Technology in Society 71 (2022) 102088

Contents lists available at ScienceDirect

Technology in Society

journal homepage: www.elsevier.com/locate/techsoc

Machine Capacity of Judgment: An interdisciplinary approach for making machine intelligence transparent to end-users

Aurelia Tamò-Larrieux [a,1,*], Andrei Ciortea [b], Simon Mayer [b]

[a] Maastricht University, the Netherlands
[b] University of St. Gallen, Switzerland

ARTICLE INFO

Keywords:
Machine Capacity of Judgment
Responsibility
Transparency
Agency
Artificial agents
Autonomy

ABSTRACT

Intelligent machines surprise us with unexpected behaviors, giving rise to the question of whether such machines exhibit autonomous judgment. With judgment comes (the allocation of) responsibility. While it can be dangerous or misplaced to shift responsibility from humans to intelligent machines, current frameworks to think about responsible and transparent distribution of responsibility between all involved stakeholders are lacking. A more granular understanding of the autonomy exhibited by intelligent machines is needed to promote a more nuanced public discussion and allow laypersons as well as legal experts to think about, categorize, and differentiate among the capacities of artificial agents when distributing responsibility. To tackle this issue, we propose criteria that would support people in assessing the Machine Capacity of Judgment (MCOJ) of artificial agents. We conceive MCOJ drawing from the use of Human Capacity of Judgment (HCOJ) in the legal discourse, where HCOJ criteria are legal abstractions to assess when decision-making and judgment by humans must lead to legally binding actions or inactions under the law. In this article, we show in what way these criteria can be transferred to machines.

1. Introduction

Digital evolution describes the "evolutionary processes embodied in digital substrates" and a myriad of creative and surprising developments of artificial agents exist: Lehman et al. [1] give an overview of surprising (and entertaining) emergent behavior of agents since the occurrence of unintended consequences of automatic callback systems. This includes the agents' ability to identify and exploit bugs in physics engines and computer games and the identification of robot control strategies that were surprising—and counter-intuitively, at first—even more stable than human-made strategies.

These examples give rise to highly relevant normative questions, as they might give us the impression that machines exhibit autonomous judgment [2]. Yet, interpreting artificial agents' actions as an autonomous judgment might be not only misleading–because it posits the notion of agency in a context where there is none–but also dangerous–"because it elides [that] humans, and the institutions within which they sit, are in fact responsible in the first instance for the data selection and programmatic choices" (p. 635 [2]; see also [3]). Addressing these issues is as central as it is difficult. It requires not only a deeper understanding of decision-making and judgment by machines but also a master plan for a responsible and transparent distribution of responsibility between all the involved stakeholders.

In this article, we explore whether the concept of Machine Capacity of Judgment (MCOJ) can help to establish a framework for a responsible and transparent distribution of accountability. To understand what this article proposes as MCOJ, we draw from the use of Human Capacity of Judgment (HCOJ) in the legal discourse: HCOJ criteria are legal abstractions to assess under the law when complex decision-making and judgment by humans must lead to legally binding actions or inactions. Based on HCOJ, we explore which criteria could establish such a capacity in artificial agents—potentially, even leading to a metric to determine when autonomous software acts in a legally binding manner.

Aside from an introduction and conclusion, this article covers the following content: In the section "Human capacity to act and judge" we discuss human decision-making and judgment processes before elaborating on the legal concept of capacity to act and judge. The description of human judgment and the capacity to act and judge under the law shows the necessity to—especially in cases of disputes—categorize abstract and complex human decision-making and judgment to determine

* Corresponding author.
E-mail addresses: a.tamo@maastrichtuniversity.nl (A. Tamò-Larrieux), andrei.ciortea@unisg.ch (A. Ciortea), simon.mayer@unisg.ch (S. Mayer).
[1] former: University of St. Gallen, Switzerland.

https://doi.org/10.1016/j.techsoc.2022.102088
Received 11 May 2022; Received in revised form 10 August 2022; Accepted 14 August 2022
Available online 19 August 2022

the legal consequences that it triggers. We discuss how, to decide whether a human is capable of legally being bound by his or her actions or inactions, the legal literature and jurisprudence have had to create HCOJ criteria. Upon this basis, we discuss how the artificial intelligence (AI) research community has similarly produced human-level abstractions for systems of artificial agents, as research within this field was highly inspired by human-oriented and societal constructs when designing complex systems [4].

Upon these foundations, we hypothesize in the section "Human-oriented abstractions for intelligent machines" that the concept of MCOJ can, similar to HCOJ, enable differentiating between whether an artificial agent is capable to act in a given context or not. One central motivation behind establishing MCOJ criteria is to allow laypersons to think, categorize, and differentiate among the capacities of artificial agents, which is tied to the need to create more granular categories to determine the (autonomous) capabilities of artificial agents. MCOJ is, however, not a general metric to measure machine intelligence but works complementarily towards the goal of establishing a framework to enable comparing different capabilities of artificial agents, injecting a legal viewpoint into discussions such as [5]. Moreover, concerning accountability, MCOJ could provide a tool to assign responsibility among stakeholders engaging with artificial agents. We then conclude and briefly exemplify and discuss the application of MCOJ criteria in the concrete context of a manufacturing system that is based on the authors' research in the domain of hypermedia-based multi-agent systems [6].

In this article, we do not discuss the topic of machine consciousness, which refers to the awareness of machines of their internal and external existence. Moreover, we are not arguing in favor or against proposals to grant machines (e.g., robots) legal personhood. We are not positioning our analysis within the discussions on the merits and demerits of creating a new concept of personhood for sophisticated electronic devices altogether [7–9], although this might be an important discussion to revisit with future developments in the field of robotics. Thus, our analysis in this article is focused merely on exploring and understanding how—in comparison to humans and, more precisely, to natural persons under the law—artificial agents' decision-making and actions could be classified to create what this article terms a Machine Capacity of Judgment.

2. Human capacity to act and judge under the law

Humans think to make decisions, form beliefs, and choose personal goals. Decision-making encompasses multiple elements. First, 'a choice of action' requires the availability of different options. Second, we need to take into account different contexts that impact the one deciding and thus impact the decision. While events cannot be influenced by the decision-making itself, the decision-maker knows (or can guess) the probability of an event which in turn impacts the overall decision [10] (see also [11]). Third, decisions have consequences as they trigger specific outcomes. The decision-maker can (reasonably) expect some of these but not others. Fourth, each human has different goals as one has different constructs of what a desirable outcome looks like. Making decisions thus requires judging situations and adapting accordingly [12,13]. Judgment-making requires the ability to evaluate based on past experiences and accumulated know-how about a situation; the ability to extrapolate those past experiences and evidence into the future; and the ability to match the expected outcome with one's preferred outcome and determine what action to take accordingly.

The ability of humans to decide and judge gives us autonomy and agency. It is also tied to the idea of allocation of responsibility since, commonly, judging human beings should be held responsible for their judgments and corresponding actions. Thus, these cognitive processes are linked to legal agency, and it is not surprising that with advancements in automated and autonomous behaviors of artificial agents, a discussion has emerged as to when legal agency can and should be attributed to these new agents [8]. These discussions center around the concept of the legal capacity to act.

In European jurisdictions, one typically differentiates between legal capacity and capacity to act. This distinction dates back to Roman Law and has evolved into an important pillar of civil law. Legal capacity means that a person is capable of making use of their rights under the law and thus can be held accountable under the rules of the law. Nowadays legal capacity is seen as a right one is born with and does not distinguish between sex, social status, or origin [14]. In contrast, the capacity to act refers to the capacity of a person to create rights and obligations through his or her actions (also understood as the ability to enter into contractual obligations). Typically, the capacity to act includes two dimensions: First, a quantitative dimension means that a person has to have reached a certain age limit to claim a capacity to act. The second is a qualitative and contextual dimension that describes the individual's capacity for judgment. For this article, the second dimension is of particular interest.

The onus of the qualitative dimension rests on the ability to judge situations rationally and freely [14]. First, the *intellectuality element* refers to the capacity to ('rationally') understand the sense, benefit, and impact of certain actions, and the ability to weigh certain decisions against each other. 'Rationally' here does not mean that an individual has to act like a perfect *homo economicus*. Such a bar would be impossible to achieve, as literature on behavioral economics has shown the many fallacies of human thinking and limits of rational behavior [13,15,16]. Instead, building upon such behavioral economics literature it must be accepted that an individual is capable of judgment even if some judgments seem unreasonable.[2]

Second, the *will element* refers to the ability to act freely from internal and external pressure. This element does not reflect on the (philosophical yet empirically grounded) arguments against the notion of 'free will' [18,19]. In other words, the legal discourse does not take into account the determinist approach that shows that free will is an illusion or an impossibility since this would be incompatible with any notion of a capacity to act. Thus, from a legal perspective, the (free) 'will' element must be understood differently to mean being free of external and internal conditions that cannot be resisted (e.g., threat, mental illness, intoxication).

Furthermore, these two elements must always be put into a *specific context*. In other words, the capacity of judgment is always relative to the circumstances of one particular action, and it cannot be claimed in abstract terms that a person is not capable of judgment. The capacity to judge is thus inherently a relative term as it relates to a concrete person, a concrete legal act (or similar legal acts), and the time of the judgment and conditions of the judgment.

This is reflected in cases of disputes in civil law where a decision must be reached by a judge as to whether a person was capable of judgment in a specific situation. For example, the Swiss Federal Court was asked whether a mental deficiency of a young woman would affect her capacity of judgment and thus her ability to marry an older man whose child she was pregnant with at the time of the decision (her family had objected to the marriage; BGE 109 II 273). The court determined that the woman was able to understand the concept and meaning of marriage and argued that the requirements for capacity of judgment should not be set at a level that might render marriage impossible to a large part of the population since few individuals can truly and fully understand the consequences of getting married. Taking the context into account–and the fact that the woman was already living with the older man–the court argued that it is in the best interest of the woman (and

[2] While some authors, notably [16,17], postulate that in order to address limits of rational behavior algorithmic decision-making system should be set in place, our analysis does not point toward supporting or contouring that claim. We show that judicial practices and social norms have accepted the 'flaws' of human judgment without it impeding the assignment of responsibility through the legal abstraction of capacity to act.

the unborn child) to permit the marriage.

3. Human-oriented abstractions for intelligent machines

The artificial intelligence (AI) community has produced abstractions for systems of artificial agents that are inspired by human-oriented and societal constructs. For instance, some of the most influential work on architectures for artificial agents draws from Bratman's theory of human practical reasoning and the Belief-Desire-Intention (BDI) model [20]. In this line of work, artificial BDI agents are designed and programmed in terms of mentalistic notions. According to Rao & Georgeff [21], these are: these are: *beliefs* (i.e., the information an agent holds about the world), *desires* (i.e., the states of affairs the agent wishes to bring to the world), and *intentions* (i.e., the states of affairs the agent has decided to work towards).

An assumption underlying agent-oriented programming [22] is that mentalistic notions provide a level of abstraction that simplifies the design and programming of artificial agents (tracing back to [23]; see also [4]). Moreover, agent-oriented programming promotes a societal view of computation—one in which multiple agents interact with one another—but focuses on the design and programming of individual agents. In more recent work, the JaCaMo meta-model for multi-agent-oriented programming [24] provides developers with a level of abstraction for programming not only artificial agents but also the environment in which the agents live (as inspired by activity theory [25]) and the organization they can form. Multi-agent-oriented programming thus extends the human-oriented level of abstraction introduced by agent-oriented programming from individual agents to systems of artificial agents. Furthermore, the separation of concerns between the three dimensions—agent, environment, and organization—promotes the independent development and deployment of software components by stakeholders in different parts of the world.

Another human-inspired notion of particular importance in research on artificial agents is *autonomy*. One well-known definition of autonomy refers to an artificial agent's ability to operate on its own, without the need for direct intervention from humans or other agents [26]. This definition provides an operationalized notion of autonomy that focuses on individual agents. In a different view, Castelfranchi and Falcone define autonomy as a relationship between three classes of entities [27]: (i) the agent whose autonomy is being evaluated, (ii) a function/action/goal that must be realized or maintained by the agent, and (iii) a secondary entity to which the agent should be considered autonomous with respect to the given function/action/goal. This definition allows for a more nuanced and multi-dimensional view of autonomy—one that goes beyond individual agents and can characterize relationships across the different dimensions introduced in multi-agent-oriented programming.

The following Table 1 illustrates this multi-dimensional view of autonomy in the context of the above-mentioned JaCaMo meta-model. Our interpretation aims to bring an engineering perspective on the autonomy of artificial agents in complex systems.

4. Conceptualizing a machine's capacity of judgment

Several reasons motivate the establishment of a Machine Capacity of Judgment (MCOJ). Most prominently, MCOJ works towards establishing a more transparent and granular understanding of autonomous capabilities. Today, typically, metrics to evaluate autonomous capabilities rely on a scale (e.g., from 0 to -5 or 1 to 10) differentiating actions by artificial agents as being fully independent of human oversight (Level 5 or Level 10) to fully dependent on human actions (Level 1 or Level 0) (e.g. see [28,29]). These metrics are one-dimensional, with the key ingredient being the need for or lack of human involvement, which is a too narrow understanding of autonomy ([30]; see Table 1). In this context, MCOJ represents a more detailed, granular metric to think about and evaluate autonomous behavior, thereby contributing to the

Table 1
Autonomy dimensions for artificial agents in complex systems illustrate the human-oriented level of abstraction that the AI research community has established to help developers efficiently design, program, and regulate systems of artificial agents.

Autonomy Dimension	Freedom	Description
Goal autonomy	Freedom from outside coercion when choosing goals	An artificial agent is endowed with its own goals or adopts a task/goal from other agents only when the agent believes that the adopted task/goal will help it move closer to its own goals. In contrast, any binding command received from the outside would infringe on the agent's goal autonomy. This characterization of goal autonomy is directly applicable to artificial agents that are programmed in terms of mental states and can reason about their beliefs and goals.
Executive autonomy	Freedom from outside dependencies regarding execution means	An agent has executive autonomy if it can achieve the task independently. Executive autonomy may imply the agent has specific abilities, such as the ability to synthesize a plan for the task at hand.
Autonomy from the environment	Freedom from coercive input from the environment	An agent is autonomous from its environment if its behavior cannot be completely determined and predicted based solely on input from the environment.
Autonomy from other agents	Freedom from coercive social interactions with other agents	Social autonomy means independence/self-sufficiency from other agents and autonomy in collaboration. The former refers to the autonomy of an agent to achieve a task/goal without the help or resources of other agents. The latter refers to how much an agent is autonomous when it is working for another agent, for instance when the agent is helping, exploited by, or delegated by another agent.
Autonomy from organizations	Freedom from permissions, prohibitions, and obligations imposed by organizational structures	Autonomy from organizations refers to deontic autonomy. Organizational structures can permit, prohibit, or oblige an agent to an action/goal/etc. If an artificial agent has the autonomy to ignore the obligation imposed by the organization then the agent has deontic autonomy.

Table 2
Machine Capacity of Judgment dimensions and sub-dimensions drawn upon from the HCOJ dimensions.

Dimension	Sub-dimension	Description
Freedom from pressure	Coercive external pressure	**Freedom from the environment**: In a complex multi-agent system, the environment in which artificial agents live and pursue their design objectives can be a standalone software component with clear-cut responsibilities in the system. Depending on the scale of the system, the application environment may be developed by independent stakeholders and potentially over long periods. In such a complex system, it is then useful to understand to what extent the application environment drives the behavior of artificial agents and which components of the environment are affecting the agents' behavior (if any). **Freedom from other agents**. An artificial agent can also be defined in relation to other (human or artificial) agents—and to what extent the agent is dependent on these relations. Depending on the scale and design objectives of the system, other artificial agents in the system may also be developed by independent stakeholders. Assessing the autonomy of agents concerning other agents is thus critical for increasing transparency in the system and distributing accountability among the different stakeholders. **Freedom from organizations**. Agents may enter or leave organizations, and organizations are first-class abstractions in the system: They are defined by a designer to coordinate participating agents towards the achievement of organizational goals. In this context, coordination is achieved through norms that restrict or incentivize the behavior of agents, such as the obligation to achieve a goal within the organization. An artificial agent may also participate in multiple organizations simultaneously where the organizations could potentially work towards cross-purposes. In such complex settings, assessing the deontic autonomy of an artificial agent with respect to the organizations it is a part of would thus be essential both for increasing the transparency of the agent's behavior and for distributing accountability among the different stakeholders.
	Coercive internal pressure	Here we distinguish between **goal autonomy** and **executive autonomy** described above. While goal autonomy makes strong assumptions about an artificial agent's internal state (e.g., the agent has an explicit representation of goals), executive autonomy is applicable more broadly.
Decision-making	Understanding the impact of a decision	If an artificial agent can synthesize a course of action, for instance using any classical approach for automated planning, the agent is essentially running a simulation of the world: Given the current state of its environment and a sequence of actions described in terms of pre-and post-conditions, the agent infers what would be the state of the environment if that sequence of actions is executed. The result of such inferences could then potentially be used to **decide a course of action** from several alternatives.
	Ability to balance options	Some types of artificial agents may have the **ability to weigh decisions**. For instance, BDI agents can decide/select among competing desires/goals or among competing courses of action for achieving a given goal. The former would relate to the agent's goal autonomy and the latter would relate to the agent's executive autonomy. In both cases, the selection functions can have trivial implementations. For instance, BDI agents programmed with the JaCaMo platform will, by default, prioritize their desires/goals in the order in which they appear: with the default goal selection function, the agent keeps a queue of goals and achieves them in a first-in-first-out manner. Similarly, the agent will, by default, select a plan applicable to a given goal based on the order in which the plans were added to the agent's plan library. The default implementations for both selection functions, however, are intended to be customized with more versatile functions as needed for specific domains or applications. Such versatile functions could also benefit from an agent's ability to understand the impact of decisions (e.g., achieving a goal, executing a plan from a pool of plans applicable to a given goal).
Rationality		For artificial agents, the ability to act **rationally** bares more weight than for humans both because it is less prevalent and because emotion-driven behavior is only studied in some areas such as affective computing [36] (see also Section 2.3). We adopt a definition of rationality from [37]: an artificial agent is rational if it will act to achieve its design objectives and will not act in such a way as to prevent its design objectives from being achieved.

discourse about machine capabilities (towards the integration of "artificial intelligence") as a whole. With this metric, it should become visible to individual non-expert users of artificial agents what capabilities and competencies these agents have. This increases the predictability of how an artificial agent might act within an environment. Thereby, MCOJ could play an important role in dismantling current 'cultural imaginaries' around the magic-like capabilities of artificial agents [31].

Building upon this increased transparency, one could enable the distribution of expectations and (potentially) even responsibilities. Towards expectations, we envision a (standardized) labeling system that might transparently communicate what MCOJ criteria are fulfilled by an artificial agent, and to what extent. Responsibilities could be better distributed because knowledge of the abilities, competencies but also limitations of an artificial agent's decision-making and acting helps to determine how far one may rely on such an agent in a particular situation. For instance, if a self-driving vehicle scores low on various MCOJ properties, using that car's self-driving mode in a crowded street would not be a sensible choice. Similarly, if human workers know that a collaborative robot they interact with is bound by external interventions by other agents, the human worker will behave more cautiously around that robot. In addition, having a better understanding of the capabilities of autonomous technology could help reduce situations in which individuals overtrust systems. Research has shown individuals tend to conform to suggestions of artificial agents (e.g. see [32]) and to believe that autonomous systems have greater capabilities than they do ([33–35]). Appropriate information on the limits of the capabilities of an artificial agent thus works towards avoiding misunderstandings and preventing unintended uses of automation (especially beyond the reasonable capabilities of a system).

With these motivations as well as the human-oriented conceptualizations of AI described above in mind, we propose the concept of MCOJ that centers around three dimensions: (1) freedom from pressure, (2) ability of decision-making, and (3) rationality. In particular, the first two dimensions are grounded within the legal categorization of the capacity to judge described above, where we distinguish between the element of intellectuality (which we further divide into the ability to understand the impact of a decision and its benefits, and the ability to compare and evaluate different options against each other concerning their contribution to reaching an individual goal); the element of will (which we further divided into the freedom from internal pressure and the freedom from external conditions that cannot be resisted upon); and lastly the context of a decision-making process, which includes an agent's environment, the timing of a judgment, and then determining the judgment. These elements are the ones that determine the Human Capacity of Judgment (HCOJ).

The following Table 2 combines the theories discussed in Table 1 and elaborates on the MCOJ dimensions.

5. Conclusions

With this article, we propose a way of determining the capacity of judgment of artificial entities—intelligent machines—while aligning with and drawing from the discourse on assessing the capacity of judgment of humans in the legal disciplines. We propose that this more granular way of determining MCOJ directly establishes more transparency about the capabilities of artificial agents and can be useful today

and in the future to solve challenges such as the distribution of responsibility in systems of autonomous agents and, thereby, better manage trust relationships among humans and automated systems. On a strategic level, for creators of intelligent machines, this higher level of granularity permits the intentional gearing of development efforts towards entities that are more (or less, depending on the organizational goals) likely to be perceived as capable of judgment under the law, with the corresponding consequences regarding the allocation of accountability and projection of (dis)trust in an artificial agent's abilities. To permit this, we have conceived the MCOJ dimensions to be quantifiable; however, how exactly this quantification could happen is the subject of ongoing research.

Our model is furthermore useful to shed light on how far artificial agents have advanced towards possibly being endowed with the capacity of judgment. As a concrete scenario,[3] consider an industrial robot arm that is controlled by autonomous agents in a system that follows the JaCaMo meta-model introduced above. This system is essentially solving a planning problem given its representation of the local context and any data it might have been trained on. We can therefore assume the system to act rationally within its design objectives. Furthermore, within its problem domain, the system can understand the impact of its decisions; however, it merely knows a very narrow planning domain and is, therefore, unable to take into account any impact beyond this domain. Also, both regarding balancing decisions and with respect to organizational pressure (see below), the robot may be required to fall back to higher-level norms that are given by its designers, where it might resort to consulting humans directly if a situation is too ambiguous or complex to decide. Finally, regarding freedom from pressure, this system can independently accomplish tasks that are delegated to it and may reject instructions as well as new goals that are sent to it based on its internal state (including any self-set goal(s)). The robot thus has goal-setting autonomy, where we consider highly abstract goals that are set by a human as the design objective of the robot and therefore as self-set goals. The environment as well as other agents might infringe on the robot's autonomy, either indirectly by modifying the environment or directly by overriding the robot's current task. This fact should be known to human users who interact with the robot to avoid overtrust in the system. Finally, regarding external influences on its behavior, the robot might be integrated with different organizations and carry different roles within these organizations which would subject the system to organization-induced external pressure and thereby reduce its deontic autonomy. Furthermore, the robot might have different objectives—for instance, if the robot forms the interface between a logistics zone of a company and a manufacturing plant, it is subjected to the (conflicting) goals of keeping stockpiles above a certain level (to control risk) vs. maximizing production (to increase revenue).

This example shows that some of the criteria we put forward for MCOJ (in particular rationality) are closer to being fulfilled in current systems than others. Regarding decision-making, today's systems are able to understand the impact of their actions, but this is restricted to their immediate problem domain and evaluating what aspects of its impact the system should be able to understand for it to be endowed with MCOJ requires a specific analysis of what can reasonably be expected of the system, for instance, compared to other similar systems on the market and expectations of other stakeholders. On the other hand, we evaluate current systems as far away from achieving the ability to balance decisions and freedom from pressure. Especially for the second, it is not clear whether this is a desirable objective in the development of AI systems in the first place.

We would like to point to further clarifications, limitations, and future work. It should be clarified that we have enlarged the scope of judgment to autonomy when talking about artificially intelligent systems. While the human criteria to evaluate judgment are in our opinion a valid starting point to assess the autonomy of artificial agents, we do not postulate that intelligent machines should (aim to) attain a (legal) capacity of judgment. Rather, we use the term MCOJ in reference to HCOJ: Similar to how HCOJ provides an abstraction for assessing human-level judgment and autonomy from a legal standpoint, MCOJ aims to provide an abstraction for assessing the autonomy of machines. Moreover, we limit the scope of our discussion considerably by not touching upon the topic of machine consciousness nor the concept of personhood [7–9], and refer to the term "responsibility" broadly without tying it to particular frameworks of legal accountability, liability provisions, or literature on culpability ([38]). Within said research stream different interesting approaches to allocating liability more specifically have been discussed, among other insurance schemes [38,39]. While these are interesting topics to potentially revisit with future developments in the field of AI and robotics, our analysis is focused merely on exploring and understanding how—in comparison to natural persons under the law—artificial agents' decision-making and actions could be classified to create what this article terms MCOJ.

[3] This scenario is inspired by a use case that is currently being investigated within a large European research project in the area of the next-generation Internet of Things (https://intelliot.eu/manufacturing) that focuses on manufacturing systems.

Funding

This work was supported by the International Postdoctoral Fellowship grant (University of St.Gallen; project number 1031564) and the Swiss National Science Foundation (project number 189474, "HyperAgents").

Data availability

No data was used for the research described in the article.

References

[1] J. Lehman, J. Clune, D. Misevic, C. Adami, L. Altenberg, J. Beaulieu, et al., The surprising creativity of digital evolution: a collection of anecdotes from the evolutionary computation and artificial life research communities, Artif. Life 26 (2) (2020) 274–306.
[2] A. Solow-Niederman, Administering artificial intelligence. 93 S, Calif. Law Rev. 633 (2019).
[3] R. Crootof, The internet of torts: expanding civil liability standards to address corporate remote interference, Duke LJ 69 (2019) 583.
[4] J. McCarthy, Ascribing Mental Qualities to Machines, Tech. Rept. Memo 326, Stanford AI Lab, Stanford, CA, 1979.
[5] C.H. Hoffmann, Is AI intelligent? An assessment of artificial intelligence, 70 years after Turing, Technol. Soc. 68 (2022), 101893.
[6] A. Ciortea, S. Mayer, F. Gandon, O. Boissier, A. Ricci, A. Zimmermann, A decade in hindsight: the missing bridge between multi-agent systems and the world wide web, in: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '19), International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2019, pp. 1659–1663.
[7] J.S. Gordon, AI and law: ethical, legal, and socio-political implications, AI Soc. 36 (2021) 457–471.
[8] A. Waltermann, On the Legal Responsibility of Artificially Intelligent Agents, Technology and Regulation, 2021, pp. 35–43.
[9] E.A.R. Dahiyat, Law and software agents: are they "Agents" by the way? Artif. Intell. Law 29 (1) (2021) 59–86.
[10] H.R. Pfister, H. Jungermann, K. Fischer, Die Psychologie der Entscheidung, Springer Berlin Heidelberg, Berlin, Heidelberg, 2017.
[11] A. Tversky, D. Kahneman, Judgment under uncertainty: heuristics and biases, Science 185 (4157) (1974) 1124–1131.
[12] A. Likierman, The elements of good judgment, Harv. Bus. Rev. 98 (1) (2020) 102–111.
[13] J. Baron, Thinking and Deciding, fourth ed., Cambridge University Press, 2008.
[14] E. Flynn, A. Arstein-Kerslake, Legislating personhood: realising the right to support in exercising legal capacity, Int. J. Law Context 10 (1) (2014) 81–104.
[15] D. Kahneman, Thinking, Fast and Slow, Macmillan, 2011.
[16] D. Kahneman, O. Sibony, C.R. Sunstein, Noise, HarperCollins UK, 2022, pp. 38–46.
[17] C.R. Sunstein, Governing by Algorithm? No noise and (potentially) less bias, Duke Law J. 71 (6) (2022).
[18] S. Harris, Free Will, Simon and Schuster, 2012.
[19] G. McFee, Free Will, Routledge, 2014.
[20] M.E. Bratman, Intention, Plans, and Practical Reason, Harvard University Press, Cambridge, MA, 1987.

[21] A.S. Rao, M.P. Georgeff, Modeling rational agents within a BDI-architecture, in: J. Allen, R. Fikes, E. Sandewall (Eds.), Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR'91), Morgan Kaufmann, San Mateo, CA, USA, 1991, pp. 473–484.
[22] Y. Shoham, Agent-oriented programming, Artif. Intell. 60 (1) (1993) 51–92, https://doi.org/10.1016/0004-3702(93)90034-9. ISSN 0004-3702.
[23] D.C. Dennett, The Intentional Stance, The MIT Press, 1987.
[24] O. Boissier, R.H. Bordini, J.F. Hübner, A. Ricci, A. Santi, Multi-agent oriented programming with JaCaMo, Sci. Comput. Program. 78 (6) (2013) 747–761, https://doi.org/10.1016/j.scico.2011.10.004. ISSN 0167-6423.
[25] B. Nardi (Ed.), Context and Consciousness: Activity Theory and Human-Computer Interaction, MIT Press, Cambridge, 1996.
[26] Michael Wooldridge, Intelligent agents, in: Gerhard Weiss (Ed.), Multiagent Systems, second ed., MIT Press, 2013.
[27] C. Castelfranchi, R. Falcone, From automaticity to autonomy: the frontier of artificial agents, in: H. Hexmoor, C. Castelfranchi, R. Falcone (Eds.), Agent Autonomy, Multiagent Systems, Artificial Societies, and Simulated Organizations (International Book Series), vol. 7, Springer, Boston, MA, 2003, pp. 103–136, https://doi.org/10.1007/978-1-4419-9198-0_6.
[28] G.Z. Yang, J. Cambias, K. Cleary, E. Daimler, J. Drake, P.E. Dupont, N. Hata, P. Kazanzides, S. Martel, R.V. Patel, V.J. Santos, R.H. Taylor, Medical robotics—regulatory, ethical, and legal considerations for increasing levels of autonomy, Sci. Robot. 2 (4) (2017) 8638.
[29] Society of Automotive Engineers (SAE), Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles J3016_201806, 2018.
[30] J.M. Bradshaw, R.R. Hoffman, D.D. Woods, M. Johnson, The seven deadly myths of "autonomous systems", IEEE Intell. Syst. 28 (3) (2013) 54–61.
[31] M.C. Elish, D. Boyd, Situating methods in the magic of big data and AI, Commun. Monogr. 85 (1) (2018) 57–80.
[32] I. Gaudiello, E. Zibetti, S. Lefort, M. Chetouani, S. Ivaldi, Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to iCub answers, Comput. Hum. Behav. 61 (2016) 633–655.
[33] J. Borenstein, A.R. Wagner, A. Howard, Overtrust of pediatric health-care robots: a preliminary survey of parent perspectives, IEEE Robot. Autom. Mag. 25 (1) (2018) 46–54.
[34] P. Robinette, W. Li, R. Allen, A.M. Howard, A.R. Wagner, Overtrust of robots in emergency evacuation scenarios, in: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2016, March, pp. 101–108.
[35] M. Salem, G. Lakatos, F. Amirabdollahian, K. Dautenhahn, Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust, in: Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction, Portland, OR, 2015, pp. 141–148.
[36] R.W. Picard, Affective Computing, The MIT Press, 2000.
[37] M. Wooldridge, N.R. Jennings, Intelligent agents: theory and practice, Knowl. Eng. Rev. 10 (1995) 115–152, https://doi.org/10.1017/S0269888900008122.
[38] S. O'Sullivan, N. Nevejans, C. Allen, A. Blyth, S. Leonard, U. Pagallo, et al., Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery, Int. J. Med. Robot. Comput. Assist. Surg. 15 (1) (2019), e1968.
[39] D. Schneeberger, K. Stöger, A. Holzinger, The European legal framework for medical AI, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, Cham, 2020, August, pp. 209–226.