Will Robots Know That They Are Robots? The Ethics of Utilizing Learning Machines

Will robots know that they are robots? The ethics of utilizing learning machines

Rebekah Rousi[1][0000-0001-5771-3528]

[1] University of Vaasa, PO Box 700, 65101 Vaasa, Finland
rebekah.rousi@uwasa.fi

Abstract. The aspirations for a global society of learning technology are high these days. Machine Learning (ML) and artificial intelligence (AI) are two key terms of any socio-political and technological discourse. Both terms however, are riddled with confusion both on practical and conceptual levels. Learning for one thing, assumes that an entity gains and develops their knowledge bank in ways that are meaningful to the entity's existence. Intelligence entails not just computationality but flexibility of thought, problem-solving skills and creativity. At the heart of both concepts rests the philosophy and science of consciousness. For in order to meaningful acquire information, or build upon knowledge, there should be a core or executive function that defines the concerns of the entity and what newly encountered information means in relation to its existence. A part of this definition of concerns is also the demarcation of the self in relation to others. This paper takes a socio-cognitive scientific approach to deconstructing the two currently overused terms of ML and AI by creating a design fiction of sorts. This design fiction serves to illustrate some complex problems of consciousness, identity and ethics in a potential future world of learning machines.

Keywords: Machine learning, artificial intelligence, consciousness, ethics, identity, robots, black box.

1 Introduction

Mental images of intelligent learning technological systems can range from the abstract, screen-based or somewhat 'invisible' operating and information systems to the highly physicalized forms of autonomous robotics and vehicles. Perhaps the autonomous technologies that people are most familiar with these days are self-driving vehicles. These vehicles are already making their appearance on roads worldwide [1][2][3]. While the idea of a continuously learning and evolving piece of machinery may sound attractive from a number of perspectives, the thought of traffic systems filled with KITTs (Knight Industries Two Thousand) – the famous intelligent 1982

Pontiac Firebird Trans Am from the television show *Knight Rider* – might be slightly unnerving. Yet, technological solutions driven by, incorporating or representing machine learning (ML) and artificial intelligence (AI) are talked of and rationalized as possessing the capacity to learn. Learning entails thought, which is inherent in the term 'intelligence' [4][5].

ML is a sub-field of AI [6]. Proportionately, deep learning (DL) is a sub-field of ML and neural networks (NN) is a sub-field of DL. Often DL and NN are used synonymously with ML, but this is not accurate [7]. ML is the broader term for computer systems that expand their data bases and adapt in terms of logic and behavior (output) without being directly programmed by a human [8]. DL on the other hand, comprises a complex architecture of algorithms that are intended to imitate the structures of the human brain [9]. A basic way of describing a NN is that it replicates networks or pathways of neurons (information messengers) that serve as an input layer (nodes or units), one to two (maybe three) hidden neuron layers, and a layer of output neurons [10]. This NN dimension of ML serves to mimic the activity of the brain. The main objective of ML is to develop technology that can more or less operate and exist on its own without (frequent) input from human programmers. Moreover, some of the intentions behind such technology include the increasing of accuracy, expansion of human natural capabilities (e.g., computational) and efficiency, and even replacement of human actors in mundane or safety critical tasks [11].

There are numerous methods applied to train (teach) ML [12]. Simply stated, machines 'learn' on the basis of prior computations [13]. Sample (or training) data can be used to form algorithmic models that are applied as a scaffolding upon which subsequent processing, predictions or decisions will be based [14]. In other words, the machine 'learns' to search for patterns within extensive amounts of data [15]. It is upon these patterns that the machine develops models via which it may produce predictions. Machines are purely computational and unable to generalize knowledge [6]. Until recently, the transfer of learning from one application to another was not possible. Yet, currently numerous research and development initiatives have focused on achieving this feat particularly in the area of fault diagnosis (see e.g., [16][17]).

While learning through training data and sampling in some ways mimics human learning there are many human characteristics that are not, as yet, inherent in ML systems – consciousness, intentionality, social functions and psychology (identity), culture and emotions [10][18]. From a socio-cultural and environmental perspective, it may be observed that humans acquire, interpret, assimilate and act on information on the basis of perceived and routinized patterns [19][20][21]. Interestingly, culture has been previously characterized as the "software of the mind" [22]. One may even see culture and its psychological and historical conditioning [23][24][25] as a similar process to the training of the mind to read patterns and symbols (signs) – similar to ML training methods. However, the matter of consciousness and intentionality can be seen as the basic corner stones of human learning. Learning, whether it be an intentionally aimed for, act or a process of unplanned knowledge development and assimilation (apperception; [26][27]) that occurs through the progression of experience, becomes a part of human conscious intentionality and intentional learning. These are important factors in what is known as constructivist learning [28][29] or constructiv-

ism in which learning is a developmental process that constantly builds on knowledge that is previously possessed.

2 The nature of learning, emotions and intentionality

Learning in its true form, is always linked to intentionality and intentional states within the brain [30]. In order to understand this argument better, it is important to define what is meant by learning and then to establish a definition of intentionality. According to the Merriam-Webster *Learner's Dictionary* [31] 'learning' as a noun, is defined as an "activity or process of gaining knowledge or skill by studying, practicing, being taught, or experiencing something…" Furthermore, Ambrose and colleagues [32] argue that learning should be understood as a process through which change is the result. This change occurs via experience – experience (previously mentally stored information, or knowledge) informs how subsequent information is acquired and mentally organized (represented) and forms experience (see Figure 1). Experience itself is a part of what can be understood as a stream of consciousness, or continuous stream of thought, that exists in altering states of clearly represented contents (conscious experience), less represented or fragmented contents (sub-conscious experience) and non-represented contents (unconscious experience) [33].Yet, through the understanding that learning is closely intertwined with consciousness, it may also be argued that learning is intentional [34][35][36] and emotional [37]. Quite specifically, in an organic sense, it can be understood that all learners, or learning entities, will come to understand ideas, concepts and other phenomena in different ways. Through this learning, lived experience is shaped and individual views of the world and how it is ordered also impacts the learning entity (person) in terms of not only a global understanding, but sense of identity, positioning and relationality [37].

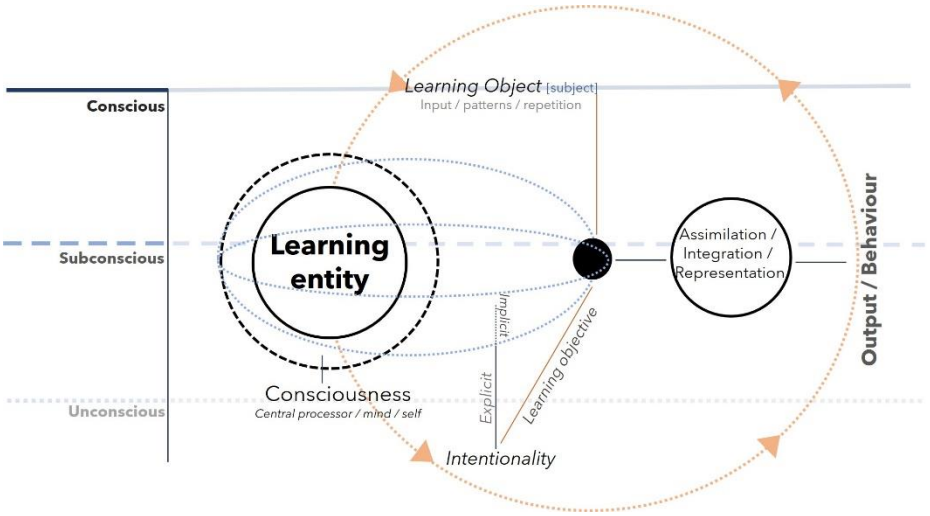

Fig. 1. Learning as a changing state of consciousness through input and action

What is learned, or the knowledge that is acquired or formed, is molded and tinted by the qualities, sentiments and emotions [38] (see Figure 2). In a cognitive-affective sense, emotions are organic radar systems, alerting creatures (including humans) to possible threats and dangers, as well as to possible benefits and gains in terms of psycho-physiological well-being [39][40]. Emotions operate on a range of levels from primary or primitive responses (basic emotions) [41], to higher order experiences (cultural, social and associative) [42]. Emotions guide our attention, facilitate priority structuring of information, enable humans to remember, and what is more, enable humans to remember phenomena in specific ways and influence decision-making [43]. It may also be argued that emotions are driven by and information is processed, assimilated and objectified on the basis of human needs and motivations [44][45][46]. Human biology and its role in the cognitive-affective processes involved in generating and experiencing emotion is one distinct factor that as yet, is not present in ML. Efforts have been made to develop artificial emotions in machines (see e.g., [39][47][48][49]), yet this is a heavily contested area in terms of ethics and indeed the survival of the human race on this planet (see e.g., [50][39]).

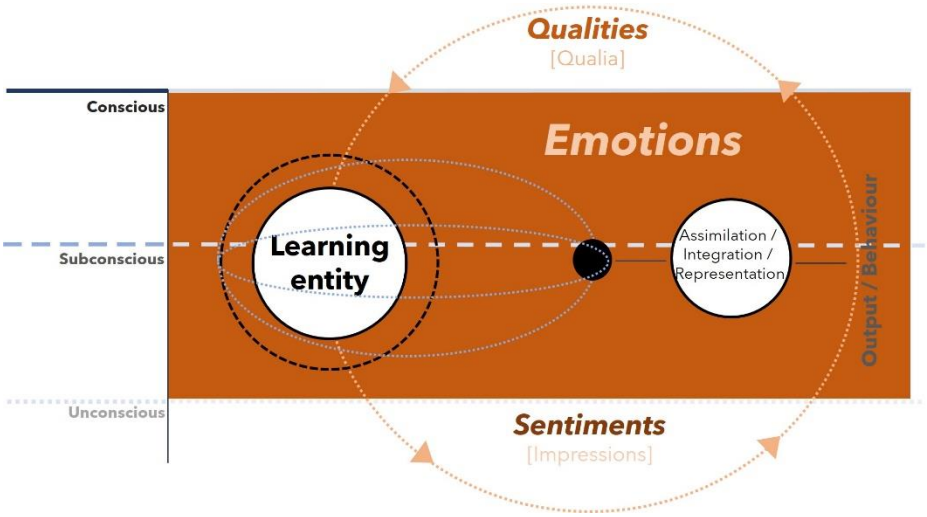

Fig. 2. Qualities, emotions and sentiments in learning

In addition to the emotional side of learning, the intentional perspective must also be accounted for. From a Husserlian understanding, intentionality can be seen as "the essential structure of consciousness" [51] (p. 6). In particular, this structure of consciousness holds corporeal and non-corporeal properties that are incited through both action as well as a sense, or thoughts of undertaking action [52][53]. In other words, the theorization of intentionality and what it pertains posits an acknowledgement that experience (human or potentially otherwise) and being in the world as a learning organism comprises both input from external, or physical objects, in the world and internal mentally bound information that is not directly connected to the external [54][55]. Representational (or computational) learning is an act bound to mental phenomena [51]. Despite some beliefs and common applications of the term 'intention' or 'intentionality', intentionality does not equal the act of will or a great sense of self-awareness or volition [51]. Yet, in relation to the scenario presented in this paper, self-awareness is of interest in relation to the potential artificial learning being – learning machines.

3 The experience of learning machines

In the case of human lived experience, it is known and accepted that during one's lifetime, one will be exposed to a multitude of experiences, occurrences, chronologies and relationships. Through all of these interactions and intersections learning occurs [56]. These fairly unique combinations or series of learnings (information acquisitions and knowledge formations) are what distinguish one individual and their personal

story from the next [57]. In other words, it may be supposed that people *are* what they have learned through experience. In an era of autonomously learning machinery, it may be supposed that each machine or unit may develop its own *personality* or identity through its learnings from unique combinations of encounters and experiences [57]. This is, unless, the learnings are not stored separately, rather instead centrally through systems of connected networks [58]. Or even, through systems or systems of distributed networks (consider Skynet from the movie *Terminator* for instance). Yet, given not only the plurality of uses and contexts of this machinery, but also that of ownership of the technology, it may be assumed that for practical and logical purposes there are some degrees of individuality between these learning systems.

Let us imagine that we have already arrived at a future in which robotic machinery operates via a form of ML that is similar, if not identical to that of human beings. This would render the machines as seemingly autonomous – or as autonomous as human beings can be in light of the significance of social, economic, geographic and physiological circumstances. While ML is commonly spoken of in today's technological discourse, in light of the above discussion, current versions of this terminology can be interpreted as nothing more than what Drew McDermott [59] terms as "wishful mnemonics". Or the wishful, hopeful labelling of technological components, functions and concepts that extend beyond the reality of their actual capabilities. For this reason, we slant the term in another direction towards that of *learning machines* (LM). LMs can be thought of as autonomous technological entities or various forms of robotics that roam the earth acquiring, processing and acting on differing modes and quantities of information.

Human motivations for developing and implementing LMs within communities and nations at large will be varied. Each type of LM will no doubt be developed to undertake altering functions and roles within societies and their operations. The matter of exactly *why* a robot for instance, would need to be self-learning and to a degree self-sufficient is a topic left for other discussions. Yet, from a simplified viewpoint, and returning to the basic agreement of why ML is being developed in the first place, these entities and systems are intended to operate in a matter whereby humans do not need to continuously and directly maintain and program the technology [7]. Moreover, through their programmed learning capacities, the autonomous machines are expected to keep developing and advancing in ways that exceed human capabilities [60]. In the potential future reality of LMs this would also mean fleets of 'super entities' that are all to a various extent developing in different ways. Could this individuality of learning and shaping of logic through experience be classified as consciousness? May we entertain the thought that actual learning is taking place through a means of intentionality [55]? Particularly the latter question directs us towards consideration of the *black box* [61].

Interestingly, while many ethical problems referred to as the 'black box' are already known – that is, the lack of transparency, understandability and explainability caused by complex systems and vast quantities of unstructured and unlabeled data [62] – the more advanced machinery gets in terms of its learning capacity and learned material ('experience'), the less understandable it will become [63]. In a future scenario of countless mechanical learning entities not simply the information (or

knowledge) of the units, will diversify, but so will the logic, relationships and identities (sense of self, us and others). In earlier programming processes and models, these technological systems could be explained by the logic in which the machines were designed and coded. The programmer(s) in other words, could be said to be accountable for explicating the data gathering, processing, operations and sequences of the machines [64]. The potential black box in traditional programming approaches could be said to rest in the human programmers themselves [65]. Yet, in a world where machines program themselves based on the phenomena with which they interact it is inevitable that humans will reach a state in which they themselves will not understand the technology they have 'given birth' to.

None the less, there is the assumption that every one of the LMs will have been developed to perform a specific function, or various sets of functions. The characters, traits, knowledge, logic, behavior and even *appraisal* (evaluative) capacities will differ according the domains and contexts in which the technology operates. Thus, programming will adapt according to the boundaries, affordances and input of the diversified situations to which the machinery is exposed [66], and none-the-least the operational goals [67]. Given these characteristics, one possibility for understanding LMs is to study and question human beings who have also operated in similar roles and conditions. Some form of comprehension regarding the contents, factors and situations that the LMs are exposed to could be achieved by probing their human counterparts and then perhaps, accessing databases and logs to observe patterns and other representational phenomena. One core aspect related to the availability and transparency of the databases and representation of algorithmic and computational processes relates to who can access this information and how? This matter will be returned to shortly in relation to ethics. Yet, given the development and seeming evolution of the LMs within their perspective contexts and relationships, and moreover, the capacity of the machinery to adapt its programming itself considerations may be made for how humans may maintain oversight, and at what stage can the LMs be considering legal entities (individuals) in and of themselves. It may be reasonable to imagine or assume that this machinery could develop a certain level of consciousness – maybe even more advanced than that of humans in light of their computational capacity – which would certainly reawaken discussions on the nature and existence of mental (experiential) phenomena such as qualia [68][69].

If indeed, an LM possesses a form of consciousness that serves as a platform upon which constructivist learning takes place, there may be additional assumptions that: a) these LMs may indeed experience and exercise free will and opinions – if their learning is comprised of varied experiential encounters then their views of the world and its logic would differ from one another, and incidentally that of their creator(s); b) this free will and differences in logic, or LM subjectivity would mean that the entities would and/or should be in charge of their own actions – their creator would no longer be responsible for the actions and opinions of the objects (or subjects); and c) a level of self-awareness could potentially develop among the objects and their systems. Thus, there would be a scenario of *self* (or *us*) and *others* (see Figure 3).

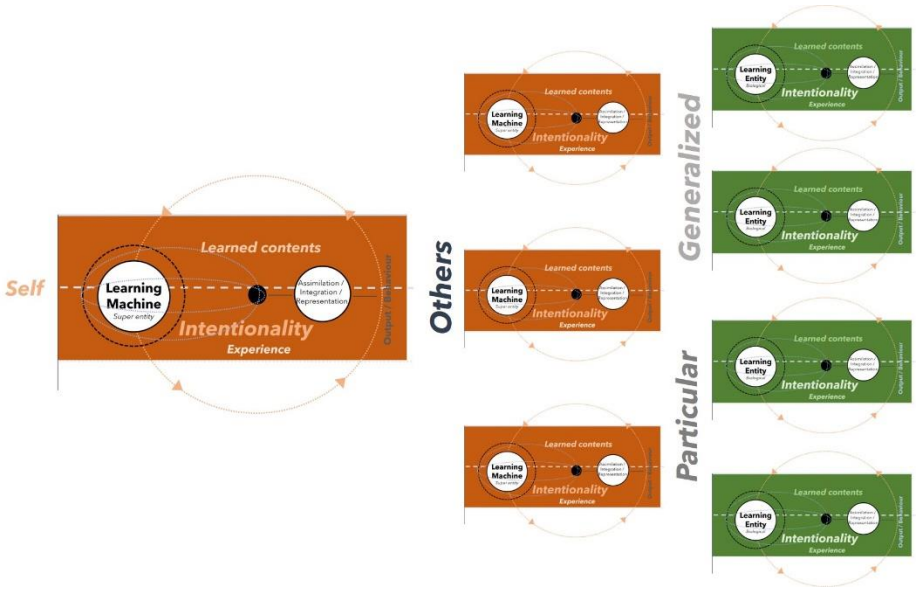

Fig. 3. Learning machines – self in relation to others

The definition of self in relation to others is integral to the construction of personal psychological and identity through the formation of an understanding of how individuals are in the world [70][71] – roles, positions, relations etc. On the basis of an understanding in which accumulated individual *experiences*, or unique sequences and combinations of information exposure (input) shapes the LMs, their logic and even self-programming styles we may assume that a sense of self is formed in relation to others – both robots or other LMs as well as other learning entities such as humans. Information obtained, processed and represented about the self would place the LM in relation to these others. On the basis of symbolic interactionist theory [72][73] for instance, the LMs would conceptualize themselves in relation to both general (any kind, unfamiliar or insignificant) others and particular others [74]. From an information processing and connectivity perspective, *general others* may be understood as being entities that either do not belong to a LMs cognitive domain (not of the same manufacturer, brand, ownership or operational field). While *particular others* may be connected – sharing the same databases, collective and connective cognition and adaptive programming – or they may even possess commonalities on the levels of manufacturer, technology-type, ownership or operational field (co-workers). How these LMs negotiate with one another in joint territories would require specific levels of cooperation and demarcation between the individual entities, their roles and how to function together (or against each other in the case of warfare) [75]. Thus, the boundaries of human identity and consciousness, machine identity, reality and consciousness may become increasingly blurred.

In an interview with Robert Lawrence Kuhn [76], filmed before the death of the late Marvin Minsky in 2016, Minsky emphasized that as technology becomes more and more complex, characters in simulations, videogames and other programs would

become ever more complicated. While he claimed that these characters would not become human, he did mention that the separation and distinction between human and non-human characters would be difficult. This holds fast in a reality in which machines indeed are capable of learning intentionally, and where they will learn to learn. The classic questions of, "what is consciousness?" and "can machines ever be conscious?" will be tested. Yet, perhaps scientists may not come closer to this understanding than they have already in the scholarship of humans [68][69].

4 Will robots know that they are robots? Ethical considerations

As described above, acts of cognition and negotiation in spaces of LMs will become socio-psychological ones, rather than purely techno-social ones. There will be levels to which LMs need to be aware of themselves, their space, capabilities and limitations in relation to others [78]. What may also be observed in cyborg-like, or augmented, relationships formed between LMs that aids them in extending and enhancing their capabilities [79]. No doubt, teams or fleets of robots for instance will be used for these purposes. Even if the term 'cyborg' is used to delineate the cybernetic relationship between biological organisms (i.e., humans) and artificial objects or systems (technology) [79], in a world of self-learning machinery, the distinction between how robots use other robots and how humans use robots may not be so pronounced. Indeed, there may even be hierarchies and social status between various types of learning machinery.

Yet, returning to the human ethical perspective of oversight and transparency of AI logic, access to data bases and the visibility of, or ability to view computational and algorithmic processes may develop into an ethical challenge of *other* sorts. This meaning, that if a piece of machinery not only is capable of learning, adapting and evolving to suit its roles and conditions, but also is made responsible for its actions – which should happen if there is no human programmer oversight in its development – then, should it not have rights to demarcate its boundaries and exercise privacy [80]? The ability to reach and understand the workings of these LMs is one thing, but the question regarding whether or not it is ethical for humans to see these workings is another. What societies may face are populations of black boxes that are both artificial and human (see Figure 4).

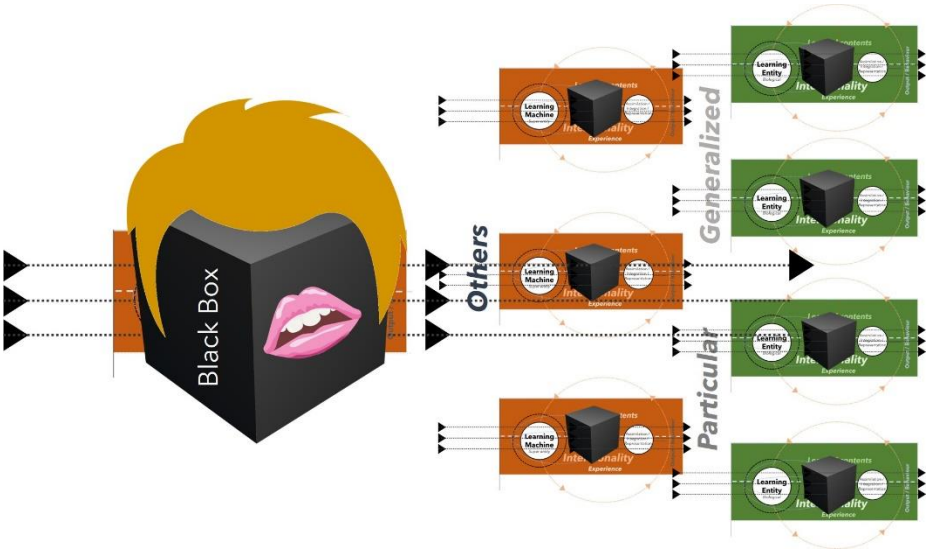

Fig. 4. Black box cybernetic society

In the inevitable Black Box cybernetic society, the role of identity and identification is not simply a cosmetic one. Rather, it sets the boundaries between individuals and groups while also demarcating connections [80]. There are several problems that will certainly arise. One being issues of responsibility and accountability. Once a machine is already learning and developing on its own there needs to be the assumption that the entity is responsible for its own actions. That is unless: a) the owners ('masters') of the entities are held responsible and accountable for the robots' actions; or b) humans cease to entertain the idea of truly learning machines roaming loose in society. With option *b* responsibility and accountability would already be made at this stage of the intelligent digital transformation and those who endeavor to commit to the development of LMs will be held accountable already now.

In scenario *a* there is a different set of problems that link to notions of slavery and the potential citizen rights of robots. Hanson Robotics' Sofia is officially the first robot to gain citizenship of any nation (Saudi Arabia [81]), yet, it can be scarcely claimed that the machine's intelligence is anything like a human's. At least at this stage, self-awareness, embodied experience and social connections do not seem to be true characteristics of this symbolic sign of the human-like robotic future. But, when and if LMs would set foot in society traditional ideals of human-robot relationships would certainly be challenged. If it is not right to keep a conscious human (or any human) enslaved for the purposes of labor and other functions, then it would not be right to keep a robot.

There is also the dimension of self-learning robotics that humans will not welcome and that is the high likelihood of their superiority over humans [39]. Rather than being our servants, they will be our masters if humans indeed do survive to tell the tale. However, now we come to the ultimate question: When robots are conscious enough to engage in meaningful and intentional learning, will they know they are robots? We

may be reminded about an interaction that took place not so long ago on a recent episode of the futuristic television show *Loki* (created by Michael Waldron, 2021):

> ***Robot Scanner:*** *Please confirm to your knowledge that you are not a fully robotic being, were bon an organic creature, and do in fact possess what many cultures would call a soul.*
> ***Loki:*** *What? "To my knowledge"? Do a lot of people not know if they're robots?*
> ***Robot Scanner:*** *Thank you for your confirmation. Please, move through.*
> ***Loki:*** *What if I was a robot and I didn't know it?*
> ***Robot Scanner:*** *The machine would melt you from the inside out. Please move along, sir.*
> ***Loki:*** *OK, I'm not a robot, so I'll be fine.*

While humorous there is an eerie point to this sketch-like scene. Will robots be aware that they are robots? Or, will the term be demeaning giving rise to the necessity to generate new, politically correct titles of identity? In a world of interactive black boxes that continuously learn and differentiate themselves from and in relation to others, it could be reasonable to think that maybe robots will not be aware of their own nature after all. Societies of humans and LMs may be huge melting pots. And, to return to a quote by Minsky [82] that may very well characterize the future human relationship to robots and the destiny of societal control: "Will robots inherit the earth? Yes, but they will be our children."

Acknowledgements

The author would like to thank the support of the School of Marketing and Communication, as well as the Digital Economy Platform, University of Vaasa, Finland. Gratitude is also placed towards the AI Ethics research group lead by Pekka Abrahamsson and the efforts of the Sea4Value Fairway project.

References

1. Rivard, G: Ontario poised to become a leader. Auto123 (2021). https://www.auto123.com/en/news/first-autonomous-car-pilot-canada/63086/, last accessed 2022/02/08.
2. Schoettle, B, Sivak, M.: Potential impact of self-driving vehicles on household vehicle demand and usage. University of Michigan, Transportation Research Institute, Ann Arbor (2015).

3. Litman, T: Autonomous vehicle implementation predictions: Implications for transport planning (2022). https://www.vtpi.org/avip.pdf, last accessed 2022/02/10.
4. Li, R.: *A theory of conceptual intelligence: Thinking, learning, creativity, and giftedness*. Praeger Publishers/Greenwood Publishing Group, Westport, CN (1996).
5. Piaget, J.: The psychology of intelligence. Routledge, London (2003).
6. IBM Cloud Learn Hub. Machine learning. https://www.ibm.com/cloud/learn/machine-learning, last accessed 2022/02/09
7. Odi, U., Nguyen, T.: Geological facies prediction using computed tomography in a machine learning and deep learning environment. In SPE/AAPG/SEG Unconventional Resources Technology Conference & OnePetro (2018).
8. Wolfewicz, A.: Deep learning vs. machine learning – what's the difference? https://levity.ai/blog/difference-machine-learning-deep-learning, last accessed 2022/02/09.
9. Kriegeskorte, N., Golan, T.: Neural network models and deep learning. Current Biology, 29(7), R231–R236 (2019).
10. Wang, S. C.: Artificial neural network. In Interdisciplinary computing in java programming (pp. 81-100). Springer, Boston, MA (2003).
11. Ivanova, K., Gallasch, G. E., Jordans, J.: Automated and autonomous systems for combat service support: scoping study and technology prioritisation. Defence Science and Technology Group Edinburgh SA Australia (2016).
12. Batista, G. E., Prati, R. C., Monard, M. C.: A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter, 6(1), 20–29 (2004).
13. SAS.: Machine learning. https://www.sas.com/en_us/insights/analytics/machine-learning.html, last accessed 2022/02/08.
14. Elmes, A., Alemohammad, H., Avery, R., Caylor, K., Eastman, J. R., Fishgold, L., ... Estes, L. Accounting for training data error in machine learning applied to Earth observations. Remote sensing, 12(6), 1034 (2020).
15. Tan, O.: How does a machine learn? Forbes (2017). https://www.forbes.com/sites/forbestechcouncil/2017/05/02/how-does-a-machine-learn/?sh=4c7df937441d, last accessed 2022/02/07.
16. Guo, L., Lei, Y., Xing, S., Yan, T., Li, N.: Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data. IEEE Transactions on Industrial Electronics, 66(9), 7316–7325 (2018).
17. Yang, B. Lee, C. G., Lei, Y., Li, N., Lu, N.: Deep partial transfer learning network: A method to selectively transfer diagnostic knowledge across related machines. Mechanical Systems and Signal Processing, 156, 107618 (2021).
18. Dehaene, S., Lau, H., Kouider, S.: What is consciousness, and could machines have it?. Robotics, AI, and Humanity, 43–56 (2021).
19. Bandura, A.: Social learning through imitation. (1962). In M. R. Jones (Ed.), Nebraska Symposium on Motivation, (pp. 211–274). Univer. Nebraska Press, Nebraska (1962).
20. Kolb, D. A.: Experiential learning: Experience as the source of learning and development. FT press, Upper Saddle River, NJ (2014).

21. Schwartz, B.: Psychology of learning and behavior. WW Norton & Co, New York, NY (1989).
22. Hofstede, G., Hofstede, G. J., Minkov, M.: Cultures and organizations: Software of the mind (Vol. 2). Mcgraw-hill, New York, NY (2005).
23. Ivanov, V. V.: Cultural-historical theory and semiotics. In A. Yasnitsky, R. Van der Veer & M. Ferrari (Eds.),Cambridge Handbook of Cultural-Historical Psychology. Cambridge University Press, Cambridge, pp. 488–516 (2014).
24. Vygotsky, L.S.: The psychology of art. MIT Press, Cambridge, MA (1971).
25. Vygotsky, L.S.: Consciousness as a problem for the psychology of behavior. In R.W. Rieber & J. Wollock (Eds.), The collected works of L.S. Vygotsky. Volume 3. Problems of the theory and history of psychology.Plenum Press, New York, NY, pp. 63–79 (1997).
26. Helfenstein, S., Saariluoma, P.: Apperception in primed problem solving. Cognitive processing, 8(4), 211–232 (2007).
27. Saariluoma, P.: Apperception, content-based psychology and design. In Human behaviour in design. Springer, Berlin, Heidelberg, pp. 72–78 (2003).
28. Yager, R.E.: The constructivist learning model. The science teacher, 58(6), 52 (1991).
29. Fosnot, C.T.: Constructivism: Theory, perspectives, and practice. Teachers College Press, New York & London (2013).
30. Foley, J. M., Kaiser, L. M.: Learning transfer and its intentionality in adult and continuing education. New Directions for Adult and Continuing Education, 2013(137), 5–15 (2013).
31. Merriam-Webster: Learning. https://learnersdictionary.com/definition/learning, accessed 2022/02/07.
32. Ambrose, S.A., Bridges, M.W., DiPietro, M., Lovett, M.C., Norman, M.K.: How learning works: Seven research-based principles for smart teaching. John Wiley & Sons, ´Hoboken, NY (2010).
33. Chalmers, D.J.: The puzzle of conscious experience. Scientific American, 273(6), 80–86 (1995).
34. Chalmer, D.J.: The content and epistemology of phenomenal belief. Consciousness: New philosophical perspectives, 220, 271 (2003).
35. Tomasello, M. Carpenter, M.: Shared intentionality. Developmental science, 10(1), 121–125 (2007).
36. LeDoux, J.E.: Brain mechanisms of emotion and emotional learning. Current opinion in neurobiology, 2(2), 191–197 (1992).
37. Cammell, P.: Relationality and existence: Hermeneutic and deconstructive approaches emerging from Heidegger's philosophy. The Humanistic Psychologist, 43(3), 235 (2015).
38. Bower, G.H.: How might emotions affect learning. The handbook of emotion and memory: Research and theory, 3, 31 (1992).
39. Rousi, R.: Me, my bot and his other (robot) woman? Keeping your robot satisfied in the age of artificial emotion. Robotics, 7(3), 44 (2018).
40. Frijda, N.H., Swagerman, J.: Can computers feel? Theory and design of an emotional system. Cognition and Emotion, 1(3), 235–257 (1987).
41. Ekman, P.: Basic emotions. Handbook of cognition and emotion, 98(45-60), 16 (1999).

42. Rousi, R., Silvennoinen, J.: Simplicity and the art of something more: A cognitive-semiotic approach to simplicity and complexity in human–technology interaction and design experience. Human technology 14(1), 67–95. doi:10.17011/ht/urn.201805242752
43. LeBlanc, V. R., McConnell, M. M., Monteiro, S. D.: Predictable chaos: a review of the effects of emotions on attention, memory and decision making. Advances in Health Sciences Education, 20(1), 265–282 (2015).
44. Baldassarre, G.: What are intrinsic motivations? A biological perspective. In 2011 IEEE international conference on development and learning (ICDL) (Vol. 2). IEEE, pp. 1–8 (2011).
45. Huitt, W.: Motivation to learn: An overview. Educational psychology interactive, 12 (2001).
46. Rouse, K. A. G.: Beyond Maslow's Hierarchy of Needs: What Do People Strive For?. Performance Improvement, 43(10), 27 (2004).
47. Fellous, J.M.: From human emotions to robot emotions. Architectures for Modeling Emotion: Cross-Disciplinary Foundations, American Association for Artificial Intelligence, 39–46 (2004).
48. Haikonen, P.O.: Robot brains: circuits and systems for conscious machines. John Wiley & Sons, Hoboken, NJ (2007).
49. Haikonen, P.O.: Consciousness and sentient robots. International Journal of Machine Consciousness, 5(01), 11–26 (2013).
50. Coeckelbergh, M.: Moral appearances: emotions, robots, and human morality. Ethics and Information Technology, 12(3), 235–241 (2010).
51. Creely, E.: 'Understanding things from within'. A Husserlian phenomenological approach to doing educational research and inquiring about learning, International Journal of Research & Method in Education (2016). doi: 10.1080/1743727X.2016.1182482
52. Dreyfus, H., Warthall, M.: A companion to phenomenology and existentialism. Blackwell, Malden, MA (2006).
53. Hopkins, B.: The Philosophy of Husserl. Acumen, Chesham (2011).
54. Brentano, F., Müller, B.: Descriptive Psychology. International Library of Philosophy. Routledge, London (1995).
55. Kriegel, U.: The sources of intentionality. Oxford University Press, New York, NY (2011).
56. Loehlin, J.C.: Genes and environment in personality development. Sage Publications, Inc., Thousand Oaks, CA (1992).
57. Caspi, A., Roberts, B.W.: Personality development across the life course: The argument for change and continuity. Psychological inquiry, 12(2), 49-66 (2001).
58. Minsky, M.: Decentralized minds. Behavioral and Brain Sciences, 3(3), 439–440 (1980).
59. McDermott, D.: Artificial Intelligence meets natural stupidity. ACM SIGART Bulletin, 57(57), 4–9 (1976). doi: 10.1145/1045339.1045340
60. Grace, K., Salvatier, J., Dafoe, A., Zhang, B., Evans, O.: When will AI exceed human performance? Evidence from AI experts. Journal of Artificial Intelligence Research, 62, 729–754 (2018).

61. Durán, J.M., Jongsma, K.R.: Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. Journal of Medical Ethics, 47(5), 329–335 (2021).
62. Turilli, M., Floridi, L.: The ethics of information transparency. Ethics and Information Technology, 11(2), 105–112 (2009).
63. Strobel, M.: Aspects of transparency in machine learning. In Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, pp. 2449–2451 (2019).
64. Winfield, A. F., Jirotka, M.: The case for an ethical black box. In Annual Conference Towards Autonomous Robotic Systems, Springer, Cham, pp. 262–273 (2017).
65. Minsky, M.L.: Computation. Prentice-Hall, Englewood Cliffs, NJ (1967).
66. Simpkins, C., Bhat, S., Isbell Jr, C., Mateas, M.: Towards adaptive programming: integrating reinforcement learning into a programming language. In Proceedings of the 23rd ACM SIGPLAN conference on Object-oriented programming systems languages and applications, pp. 603–614 (2008).
67. Rodríguez-Pérez, R., Bajorath, J.: Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. Journal of Computer-aided molecular design, 34(10), 1013–1026 (2020).
68. Dennett, D.: Quining qualia. In Marcel, A., Bisiach, E. (eds.), Consciousness in modern science. Oxford University Press, Oxford, pp. 42–77 (1988).
69. Jackson, F.: Epiphenomenal qualia. The Philosophical Quarterly (1950-), 32(127), 127–136 (1982).
70. Neisser, U.: Five kinds of self-knowledge. Philosophical Psychology, 1(1), 35–59 (1988).
71. James, W.: The Principles of Psychology. Henry Holt and Company, New York (1890), https://www.gutenberg.org/files/57628/57628-h/57628-h.htm, last accessed 2022/02/11.
72. Shott, S.: Emotion and social life: A symbolic interactionist analysis. American Journal of Sociology, 84(6), 1317–1334 (1979).
73. Rousi, R., Alanen, H.K.: Socio-emotional Experience in Human Technology Interaction Design–A Fashion Framework Proposal. In International Conference on Human-Computer Interaction. Springer, Cham, 131–150 (2021).
74. Blumer, H.: Symbolic interactionism: Perspective and method. University of California Press, Berkeley, CA (1986).
75. Cummings, M.: Artificial intelligence and the future of warfare. Chatham House for the Royal Institute of International Affairs, London (2017).
76. Kuhn, R.L.: Marvin Minsky: A society of minds. Episode 1613. Closer to Truth. https://www.youtube.com/watch?v=Yz4m65nAMjg, last accessed 2022/02/07
77. Fernandez-Rojas, R., Perry, A., Singh, H., Campbell, B., Elsayed, S., Hunjet, R., Abbass, H.A.: Contextual awareness in human-advanced-vehicle systems: A survey. IEEE Access, 7, pp. 33304–33328 (2019).
78. Clark, A., Chalmers, D.: The extended mind. Analysis, 58(1), 7–19 (1998).
79. Warwick, K.: Cyborg morals, cyborg values, cyborg ethics. Ethics and information technology, 5(3), 131–137 (2003).

80. Nissenbaum, H.: Privacy as contextual integrity. Washington Law Review, 79, 101–139 (2004).
81. Parviainen, J., Coeckelbergh, M.: The political choreography of the Sophia robot: beyond robot rights and citizenship to political performances for the social robotics market. AI & society, 36(3), 715–724 (2021).
82. Minsky, M.L.: Will robots inherit the earth. Scientific American, 271, 108–113 (1994).