An Inclusive Ethical Design Perspective for a Flourishing Future with Artificial Intelligent Systems

Anne GERDES*

Abstracts

The article provides an inclusive outlook on artificial intelligence by introducing a three-legged design perspective that includes, but also moves beyond, ethical artificial systems design to stress the role of moral habituation of professionals and the general public. It is held that an inclusive ethical design perspective is essential for a flourishing future with artificial intelligence.

I. Introduction

The paper presents an inclusive outlook on the ethical design of artificial intelligent systems (AIS),[1] which emphasises the importance of a three-legged design perspective that includes but also moves beyond value alignment in ethical AIS design, to discuss the role of moral habituation of systems developer professionals[2] and the general public. Hence, within the community of AI-professionals, it is pivotal to facilitate ethical excellence cultivated in a professional practice that advance ethics as second nature. Moreover, to improve professionals' opportunities to bring ethics by design onboard, it is crucial to promote the use of value based design methods, enriched with a global outlook based on participatory design traditions. Equally, to empower citizens, AI literacy is a prerequisite for active democratic citizenship in tandem with a demand for explainable AI. It is argued that an inclusive ethical design perspective is necessary for a flourishing future with AIS.

To anchor these three perspectives, the paper fleshes out foundational historical ideas within the field of AI, as a precondition for subsequent reflections on key ethical challenges related to present and near future AI. Next, the design of ethical AIS is discussed, by pointing to value alignment strategies within machine ethics and machine learning. In continuation thereof, the paper explores how facilitation of moral self-cultivation in professional communities of practices may take place, and, furthermore,

* Associate Professor, The University of Southern Denmark, Department of Design and Communication.

[1] AIS broadly covers artificial intelligence in artificial intelligent systems and advanced intelligent robotics.

[2] The term "systems developer" refers to engineers, SW-developers, computer scientists, and practitioners involved in developing AIS.

European Journal of Risk Regulation, 9 (2018), pp. 677–689
doi:10.1017/err.2018.62

emphasises the role that participatory and value sensitive design methods may play in addressing ethical challenges in AIS development. In particular, it is argued that to prevent dual-use of AIS, contemporary system design methods (namely, value sensitive design and Privacy by Design) need to be enriched with a global outlook, backed up by international institutions. Finally, the paper emphasises the importance of empowering the public through educational, or *Bildungs*, initiatives to ensure that the general public achieves mastery-oriented strategies towards AIS, viz grows techno-moral wisdom, which facilitates navigation in a technology-mediated world changing apace.

II. Foundational ideas within the field of AI

The historical development of AI tells the story of a field, which from its start with the Dartmouth Conference 1956, where McCarthy coined the term "AI", until now has been passing through cycles of AI optimism and AI winters.

Initially, Turing's contribution with his two milestone papers 1937, "On Computable Numbers, With an Application to the Entscheidungsproblem"[3] and 1950, "Computing Machinery and Intelligence",[4] played a pivotal foundational role in shaping the field. Turning's 1937 paper was on abstract problems in mathematics, and represented a response to the decision problem as posed by Hilbert in asking whether there exists a step-by-step procedure enabling us to determine the truth status of any given statement in mathematics; obviously, $3 + 3 = 6$ is true, however for complex statements, the decision problem manifests itself. Similarly, Gödel's incompleteness theorem had shown that in systems beyond simple proposition logic, there are true statements, which cannot be proved in the given system (their status cannot be settled). In addressing Hilbert's problem of decidability in mathematics, Turing presented the idea of a step-by-step procedure, which turned out to be a conceptualisation of an abstract computer, the stored-program general-purpose computer, namely a Turing Machine. At the same time, at Princeton, Alonzo Church was working on a similar problem and came up with a response in his formulation of the Lambda calculus, ie an idealised programming language whereby a function is intuitively computable if and only if it can be written as a lambda term. Functional programming languages (eg Lisp) rest on the lambda calculus. Henceforth, the answer to the decision problem is referred to as the Church-Turing thesis.

Against this backdrop, Turing presented the idea of an algorithm, illustrating that there is a procedure for deriving correct descriptions of a system's outputs from its inputs. Consequently, the Church-Turing thesis states that a universal symbol system can simulate any algorithmically calculable system, ie a universal computer with the right program can simulate anything computed by any other computer. So, if our cognitive processes are algorithmically calculable, then, *in principle*, a computer could simulate the mind in an idealised world with no limits on computer time or memory (although in practice this might not be feasible, due to limited time resources, for example). Turing

[3] AM Turing, "On Computable Numbers, With an Application to the Entscheidungsproblem" (1937) XLII Proceedings of the London Mathematical Society, Series 2, 236.

[4] AM Turing, "Computing Machinery and Intelligence" (1950) 59 Mind 434.

continues this line of argument in his paper on "Computing Machinery and Intelligence" (1950) emphasising the theoretical importance of so-called "infinitive capacity computers":

> "Most actual digital computers have only a finite store. There is no theoretical difficulty in the idea of a computer with an unlimited store. Of course, only a finite part can have been used at any one time. Likewise only a finite amount can have been constructed, but we can imagine more and more being added as required. Such computers have special theoretical interest and will be called infinitive capacity computers".[5]

Here, Turing presents his famous Turing test. He proposes an imitation game, suggesting that rather than asking what intelligence is, one ought to address this issue by asking what verbal behaviour an intelligent system needs to display for us to judge it as intelligent? Hence, a computer system may pass the Turing test if one is unable to distinguish whether a computer or human produces utterances. Moreover, as a mean to reveal human cognition, Turing presented what has later become known as knowledge acquisition engineering, ie the idea that one can extract and externalise expert knowledge and translate it to symbols as preparation for programming it into a computer:

> "The book of rules which we have described our human computer as using is, of course, a convenient fiction. Actual human computers really remember what they have got to do. If one wants to make a machine mimic the behavior of the human-computer in some complex operation one has to ask him how it is done and then translate the answer into the form of an instruction table. Constructing instruction tables is usually described as 'programming'. To 'programme a machine to carry out the operation A' means to put the appropriate instruction table into the machine so that it will do A".[6]

This assumption is also reflected in the idea of the physical symbol system hypothesis, which dominated the field of AI, arguing that formal knowledge representation via symbol manipulation would provide for AI.[7] However, the physical symbol system hypothesis failed to deliver results, and so did the parallel distributed paradigm, which was introduced in the first work on a neural network model by McCulloch and Pits (1943). Subsequently, Marvin Minsky and Dean Edmonds made SNARC, the first neural network computer in 1950. Unfortunately, Minsky and Papert's book *Perceptrons – An Introduction to Computational Geometry* (1969) illustrated the shortcomings of neural networks, which could not be trained to deliver significant results. This pessimistic message has been said to have influenced research funding in favour of the physical symbol system hypothesis.[8] As is well-known, formalisation of common-sense and tacit knowledge could not escape the frame problem and seemed to be an intractable computational task, independently of whether one had available sufficient computer processing and storage power.

Consequently, only small successes were seen with models working on limited microworlds (eg IBM's Deep Blue beat Garry Kasparov in chess. However, this success could not easily be transferred to other domains). Hence, the symbol reasoning paradigm

[5] ibid, 438.

[6] ibid, 438.

[7] A Newell and HA Simon, "Computer science as empirical inquiry: symbols and search" (1976) 19(3) CACM 113.

[8] The historical outline in this section also draws on SJ Russell and P Norvig, *Artificial Intelligence – A Modern Approach* (Pearson 2015) pp 16–27.

ruled until the early 1980s, when the connectionist paradigm was revitalised. However, lack of sufficient amounts of training data sets and computer power led to sparse successes within limited domains (today, however, "… the current view is that connectionist and symbolic approaches are complementary, not competing"[9]). The continuing lack of a breakthrough in AI kept Dreyfus busy over the years, retelling versions of his joke on AI research:

> "According to this definition, the first man to climb a tree could claim tangible progress toward flight to the moon".[10]

Or:

> "… claims and hopes for progress in … making computers intelligent are like the belief that someone climbing a tree is making progress towards reaching the moon".[11]

The AI optimism was followed by the worst AI winter ever in the late 1980s. Ironically, it was not until the AI research field was almost dried out that results started to emerge. Challenges in pattern recognition, language understanding, speech and vision systems, and decision systems were overcome by a combination of large-scale data sets, statistics, computer power, and cloud storage in tandem with use of deep learning algorithms in the field of machine learning:

> "Google can translate with reasonable competence between 135 written languages … word by word, these AIs simply calculate the likelihood of what comes next… it is just a matter of probabilities".[12]

Likewise, there are obvious reasons why we should apply AIS with the purpose of fighting the big challenges of our times, such as cancer, epidemics, and global warming. For instance, AIS may play a key role in accelerating knowledge discovery and scientific breakthroughs, as illustrated below, in the quotation from a global Mission Innovation initiative on clean energy innovation:

> "The main recommendation coming from the workshop participants is the need to develop the materials acceleration platform(s) (MAPs), which integrate automated robotic machinery with rapid characterization and AI to accelerate the pace of discovery. The deployment of the proposed acceleration platforms will unleash a 'Moore's law for scientific discovery' that will speed up the discovery of materials at least by a factor of ten – from 20 years to 1 to 2 years. This will catalyze a transition from an Edisonian approach to scientific discovery to an era of inverse design, where the desired property drives the rapid exploration, with the aid of advanced computing and AI, of materials space and the synthesis of targeted materials. The inverse design of materials allows for their accelerated scale-up into installed technologies, accelerating energy technology innovation. This, in turn, will benefit all seven Innovation Challenges of Mission Innovation".[13]

---

[9] ibid, p 25.

[10] HHL Dreyfus, *Alchemy and Artificial Intelligence* (T Belvoir: Defence Technical Information Center 1965) p 17.

[11] HHL Dreyfus, "Why Heideggerian AI failed and how fixing it would require making it more Heideggerian" (2007) 171(18) *Artificial Intelligence* 1142.

[12] N Christianini, "Machines that learn: the mechanics of artificial minds" in D Heaven (ed), *Machines that Think* (New Scientist 2017) pp 36–37.

[13] See A Aspuru-Guzik et al, *Materials Acceleration Platform – Accelerating Advanced Energy Materials Discovery by Integrating High-Throughput Methods with Artificial Intelligence* (January 2018) p ii, available at

Visions for the envisaged acceleration platform for discovery of clean energy materials have not attracted much attention in the public debate, despite being highly relevant. On the other hand, a more audience-friendly, and at the same time the most notable breakthrough in AI, took place in 2016 where DeepMind's (Google Alphabet Group) unsupervised deep reinforcement learning algorithm, AlphaGo, defeated Lee Sedol, the South Korean master of the ancient game GO. Basically, AlphaGo learns by experience, starting out mimicking data on human amateur GO players' style before moving on to the expert level by learning from expert play data streams. After having beaten experts, it plays against itself to move beyond human level Go skills. Using reinforcement learning, it learns to improve itself incrementally and to side-step errors. The reinforcement model AlphaGo uses is model-free, implying that AlphaGo does not need a pre-programmed structure or theory to learn. Because its reinforcement learning is model-free, it may carry over to other domains. Such model-free learning holds promises that artificial general-purpose intelligence or super intelligence (beyond human-level intelligence) might be lurking in the horizon.[14]

III. The quest for value alignment

As such, we are increasingly faced with ethical issues; particularly we need to address the challenges following the cocktail of a presumed intelligence explosion and a self-modifying AI that might change its optimisation goals without humans being able to interfere or even discover that this was the case.[15]

The most significant difference between early days AIS, of the physical symbol system paradigm or the classic connectionist paradigm, is that these AIS were not model-free, but involved engineering by hand.[16] Also, conventional machine learning techniques required domain expertise to prescribe which features the system should be paying attention to when confronted with raw data (which were not big in the 1980s). Hence, all thought their efficiency was limited to narrow domains, formal verification and validation processes could successfully be applied. As such, these AIS produced results, which could be scrutinised by applying techniques, such as reverse engineering. However, this is not the case with reverse reinforcement learning systems. For instance, with the revival of back-propagation learning algorithms for multilayer networks, in tandem with big data training sets, we are dealing with AIS that work by sending error outcomes backwards in the network for a "tweak" of the weights to change the neuron activation trigger up to a point at which the network fits the smallest possible error rate.[17]

(F'note continued)
< mission-innovation.net/wp-content/uploads/2018/01/Mission-Innovation-IC6-Report-Materials-Acceleration-Platform-Jan-2018.pdf > accessed 14 November 2018.

[14] Heaven, supra, note 12, p 75.

[15] E Yudkowsky, "Artificial Intelligence as a Positive and Negative Factor in Global Risk" in N Bostrom and MM Ćirković (eds), *Global Catastrophic Risks* (Oxford University Press 2011) p 308.

[16] Y LeCun et al, "Deep Learning" (2015) 521(28) Nature 436.

[17] P Vossen, 'How AI detectives are cracking open the Black Box of deep learning' (2017) Science, 6 July < www.sciencemag.org/news/2017/07/how-ai-detectives-are-cracking-open-black-box-deep-learning >, accessed 14 November 2018.

Consequently, as learning starts to become the algorithm's responsibility, AIS increasingly function in mysterious ways. Some argue that we have to trade off understandable AI for AI that works:

> "Yet the reasoning made by a data-driven artificial mind today is a massively complex statistical analysis of an immense number of data points. It means that we have traded 'why' for simple 'what'. Even if a skilled technician could follow the maths, it might not be meaningful. It would not reveal why it made a decision, because the decision was not arrived at by a set of rules that a human could interpret, says Microsoft's Chris Bishop. But he thinks this is an acceptable trade-off for systems that work. Early artificial minds may have been transparent, but they failed. (..) 'Explainability is a social agreement' says … Christianini. 'In the past we decided it mattered. Now we've decided it doesn't matter'".[18]

Nevertheless, it is reasonable to say that black box AIS give rise to both ethical as well as epistemic discomfort. One may seriously doubt that we will trust AIS with important decision-making without explainability working as an interface between AIS and humans. Fortunately, not all agree that explainability comes at the cost of AIS that works. Currently, finding ways to build predictability into AIS is a highly prioritised research area.[19] For instance, the Darpa Explainable AI program has as its main goal to turn black-box models into glass-box models, implying that humans are kept in the loop and able to understand the AIS reasoning and cognitive models.[20] Likewise, some seek to develop algorithms that, by probing, can look inside the black box of a trained "self-made" neural network and reveal the logic behind a subset of results, but not the overall logic of the system. Others "embrace the darkness" and rely on neural networks to explore other neural networks.[21]

Against the backdrop of these reflections, we need to safeguard AIS to ensure transparency by explainable AI (output that is meaningful and understandable to us) and interpretable AI (technical insight into how neural networks operate). We must make sure that a deep learning data-driven AI is capable of navigating in complex domains with moral ambiguity, while acting in coherence with our value systems and at the same time reason about its knowledge discovery processes in a manner understandable to us.

Different approaches are suggesting how to tame AIS to ensure that their strategies are aligned with human ethical values and norms. However, pure machine ethics strategies, that seek to top-down wire ethical guidelines into the system design (by relying on formalisation of selected moral theories and explicit specification of preferences), might not be a feasible strategy for (near) future AIS, which will presumably be acting with high degrees of autonomy in open-ended domains.[22] Rather, it seems to be more fruitful to investigate bottom-up machine learning approaches, and apply reinforcement learning as a strategy to let systems learn ethics by doing, ie by being capable of inferring value preferences via observations of human ethical behaviour:

---

[18] Supra, note 14, p 44.

[19] S Russell, D Dewey, and M Tegmark, 'Research Priorities for Robust and Beneficial Artificial Intelligence' (2015) AI Magazine, Winter, 105.

[20] D Gunning, 'Explainable Artificial Intelligence (XAI)' (2018), available at <www.darpa.mil/program/explainable-artificial-intelligence>, accessed 14 November 2018.

[21] Supra, note 17.

[22] Supra, note 19, p 110, also see A Gerdes and P Øhrstrøm, "Issues in robot ethics seen through the lens of a moral Turing Test" (2015) 13(2) Journal of Information, Communication and Ethics in Society 98.

> "Reinforcement learning raises its own problems: when systems become very capable and general, then an effect ... is likely to occur, in which sophisticated agents attempt to manipulate or directly control their reward signals ... this motivates research areas that could improve our ability to engineer systems that can learn or acquire values at run time ... For example, inverse reinforcement learning may offer a viable approach, in which a system infers preferences of another rational or nearly rational actor by observing its behaviour".[23]

However, we cannot be sure that learning ethics by experience will result in morally right-doing. Reinforcement learning networks must be modified to avoid human bias.[24] Hence, Wallach and Allen suggest a hybrid model, which uses bottom-up learning models combined with a top-down virtue ethical architecture to facilitate and scaffold a process that might lead to systems which acquire ethical excellence aligned with human values.[25]

IV. Professional communities – cultivating for ethics as second nature

It is a complicated matter to settle what shared human value systems, preferences, and norms are significant, ie which and whose ethical and cultural values do we have to pay attention to in order to form the base for the kind of value alignment needed to bring about beneficial AI? For that reason, the quest for value alignment must be seen in tandem with moral habituation into social roles and responsibilities within systems developers' professional communities.

Here, the Aristotelian concept of phronesis[26] may inform us by pointing to the importance of the kind of situated practical wisdom that we eventually grow through life experience, via influence from upbringing, mentors, and education. Accordingly, a theoretical guideline to ethically wise decision making does not exist. Most of us have struggled with the trolley problem in philosophical classes. What the trolley problem illustrates is, as, pointed out by Rennix and Robinson, that "(..) it's very easy to temporarily become a psychopath if your professor says doing so will be intellectually useful".[27]

Instead, Aristotle emphasises practical experience and the situated basis of phronetic enactment and argues that young people may be good at reasoning while lacking ethical skills, which the phronimos has, since he is capable of reflecting upon how to bring about good actions against the backdrop of his life experience.[28]

---

[23] Supra, note 19, p 110.

[24] As in the epic case of Microsoft's twitter bot TAY, which was supposed to learn to communicate via real-time data streams of tweets. However, alas, overnight, Tay's tweets turned evil-minded. This could have been foreseen if Microsoft had "tweaked" Tay to prevent influence from biased data sets (eg from trolls and people making bad jokes), which would let Tay digest data in a morally sound manner.

[25] W Wallach and C Allen, *Moral Machines – Teaching Robots Right from Wrong* (Oxford University Press 2009).

[26] See NE Book 2, Aristotle, *Nicomachean Ethics* H Rackman (trans) (Harvard University Press, William Heinemann Ltd 1934).

[27] B Rennix and NJ Robinson, "The Trolly Problem Will Tell You Nothing Useful about Morality" (2017) Current Affairs, available at < www.currentaffairs.org/2017/11/the-trolley-problem-will-tell-you-nothing-useful-about-morality > accessed 14 November 2018.

[28] Supra, note 26, NE, Book 6.

By itself, experience does not automatically bring about phronesis: "according to Aristotle ... excellence of character and intelligence cannot be separated".[29] Aristotle holds that standards of good practice need also be anchored in rational knowledge, implying that a flourishing life requires rational reasoning activities in accordance with virtue or excellence.[30] As such, Aristotle distinguishes between intellectual virtues acquired by education, and virtues of character, and ethics grown by habitual exercise.[31] When facing ethical challenges, phronesis enables us to rely on situational awareness informed by both particular as well as general knowledge (*nous*). Phronesis can then be understood as a special kind of intellectual virtue, which is at the same time an ethical virtue.[32]

With Aristotelian insights, it becomes essential to facilitate phronetic enactment by fostering a kind of ethical preparedness, which is cultivated in educational as well as organisational practices with the purpose of empowering the individual systems developer as well as systems developers' communities to enact moral engagement and willingness to take a stance to hinder moral wrongdoing. Similarly, Vallor highlights the role of *relational understanding*[33] as important to our moral self-cultivation, and furthermore notices that both Western and Eastern (Confucian and Buddhist ethics) virtue traditions emphasise relationships as focal to our moral development and practice. This perspective is opposed to the Kantian notion of the autonomous agent, who undertakes moral deliberation in isolation while referring to universal principles disconnected from the unique context in which she is situated:

> "... the ideal moral agent is *not* the person who most successfully detaches her deliberations from her own relational context, but rather the one who understands and responds to that context most fully – that is, a person of practical wisdom".[34]

With departure in systems developers' professional relational context, it may be possible for systems developers to become empowered and capable of prioritising, and negotiating ethics in design to ensure the design of beneficial and ethical AIS. This will not guarantee that systems developers can anticipate all kind of ethical problems. Nevertheless, moral self-cultivation by habituation in practices of excellence might build characters prone to carry out phronetic activities based on situated negotiation about moral choices followed by phronetic enactment.

As such, we need to foster community practices that advance ethics as second nature. How this is done is a complex question. For now, I will narrow it down to focus on a key criterion, namely the importance of being capable of critical reflection through dialogue, as suggested by Habermas in his notion of communicative actions. Here, Habermas points to the importance of dialogue among equal participants aiming towards a mutual

[29] A MacIntyre, *After Virtue – a Study in Moral Theory* (Oxford University Press 2000) p 154.

[30] J Dunne, *Back to the Rough Ground – Practical Judgment and the Lure of Technique* (University of Notre Dame Press 1993) p 275.

[31] Supra, note 29, p 154.

[32] O Eikeland, *The Ways of Aristotle – Aristotelian Phrónesis, Aristitelian Philosophy of Dialoge, and Action Research* (Peter Lang AG 2008) p 53.

[33] S Vallor, *Technology and the Virtues – a Philosophical Guide to a Future Worth Wanting* (Oxford University Press 2016) p 76.

[34] ibid, p 77.

understanding, which is based on critical reflection, freed from power relations.[35] Similarly, Arendt argues that good judgment is not anchored in subjective or objective knowledge. Instead, moral judgement is related to intersubjectivity and responsiveness towards others' perspectives. "To think with an enlarged mentality means that one trains one's imagination to go visiting".[36] As such, professional systems developers must foster a "visiting imagination", enabling them to come to grasp with the manifold and complex value perspectives at play in a given design context.

To achieve this, it is important to foster room for cross-disciplinary dialogue, and for dialogue between systems developers and society stakeholders to reflect upon and address questions of what factors are more significant in light of which specific norms, practices, values, and cultures.

These ideas resonate well with ideas in the European approach to participatory design, most notably the Scandinavian Tradition for Democratic System Development. As a historical example of excellence, one of the early Scandinavian research projects, UTOPIA (1981–84) developed technology for graphical workers, who at that time felt insecure and afraid that their jobs might be taken over by technology. Yet by joint forces among researchers/systems developers, workers, and the trade union, the UTOPIA project managed to provide IT-solutions, which gave quality graphical products, but without diminishing the workers' skills and the democratic organisation of work processes.[37]

Thus, to improve professionals' opportunities to bring ethics by design onboard, it is important to promote the use of value based design methods, enriched with a global outlook and anchored in participatory design traditions.

For now, methods, such as Privacy by Design, and Value Sensitive Design should be praised for pro-actively seeking to do ethics by design, motivated by philosophical clarification of important values, combined with stakeholder analysis and investigation of the user context. However, although these methods pay attention to cultural values, they are not appropriate for tackling dual-use challenges. Consequently, they provide no answers to systems developers who carefully apply these methods, only to discover that systems developed with the best of intentions and adherence to ethical principles, shielded by legislation in democratic states, might give rise to moral wrongdoing due to dual-use in non-democratic states.

In the report "The Malicious Use of AI – Forecasting, Prevention, and Mitigation",[38] four key recommendations are pointed out, one of which stresses that: "researchers and engineers in AI should take the dual-use nature of their work seriously, allowing misuse-related considerations to influence research priorities and norms, and proactively reaching out to relevant actors when harmful applications are foreseeable".[39] Hence, on a global scale, there is a need to encourage international codes of conduct for systems

[35] J Habermas, *The Theory of Communicative Action. Volume 1: Reason and the Rationalization of Society* (Beacon 1984).

[36] H Arendt, *Lectures on Kant's Political Philosophy* (The University of Chicago Press 1992) p 43.

[37] S Bødker et al, "A UTOPIAN Experience: 'On Design of Powerful Computer-Based Tools for Skilled Graphical Workers'" in G Bjerknes et al (eds), *Computers and Democracy – a Scandinavian challenge* (Gower Publishing 1987) p 251.

[38] M Brundage et al, *The Malicious Use of AI – Forecasting, Prevention, and Mitigation* (Future of Humanity Institute 2018), available at < arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf > accessed 14 November 2018.

[39] ibid, pp 4–5.

developers, as reflected in first principle *of The ACM Code of Ethics and Professional Conduct*:

> "This principle concerning the quality of life of all people affirms an obligation to protect fundamental human rights and to respect the diversity of all cultures ... When designing or implementing systems, computing professionals must attempt to ensure that the products of their efforts will be used in socially responsible ways, will meet social needs, and will avoid harmful effects to health and welfare".[40]

Similarly, initiatives, such as the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems: "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems", emphasise that: "(...) human rights as defined by international law, provide a unilateral basis of creating any A/IS system as they affect humans, their emotions, data, or agency…. human rights should be part of the ethical risk assessment of A/IS".[41]

Professional bodies like the IEEE, which has around 400,000 members worldwide,[42] play a key role by working on establishing industry standards and policy guidelines for beneficial AIS, and raising ethical awareness among systems developers, as well as undertaking obligations to reach out in the public sphere.

Increasingly, systems developers within the field of AIS have felt a responsibility to take an active part in the public debate on AIS through different kinds of outreach activities with the purpose of empowering citizens and informing policy-makers. We also welcome efforts on how to provide transparency by explainable AI as a mean to uphold social rights and the right to an explanation. The notion of transparency is reflected in the European General Data Protection Regulation as a key requirement to avoid black box decisions (see Arts 4, 13, 14, 15, 22).

Also, Boddington points out that "in the case of AI, the fact that professionals are themselves having directly to grapple with how to control the behavior of machines surely gives end users, and the general public, considerable reason to be kept informed at least, and actively involved at best".[43] However, the scale of this should not be underestimated: information by itself must be backed up by enlightening initiatives to ensure that public enactment becomes grounded in knowledge obtained via skill-building within fields such as computational thinking and AI literacy.

V. Public empowerment– cultivation of techno-moral wisdom

In her book *Technology and the Virtues – a Philosophical Guide to a Future Worth Wanting*,[44] Shannon Vallor calls for deliberation about which educational and cultural concepts we need to cultivate to control rather than passively adapt to emerging

[40] See ACM Code of ethics and professional conduct at < www.acm.org/about-acm/acm-code-of-ethics-and-professional-conduct#imp1.3 > accessed 14 November 2018.

[41] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, Version 2. IEEE, 2017 < standards.ieee.org/industry-connections/ec/autonomous-systems.html > accessed 14 November 2018.

[42] P Boddington, *Towards a Code of Ethics for Artificial Intelligence* (Springer 2018) p 64.

[43] ibid, 65.

[44] Supra, note 33.

technologies.[45] Moreover, she notes that her proposals about how we might do that, by eg bringing ethics to design of emerging technologies (such as self-surveillance technologies), of course, require "cultural and technical feasibility".[46] On this background, it makes sense to go back to 1980, when Seymour Papert wrote his milestone book, *Mindstorms – Children, Computers and Powerful Ideas*,[47] in which he presented a Piaget-inspired constructivist approach to learning.[48] In particular, he was interested in providing children with "objects-to-think-with" to activate and engage them in discovery learning processes. His Logo programming language was designed in a way which encouraged children to instruct a computer-controlled turtle, a design idea which was also inherent in the LEGO Mindstorm robots that Papert took an active part in developing. Papert wanted children to become familiar with computers and programming, since "Difficulty with school math is often the first step of an invasive intellectual process that leads us all to define ourselves as bundles of aptitudes and inaptitudes ...".[49] He was frustrated by the negative consequences of the cultural "math-phobia", namely that people, by distancing themselves from math, science, and computers, would miss out on important learning and knowledge opportunities, which would lead to a socio-technological development in which only engineers played a role. It is fair to say that Papert's concerns hold merits today as well:

> "This book is about how computers can be carriers of powerful ideas and of the seeds of cultural change, how they can help people form new relationships (…) it is about whether personal computers and the cultures in which they are used will continue to be the creatures of 'engineers' alone or whether we can construct intellectual environments in which people who today think of themselves as 'humanists' will feel part of, not alienated from, the process of constructing computational cultures".[50]

Papert realised the importance of educating children to embrace science via computational thinking activities. However, the public in general needs to be brought into the loop as well, since AIS increasingly influence important societal decision-making processes with a decreasing degree of human oversight.[51] Consequently, we need to find ways to ensure that the public becomes empowered to negotiate and take an active part in the shaping of AIS' roles in society.

Vallor outlines a taxonomy of 12 universal techno-moral virtues (present in Aristotelian, Confucian, and Buddhist ideas on virtuous development) that will

---

[45] Supra, note 33, p 207.

[46] Supra, note 33, p 206.

[47] S Papert, *Mindstorms – Children, Computers and Powerful Ideas* (1980).

[48] Later, Seymour Papert and Idit Harel wrote *Constructionism Research Reports and Essays 1985–1990*, Epistemology and Learning Research Group, the Media Lab, Massachusetts Institute of Technology (Ablex 1991). Here, they emphasised ideas of playful learning, viz learning by making, in order to build mental models: "Constructionism – the N word as opposed to the V word – shares constructivism's connotation of 'learning as building knowledge structures'. Irrespective of the circumstances of the learning. It then adds the idea that this happens especially felicitously in a context where the learner is consciously engaged in constructing a public entity, whether it is a sand castle on the beach or a theory of the universe": < namodemello.com.br/pdf/tendencias/situatingconstrutivism.pdf >, accessed 14 November 2018.

[49] Supra, note 47, p 8.

[50] Supra, note 47, p 4.

[51] See eg C O'Neil, *Weapons of Math Destruction* (Penguin Books 2016).

facilitate human flourishing in an unpredictable world with emerging technologies.[52] For example, the virtue of techno-moral justice will allow us to negotiate fair surveillance practices on a global scale. Likewise, we need to consider how to balance surveillance technology with self-control to resist the temptation to passively adapt to recommendations from, eg new self-surveillance and nudging technologies.[53] Instead of walking through Vallor's taxonomy of techno-moral virtues, I shall be paying attention to the virtue that she singles out as a unifying techno-moral virtue, namely the virtue of techno-moral wisdom, which, closely related to the Aristotelian notion of phronesis, grants its owner a "well-cultivated and integrated moral expertise that expresses successfully … each of the other virtues of characters that we, individually and collectively, need in order to live well with emerging technologies".[54]

Hence, we must ensure that the individual's moral self-cultivation, towards becoming an independent and techno-morally wise citizen, can take place against a backdrop of techno-moral wisdom cultivated in family settings, educational settings, and communities, as well as on a global scale in institutions and policy forums. As such, it takes collaborative efforts to develop techno-moral wisdom that will foster strong characters with the moral robustness needed to live well in a world of rapid technological change.

Our being in the world is mediated by technology to a degree where we can be said to extend our minds with technology.[55] Moreover, Vallor is not blind to the fact that our lives are entangled with technology, and acknowledges that technology might:

> "… reinforce our efforts of moral self-cultivation, forming a virtuous circle that makes us even more ethically discerning in techno-social contexts as a result of increasing moral practice in those domains. This growing moral expertise can enable the development of still better, more ethical, and more sustainable technologies".[56]

Still, one might question the power of virtues. For instance, Boddington doubts whether the classical virtues are "suited to the new world that we are creating. We are rewriting the world in which we live, as we live in it, and our ideas about how to live in it are shaped partly by the very changes we are making".[57]

But Vallor's ideas are not motivated by wanting to point to quick fixes of shallow "virtue signalling", or guidelines for implementing prescriptions that will bring about a public enriched with techno-moral wisdom. Rather, and far more usefully, her contribution provides a virtue ethical reflective framework for addressing empowerment of the public without falling into paternalism or over-simplification: a framework which convincingly illustrates that it is about time we faced the tremendous challenges we need to overcome to live well with technology in the future. Hence, to balance global technological power and techno-moral wisdom, it will take orchestrated efforts, both on a local level in educational settings and communities, as well as via

[52] Supra, note 33, p 117.

[53] Supra, note 33, p 206.

[54] Supra, note 33, p 154

[55] A Clark and D Chalmers, "The Extended Mind" (1998) 58(1) Analysis 7.

[56] Supra, note 33, p 11.

[57] Supra, note 42, p 68.

global collaboration in existing and new institutions. However, it is a feasible project that may take its departure from human creativity, and seek to strengthen virtues that are already there, and shared across cultures.

VI. Concluding remarks

The AI-powered world is emerging, and, presumably, we are facing a future that will increasingly be invented by humans and AIS in collaboration. Therefore, this paper has emphasised the importance of a three-legged inclusive ethical design perspective, which covers ethical AIS design to provide for understandable and technical interpretable AIS, which values are aligned with human ethical norms and values. On this background, it becomes important that professional communities facilitate moral self-cultivation by enhancing opportunities for AIS-professionals to engage in phronetic activities while adhering to value based design methods. Moreover, such methods could benefit from being enriched with a global outlook (to prevent dual-use) and insights from Scandinavian participatory design traditions. Finally, due to the huge impact of AIS on society, we need to promote enlightening activities to encourage public empowerment, driven by the cultivation of techno-moral wisdom in communities and on a global scale.