticized conception of mindedness (as an internal property of the agent that would only be directly 'observable' by the agent oneself), the 'verification' whether an other agent is minded, and its results would need to be indirect. The verification that could be directly made by each of us about oneself could and would need to be indirectly made by each of us about each other. When applied to the case of machines, the verification of the mindedness of machines would thus both be required to be achieved necessarily indirectly (by us by means of interactions with machines). For, the direct verification of the mindedness of a machine would need to be directly impossible – as we are not machines – except in a 'metaphysically' analogical sense that could not be relevant to reply to the initial question in any sense. And yet such verification would also have to be necessarily indirectly impossible (for the mindedness of machines could not result from our actions without contradicting our expectations of the autonomy of their doings). Thus the sense of inescapability that would be required for the disjunctive entrapment to start presenting some relevance is at best an illusion, but more probably, not even that much.

The criticism of private language thus also presents significance to pose and address the problem of the responsibility of automated artificial agents, to think automative agency. Once dispelled the illusion that we could not but have to presuppose that:

> 1. Either we are (indirectly) responsible for every action of (automated) artificial agents, or we are responsible for none of the actions of (automated) artificial agents.
> 2. Automated artificial agents either are (directly) responsible for each and every of their actions, or artificial agents are not responsible at all for any of their actions.
> 3. Human and artificial agents cannot both directly and indirectly be responsible of their actions in different senses during one single and only event.

Then we start reaching a better position to pose and address the problem of the intelligibility of automative agency. Interestingly enough, the problem of artificial

intelligence, in the sense of the problem of the agency of automated artificial agents, does not reduce to the problem whether we can hold an (automated) artificial agent legally accountable for one's action, and turns out to be equivalent both with a form of the philosophical problem of other minds, and with a version of the problem posed by avoidance (Cavell, 1979; Sartre, 2003; Wittgenstein, 2009).

3.3 Responsibility, artificial agents, and us: The distinctiveness of automated artificial agents

A way to recover, recognize and further disentangle the various threads of the knot constituted by the oppositive conceptions of the direct and indirect conceptions of the liabilities of artificial agents, is to distinguish the elements of the confusion of artificial agents with automated artificial agents. For, necessarily misleading is to suppose that artificial agents are necessarily equivalent to automated or automatized artificial agents in every case. Hallevy (2012) effectuated a 'reduction' to establish the similarity of the action realized by means of an artificial agent and a tool. But such 'reduction', which renders achievable the isolated consideration of an aspect, can have for problematic result the confusion of artificial agents with automated artificial agents. Obviously, in the case in which we compare a robot to a tool (as a hammer), the action of the tool, as the hammering action of the hammer, or the hammering action of a robotic-hammer, or the hammering action of a robot which uses a hammer, can be correctly imputed or ascribed to the user. However, this precisely involves an asymmetry between the artificial agent (the agent whose action is considered in diverse senses) and the automated artificial agent. Not only that by contrast with artificial agents the action of an automated artificial agent requires prior *activation*, but also, disposes of a margin of decision or action which results at least partially from some of the past actions of its creator/s, producer/s, and user/s. Indeed, creators conceived and producers made automated artificial agents, such as AIUVs, that can take decisions according to ranges of decisions which are predetermined. Such automated artificial agents can for example be programmed to take the shortest or the quickest path to reach a set place. Further, such agents can take into consideration changing sets of possibilities correlative to informations obtained by means of their sensors, or, their connexions with remote sensors, or, to informations available online. Automated artificial agents can thusly realize exactly the aims of the (humans) these agents have been designed for, without any additional intervention of a non-artificial agent. But importantly enough, artifacts and humans could not have exchanged positions in each and every sense, and correlatively be mutually reduced to each other. Even the practice of the humanization of artifacts display limits. Such exchanges of positions do not necessarily belong to our world-conceptions, even to the ones of the most ardent defenders of the idea that machines, robots, artificial intelligence have rights that should be defended. Many public debates are dedicated to the discussion of the evolutions of our world-conceptions according to technological aims and results. But some are at least momentarily already debased, already obsolete, and some

illegitimately excluded by practices internal to (military) researches dedicated to the conception of ways for humans to destroy other humans.

In that, *derivative* autonomy characterises automated artificial agents in a way that does not, and could not, apply to every artificial agent. To this extent, although the indirect conception of the liability of artificial agents might seem intellectually more satisfactory or felicitous, we need also to consider that the supposition according to which automated artificial agent *cannot* be *morally* responsible is *also* inherently problematic. The would-be idea that a novel sort of agency or rationality that would in a sense be proper to automated artificial intelligence could be discovered could not be *neutral* in any desirable sense, and its sensicality is at least necessarily questionable. That is the assumption whose idleness and confusedness was elucidated by Wittgenstein's criticism of private language by the dissolution of the misleading presupposition of the need for a startpoint from nowhere. Surely, both artifacts and humans can be characterized as autonomous. But a similarity of their characterizations could not imply the equivalence of their concepts. By principle humans can autonomously act, set to themselves and achieve by themselves interrelated objectives whose realizations contribute to their own developments, liberations, and aspirations.

Undeniably, automated artificial agents can similarly autonomously act, set to theseselves and achieve by theseselves interrelated objectives – which are those of the humans which conceived or for which these agents were conceived and produced. But by contrast with automated artificial agents, human agents can, deliberately yet unsuccessfully, fail their own autonomy. That is a non-relevantly negligible asymmetry between automated artificial agents and human agents. For there is a sense in which automated artificial agents could not fail their own autonomy, and that their autonomies are necessarily derivative from those of humans. Some humans, willing, among other tensions, to get rid of the (would-be) angst of being unconsciously reproducing with respect to machines, robots, artificial intelligences discriminations which were or are being done with and to women, minorities, children, strangers, animals, differents, others, might be more prone to ascribe rights to automated artificial agents. However, unhelpful would be to exclude that the ascription of such rights could become, a way for ancient oppressions to be maintained, if these are. We need to find the source of the fascination. In the case of recent automated artificial agents equipped with or constituted by artificial intelligence systems, the source of the fascination is that such 'entities' form sentences in the grammatical sense, and achieve others actions which were, as much as we know, realized only by humans. Such 'entities' (of which whether one can truly wonder whether these exist, as corroborated by ongoing legal debates) are therefore potentially autonomously enacting the disposition that is, according to the tradition, the uniquely identifying disposition of mind, consciousness, or intelligence, that is, *articulated language*. Consequently, such 'entities' are apparently in a much better position than others (notably by comparison with non-human animals) to be considered as autonomous individual persons. Indeed, while in the case of artificial intelligences, although it is at best uncertain that the 'entity' thinks in a non-derivative sense, automated

artificial agents form phrases, sentences in a grammatical sense. While in the case of non-human animals, although the 'entity' does not form phrases, it is at best uncertain that the 'entity' does *not* un-derivatively think. We just need to be aware of the oversimplification implied by the traditional conception: animals would not speak, and thus would not think, and thus would not be minded. But it is remarkable enough that by contrast, there is no doubt when the minded sees the cat willing to catch the mouse, that the cat wills to catch the mouse. The absence of articulated language does not imply the absence of thought and the absence of mind. We just need not to deprive ourselves from our intellectual means *by* or *when* anthropomorphizing animals or artifacts. Surprisingly enough, the occurrence of instances of intelligent behaviour could not necessarily imply the existence of intelligent agents. The problem of artificial intelligence in fact turns out to consist in a form of the philosophical problem of other minds, which *was* the quintessence of the problem of the existence of the external world.

3.4 Recovering expressive resources: Disentangling agenthood from personhood

The difficulty we confront is less related to the structures involved by our thoughts and responsibility-ascriptions to automated artificial agents, than to the linguistic strictures resulting from the overimposition of the oversimplified account of our relations among ourselves, and eventually with (automated) artificial agents, over aspects of these relations, practices, and forms of life. The correlative task then is first the one of a linguistic recovery, rather than a linguistic simplification. And that recovery can be achieved through a conceptual disentanglement, which involves the one of agenthood from personhood. This disentanglement does not reduce to the elucidation of a conceptual distinction between agenthood and personhood, but further to the extraction of the concept of agenthood from the set of difficulties where it was plunged due to its confusion with the concept of personhood. Indeed, personalism (by contrast with egoism, egotism, and expressions of mere hypocrisy), tends to render confuse that personhood implies agenthood in a way in which agenthood does not, and could not, imply personhood.

To address and solve this problem, let us first remark that our concept of person needs to account for the wide variety of uses that this concept admits in law and courts of justice. Thus, human beings, women, children, men, but also corporations as other entities which can directly or indirectly engage in a legal system, which can own, sue or be sued, contract, are persons. In such contexts, persons are agents in the sense that persons are possibly the authors of the actions due to the realization of which their legal responsibilities, their liabilities, are assessed. That is to say, persons are those who have or might eventually have individually or collectively realized the actions whose realizations are judged at the court. But agenthood does not apply to, and is not characteristic of, *only* those who have or could have realized actions similar to the ones for which some persons are being judged at a given occasion at a court. The implicit assimilation of agenthood and personhood, which is sufficient in many ordinary

contexts to think and assess actions of persons, can also turn to be its own obstacle. Objects and substances can also relevantly and correctly be considered as agents, by contrast with patients. In most cases, concepts of agenthood involve some contrast between the agent and the patient, the active and the passive. But the same does not apply to concepts of persons, which do not necessarily involve a structurative duality of opposables – although the dichotomy between the human and the non-human is often assumed to constitute such duality. Thus, agenthood could not be equivalent with personhood. Surely, not every agent is a person, but the opposite grammatical sentence, is also true, i.e. not every person is an agent. We saw that Floridi and Sanders (2004) addressed the difficulty by distinguishing the causal source and moral source or responsibility of moral actions. Thereby the compatibility of the indirect liability of artificial agents and the direct liability of human agents was shown. But if we consider a range of cases that is central for the practices of justices, we need also to consider that action in metaphorical and derivative senses[11] are both relative to and independent from our actions in ways which are compatible and required both for scientific and legal enquiries. That is to say, the agenthood, in the sense of the eventual *causal action* of, say, a substance over a surface, which can be studied[12], and the agenthood of a person, in the sense of the eventual *moral responsibility* of a moral agent in the realization of an action, may well be related. The moral responsibility of an action of spilling a corrosive agent over a surface is not as such relevant to the outcome of the scientific enquiry about the causal relations between agents and patients. But, that is the exact sense which is relevant to judge, in a court of justice, the responsibility of a similar action on a human body. And those senses, although distinct, can also, by principle, be related, in unproblematic ways. The latent anthropomorphism which consisted in the presumption of both the necessity and the impossibility of the application to automated artificial agents and humans of the same standards of liabilities can thus be left aside. The model of the search for a uniquely corresponding condemnable state of mind (the state of mind outside 'in' the mind of the other) in the paralleled series of states of mind (the 'matching' state of mind) to judge the action of the automated artificial agent does not display relevance as such. Correlatively, constitutive asymmetries between humans and automated artificial agents could not determine reciprocal false-impossibilities. We could not unavoidably have to find in the case of automated artificial agents the lack of the counterpart that we could not but have unavoidably have to dismiss in the case of other minds. In order to avoid mistakenly uniformizing our expectations with respect to differences internal to ordinary contexts and less frequent contexts of juridical assessment, we need to be critical and reject the surreptitious remanences of simplistic representationalism.

[11] As when we speak of the corrosive action of a solvent.
[12] As different solvents have differents actions over differently composed surfaces.

3.5 Accounting for responsibilities involving automative agency

We can thus account for responsibilities concerning or involving automative agency without presupposing the unavoidability of the opposition of the direct and indirect (conceptions of) responsibilities of automated artificial agents. Our account rather should integrate the circumstancial applicability of the distinction between the direct and the indirect, the mediate and the immediate, both to the legal and the causal responsibilities of the agents' actions at the occasion of an event. But the distinction between responsibility, as an ineliminable dimension of human action, and legal liability, as an ineliminable dimension of human practices, needs to be preserved to achieve such distinction. The search and establishment of *legal* responsibility, that is liability, remands to contexts in which elevated epistemic standards are applied.[13] But, human responsibility could not conceivably reduce to legal liability. And while some contexts involve the evaluation of causal responsibility to achieve the evaluation of legal responsibility, the opposite could not have held.

That is to say, a (human) agent, as a human person, can be directly or indirectly responsible for the realization of an action, and can be such both legally and causally, although the legal and the causal responsibilities involved by one's action are unlikely to be thusly evaluable and evaluated in most cases. Ordinary actions, such as the one of buying, or the one of having someone buy some item from someone else at some market, provide us with appropriate examples. But, to better distinguish legal and causal dimensions of responsibility-ascriptions, we need to consider cases in which the causal or legal responsibilities involved by the occurrence of an action are evaluated. Some ascriptions of responsibility indeed involve a distinction between the direct or immediate, and the indirect or the immediate. For this motive, let us consider the case of someone who makes inadvertently fall and break an item in a shop. That person would be considered both legally and causally directly responsible for one's (inadvertent) action in many cases. But now, consider the case in which someone inadvertently makes someone else unintentionally break an item in a shop by falling on that person. In such cases, many ways for responsibilities to be assessed are envisageable. But misleading would be to neglect that the person who initially fell was, both legally and causally, indirectly responsible for the other's 'inadvertent' action in many cases.

By contrast, an artificial agent, as a tool, can be held to be causally responsible for an action and for its consequences, but would not be held legally liable for the realization of an action. That is to say, it is common practice to conceive, design, produce, evaluate, and, eventually identify and verify objects which render aims achievable. Artificial agents such as tools can be conceived as externalized ways for aims to be rendered (more easily) achievable. For example, a hammer can be held responsible for the hammering of a nail. And different sorts of hammers could be used and would be better appropriate to plant different sorts of nails. But even in the absence of any corresponding forbiddance,

[13] Such as contexts in which future liabilities are considered in preventive ways, or past liabilities are considered in judicial ways.

a hammer would not be legally held responsible for the realization of an action. Although the nail was knocked in by a hammer, we would not expect hammers to knock in nails. Artificial agents can thus be held to be *directly* responsible for their actions and consequences, although we do not hold these accountable for actions achieved by their uses. And furthermore, artificial agents can also be held to be *indirectly* responsible for their actions and consequences. Suppose for example that during one of its uses, a hammer hit the frame of a picture which then fell and broke. Or suppose that, at your return at your flat, after an earthquake, you discovered that the hammer you had left on a row fell on a table, which also fell, and broke a cup you had left on the table. In such cases, the artificial agent could be held to be causally and *indirectly* responsible for the occurrence of another event. To this extent, although legal responsibility does not concern such artificial agents, direct and indirect causal responsibilities can be adequately attributed to such agents. Finally, automated artificial agents, as AIUVs or robots or artificial intelligence systems can be held to be causally responsible of an action, and, it is not inconceivable that these will also become legally liable for the actions of which these can be directly or indirectly responsible. In that sense, this range of cases is the one that is central to account for the responsibilities of automative agency. Such artificial agents have a certain margin of action, as these agents can derivatively decide to realize or not predetermined actions, and can even redetermine and make evolve their ranges of conceivable actions relatively independently from further actions of (human) agents. The initiatives of the realizations of their actions are derivatively taken by these agents theseselves, and that is the central motive for the characterization of such agents as 'autonomous'. Such agents can also be directly or immediately responsible of an event. Consider, for example, the case in which a vehicle equipped with an artificial intelligence system successfully transports someone from one place to another. Or the case in which such a vehicle crashes and kills its driver-user by taking an unadapted decision. In such cases, automated artificial agents are *directly* responsible for the occurrence of an event, whether planned or not. By contrast, there are cases in which automated artificial agents are *indirectly* responsible for the occurrence of events. Consider for example the case of an AIUV which chose to destroy itself rather than its human user-driver or another human person. Or consider the case of an AIUV which provoked an accident by having failed to detect a rock detached from a cliff and rolling on a road which it then hit. To this extent, although only some selected aspects of intentional actions have been considered, events whose occurrence involve automative agency can be analyzed to explicit direct or indirect, legal or causal, responsibilities and liabilities of agents.

4 Conclusion

The study of the opposition of direct or immediate and indirect or mediate conceptions of the liabilities of artificial agents elucidated central requirements to address the problem posed by some artificial agents to accounts of liability. Indirect or mediate conceptions of the liabilities of artificial agents recall the *instrumental* place that artificial agents have in our lives. While direct or immediate conceptions of the liabilities of artificial agents recall that we need to acknowledge that some artificial agents can somehow be *responsible* for actions, even if only to be able to count as the instruments of the actions of non-artificial agents. The depth of this opposition is further realized if the contextuality of the relevance of liability-assessments standards is considered. Indeed, legal or technical concerns underdetermine elevations of epistemic criteria of assessment. Ways for practices with unwanted consequences to be forbidden, for technical problems to be solved, and for injustices to be judged, involve the determination of the liabilities of involved agents and persons in determinate events. However, reductionist and functionalist approaches could neither enable nor suffice to address this problem, which can be failed to be posed not only by supposing that artificial agents *can* be morally responsible, but also by supposing that artificial agents *cannot* be morally responsible. For, asymmetries between agents and possibilities of action are not equivalent with impossibilities which could be have been obtained from the comparisons of humans and artificial agents. In this sense, Turing rightly drew our attention to the fact that would-be attempts to verify whether machines *can* think are more significative of human tendencies than scientific interests, although concerns with verification often testify of a concern with scientific and objective truth. However, the problem of the apparent unavoidability of the opposition between the direct and indirect conceptions of the liabilities of artificial agents can be dis/solved by means of the study of Wittgenstein's criticism of 'private language'. This criticism can indeed be applied to show that nothing could have implied that human and artificial agents cannot both directly and indirectly be responsible of their actions in different senses during one single and only event. Then it appears that the (legal) problem of the liability of artificial agents in fact is the (legal) problem of the liability of *automated* artificial agents, those artificial agents which can, once activated, realize some actions independently from further actions of their creators, producers, or users. Personhood needs to be further distinguished from agenthood, inasmuch as some agents are not persons, and that persons are not necessarily agents. Then we can think anew automative agency in its relation to responsibility. Direct and indirect, causal, and eventually legal and derivative responsibilities of automated artificial agents can be distinguished in their relations to the similar responsibilities of human agents.

Disclosure of Interests. The author has no competing interests to declare that are relevant to this article.

References

1. Cavell, S.: The Claim of Reason. Oxford University Press, New York (1979)
2. Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment, https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-against-torture-and-other-cruel-inhuman-or-degrading, (1984)
3. Denett, D.: The Intentional Stance. MIT Press, Cambridge (1987)
4. Descartes, R.: Œuvres, t. V.. Vrin, Paris (1996)
5. Forrest, K.: The Ethics and Challenges of Legal Personhood for AI. The Yale Law Journal **133**, (2024)
6. Floridi, L.: The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities. Oxford University Press, Oxford (2023)
7. Floridi, L. Sanders, J. W.: On the Morality of Artificial Agents. Minds and Machines **14**(3), 349-379 (2004)
8. Hallevy, G.: Unmanned Vehicles – Subordination to Criminal Law Under the Modern Concept of Criminal Liability. Journal of Law, Information and Science **21**(200), (2012)
9. McAllister, A.: Stranger than Science Fiction:: The Rise of A.I. Interrogation in the Dawn of Autonomous Robots and the Need for an Additional Protocol to the U.N. Convention Against Torture. Minnesota Law Review **180**, (2017)
10. Pagallo, U.: From Automation to Autonomous Systems: A Legal Phenomenology with Problems of Accountability. In: 26th International Joint Conference on Artificial Intelligence, pp. 17-23. International Joint Conference on Artificial Intelligence, Torino (2017)
11. Putnam, H.: Minds and Machines. In: Mind, Language and Reality, pp. 342-361. Cambridge University Press, New York (1975a)
12. Putnam, H.: Robots: Machines or Artificially Created Life?. In: Mind, Language and Reality, pp. 386-407. Cambridge University Press, New York (1975b)
13. Rome Statute of the International Criminal Court, https://www.icc-cpi.int/sites/default/files/RS-Eng.pdf, (2021)
14. Sartre, J.-P.: Being and Nothingness. Routledge, London (2003)
15. Turing, A.: Computing Machinery and Intelligence. Mind **LIX**(236), 433-460 (1950)
16. Uçan, T.: Machines and Us: The Comparison of Machines and Humans at the Test of the Problematic of Solipsism. In: Balkan Analytic Forum, pp. 87-126. Balkan Analytic Forum: Normativity and Normativity of Art, Belgrade (2024). https://doi.org/10.18485/baf.2024.1.ch4
17. Wittgenstein, L.: Philosophical Investigations. 4th edn. Blackwell, Oxford (2009)