Ethics of Artificial Intelligence Demarcations

Anders Braarud Hanssen and Stefano Nichele

Oslo Metropolitan University, AI Lab, {`anbh, stenic`}@oslomet.no

Abstract. In this paper, we present a set of key demarcations, particularly important when discussing ethical and societal issues of current AI research and applications. Properly distinguishing issues and concerns related to Artificial General Intelligence and weak AI, between symbolic and connectionist AI, AI methods, data and applications are prerequisites for an informed debate. Such demarcations would not only facilitate much-needed discussions on ethics on current AI technologies and research. In addition, sufficiently establishing such demarcations would also enhance knowledge-sharing and support rigor in interdisciplinary research between technical and social sciences.

Keywords: Artificial Intelligence · Ethics · Narrow AI · Artificial General Intelligence · Bias.

1 Introduction

The original goal of Artificial Intelligence (AI) research was to create an artificial (electronic) brain. This idea was explored in the seminal work by McCullock and Pitts [14], where they proposed a network of simplified abstract versions of biological neurons. The goal of creating a full artificial brain with the same degree of intelligence of a human brain is still an open challenge. From the idea of a brain capable of general (human) intelligence, the interest of the AI community quickly moved towards simplified (narrow) versions of artificial intelligence, to solve specific tasks.

The state-of-the-art in (narrow) AI was described by D. Waltz on the Scientific American back in 1982 [18] as *"Computer programs that not only play games but also process visual information, learn from experience and understand some natural language"*. In addition, he added that *"The most challenging task is simulating common sense"*. The current state-of-the-art in AI has not changed radically from Waltz's definition. Today the most compelling and less understood aspect is still the simulation of common sense, i.e., reasoning and cognition. The scaling of computational resources has allowed advances in playing computer games, computer vision, and natural language processing, pretty much with the same methods used in the '80s. While the initial AI inspiration was the human brain, in the meanwhile several methods to simulate intelligence without neural-based systems emerged, e.g., symbolic AI. Such methods had a certain degree of success thanks to the less need for computational resources. The recent availability of massive computational resources has allowed scaling of neural systems with results that surpassed non-neural systems in most application domains.

In December 2018, the European Commission's High-Level Expert Group on Artificial Intelligence has proposed the following updated definition of AI [7]:

"Artificial Intelligence (AI) refers to systems designed by humans that, given a complex goal, act in the physical of digital world by perceiving their environment, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best action(s) to take (according to pre-defined parameters) to achieve the given goal. AI systems can also be designed to learn and adapt their behaviour by analyzing how the environment is affected by their previous actions. As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)."

The current understanding of AI ethics is rather vague, due to the broad definitions of AI used in the literature, and do not necessarily reflect the aspects and demarcations within the research community, the algorithms and methods, the computing substrates [11], and the target applications.

In the remainder of this paper, we will outline and discuss some important AI demarcations which have strong implications for the ethical aspects and possible reflections to address key issues in research on societal impacts such as Responsible Research and Innovation (RRI).

2   AI Demarcations

2.1   Weak AI vs AGI

The first and perhaps the most well known AI demarcation is the one between Weak AI (also known as Narrow AI, Applied AI) and Artificial General Intelligence (also called Strong AI or Full AI). While Weak AI aims at making a machine learn to solve a specific task, AGI targets machines that can learn and perform any intellectual task. This implies that AGI has the ability to "learn to learn", as well as the ability of problem-solving, reasoning, modelling and planning. G. Marcus and Y. LeCun, two prominent AI researchers, while disagreeing in many aspects of the future of AI, agree on a list of seven points [13]:

- AI is still in its infancy
- Machine learning is fundamentally necessary for reaching strong AI
- Deep learning is a powerful technique for machine learning
- Deep learning is not sufficient on its own for cognition
- Model-free / Reinforcement learning is not the answer, either
- AI systems still need better internal forward models
- Commonsense reasoning remains fundamentally unsolved
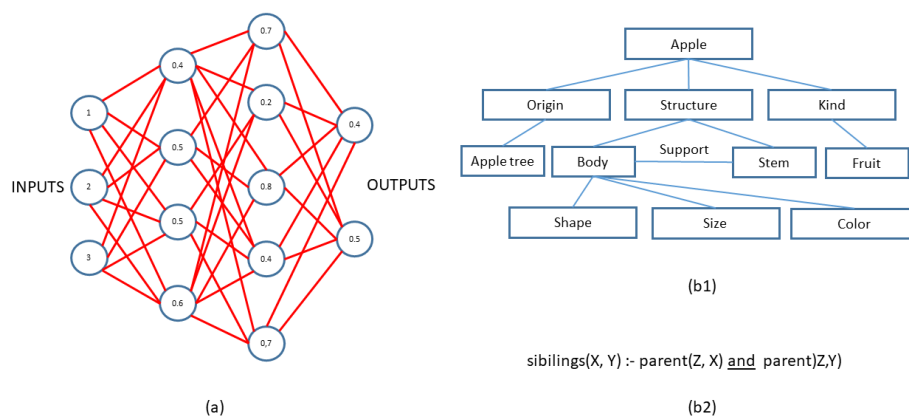
Fig. 1. (a) connectionist representation where information is represented by synapses (red lines) between neurons (blue nodes). (b) two examples of symbolic representations, b1 with a tree representation and b2 with logic expression.

It is evident that the demarcation between weak AI (all AI today and in the near future) and AGI implies that all current methods do not incorporate any form of commonsense reasoning, and the most used method of deep learning is not sufficient for a truly cognitive system. In addition, there is no current understanding or scientific theory on how commonsense reasoning could be achieved.

2.2   Symbolic AI vs Connectionist AI

Another important demarcation for AI systems is represented by the way information and relations are represented and encoded. In symbolic AI (also called algorithmic AI), knowledge is encoded in a symbolic form, together with rules to manipulate symbols and their relations. While symbol representation and manipulation makes it possible for a more rigorous study and explanation of weak AI systems, there is no evidence that the human brain is programmed as a symbolic machine. On the other hand, connectionist AI refers to a large network of units (neurons) that are interconnected together and encode/process information in a distributed way. While such models are more biologically plausible, they are typically data and compute hungry. Examples of symbolic and connectionist AI representations are depicted in Figure 1.

2.3   AI method vs Data

One demarcation that is often confused, especially in the context of AI bias, is the dataset used to train the AI model vs. the learning algorithm used to train the AI model (Note: the result of the training process using a specific dataset is an actual trained model, see next subsection). The fact that a trained model is

biased is a feature of the AI model and not a bug. In fact, if one wants to model a real-world system, the actual real-world model may be biased. The training algorithm is transparent to bias and therefore should not be attributed for the AI system being biased. If the intention of the AI model is to be unbiased, then the used dataset (the sole source of bias) should be corrected. One example of training algorithm for neural networks is backpropagation. Backpropagation involves mathematical operations such as calculating the derivative of the squared error function with respect to the weights of the network. This type of mathematical operations does not allow for algorithmic bias.

2.4   AI method vs application

The actual trained AI model, and therefore the application in which the AI is used, is not to be confused with the AI algorithm or method used for training. This demarcation is very important as restrictions have to be considered at the application level rather than at the AI method and algorithmic level (i.e., the method may be the same in very different domains, and obviously with very different sets of data). An example: regulating databases. It is the wrong level of abstractions. What is regulated is the use cases of databases (e.g., credit card companies, or insurance companies). We regulate lawyers, not Word. We regulate financial companies, not Excel. We do not regulate steel companies, we regulate guns. And we do not ask steel companies to regulate guns. AI is not an application, it is a general set of building blocks.

2.5   AI vs humans

Would regulators approve and consider ethically acceptable human-driven cars if they were invented today? Probably not. They are very dangerous by today's ethical standards. One important demarcation is human intelligence vs. artificial intelligence. Many of the issues with artificial intelligence are also present in human intelligence, e.g., black box. Can our intelligence be inspected when we drive a car? Is human intelligence open-source? Is the intelligence architecture known? Is the data used to train us for driving biased? Yes indeed. As we are given different datasets when we learn to drive. Is human intelligence deterministic? Can human intelligence be evaluated under different environmental conditions or noise? Are experiments repeatable? Those are all relevant questions to better understand the demarcation between human intelligence and AI.

2.6   Embodied AI

Features of the human cognition are shaped by aspects of the body (beyond the brain) [16]. Intelligence and cognition include high-level mental constructs (concepts and categories) and human performance on various cognitive tasks (reasoning / judgment), as a result of embodiment. Aspects of the body that shape cognition include the motor system, the perceptual systems, as well as the

body interactions with the environment. It is therefore expected that artificial general intelligence requires embodied agents living in an environment. One may argue that weak AI lacks often embodiment and a reactive environment.

2.7 Compelling questions

Through the description of the AI demarcations above, a list of relevant compelling questions for AI ethics emerged, and is listed below:

- What do we consider artificial intelligence?
- Are intelligent machines considered living machines?
- Can we demonstrate the emergence of intelligence and mind in an artificial living system?
- What ethical principles should be established for artificial general intelligence and weak artificial intelligence?
- What role do societal and ethical perspectives play in understanding the difference between human and machine intelligence?

3 Analysis

Based on the above synopsis of demarcations, we turn to how ethical and societal considerations are addressed within the generic field of AI. Current ethical and social science issues in and around artificial intelligence may benefit from a more rigorous articulation of demarcations within AI research. However, an overview of such issues would first benefit from situating applied ethics in new and emerging technologies.

Applied ethics as a discipline could be understood as relating to various practical applications of moral thought and principles and has a longstanding role within such fields as medicine, law and within various processions. In recent years, a range of approaches within ethics has addressed applications and implications of various ethical concerns within AI and machine learning [12], [6]. Nevertheless, scholars have argued that merely addressing specific technical and ethical concerns in isolation may not be the most viable approach to the legitimacy of the future of AI research and applications [1]. The humanities and social sciences have gained validity when facing the many uncertainties related to societal impacts of new and emerging technologies. Arguably, AI research poses unprecedented societal and ethical questions both to the nature of such research and its outcomes. Thus, specific ethical questions and implications could also be seen in a broader context. Among such broader considerations are the importance of stakeholder involvement, transparency and accountability. Such issues engage considerations beyond the discipline of applied AI ethics and involve questions of governance and policy. As a consequence, AI research benefits from research addressing these concerns in particular. But what kinds of research incorporate these broader concerns and how does such research sufficiently incorporate the necessary rigor related to technical demarcations within various sub-fields of AI?

Within new and emerging technologies, ethical and societal considerations gained prominence after the surge in genomic research in the U.S through the Human Genome Project (HGP) under the label of ethical, legal and social implications (ELSI)[8] and later through its European counterpart ELSA. ELSI Research was seen as a necessary component of addressing potential social and ethical implications of the vast uncertainties related to genomic research, particularly through its commercialization. These avenues of research have in recent decades been applied to new and emerging technology areas such as nanotechnology [10], synthetic biology [3] and various ICT research [15]. After 2010 these research areas, through increased awareness of policy considerations combined through the term responsible research and innovation (RRI), which soon was adopted by the European Unions Framework Programmes. From both a research and policy perspective RRI emphasized the need to take the societal, ethical and environmental impacts of emerging science and technology into account. Simultaneously RRI has emphasized the need to align research and innovation with societal challenges. Nevertheless, research on applied ethics on AI and also RRI-informed research on AI is still in its infancy and is fraught with many shortcomings. Among these are establishing necessary demarcations and distinctions that are both epistemic and normative in nature. In fact, very few examples exist in current RRI-literature where a clarification between key concepts and issues in both research and applications are undertaken in a systematic manner. Such clarifications and demarcations would inform the trajectory of various discussions around ethics and societal impacts. Arguably they would also contribute to establishing a better understanding and learning outcomes between AI researchers, ethicists and social scientists. To further illustrate the importance of such demarcations in the context of RRI- or other forms of social science-based research, a few key examples will be presented in the following paragraphs.

Sufficiently distinguishing between weak AI and AGI underscores the need to separate broad debates on AGI from timely and necessary reflection on societal embedding of weak AI. Although compelling, AGI debates are marked by both dystopian and utopian narratives and based on probability and hype [2]. Moreover, obvious knowledge gaps in the current research frontier seem to under-emphasize the limitations of the current understanding of common sense reasoning and cognition in humans. Such limitations are currently making the realization of superintelligent AGI unforeseeable [19] . Nevertheless, the recurring worst-case scenarios and hype of AGI threaten the legitimacy of various weak AI applications and research in the general public. These debates may also overshadow the need to address pressing questions in relation to governance and regulation or areas where weak AI already is being implemented. Moreover, weak AI-based research frequently lacks the presence of integrated social science and ethical perspectives in their design. Such perspectives may contribute to a broader understanding of challenges within areas such as machine learning. Designed on the semblance of human learning, it draws from cultural and social structures and extrapolates from them. However, a better understanding of how

algorithms build on such structures would also inform our understanding of what they cannot do, i.e such as present solutions for any scenario.

The nature of algorithmic design also shows that ethical and societal issues may be of very different natures with regard to current symbolic and connectionist AI and thus also provide very different ethical and societal scenarios. Symbolic AI may have vulnerabilities related to the quality of the design and/or hidden bias embedded into the algorithm itself, i.e bias related to relationships or symbols within symbolic language such as representing 'nurse' as 'female'. Although easily correctable it shows that ethical considerations such as gender equality point back to tacit linguistic biases and cannot be seen in isolation. At the same time, symbolic AI yields greater transparency towards such bias. Connectionist AI, as in deep neural networks, represents concerns more related to accountability and lack of transparency of what now seems to be 'black box-issues'. These may involve biases embedded in the data sets, such as societal, linguistic, cultural and heuristic biases that are embedded in data while at the same time present correlations that are context-sensitive, such as deep neural networks that may successfully predict sexual orientation by image analysis [20]. Further, by being 'data-hungry', connectionist AI seems vulnerable to error if data sets are not sufficiently substantial. Thus, in this regard ethical discourse around symbolic AI may yield results swiftly while in various connectionist AI context-of-use scenarios may be the most viable area of study.

The demarcation between the trained (applied) AI model and the training algorithm should inform what forms of ethical considerations are addressed; considerations that may often be misplaced. Some scholars argue regulation and law should primarily focus on the use of the model while the training algorithm itself could be considered as merely a tool [17]. Others argue that regulation and standardization are equally important in both [9]. Nevertheless, arguments that the training algorithms themselves are biased could be resolved by a proper demarcation between the training algorithm and application of the model. However, more research on the value-assumptions embedded in algorithmic training is needed, particularly in the discrepancy of the data embedded in the training algorithm and real-life scenarios [5]. Beyond these demarcations, there are different ethical considerations to be accounted for in the role that certain data sets play in the model and how an AI-tool is applied to various decision-making situations. In particular, if the bias is unknown or unidentified before the model is implemented it may have downstream impacts. Thus, in a range of scenarios, considerations such as distributive justice and or privacy may engage concerns related to both the training algorithm and the applied model. The data used in the training algorithm and the context of which the trained model is applied may in different ways combine to produce a complex set of urgent ethical and societal considerations. Nevertheless, distinguishing between the algorithm itself and the application of the model may itself resolve a range of unnecessary discussions about AI bias.

The distinction and/or similarities between human and artificial intelligence in relation to autonomous systems points to a discussion about the role of 'ethical

algorithms' that in many situations may be a misplaced concern. Albeit debates of whether the 'Trolley-problem' poses a contrived and unrealistic dilemma between utilitarianist and deontological reasoning in real-life scenarios [4], its resolution may reside elsewhere. Both humans and autonomous systems opaque reasoning may in decision-making situations pose unreasonable risks and uncertainties. However, as of yet, the unforeseen consequences of outsourcing legal agency and responsibility from humans to autonomous systems may potentially be a more considerable societal risk. The incommensurability of legal, ethical and scientific reasoning may here be a more pressing subject than making algorithms 'ethical' and should address such issues as the problem of legal accountability. To what extent may we accept bias in humans while not in autonomous systems if we consider autonomous systems bias to be a liability? Some may argue that ethical considerations for humans and machines should be considered distinct and separate. Humans are by default prone to error while legally accountable. Machines, who we seek to error-correct all the time, may be equally imperfect while in particular scenarios present algorithmic decision-making that seems ethically 'superior' to human action. While the question remains if we should base moral judgments on the outcomes or intention of an action, demarcating human and machine 'ethics' is a pressing concern. It would at least seem important to define what moral status human and machine action have when they are equally nontransparent. However opaque, human intelligence is a product of adaptation to the environment. This embodied aspect of intelligence may at least provide us with a demarcation between machine and human intelligence (embodied vs disembodied). Weak AI, by virtue of lacking embodiment, substantially differs in nature from human intelligence. It would thus follow that there should be other ethical considerations (i.e moral rights and obligations) for weak AI systems than embodied systems.

4   Conclusions

For the purpose of our discussion, a substantial part of the current debate about AGI evolves around threats and promise based on speculation. However, such concerns are less pressing and bound by considerable uncertainties and unresolved scientific challenges to develop a fully cognitive system. We argue that a proper demarcation between AGI and weak AI would facilitate a more informed debate about pressing concerns related to challenges and opportunities. Such scenarios are worthy of discussion, not only for ethicists but for AI research institutions, policy-makers and for society-at-large. Further, when ethical and societal impacts are discussed, a distinction between issues related to connectionist and symbolic AI is needed to be able to identify both vulnerabilities, risks and countermeasures. Similarly, in discussing AI bias, distinguishing between the AI method and the bias embedded into the data, the training algorithm itself and the trained model would resolve uncertainties and identify how and to what extent bias should play a role in training algorithms. Further, it would clarify where ethical and societal issues may most adequately be identified. Sufficiently

clarifying the difference between human and machine ethics points back to the insufficient demarcation between human and machine reasoning which is often equally opaque. Such a clarification would inform the debate on how to proceed with developing a more feasible 'machine ethics' and 'ethical algorithms', and potentially point the attention to issues of legal accountability.

It has been the overarching objective of this paper to illustrate the need for more rigor in the discourse on ethical and societal impacts of AI research in relation to a lack of sufficiently demarcating between key features of AI methods and tools. By providing illustrations of key demarcations, we have also suggested to what extent this may inform research on ethical and societal issues, both on weak AI and AGI. Further, through establishing the relevance such demarcations in the context of ethical and societal impacts, a range of ongoing discussions about AI could adopt them to provide a more nuanced and more elaborate dialogue across disciplinary boundaries. In particular, such a broader approach to ethical and societal impacts of AI should shift focus from isolated and narrow ethical questions to include governance and regulatory considerations and facilitate knowledge-sharing among stakeholders. However, approaches such as RRI may not successfully engage in knowledge-sharing with AI researchers if the aforementioned demarcations are not taken sufficiently into account.

References

1. Cath, C.: Governing artificial intelligence: ethical, legal and technical opportunities and challenges (2018)
2. Cave, S., Craig, C., Dihal, K.S., Dillon, S., Montgomery, J., Singler, B., Taylor, L.: Portrayals and perceptions of ai and why they matter (2018)
3. Coenen, C., Hennen, L., Link, H.: The ethics of synthetic biology. Contours of an emerging discourse. Technikfolgenabschätzung–Theorie und Praxis **18**(2), 82–87 (2009)
4. De Freitas, J., Anthony, S.E., Alvarez, G.: Doubting driverless dilemmas (2019)
5. Dent, K.: Ethical considerations for ai researchers. In: 2018 AAAI Spring Symposium Series (2018)
6. Etzioni, A., Etzioni, O.: Incorporating ethics into artificial intelligence. The Journal of Ethics **21**(4), 403–418 (2017)
7. EU: A definition of AI: main capabilities and scientific disciplines. High-level expert group on artificial intelligence - Reports and Studies (2018), https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines
8. Fisher, E.: Lessons learned from the ethical, legal and social implications program (elsi): Planning societal implications research for the national nanotechnology program. Technology in Society **27**(3), 321–328 (2005)
9. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a right to explanation. AI Magazine **38**(3), 50–57 (2017)
10. Hullmann, A.: European activities in the field of ethical, legal and social aspects (elsa) and governance of nanotechnology. DG Research, Brussels: European Commission (2008)
11. Konkoli, Z., Stepney, S., Broersma, H., Dini, P., Nehaniv, C.L., Nichele, S.: Philosophy of computation. In: Computational Matter, pp. 153–184. Springer (2018)

12. Luxton, D.D.: Artificial intelligence in behavioral and mental health care. Academic Press (2015)
13. Marcus, G., LeCun, Y.: Does AI need a more innate machinery? Debate between Yann LeCun and Gary Marcus at NYU, October 5 2017. Moderated by David Chalmers. Debate sponsored by the NYU center for Mind, Brain, and Consciousness. URL: https://www.youtube.com/watch?v=vdWPQ6iAkT4 (2017)
14. McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics **5**(4), 115–133 (1943)
15. Nydal, R., Myhr, A.I., Myskja, B.K.: From ethics of restriction to ethics of construction: Elsa research in norway. Nordic Journal of Science and Technology Studies **3**(1), 34–45 (2015)
16. Pfeifer, R., Bongard, J.: How the body shapes the way we think: a new view of intelligence. MIT press (2006)
17. Vayena, E., Blasimme, A., Cohen, I.G.: Machine learning in medicine: Addressing ethical challenges. PLoS medicine **15**(11), e1002689 (2018)
18. Waltz, D.L.: Artificial intelligence. Scientific American **247**(4), 118–135 (1982), http://www.jstor.org/stable/24966706
19. Wang, P., Liu, K., Dougherty, Q.: Conceptions of artificial intelligence and singularity. Information **9**(4), 79 (2018)
20. Wang, Y., Kosinski, M.: Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. Journal of personality and social psychology **114**(2), 246 (2018)