December 2023

Between Hammer and Terminator

Giovanni SILENO [a,1], Tomasz ZUREK [a]
[a] *Informatics Institute, University of Amsterdam, The Netherlands*

Abstract. This position paper aims to create a frame of discussion across contemporary debates concerning artificial systems and their relationship with humans. Various misunderstanding are occurring between humanities- and computational-oriented perspectives, and between different positions within. Can machine be like humans? Can they never be like humans? Are machines just like hammers? Can they become terminators? The paper decomposes these questions sketching a conceptual framework, elaborating on notions like moral agency and patiency to identify key components that may conduct to a weaker form of agency, offering a more neutral grounds to discuss responsibility, accountability, and liability.

Keywords. Responsibility, Artificial Systems, Morality, Agency, Patiency

1. Introduction

With the increasing uptake of AI, strong discussions are being held concerning the role of AI-driven systems (ranging from decision-support tools to fully autonomous devices) with respect to humans and society at large. These discussions involve experts from various disciplines, and becomes particularly harsh when it concerns problems of *moral agency* and responsibility (cf. literature on the *responsibility gap* [1,2]). Who is responsible, when AI takes a decision? Let's make clear our position: we firmly believe that humans are ultimately responsible and there is no escape from that responsibility. But in what sense humans are responsible?

Unfortunately, expressed positions are often based on unspoken assumptions and generalizations, usually related to the disciplinary background of the speaker or to common-sense interpretations of terms which may be presented instead in a technical sense. Responsibility, for instance, has a wide spectrum of rather different meanings: in law (legal responsibility), in moral philosophy (moral responsibility), in politics (political responsibility), as well as in engineering and in natural sciences (as functional or causal responsibility). To complicate further, daily unqualified responsibility attributions typically cover more than one of these meanings, and possibly others. In a way, this cannot be avoided; natural language is figurative, and meaning is based on context, most of which is not expressed by speakers. Yet, this setting leads to many misunderstandings, and is not particularly fruitful, nor to advance our understanding as individuals, nor to find some valuable synthesis on which to act upon. Our aim is then to take a step back, and elaborate and decompose relevant underlying concepts, armed with gool-old-fashioned tools of conceptual inquiry.

The current debates on AI can be roughly summarized in two distinct axes. The first axis concerns the relationships between machines and humans, and can be described by

[1]Corresponding Authors: g.sileno@uva.nl, t.a.zurek@uva.nl.

two opposite stances: (i) machines are never like humans; (ii) machines can be just like humans. But in what sense machines may or may not be like humans? We will consider these stances under the agent/patient distinction, highlighting the need for weaker notions of agency/patiency applicable to artificial systems for engineering purposes. References that are relevant here are explicit positions against the antromorphization of artificial entities (eg. Joanna Bryson [3]), the ban against synthetic phenomenology (eg. Thomas Metzinger [4]), and on the opposite side, the various claims on "sentience" in AI chatbots that since a year ago created some commotion.[2]

The second axis concerns more specifically the capabilities[3] of artificial systems, and can be expressed by these two stances: (a) machines are just hammers; (b) machines can be like terminators.[4] References relevant here, in particular for the second position, are the various contemporary discussions, research[5] and policy efforts[6] going under the term AI safety, and AI alignment.

In order to investigate these two axes, the paper introduces in section 2 a common conceptual framework. Section 3 and section 4 apply this framework on the questions whether machines can be like humans, and whether they can be like terminators. A short conclusion summarizes the main points of the paper.

2. Relevant concepts

2.1. Action: main roles, and characterizations

Central to all moral evaluations, there is an idea of action. Linguistic thematic roles tells us that an action is performed by some entity (the *agent*), and involve at least another entity (the *patient*) which suffers some change (external or internal) because of this specific intervention. Furthermore, human conceptualization of action is known to exhibit three levels of abstraction (see eg. [6], "Brutus stabbed/killed/murdered Caesar"). The procedural or *behavioural* characterization describes the specific behaviour which is performed. The *productive* characterization describes the outcome which that behaviour produces. The *purposive* characterization provides further information of what drives the entity leading behaviour. Note that here we are capturing the interpretative frame used by the observer; the purposive characterization in particular is about mental states *ascribed* to the acting entity. However, exhibiting some mental state does not mean necessarily to be sentient or have consciousness; purposes can be hard-coded in living entities by evolution, in artifacts by design.

Acknowledging the purposive characterization opens up to an *intentional* interpretation of the agent, ie. evaluating its behaviour through the ascription of beliefs, desires,

[2] See eg. Scientific American, 12 July 2022, `https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters/`

[3] We distinguish *capacity*, from *capability*, and from *ability*. Capacity captures contingent limitations/boundaries of the entity. Capability includes potentialities, covering adaptations and modifications of capacities. Ability is a capability in a context of *intentionality*.

[4] "Terminator" is a fictional entity made famous by a series of films—the first two (1984, 1991) particularly popular, a robotic assassin programmed to hunt and kill people (as well as to find and protect them, in the second episode). We take it here as a placeholder for some entity capable of "terminating" humankind.

[5] See eg. the recent survey provided in [5].

[6] See eg. the contemporary AI Safety Summit: `https://www.aisafetysummit.gov.uk/`.

and intentions (the so-called BDI theory of mind). This allows us to further distinguish mental attitudes as *epistemic*, *volitional*, and *deliberative*. Note that intentionality may be ascribed also to non-living entities (cf. cargo cults). Exhibiting some mental state does not mean having some mental state. Yet, nothing limits us to design artificial systems that explicitly abide by a BDI computational architecture, ie. systems that are intentional by design, whose decisions can be transparently explained following a theory of mind.

2.2. Responsibility, Accountability, Liability

To reduce the linguistic ambiguity, from now on we ascribe **responsibility** to the agent *who determines the action to occur* (or at least, it is deemed so, from a descriptive point of view).[7] As any action can be defined at three levels of abstraction, we can correspondingly distinguish: *operational responsibility* (related to execution or performance), *tactical responsibility* (related to outcome), and *strategic responsibility* (related to policy). If the responsible agent only contributed to some extent to the action, we are in the context of *distributed responsibility.*

We will ascribe **accountability** to the agent *who justifies the action to occur*. Determination (eg. performance) can result from justification (eg. explanation), but generally the two are contingent processes, and can be ascribed to different agents. Justification may concern what (outcome) and how (performance) it has been done, and why (purpose, policy); there are therefore several levels of accountability, including the various strata of why. As above, if the accountable agent can contribute only to part of the picture, we need to introduce *distributed accountability*. We can further distinguish between *ex-ante* accountability, generally taking a process view over the responsible entity, and so related to *auditability* (eg. for *compliance*); from *ex-post* accountability, focusing on specific behavioural instances, typically associated to *forensics* efforts on incidents.

Finally, we ascribe **liability** to the agent *who has to be blamed for the action* (or praised, in an impact-wise dual setting). Note that liability requires the presence of an evaluative framework (e.g. a preferential ordering or a value structure). If actions cannot be evaluated as good or bad, blaming or praising the agent would not make sense. An implicit assumption necessary for moral systems to work is that the agent being liable is somehow susceptible to the action of blame (or praise), so that liability becomes a dissuasion (or persuasion) mean for behavioural guidance. Key differences between *moral* and *legal judgments* about liability concern the evaluative framework (in the case of law it has to be based on jurisprudence), and descriptive mechanisms (everybody should be equal before the law), but for the rest they seem to follow similar mechanisms [7].[8]

A prototypical concrete example of this role decomposition is offered by corporate organizations. Employees may be responsible (intentionally or not) of misbehaviour according to the law. Auditors are meant to minimize this possibility by analyzing compliance of organizational processes to regulations and best practices. Accountants are meant to keep adequate traces of economic transactions in case of further investigations (eg. for

[7] For simplicity, we neglect here negative actions, ie. omissions, although they have a central role in moral and legal judgments. In principle, the proposed definitions may be extended to cover them.

[8] Note that accountability and liability can also be seen through the lens of responsibility of action, *albeit* for specific types of action: accountability would be responsibility of justifying, and liability would be responsibility of taking the blame (or praise). If these specific actions fail, that would trigger accountability/liability at a meta-level.

supposed incidents). Corporate organizations (in some circumstances their executives) are the liable entities from a legal point of view.

With these definitions, we can also reframe the various uses of the terms of responsibility in diverse disciplines. Natural sciences and engineering refer primarily to operational and tactical responsibility (respectively denoted as causal and functional responsibility). Political responsibility is captured at best by strategic responsibility. Moral responsibility is primarily a matter of (moral) liability. Legal responsibility is a combination of (legal) liability, and of accountability, both in *ex-ante* (applying the rule of law) and *ex-post* senses (justify one's action before the law).

2.3. Moral agency and moral patiency

Morally relevant actions are actions that determine a morally relevant outcome. The moral relevance of an outcome always depends on some evaluative framework, consisting of (i) some value or preferential structure, and (ii) some heuristics or procedure to evaluate what is or may be the case depending on this structure. Note that an entity does not need to be a moral agent to produce morally relevant actions. The moral relevance is in the mind of the observer (the evaluator).

Moral patiency is generally assigned to agents that are deemed to suffer from the outcomes of an action, and such sufferance is deemed unacceptable. As every individual has some experience of suffering, the ascription of patiency can be seen intuitively as reifying a proximity to the self (and thus being related to empathy). Children and people with mental infirmities are given moral patiency, even without ascribing them moral agency. Dehumanizing stances instead deny moral patiency to humans. In the western world, traditionally animals were not deemed to suffer (eg. the general position of Christian religion in medieval times), but this has been changing, and today there are voices asking to reducing suffering also to living forms which are deemed more distant from us (eg. insects, possibly even plants). Moral patiency is generally considered to be a sufficient basis for protective rights to emerge, although in some cases rights are also ascribed to whole ecological systems, eg. rivers[9] or forests. On the opposite side, some authors[10] critique this tendency, suggesting that welfare regulations are more appropriate. Artificial agents are instead usually not treated as moral patients [8], mostly because of lack of pain or self-preservation instinct.

Moral agency is ascribed to agents which are in a position to take and apply a moral decision. This entails both mental and practical requirements. As argued in [9], at the very least, on a mental side, the agent needs *foreesability*: to be able to foresee the outcomes of an action (and for this it also needs to have an adequate picture of the current situation of the environment); and *evaluative ability*: ie. be able to maintain and apply a moral evaluative framework to this inferential outcome. On a practical side, the agent needs *control*, i.e. to be in adequate control of the environment with respect to this action. Note that these practical (control) and mental (foreseeability, evaluative ability) requirements in principle can be reproduced in computational sense, at least to some extent. As they are necessary categories, they can be seen as associated to a weaker form of moral agency.

[9] `https://www.rightsofrivers.org/`

[10] e.g. Joanna Bryson (November 2017), "Why robots (and animals) never need rights": `https://joanna-bryson.blogspot.com/2017/11/why-robots-and-animals-never-need-rights.html`

3. Can machines be like humans?

We now go through our initial questions on the basis of the conceptual framework outlined above. When discerning whether machines can or cannot be *like* humans, we can consider a number of dimensions through which to attempt a comparison ground. We will refer here to a few of them only.

3.1. Sensory and perceptual machinery

The sensory and perceptual system of machines is radically different (at least today) from that of humans and animals. Even if some modality (eg. vision) could compare with and even outperform humans in some circumstances, there are still great architectural distinctions. Multi-modality, if present, is very different from what experienced by humans; besides that, the perceptual processing build upon completely different principles. For instance, large language models (LLMs) are systems that experience the world only through a modality (text), which is actually not directly accessible to humans (it is always mediated through another sense, eg. vision, or hearing).

3.2. Conceptualization

As concepts learned by experiencing the world (synthetic truths?) are grounded in perceptions, and machines have different sensory and perceptual machinery, the conceptualization they form are necessarily different. By definition, they cannot share an experience of the world similar to humans. Additional challenges come from cultural and linguistics factors that build on top or interact with what is grounded on a perceptual level. Returning to the example of LLMs, the embeddings resulting from their training depends on patterns observable at a discourse level on a gigantic, generally global corpus: it is in that space that they define e.g. what an apple is. The concept of apple, and its relationships with other concepts in that space, match the human corresponding mental attitudes only insofar as humans keep a general consistency of using that concept in language. However, by varying culture and linguistic structure, usage may change. This entails that, because minority languages reside only in the very long tail of that gigantic corpus, LLMs may fail to capture the "conceptualization" that these languages expresses, if adequate measures are not taken to protect locality. The way by which humans learn is rather the opposite: we start from local experience, and we generalize them further by putting ourselves in novel situations, becoming in contact with other cultures, etc.

Formalized conceptualizations, like ontologies or logic programs provide an opposite approach to embeddings constructed by induction on some perceptual source. These artifacts are hand-crafted to logically reproduce only a portion of a human conceptualization, modular to a specific domain of application, but do not have (generally) the pretension to be autonomous from humans.

3.3. Moral patiency

Above, we observed that artificial agents are usually not deemed moral patients [8], mostly because of lack of pain or because they do not possess self-preservation instinct. In general, lower-level emotions as pain and pleasure are fundamental in all living entities for e.g. keeping body integrity, individual sustenance and species continuation. Fol-

lowing Damasio [10], we can also distinguish *emotions* (reaction to stimuli that causes observable external changes in the organism) from *feelings* (arising when the agent becomes conscious of the changes it is experiencing). Some authors [11,12] narrow down the possibility to be a moral patient to biological entities or sentient entities only.

It is however interesting to note that, from a functional point of view, very simplified versions of pain and pleasure are *emulated* in computational settings, e.g. in reinforcement learning via punishments and rewards. The maximization of reward can be associated to a self-preservation instinct hard-coded by design rather than by evolution. Even if we agree that it is nonsensical to deem a hammer or a computational device a moral patient, we argue here that there is a weaker version of patiency that can be inflected in computational terms, and which is in practice utilized in all artificial forms of learning.

3.4. Moral agency

As we argued in the previous section, any moral agent should be capable to:

- *control* its own actions,
- *foresee* (predict) the outcomes of the actions,
- *evaluate* the actions and their outcomes according to an evaluative framework.

These practical and mental requirements can be reproduced to some extent in artificial systems. Yet, the extent to which they are aligned to that of humans is determined by design. In principle, the capability of AI-driven systems to perform actions on the environment or predicting the outcomes of such actions does not seem to be controversial or impossible to reach. The evaluative component of the moral agency, however, is the most challenging one, especially because it requires a kind of moral competence, both in terms of settling upon the moral evaluative framework, and of applying it.

Evaluation process De Sio and van den Hoven [2] point out that some researchers assume that "*humans possess a special kind of autonomy, a 'contra-causal' power which gives them a special metaphysical status and makes them morally responsible for their actions in a sense in which no other creature is (or can be).*" There are also opposite opinions (so-called *incompatibilists* or free-will skeptics) who do not believe in any specific type of autonomy humans have. For our purposes, we do not need taking a position in this respect. We just observe that we can decompose the evaluation process into two levels of abstraction (similarly to what presented in [2]):

- **Basic level**: the actual framework of evaluation, containing definitions of:
  - *objects of evaluation*, ie. what is evaluated: performances, outcomes, intentions,
  - *criteria of evaluation*, ie. the basis on which the evaluation is performed,
  - *evaluation process*, ie. how (heuristics, or procedure) to evaluate a given action,
  - *acceptance conditions*, ie. when a particular behaviour is acceptable to agent, or said differently, how an agent accepts or rejects a particular behaviour.
- **Meta-level**: how, and on the basis of what we define components at the basic level.

The meta-level is what ethics (for morality) and jurisprudence (for legality), as research directions, are usually focused on.

We refer to the above distinction as it allows us to distinguish "applied morality" (basic level, already deliberated) from "volitional morality" (meta-level, used for delib-

eration), and can be put in correspondence to *regulated* vs *regulative* dimensions in normative systems.[11] This could be a basis for a discussion whether and to which extent it is feasible and needed to apply the functional equivalents of "moral" (or "legal") reasoning to technical devices. We can say that in principle a device will reason like a moral agent if it will fully operate on both levels, in any other case it will not. At architectural level, basic level and meta-level can be seen as associated to two different roles. In a controlled configuration, the basic level role may be taken by the machine, and the meta-level by a designer. Consequently, a device can have some moral reasoning capabilities (basic level) without pretending to be a moral agent (meta-level). This can be also understood as a strategic choice—whose morality a device will use—one which clarifies the policy responsibility of humans in the whole construction.

4. Can machines become terminators?

Before we answer the above question, we need to clarify more what we will consider here as a "terminator". The fictional movie character is a robotic soldier, with exceptional sensorimotor skills and body resistance, whose only objective is to terminate a certain individual provided as target. It stands as the ultimate "intelligent" weapon, with all the moral considerations that it entails. In a wider sense, however, we could take terminator as an artificial entity capable (intentionally or not) to terminate our whole civilization. Under this extension, we can cover additional cases as eg. the *paperclip maximizer* [13], one of the most known thought experiments introduced to elaborate on the *existential risk* posed by AI: a super-intelligent device whose only goal is to produce as many paperclips as possible, ending up depleting all resources on Earth, including destroying the human species. This line of thoughts is followed very recently in discussions, research and policy efforts concerning AI safety and AI alignment[12]. LLMs and other transformers are for instance seen as potential disruptors of current societal processes as they can generate false, yet perfectly sound instances of images, texts, sound, images, opening up in principle to manipulation, extreme polarization, and possibly escalations of violence.

We can revisit these examples under the three requirements (control, foreesability, and evaluative ability) we considered before while analyzing moral agency. The robotic soldier has a better control and better foreseeability than humans, yet its moral evaluation provides very limited behavioural boundaries, making it indeed a very dangerous entity in a social setting. This is even more the case for the paperclip maximizer. The robotic soldier is a single entity which is localized in time and space. The paperclip maximizer may be instead realized as a collective "body" of orchestrated autonomous entities which could operate in parallel in several points in time and space. Its evaluation framework is evens further limited: a single concrete objective. Nothing good for humans may be expected from such a "god-like" and thick-headed entity.

These examples start suggesting that intelligence is not *per se* the source of the problem. We humans have general intelligence, yet we are not terminators, because the scopes

[11] The basic vs meta-level separation can be also related to the Kahneman's distinction between slow/fast decisions: fast decisions (even if they require a kind of moral evaluation) are based on the existing basic framework and can be streamlined in automatic processes, while slow decisions require also the intervention of a meta-level to carefully analyse the problem from different positions.

[12] See eg. `https://time.com/6295879/ai-pause-is-humanitys-best-bet-for-preventing-extinction/`

of our abilities and knowledge are limited and definitively bounded. Our interventions and observations in the world are *local*. Our communicative acts, before the Internet, and even more before printing, were much more reduced in scope. Even now, successful speech acts require acceptance, trust, or persuasion, so they are always mediated by further social mechanisms. Yet, we humans may still approach the terminator role when our interventions become much more impactful than what we were evolutionary selected to be, either at individual level (eg. atomic bombs) or at collective level (eg. climate warming). These two examples show that the underlying mechanisms concerning human ad artificial governance are more similar than it is generally assumed. The more the entity has control (ie. it is able to be successful in performing impactful actions), the more it requires foreseeability (ie. to be able to predict the impact it may produce)[13], the more it requires an adequate evaluation structure (eg. socially acceptable and sustainable).

From this perspective, online systems like ChatGPT raise concerns due to the breadth of their interventions (it interacts with many people simultaneously), and monitoring (besides all these communications, additional resources are provided for training), and even more, very little is known about their actual evaluation framework. From a technical point of view, we know that systems of this type are fine-tuned via Reinforcement Learning from Human Feedback (RLHF), relying on human feedback to e.g. minimize harmful or untruthful outputs. Yet, it is rather arguable whether mimicking human preferences is the best way to achieve moral behaviour. Human societies are based on two main mechanisms: on the one hand, normative systems, shared collectively, based on hierarchy and taking a top-down perspective towards the social system; on the other hand, decentralized, local autonomy, which, taking a bottom-up perspective, determines eventually the legitimacy of certain norms. Online systems like ChatGPT do not respect such an architecture, whereas locally trained transformers may offer a similar decentralization. Even those, however, do not offer an easy way to have access to an interpretable form of their underlying ”basic level”.

This issue unveils a more general principle. If we treat computational devices as mere tools (“hammers”) we would lack the methods to correctly deal with this problem. Instead, promoting computational architectures (eg. intentional agents) providing a sounder modularization may be the only way by which we can guarantee more alignment to humans.

5. Conclusion

Let us consider a simple analogy to illustrate the principles discussed in this paper. Suppose certain dog owners train their dogs to be particularly aggressive to trespassers to their estates. Suppose that their gate is left inadvertently open or is open due to a technical failure, one of the dog escapes and kills a person. Even if we correctly traced back to the dog’s actions the causes of such a dramatic outcome, we would still deem the owner fully liable (morally and legally responsible) for what happened. Whether allowing owners to train dogs to be aggressive (or even to breed super-dogs) is a social question which reflects in policy choices, and it is not a technical problem. Renouncing (to learn how) to train dogs is however only an indirect response to the issue, sub-optimal also because

[13] Interestingly, increasing foreseeability increases also the possibility of control, as it gives access to second-order actions, including manipulation.

there may be other circumstances in which training dogs is legitimate and of benefit for humans (eg. as guide dog for the blind).

This position paper argues that humans are always eventually responsible (at least on a policy level), and certainly are the only ones that can be liable. This is, at the very least, because machines can only cover lower-levels of responsibility and accountability. Yet, we also observed that machines should be deemed responsible and accountable of these lower-levels: whenever it would make sense so, they should not taken as mere "hammers". Not satisfying this requirement would deny developers of necessary means to improve their functioning. Furthermore, we argued that machines, as well as humans, may be "terminators" (in several senses), but it depends eventually on us to make in sort that to any increase of control, and foreseeability skills, there are adequate modifications on the evaluative framework side. Failing to do so would open up to improvident, possibly fatal consequences.[14] If we cannot guarantee this last part, then better not to allow any increase in the first two dimensions.

References

[1] Matthias A. The responsibility gap: Ascribing responsibility for the actions of learning automata. thics Inf Technol. 2004;(6):175-83.

[2] Santoni de Sio F, van den Hoven J. Meaningful Human Control over Autonomous Systems: A Philosophical Account. Frontiers in Robotics and AI. 2018;5.

[3] Bryson J. Robots should be slaves. In, Vol. 8. Close engagements with artificial companions: Key social, psychological, ethical and design issues (pp. 63–74); 2010.

[4] Metzinger T. Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology. Journal of Artificial Intelligence and Consciousness. 2021;1(8):1-24.

[5] Ji J, Qiu T, Chen B, Zhang B, Lou H, Wang K, et al.. AI Alignment: A Comprehensive Survey; 2023.

[6] Sowa JF. Knowledge Representation: Logical, Philosophical and Computational Foundations. USA: Brooks/Cole Publishing Co.; 1999.

[7] Sileno G, Saillenfest A, Dessalles JL. A computational model of moral and legal responsibility via simplicity theory. In: 30th international conference on Legal Knowledge and Information Systems (JURIX 2017). vol. 302. Luxembourg, Luxembourg: IOS Press; 2017. p. 171-6.

[8] Sharkey A. Can robots be responsible moral agents? And why should we care? Connection Science. 2017;29(3):210-6.

[9] Sileno G, Boer A, Gordon G, Rieder B. Like circles in the water: Responsibility as a system-level function. In: AI Approaches to the Complexity of Legal Systems, Proceedings of 3rd XAILA workshop: Explainable and Responsible AI and Law, in conjunction with JURIX 2020. vol. LNCS 13048. Springer; 2021. p. 198-211.

[10] Damasio AR. The feeling of what happens: Body and emotion in the making of consciousness. Houghton Mifflin Harcourt; 1999.

[11] Wallach W, Allen C, Franklin S. In: Wallach W, Asaro P, editors. Consciousness and Ethics: Artificially Conscious Moral Agents. London, UK: Routlege; 2017. p. 301-20.

[12] Torrance S. Ethics and Consciousness in Artificial Agents. AI and Society. 2008;22(4):495-521.

[13] Bostrom N. Ethical issues in advanced artificial intelligence. Science fiction and philosophy: from time travel to superintelligence. 2003:277-84.

[14] Echoes of this mechanism can be found in the climate warming problem. Our contemporary socio-economic practices increase our side-effects on a global level, and we are collectively impacting the world. Although we are possibly acknowledging more and more the consequences of our actions, we are much slower in implementing normative structures that effectively will reduce such a negative impact.