The Good, the Bad and the Robot: Projecting Attitudes into A.I.

Joan Casas-Roma [1]

Abstract. Artificial intelligent (AI) systems making autonomous decisions are present in many areas of our everyday lives. Ideally, and in order to facilitate the integration of these systems into our society, citizens should avoid thinking neither that AI resembles those rogue systems often found in fictional works, nor that it has an intrinsic understanding of human well-being, as both cases would draw a biased picture of what AI actually is. This position paper argues, through some examples, how the terminology used in certain articles aimed for the general public often depict AI as one of those two aforementioned images, which attribute intentions and moral values that those systems do not have and which could, in turn, have a detrimental effect on the way the general public understand (and are willing to accept) those systems into our everyday lives.

Keywords. artificial intelligence, artificial morality, projection, moral agency

1. Introduction and Motivations

As artificial intelligent (AI) systems become more autonomous, the scenarios in which they have to make morally-relevant decisions grow. Typical examples include military robots, self-driving cars, or robots for elderly care [6], to mention a few. The field of artificial morality studies whether and how AI systems can be furnished with the ability to identify morally-relevant situations and act accordingly.

Central to this topic is the debate on how to define the concepts of *moral agent* and *moral patient* and, furthermore, how to determine whether AI systems can be included in any of such groups (see [5] and [3], for example). Although, as pointed out in [5], technological tools are still often seen under an instrumentalist approach which does not usually take agency into consideration, works such as [3] or [6] define a different set of criteria that allows to understand agency at a functional level, and which do not require the agent to engage in genuine moral reflection, but only to act *as if* it did. Nevertheless, the question of moral agency in AI often takes into account issues concerning consciousness or intentions, among others. Authors such as [2] argue that, as artificial agents will never be able to engage in genuine moral reflection, they will never be proper moral agents —nor deserve any moral consideration as moral patients.

Regardless of that, we can already see examples of artificial agents acting in scenarios involving moral consequences, as pointed out by [1]. The way those events are perceived and sometimes communicated to non-specialized audiences often relies on using concepts that are heavily intentional and loaded with moral connotations, such as

[1] Falmouth University, Cornwall (United Kingdom); `joan.casasroma@falmouth.ac.uk`

"attack", or "save". However, and considering that most experts agree on the fact that AI systems cannot engage in genuine moral reflection, why are those terms used to talk about such systems, then?

This position paper points towards how the use of a certain kind of terms in mainstream media usually attribute intentions and moral values to AI systems that, more often than not, they do not have. Section 2 briefly shows examples of intentional attitudes attributed; Section 3 discusses how those "morally-loaded" terms would be applied to other entities such as inanimate objects, non-human animals or human beings; finally, Section 4 points towards further directions of inquiry on this research line.

2. Intention, Fiction and Reality

The way AI systems have been depicted through the lens of fictional stories plays an important part in this. As pointed out in [7], robots and AI systems in popular imagination are generally seen with certain fear and anxiety. From anthropomorphic, resilient robots that will not hesitate to harm a human being, such as *The Terminator*, to cold, goal-driven AI systems looking to fulfill their mission objectives at all costs, such as HAL 9000 in *2001: Space's Odyssey*, fiction has given us lots of stories about rogue AIs going astray and endangering human lives in the process.

Almost echoing those fictional stories, it is not uncommon to see newspapers' articles about malfunctioning AI systems that appeal to similar concepts as the ones we can find in those fictions, such as an article from *The Telegraph* featuring the headline "Robot vacuum cleaner 'attacks' South Korea housewife's hair"[2], or the *Daily Mail*'s article with the headline "Rise of the Terminator-style robots that can decide when and who to kill, warns expert"[3], for instance. Similarly, one can also find articles in the opposite direction of moral behavior, such as an article from *Mirror* with the headline "Robot 'uses initiative' to save little girl's life"[4].

Once one reads the articles, though, the first one points out to the fact that the housewife was sleeping on the floor, and so the robot *accidentally* started vacuuming her hair and got stuck in it, whereas the latter explains that the robot's action that resulted in saving the girl's life was pure coincidence, as the robot was simply mimicking the girl's movements, but with no awareness of neither the girl being in danger, nor of its actions preventing her from being hurt. Nevertheless, both the terms "attacks" and "saves" convey a sense of awareness and intention that, in neither case, the robots had. This kind of concepts ascribes intentional behaviors and moral values that, more often than not, only exist as a projection of our own human nature into systems that do not have them[5], and which may not, in many cases, be even able to recognize morally-relevant situations —let alone act in accordance.

[2] https://www.telegraph.co.uk/news/worldnews/asia/southkorea/11399713/Robot-vacuum-cleaner-attacks-South-Korea-housewifes-hair.html

[3] https://www.dailymail.co.uk/sciencetech/article-1204072/Warning-Rise-Terminator-style-robots-decide-kill.html

[4] https://www.mirror.co.uk/news/world-news/robot-uses-initiative-save-little-10746990

[5] In [4], the authors already argue how this anthropomorphization of technology could lead to an abrogation of human responsibility.

3. Moral Attitudes and Intentions: Tools, Animals and Beyond

What happens when we use this kind of language to talk about AI-driven systems? Take the vacuuming robot case, in which the headline of the article features the word “attacks”. Although it is only fair to note that the word is written under quotes, one’s thoughts are easily directed towards all those fictional stories of a robot uprising against humanity. Because, even if the word is (rightly) quoted in the headlines in order to give it a metaphorical sense, it still embeds a notion of intentionality that, even if only in a fictional sense, implies a will (or at least an awareness) of the robot to harm that person.

Suppose that we are writing a headline for an article saying that an unfortunate pedestrian was walking under a construction site, when a hammer accidentally fell off and hit that person. Would we say that the hammer “attacked” that pedestrian, even if we quoted the relevant word? Suppose now that the hammer did not fall on its own, but that a distracted construction worker accidentally pushed it with their foot and the hammer fell off the scaffolding. Would we write that the worker “attacked” the unfortunate pedestrian? Probably not, because we instantly recognize a lack of (at least) intention and awareness in both cases; nonetheless, these lacks seem to be often overlooked when the article is about AI systems like a vacuuming robot, even though, as pointed out by [5], machines are often understood as being instruments —like a very complex hammer.

Conversely, if our unfortunate pedestrian was not hit by a brick, but rather bitten by a dog, we would probably write a headline saying that a dog “attacked” that person (most likely without any kind of quotation). If our unfortunate pedestrian was, instead, harmed by another person, the headlines could change depending on the relevant situation. Say that the pedestrian was attacked by a burglar: for sure, the newspaper headline would feature the unquoted “attack”. Say now that the poor pedestrian gets kicked by a football player who unwillingly hit the pedestrian while aiming for the ball: the headlines would not probably feature the word “attacked” (either quoted or unquoted), and the nuance with respect to the previous case lies in the “unwillingly” bit. In order to feel that the word “attack” is being used in a legit way, we require the entity that is causing harm to fulfill some conditions of agency, such as a certain form of autonomy, choice, or intention —or at least some sort of awareness of that action harming the patient.

Although the previous paragraphs focus mainly on “negative” attributions, the same dynamics can be seen when attributing “good actions” to AI as a result of some sort of altruistic behaviors and features that robots may not have —namely, intrinsically caring about human lives, anticipating dangers and acting on their own initiative to prevent them. Even if some AI systems may be aimed towards that, such as robots in the field of healthcare that are explicitly aware of the well-being of their human patients, it could lead to the wrong impression that AI systems in general can (and should) do that, and thus may result in people tending to rely on those systems for things they are not meant to do, and which may not even be able to acknowledge.

This takes us right back to the question of AI and agency; because, if, by using those terms, we *can* attribute intentions to AI systems, then where does this leave AI in terms of its agency, and in relation to other entities, such as inanimate objects like tools, or non-human animals? By having briefly considered how those terms would apply to other kinds of entities, it can be seen how the terminology often used in communications aimed towards the general public draws a picture of AI that distances it from the way it is characterized by specialists in the field, and with respect to other kinds of entities like tools, or other living beings. Even though people working in the field may be aware of the actual state of affairs in AI, using this kind of terms in order to provide an appealing "punch-line" for the news can convey the wrong message to non-specialized audiences and draw a biased picture of what AI is, how it relates towards human beings, or whether it is even aware of human values and our general well-being.

4. Conclusions and Future Work

In this position paper it is shown through some examples how some newspapers' articles aimed for the general public often use certain terms that ascribe intentions and moral weight to AI systems and robots that challenge the status that machines have with respect to other entities; in particular, it can be seen how the use of those concepts situate AI in a fuzzy position oscillating between inanimate object, non-human animal, and often even reaching beyond that. Furthermore, it is suggested how the use of those terms for non-specialized audiences can have a detrimental effect on the integration of these technologies in our society by drawing the wrong picture of what AI actually is. This fact can easily become a double-edged sword: on the one hand, it can be a way of making certain AI systems, such as assistants, easier to relate to by projecting in them an intrinsic care towards human life; on the other hand, it can easily attribute to those systems non-existent intentions to either harm or help other human beings, which could make some people reluctant to accept them, or create expectations they cannot yet fulfill.

Those intuitions are laid out in this position paper as pressing issues, but they would need to be assessed by further research. Audience studies would need to be carried out to confirm, or disprove, the picture of AI the audience perceives from the media. Additionally, a comprehensive comparative between the way media aimed for the general public ascribe intentional and moral attitudes to inanimate objects, non-human animals, and human beings (the latter in cases of both intentional and accidental action) could provide the measurement scale to see where AI systems fit in this continuum; this could then be used to define some guidelines for journalists when writing about this topic. In addition, a challenging, but interesting future research would involve understanding how mistrust, or overconfidence in AI could hinder the effective integration of automated systems in our world, as well as foresee how those mislead perceptions could be exploited with nefarious purposes by people or organizations wanting to take advantage from that.

References

[1] I. Allen, C.; Smit and W. Wallach. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 7(3):149–155, 2005.

[2] J. J. Bryson. Robots should be slaves. *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*, pages 63–74, 2010.

[3] L. Floridi and J. W. Sanders. On the morality of artificial agents. *Minds and Machines*, 14(3):349–379, 2004.

[4] B. Friedman and P. Kahn. Human agency and responsible computing: Implications for computer system design. *Journal of Systems and Software*, pages 7–14, 1992.

[5] D. J. Gunkel. *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. MIT Press, 2012.

[6] C. Misselhorn. Artificial morality. concepts, issues and challenges. *Society*, 55(2):161–169, 2018.

[7] M. Szollosy. Freud, frankenstein and our fear of robots: projection in our cultural perception of technology. *AI & Society*, pages 433,439, 2017.