Of course, one must not answer this question, which in a way is a radical one. However, there are certainly sufficient arguments to give time and energy for reflection about anthropological foundations as well as about moral decision and action in a digital world with its artificial intelligences. This can be seen as an important basic step in assuming *responsibility* in its proper sense, with both ethical and practical intent, i.e. trying to *respond* to sincere questions and challenges precisely for the sake of indispensable *critical discernment* – namely carefully weighing up the relationship between technical feasibility and moral acceptance. According to experts, the question in regard to AI is not whether we can do it; the question is what exactly we *want* to do, what we *should* do, and *why*. What may sound like a simple question upon closer look will require major efforts in pointing out new opportunities to foster the ability and willingness to take responsibility, to promote attention for formation of conscience,[54] and thus to develop ethical thought and moral competence that could contribute to responsibly dealing with AI without, at least, contradiction to Integral Human Development.

[54] Cf. Alois Joh. Buch, "Vergewisserung des Gewissens. Zu Bedeutung und Deutung des sittlichen Urphänomens," in: J. Schmidt et. al. (ed.), *Mitdenken über Gott und den Menschen* (= FS Jörg Splett), (Münster: Lit, 2001,) 121–135; also id., "Gewissensentscheidung im Kontext von Prinzipienethik und Kasuistik," in: Bormann, Franz-Josef Wetzstein, Verena (ed.), *Gewissen. Dimensionen eines Grundbegriffs medizinischer Ethik* (= FS Eberhard Schockenhoff) (Berlin: de Gruyter, 2014), 283–309.

Peter McCormick

AI and Ethical Responsibility[1]

"a new spirit… a new heart" (*Ezek*. 36:26)

The intelligence of persons – the human capacity to know, to comprehend, to understand, and to judge – remain today unsatisfactorily explained.[2] One significant consequence is the still-widespread confusion between machine intelligence and human

[1] This text is a revised version of an invited paper presented in shorter form at the International Institute for Ethics and Contemporary Issues of the Ukrainian Catholic University's Second Annual International Conference Series on Integral Human Development in the Digital Age on the particular theme, "Informational Overload, Artificial Intelligence, and Responsibility," held at the Ukrainian Catholic University in Lviv from 26 to 28 February 2020. My thanks to Dean V. Turchynovskyy for his kind and generous invitation and to participants for their constructive comments and criticisms. Please note that more than the usual number of references are included for the interests of advanced students. Copyright C 2020 by Peter McCormick. All rights reserved. pjmccormick@gmx.com.

[2] See for example *Science Advances* (14 February 2020) and D. Drenckham and J. Farago, "*L'IA, super-physicienne?*" *Le Monde: Science et Médecine,* 19 February 2020, p. 7. See also B. Cantwell Smith, *The Promise of Artificial Intelligence: Reckoning and Judgment* (Cambridge, MA: MIT Press, 2020), P. Bartolomeo, *La pensée*

intelligence. We can see this and related matters more clearly when we try to reflect freshly on artificial intelligence in the specific contexts of an international, multi-disciplinary conference.

In their thoughtful invitation letter, our hosts have asked speakers to reflect briefly on several aspects of artificial intelligence (AI) today.[3] Here, I stick closely to the terms of that letter in offering several questions for further critical discussion.

1. Artificial Intelligence

"[T]he rapid advance of the digital technologies at the beginning of our present century," the Information Letter begins, "presents unprecedented challenges."

As the conference title "Informational Overload, Artificial Intelligence, and Responsibility" implies, the main digital technology for discussion is artificial intelligence (AI).[4] But if we are to understand how advances in AI present "unprecedented challenges," then we need to be using the polyvalent expression "artificial intelligence"

droit (Paris: Flammarion, 2019), and T. Crane, "Review of Cantwell Smith," *TLS [Times Literary Supplement],* 15 May 2020, pp. 4–5.

[3] Please note that, in accordance with the general contexts specified in the Letter of Invitation, the particular contexts of this article may be found in, among other places, the May 2019 publication of the *Conseil Permanent de la Conférence des évèques de France, Qu'est-ce que l'homme pour que tu penses à lui? Eléments d'anthropologie catholique* (Paris: Bayard, 2019), *passim*, and in the papal encyclical letter *Laudato Si* "Praise Be to You" (24 May 2015) of Pope Francis to be found in full at, http://www.w2.vatican.va/content/francesco/en/encyclicals/documents/papa-francesco_20150524_enciclica-laudato-si.html. See also "Rome Calls for AI Ethics," Vatican 28 February 2020.

[4] Generally, see the pertinent articles in the journal *Nature Machine Intelligence* at nature.com/natmachintell. A special issue of its first volume published in 2019 includes ten selected papers on machine intelligence.

in the same main senses. Consequently, an initial concern might be just what AI today is.[5]

When we talk about AI today, we are talking about the current state of advanced hybrid computer systems[6] over the last ten years.[7] These systems include deep learning neural networks[8] and multi-agent

[5] As accessed on 14 November 2019, the online Techopedia article on definition of AI continues (my underlines for separating topics): "Machines can often act and react like humans only if they have abundant information relating to the world. Artificial intelligence must have access to objects, categories, properties and relations between all of them to implement knowledge engineering. Initiating common sense, reasoning and problem-solving power in machines is a difficult and tedious task. Machine learning is also a core part of AI. Learning without any kind of supervision requires an ability to identify patterns in streams of inputs, whereas learning with adequate supervision involves classification and numerical regressions. Classification determines the category an object belongs to and regression deals with obtaining a set of numerical input or output examples, thereby discovering functions enabling the generation of suitable outputs from respective inputs. Mathematical analysis of machine learning algorithms and their performance is a well-defined branch of theoretical computer science often referred to as computational learning theory. Machine perception deals with the capability to use sensory inputs to deduce the different aspects of the world, while computer vision is the power to analyze visual inputs with a few sub-problems such as facial, object and gesture recognition. Robotics is also a major field related to AI. Robots require intelligence to handle tasks such as object manipulation and navigation, along with sub-problems of localization, motion planning and mapping."

[6] Some hybrid AI, some social scientists argue, would seem to be able "to embed human intentions in material infrastructures" and even "foresee AI agents' societal outcomes" (E. Moss *et al.*, "AI Behaviour," in Correspondence, *Nature,* 574 (10 October 2019), p. 176.

[7] See the former Turing Prize professor at New York University and the Collège de France, Yann Le Cun's recent book, *Quand la machine apprend. La revolution des neurons artificiels et de l'apprentissage profond* (Paris: Odile Jacob, 2019).

[8] D. Castelvecchi, "AI Copernicus 'Discovers' that Earth Orbits the Sun," *Nature,* 575 (14 November 2019), 266–267. Deep Neural Networks (DNNs) are " . . . software structures made up of large numbers of digital neurons arranged in many layers. Each neuron is connected to others in layers above and below it. . . . in 2013 . . . [a research team] showed that it was possible to take an image – of

reinforcement learning[9] that either aim to simulate human reasoning or to simulate both human reasoning[10] and human acting.[11]

Note however that many definitions of AI today continue to draw on the widespread assumption that the intelligence at issue in artificial intelligence is human intelligence. The basic idea appears to be that AI aims to simulate in machines the very same capabilities embodied in human intelligence itself.[12]

By the beginning of our present century, digital technologies rapidly advanced in at least two respects. According to Andy Clark, an Edinburgh philosopher working in philosophy and AI since 1984 and interviewed in the international weekly science journal *Nature* in

a lion, for example – that a DNN could identify and, by altering a few pixels, convince the machine that it was looking at something different, such as a library. The team called the doctored images 'adversarial examples.' A year later . . . [another research team] showed that it was possible to make DNNs see things that were not there, such as a penguin in a pattern of wavy lines" (D. Heaven, "Deep Trouble for Deep Learning," *Nature,* 574 [10 October 2019], p. 164).

[9] O. Vinyals *et al.*, "Grandmaster Level in StarCraft II Using Multi-Agent Reinforcement Learning, *Nature,* 575 (14 November 2019), 350–354.

[10] For notions of reasoning in artificial intelligence see F. Portoraro, "Automated Reasoning", in *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition), ed. E. N. Zalta (ed.), https://plato.stanford.edu/archives/spr2019/entries/reasoning-automated/. Cf. Rahwan *et al.,* "Machine Learning," *Nature*, 568 (25 April 2019), 477–486, where most of the key previous articles are annotated in the very complete Endnotes. Especially important are the questions the authors outline in Fig. 1 on p. 479.

[11] On one basic kind of rationality see N. Kolodny and J. Brunero, "Instrumental Rationality", in *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition), ed. E. N. Zalta (ed.), https://plato.stanford.edu/archives/win2018/entries/rationality-instrumental/.

[12] See S. Bringsjord and N. V. Govindarajulu, "Artificial Intelligence," in *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition), ed. E. N. Zalta (ed.), https://plato.stanford.edu/archives/fall2018/entries/artificial-intelligence/, citing *AIMA*, p. 2 (see note 14 below for full reference of *AIMA*).

July 2019, the first main advance was "the development of artificial neural networks." The second was the development of a theory of the brain as "a probabilistic-prediction device."[13] Ten years later came another main advance in AI, the inaugural work in 2010 on machine learning.[14] Every year since, significant advances have continued, especially with respect to AI and deep neural networks (DNNs).[15]

[13] M. Segal, "Interview with A. Clark on "A Philosopher's View of Robots," *Nature,* 571 (25 July 2019), p. S 18; cited hereafter as "Clark 2019"). See his book, *Surfing Uncertainty: Prediction, Action, and the Embodied Mind* (Oxford: OUP, 2016).

[14] See N. Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: OUP, 2016). "Machine learning," Wikipedia records in November 2019, "is a field of study of artificial intelligence that relies on mathematical and statistical approaches to give computers the ability to "learn" from data, i.e. improve their performance in solving tasks without being explicitly programmed to 'learn' from data. . . . More broadly, it concerns the design, analysis, optimization, development and implementation of such methods. Machine learning usually has two phases. The first is to estimate a model based on data, called observations, which are available and in finite numbers, during the design phase of the system. Estimating the model is to solve a practical task . . . a probability density, recognizing the presence of a cat in a photograph, or participating in the operation of an autonomous vehicle. . . . This so-called 'learning' or 'training' phase is usually carried out prior to the practical use of the model. The second phase corresponds to the start-up: the model being determined, new data can then be submitted in order to obtain the result corresponding to the desired task. In practice, some systems can continue to learn once in production, provided they have a way to get *feedback* on the quality of the results produced [my emphasis].[“]

[15] Deep Neural Networks (DNNs) are " . . . software structures made up of large numbers of digital neurons arranged in many layers. Each neuron is connected to others in layers above and below it. . . . in 2013 . . . [a research team] showed that it was possible to take an image – of a lion, for example – that a DNN could identify and, by altering a few pixels, convince the machine that it was looking t something different, such as a library. The team called the doctored images 'adversarial examples.' A year later . . . [another research team] showed that it was possible to make DNNs see things that were not there, such as a penguin in a pattern of wavy lines" (D. Heaven, "Deep Trouble for Deep Learning," *Nature,* 574 [10 October 2019], p. 164).

Thus, AI today uses greatly developed digital systems as artificial neural networks.[16] These networks are "computer systems inspired by the way that neurons interconnect in the brain." [17] Some digital systems have also continued to develop on the theory of the brain as a probabilistic–prediction system, the brain as a computer program in which a "set of predictions is sent to a user."[18] Here however we are not concerned with issues regarding digital systems generally; we are concerned with AI in particular. But just what is AI anyway? Today, how do we use this familiar expression?

In common British English (BE), the expression "artificial intelligence" denotes "the capacity of a machine to simulate or surpass intelligent human behaviour." And in quotidian American English (AE), the expression "artificial intelligence" denotes something quite similar, namely "the ability of a computer or other machine to perform those activities that are normally thought to require intelligence."[19]

Note that the idea of intelligence appears in both current usages but in different forms.[20] In the case of BE, AI is roughly defined with respect to a machine that is able "to simulate . . . intelligent human behaviour." By contrast, in AE, AI is roughly defined with respect to a machine able "to perform . . . activities normally thought to require intelligence."

[16] For essays on the advances of digital systems see for example *The Cambridge Handbook on Artificial Intelligence,* ed. K. Frankish and W. Ramsey (Cambridge: CUP, 2014. For an overview, see M. A. Boden, *Artificial Intelligence: Its Nature and Future* (Oxford: OUP, 2016).

[17] *Conference Letter 2019.*

[18] "A prediction in the set of predictions is a probability of an outcome of an event. The probability is computed using a prediction model trained" automatically. This definition is from 2013 and cited on line (accessed 5 November 2019).

[19] For current British English (BE) usage I use here *The Shorter Oxford English Dictionary,* 2 vols. , 6th ed. (Oxford: OUP, 2007) cited as *SOED*, and for current American English (AE) I use here *The American Heritage Dictionary of the English Language,* 4th ed. (Boston: Houghton Mifflin, 2000) cited as *AHDE.*

[20] For an excellent general overview, see I. J. Deary, *Intelligence: A Very Short Introduction* (Oxford: OUP, 2001).

Simulating intelligent human behaviour however is not identical with performing activities normally thought to require intelligence. Evidently, these common uses of the expression AI do not call attention to any differences between the two working definitions of what exactly is to be simulated, whether some actual human intelligent behaviour such as expressing sympathy in words and gestures or some merely virtual human activity like playing a game. Thus, just what the expressions "intelligent" and "intelligence" denote here remains vague.[21]

Besides these lexicographical indications, some online reference works provide other definitions of AI. Thus, for Techopedia, "artificial intelligence is a branch of computer science that aims to create intelligent machines. . . .The core problems of artificial intelligence include programming computers for certain traits such as: knowledge, reasoning, problem solving, perception, learning, planning, [and] ability to manipulate and move objects."[22] Here of course we

[21] Much of the kind of intelligence at issue in AI concerns reasoning and thinking. On these basic topics see J. St B. T. Evans, *Thinking and Reasoning: A Very Short Introduction* (Oxford: OUP, 2017), and, for another perspective, B. Saint-Sernin, *La Raison* (Paris: Presses universitaires de France, 2003), and R. Boudon, *La Rationalité* (Paris: Presses universitaires de France, 2009). See also P. N. Johnson, *How We Think* (Oxford: OUP, 2006), and *The Cambridge Handbook of Thinking and Reasoning,* ed. K. Holyoak and R. G. Morrison (Cambridge: CUP, 2005).

[22] As accessed on 14 November 2019, the online Techopedia article on definition of AI continues (my underlines for separating topics): "Machines can often act and react like humans only if they have abundant information relating to the world. Artificial intelligence must have access to objects, categories, properties and relations between all of them to implement knowledge engineering. Initiating common sense, reasoning and problem-solving power in machines is a difficult and tedious task. Machine learning is also a core part of AI. Learning without any kind

find again the notion of intelligence. But unlike previously we have as well a partial list of the kinds of intelligent activities computers are supposed to be capable of performing. Wikipedia in late 2019 draws on several standard works to define AI as follows: “In computer science, artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans. Further, some leading AI textbooks define artificial intelligence as the study of ‘intelligent agents’: any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals.[23]

Colloquially, the term “artificial intelligence” is often used to describe machines (or computers) that mimic “cognitive” functions that humans associate with the human mind, such as “learning” and “problem solving”.[24] Here we find still more elaboration on the vague notion of intelligence, including the very important distinction at last between “human intelligence” and “machine intelligence.” We might put this distinction in other words by saying that intelligent machines are always instrumentally intelligent or instrumentally rational, whereas intelligent human beings are only sometimes instrumentally intelligent and other times are rational in many different ways.

In short, many common definitions of AI today draw on the widespread basic assumption that the intelligence at issue in artificial intelligence is human intelligence. And the basic idea is that AI aims to simulate the very same thing as embodied human intelligence itself. But both this basic assumption and basic idea are, as we shall see, mistaken.

Early specialists in the discipline already recognized this basic idea when they met at Dartmouth College in the USA in 1956. But they hesitated about which of the two options they then had for naming what they were already experimenting with. The two options were the name “artificial intelligence” with its problematic ambiguities and the alternative name “augmented intelligence” supposedly without those ambiguities. In the end, even though many specialists conceded that the name “augmented intelligence” was a more accurate description of what they were doing, they decided nonetheless to use the expression “artificial intelligence” for their nascent discipline. In short, the founders settled this basic issue misleadingly.

Surprisingly, today’s specialists also do not agree on exactly what AI is. One useful and now standard approach, however, is to define AI in terms of its goals.[25] On this approach AI is understood as

of supervision requires an ability to identify patterns in streams of inputs, whereas learning with adequate supervision involves classification and numerical regressions. Classification determines the category an object belongs to and regression deals with obtaining a set of numerical input or output examples, thereby discovering functions enabling the generation of suitable outputs from respective inputs. Mathematical analysis of machine learning algorithms and their performance is a well-defined branch of theoretical computer science often referred to as computational learning theory. Machine perception deals with the capability to use sensory inputs to deduce the different aspects of the world, while computer vision is the power to analyze visual inputs with a few sub-problems such as facial, object and gesture recognition. Robotics is also a major field related to AI. Robots require intelligence to handle tasks such as object manipulation and navigation, along with sub-problems of localization, motion planning and mapping.”

[23] R. Poli, W. B. Langdon, and N. F. McPhee, *A Field Guide to Genetic Engineering* (NY: Lulu Press, 2008).

[24] See the standard text of S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach* 3rd edition (Saddle River, NJ: Prentice Hall, 2009), cited usually as “*AIMA*.” Wikipedia cites the 2003 edition.

[25] Descriptions of the main senses of technical terms in this definition can be found through the index and an on-line glossary. Russell modifies somewhat his goals-determined definition here in his very positively received new book, *Human Compatible: Artificial Intelligence and the Problem of Control* (NY: Viking, 2019). Note however that Russell’s very influential work in his textbook with Norvig and now in his new book continues to understand rationality almost exclusively as instrumental rationality only. But, as D. Leslie, the Ethics Fellow at London’s Turing Institute, has argued recently in *Nature,* “instrumental aptitude is not

a field that aims at building systems whose goal, along one dimension, is matching human performance or some ideal rationality, and, along another dimension, systems that reason or simply act. In tabular form then AI looks something like this.

[Defining AI in Terms of Possible Goals][26]

	Human-Based	**Ideal Rationality**
Reasoning-Based:	Systems that think like humans.	Systems that think rationally.
Behavior-Based:	Systems that act like humans.	Systems that act rationally.

If this description, or something very much like it, is what many informed persons today mean by AI, then how do professionals themselves use the term AI? They rely mainly not on the idea of any intelligent machine but on that of an "intelligent agent." AI is not the study of machines, they think, but the study of agents.[27]

The main text in the field for some years remains the massive book of more than a thousand pages by S. Russell and P. Norvig, now in its third edition.[28] In their Preface, the authors define AI as follows:

enough to account for the full gamut of intelligence capability . . .[Russell] ignores the strain of twentieth-century thinking whose holistic contextual understanding of resoning has led to a humble acknowledgement of the existential limitations of intelligence itself. . . . [intelligence cannot be treated solely] "as an engineering problem, rather than [as] a constraining dimension of the human condition that demands continuous, critical self-reflection" (D. Leslie, "Raging Robots, Hapless Humans: the AI Dystopia," *Nature,* 574 (3 October 2019), p. 33).

[26] See S. Bringsjord and N. V. Govindarajulu, "Artificial Intelligence", in *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition), ed. E. N. Zalta (ed.), https://plato.stanford.edu/archives/fall2018/entries/artificial-intelligence/, citing *AIMA*, p. 2.

[27] Cf. L. Drew, "Agency and the Algorithm," *Nature,* 571 (July 2019), S 19–S 21.

[28] See *AIMA.*

> The main underlying theme is the idea of *intelligent agent*. We define AI as the study of agents that receive percepts from the environment and perform actions. Each such agent implements a function that maps percept sequences to actions, and we describe different ways to represent these functions such as reactive agents, real-time planners, and decision-theoretic systems.[29]

This definition however remains partly misleading. Perhaps a less misleading name for what many are still calling artificial intelligence might be machine intelligence, as in the title of the new *Nature* journal founded in 2019, *Nature Machine Intelligence.* Thus a recurring problem today remains reaching both informed and professional consensus about any satisfactory solution to the problem of defining AI satisfactorily.[30]

This contemporary professional approach to definition remains attractive. But in fact it fails to dissipate the pernicious ambiguity between human and machine intelligence. Moreover, this approach overlooks a quite basic distinction.

As for the ambiguity, the professional description's two key expressions, "agent" and "agency," do not denote the capability of humans to act according to human intelligence. Rather, quite significantly, in this field these expressions denote the capability of machines to act

[29] *Ibid.,* p. viii.

[30] For a very good general overview although a bit dated see I. J. Deary, *Intelligence: A Very Short Introduction* (Oxford: OUP, 2001). Much of the kind of intelligence at issue in AI concerns reasoning and thinking. On these basic topics see J. St B. T. Evans, *Thinking and Reasoning: A Very Short Introduction* (Oxford: OUP, 2017), and, for another perspective, B. Saint-Sernin, *La Raison* (Paris: Presses universitaires de France, 2003), and R. Boudon, *La Rationalité* (Paris: Presses universitaires de France, 2009). See also P. N. Johnson, *How We Think* (Oxford: OUP, 2006), and *The Cambridge Handbook of Thinking and Reasoning,* ed. K. Holyoak and R. G. Morrison (Cambridge: CUP, 2005).

according to machine intelligence.[31] Yet these two seriously differ. Machine intelligence, some might argue, is constructed,[32] whereas human intelligence is embodied. Unfortunately, in much talk of AI today, this ambiguity persists.

As for the missing distinction, this quasi-standard professional description fails to differentiate between what AI was before 2010 and what it has become since. That distinction is between symbolic AI and hybrid AI. Symbolic AI uses hard-coded rules based on techniques[33] derived from deductive reasoning to recognize patterns in discrete 3-dimensional objects and their interrelations. By contrast, hybrid AI mixes symbolic AI with reinforced machine learning to enable hybrid systems to recognize not just 3-dimensional objects

[31] See L. Venema, “Review of W. Gibson’s *Agency,*” *Nature*, 577 (9 January 2020), 164–165. Venema is the chief editor of the significantly entitled and very recently founded journal *Nature Machine Intelligence.*

[32] See N. Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: OUP, 2016). “Machine learning,” Wikipedia records in November 2019, “is a field of study of artificial intelligence that relies on mathematical and statistical approaches to give computers the ability to “learn” from data, i.e. improve their performance in solving tasks without being explicitly programmed to ‘learn’ from data. . . . More broadly, it concerns the design, analysis, optimization, development and implementation of such methods. Machine learning usually has two phases. The first is to estimate a model based on data, called observations, which are available and in finite numbers, during the design phase of the system. Estimating the model is to solve a practical task . . . a probability density, recognizing the presence of a cat in a photograph, or participating in the operation of an autonomous vehicle. . . . This so-called ‘learning’ or ‘training’ phase is usually carried out prior to the practical use of the model. The second phase corresponds to the start-up: the model being determined, new data can then be submitted in order to obtain the result corresponding to the desired task. In practice, some systems can continue to learn once in production, provided they have a way to get *feedback* on the quality of the results produced [my emphasis].[“]

[33] See R. Thomason, “Logic and Artificial Intelligence”, in *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition), ed. E. N. Zalta (ed.), https://plato.stanford.edu/archives/win2018/entries/logic-ai/.

but also many other things besides objects only.[34] More needs to be said, but for now we need to simplify.

Perhaps we may then say that when we are talking here about AI, we are talking about advanced hybrid computer systems[35] developed over the last ten years that are either aimed at simulating human reasoning or human rationality. And perhaps we may also say that we are talking about advanced hybrid computer systems that are aimed at either matching human thinking/reasoning or matching human or rational acting in machine-like ways. Thus, in machine-like and not in human-like ways, AI today either tries to simulate human reasoning or tries to simulate both human reasoning and human acting.

Here, a first question for further discussion arises.

> Q1. Is there sufficient rational warrant for accepting the basic criticism that the “intelligence” in AI is no more than a machine-like capacity and not finally a human one for simulating but not creating human intelligence?

Consider now a second and shorter point.

2. AI and Vulnerability

The organizers write that “… digital reality has introduced a new… kind of vulnerability. [This new vulnerability] … prevents us from detecting how, through the power of invisible digital

[34] These digital system built into some advanced robots today are even claimed to be able to pick out “what matters”; they “manipulate the world and create their own data through their own actions” (Heaven 2019, p. 165).

[35] For essays on the advances of digital systems today see for example *The Cambridge Handbook on Artificial Intelligence,* ed. K. Frankish and W. Ramsey (Cambridge: CUP, 2014. For an overview see M. A. Boden, *Artificial Intelligence: Its Nature and Future* (Oxford: OUP, 2016).

algorithms,[36] our thought and decision-making processes are influenced…."[37]

This particular vulnerability is, surprisingly, not what many might expect. It is not the new military and geopolitical vulnerability that has become so familiar in Ukraine, especially after the almost annual cyberattacks since the spring of 2014. This kind of vulnerability both in Ukraine and elsewhere is a country's vulnerability to remotely controlled AI sabotage of its critical infrastructure. In other words, this vulnerability is a country's susceptibility to unfriendly countries eventually shutting down completely its entire industrial control digital systems. These systems govern among other things air defense, nuclear power plants, electrical power grids, water supplies, transportation, ATM banking, and information and communication systems.[38]

[36] For the notion of algorithm see Wikipedia (accessed 16 November 2019). "As an effective method, an algorithm can be expressed within a finite amount of space and time, and in a well-defined formal language for calculating a function. Starting from an initial state and initial input (perhaps empty), the instructions describe a computation that, when executed, proceeds through a finite number of well-defined successive states, eventually producing "output" and terminating at a final ending state. The transition from one state to the next is not necessarily deterministic; some algorithms, known as randomized algorithms, incorporate random input" (Wikipedia, accessed 16 November 2019.)

[37] Invitation Letter, my italics. See also in the same place, "Often times we make the very decisions that are not good for us, unaware of the fact that we have been manipulated. At other times, our genuine needs are not being met because our illegitimate wants have been transformed into needs through digital deception. *This confusion of wants and needs* in turn changes the very meaning of what it means to be human."

[38] See A. Greenberg, *Sandworm: A New Era of Cyberwar and the Hunt for the Kremlin's Most Dangerous Hackers* (NY: Doubleday, 2019). See also the reviews by S. Halpern, "The Drums of Cyberwar," *The New York Review of Books,* 19 December 2019, pp. 14–20, and by B. Nussbaum, "The Growing Rumblings of Cyberwar," *Nature,* 575 (14 November 2019), 280–281. Halpern recalls the helpful distinction between cyberwarfare and cyberwar. "The first is a tactic," she writes,

Rather, the main senses of the expression "vulnerability"[39] here have to do with the susceptibility not of countries but of persons, specifically individuals' vulnerability to ethically injurious attacks.[40] And what is ethically injurious to persons is what substantively undermines persons' normal capacities to act in accordance with their most well-considered ethical values. (This vulnerability concerns the political realm as well.[41])

For example, most reflective persons think that preserving their own privacy and the privacy of those who are close to them is a basic ethical value for their normal capacities to maintain proper relations both with themselves and with others.[42] Privacy in this ethical

"the second is either a consequence of that tactic, or an accessory to conventional armed conflicts" (p. 16). She also details the cyberattack in 2007 of Estonia, of Georgia in 2008, and then the first of the attacks of Ukraine in 2014.

[39] See the *The Shorter Oxford English Dictionary* (*SOED*), 2 vols.; 6th ed. (Oxford: OUP, 2007) and the *The American Heritage Dictionary of the English Language* (*AHDE*), 4th ed. (Boston: Houghton Mifflin, 2000) respectively.

[40] Cf. various works in the neuroethics movement, a subfield of bioethics, especially P. Churchland, *Conscience: The Origins of Moral Intuitions* (NY: Norton, 2019).

[41] See the general overview of C. T. Bergstrom and J. B. Bak-Coleman, "Gerrymandering in Social Networks," and, for details, A. J. Stewart, "Information Gerrymandering and Undemocratic Decisions," both in *Nature,* 573 (5 September 2019), 40–41 and 117–121 respectively. Generally, the expression "gerrymandering" denotes "the drawing [often redrawing] of district boundaries so as to favour [indeed to maximize] one's own chances in future elections" (G. W. Brown *et al., The Oxford Concise Dictionary of Politics and International Relations,* 4th ed. [Oxford: OUP, 2018]). On the somewhat technical idea here of person as contrasted with human being see P. McCormick, *Relationals: On the Nature and Ground of Persons* (Krakow: Copernicus Institute Press, 2020).

[42] Take the personal privacy of one's medical records. See for example A. Piquard, "*Accord contesté entre Google et 150 hôpitaux aux Etats-Unis,* » *Le Monde*, 14 November 2019, p. 15. Cf. the article "Google's 'Project Nightingale' Gathers Personal Health Data on Millions of Americans" in *The Wall Street Journal,* 14 November 2019. See H. Ledford, "Millions Affected by Racial Bias in Health-Care Algorithm," *Nature,* 574 (31 October 2019), 608–609.

sense, the sense of having and acting with the fundamental freedom from secret or unwanted disturbance or intrusion, is a specifically ethical value that everyone must continually respect in their daily actions.[43] This is so because privacy is intrinsically and profoundly linked to quite fundamental issues concerning the very nature of personhood and self-identity.[44]

Generally speaking, a person's vulnerability is his or her liability to be "physically or emotionally hurt." According to this British usage vulnerability is either "the state or quality" of a person to be harmed (BE). This main sense of vulnerability is echoed in American ordinary usage of a person being vulnerable denoting mainly the person's susceptibility "to physical or emotional injury or attack." American usage adds the further notion however of vulnerability as the likelihood of a person "to succumb, as to persuasion or temptation" (AE).[45] In phenomenological philosophy some important work on ethics and vulnerability goes back to the early work of the Danish philosopher, Peter Kemp (1937–2018).[46]

[43] What I am calling here the ethical sense of privacy needs to be distinguished from several other senses of privacy, such as what Wikipedia calls "the ability of an individual or group to seclude themselves or information about themselves and thereby reveal themselves selectively. Examples include: financial privacy, privacy relating to the banking and financial industries; information privacy, protection of data and information; internet privacy, the ability to control what information one reveals about oneself over the Internet and to control who can access that information; medical privacy, protection of a patient's medical information; [and] political privacy, the right to secrecy when voting or casting a ballot" (accessed 15 November 2019).

[44] See McCormick, 2020.

[45] See the *SOED* and the *AHDE* respectively.

[46] P. Kemp, *Théorie de l'Engagement,* 2 vols. (Paris : 1973), especially vol. 1, *Pathétique de l'Engagement,* and his 1991 book, *The Irreplaceable: A Technology Ethics,* translated from Danish into German, French, and Norwegian.

Let me but highlight here three specific ways only of just how AI today, advanced hybrid AI, trades on the specific ethical vulnerability of many persons. Besides hybrid AI's capacities to invade the most intimate corners of persons' privacy,[47] a second exploitation of persons' lives is hybrid AI's capacities to introduce major bias into how many persons are treated.[48] And the third is the use of facial recognition techniques to track persons' movements without their knowledge or consent.[49]

Privacy is "the ability of an individual or group to seclude themselves, or information about themselves, and thereby express themselves selectively. The boundaries and content of what is considered private of course differ among cultures and individuals. But when something is private to a *person*, it usually means that something is inherently special or sensitive to them." [50] Now, the ethical vulnerability to invasions of privacy that may be violated here by hybrid AI systems is preeminently the general right all persons have to determine for themselves just what they are willing to share with others.[51]

[47] On this topic see the excellent series of related entries in Wikipedia (November 2019) which I mainly draw on here.

[48] See R. Benjamin, *Race After Technology* (London: Polity Press, 2019), especially pp. 49–96.

[49] See the Editorial in *Le Monde,* 17–18 November 2019, p. 30. Note that the same weekend issue of *Le Monde* includes on pp. 26–27 four IA specialists' articles on the dangers of GAFA's increasing control of AI, on IA's impoverishing effect on French judicial culture, on the important debate on 29–31 August 2019 between the Chinese founder of Alibaba Jadk Ma and Elon Musk the fonder of Tesla, and the urgency of AI developers insisting on respecting ethical and social principles.

[50] Wikipedia; accessed 16 November 2019.

[51] This privacy concern is at the basis of many objections to for example Google's "Project Nightengale" that is amassing huge number of patients' digitalized medical records without the knowledge of the patents themselves. See the articles in *The Guardian* and *The Observer* together with the report on BBC World

This kind of vulnerability is importantly different from what persons may undergo all unknowingly as in many public health systems when AI algorithms are applied to for example their health records.[52] An AI algorithm “is a finite sequence of well-defined, computer-implementable instructions, typically to solve a class of problems or to perform a computation. Algorithms are unambiguous specifications for performing calculation, data processing, automated reasoning, and other tasks.”[53]

Major and still-unresolved problems with AI algorithms however are present in almost all AI algorithm developers’ lack of diversity and lack of training in the historical and social aspects of AI uses. In contrast with human decision-making processes, which have their own biases, AI algorithms have many more biases. And appropriate codes for developing minimal biases in AI algorithms have yet to win effective consensus. Persons’ vulnerability to hybrid AI systems here might be called the ethical vulnerability to algorithmic biases.[54]

for 11 and 12 November 2019. On AI and its uses in health care generally see the seven short articles in “*Nature* Outlook Digital Health,” *Nature,* 573 (26 September 2019), 97–116.

[52] See for example A. Piquard, “*Accord contesté entre Google et 150 hôpitaux aux Etats-Unis*, » *Le Monde*, 14 November 2019, p. 15. Cf. the article “Google’s ‘Project Nightingale’ Gathers Personal Health Data on Millions of Americans” in *The Wall Street Journal,* 14 November 2019.

[53] Wikipedia, accessed 16 November 2019. The article continues: “As an effective method, an algorithm can be expressed within a finite amount of space and time, and in a well-defined formal language for calculating a function. Starting from an initial state and initial input (perhaps empty), the instructions describe a computation that, when executed, proceeds through a finite number of well-defined successive states, eventually producing “output” and terminating at a final ending state. The transition from one state to the next is not necessarily deterministic; some algorithms, known as randomized algorithms, incorporate random input.”

[54] See H. Ledford, “Millions Affected by Racial Bias in Health-Care Algorithm,” *Nature,* 574 (31 October 2019), 608–609.

A third and here final kind of vulnerability to today’s advanced AI systems is persons’ ethical vulnerability to unwanted identification through AI facial recognitions. “A facial recognition system,” we can say, “is a technology capable of identifying or verifying a person from a digital image or a video frame from a video source. There are multiple methods in which facial recognition systems work, but in general, they work by comparing selected facial features from given image with faces within a database.”[55]

Of course, just as in the cases of persons’ privacy and unauthorised AI uses of their persona data and in that of persons’ rights to decisions health care decisions and the uses of biased algorithms, not all AI uses of big data and algorithms are exploitations of persons’ vulnerability. So too in the uses of hybrid AI facial recognition systems. Some uses for example in demonstrable security contexts are unobjectionable. But many other uses, for example tracking individual students participating in legally authorised demonstrations against some government university or government policies, seem clear violations of persons’ vulnerability.

In short, while persons exhibit many different kinds of vulnerability either with respect to diseases or to recurring natural disasters or to increasing climate change, persons’ different kinds of vulnerability especially with respect to the misuse today of hybrid AI systems raise particularly acute issues about ethics and social justice.[56] Among these ethical vulnerabilities to advanced hybrid AI systems are the ethical vulnerability to invasions of privacy, the ethical

[55] Wikipedia; accessed 16 November 2019. For the history and the technology see the rest of this quite extensive article. For some distinct kinds of facial recognition technology see M. Untersinger, «*La CNIL [France’s Commission nationale de l’informatique et des libertés] s’empare de la reconniassance faciale*», *Le Monde,* 16 November 2019, p. 15.

[56] Cf. for example the 1978 French law entitled *“Informatique et libertés”* and the various EU laws on information technology deriving from advanced AI today.

vulnerability to algorithmic biases,[57] and the ethical vulnerability to unwanted identification through AI facial recognitions.

So a second question for further critical discussion might be:

> Q2. How do AI developers today bear a fundamental ethical responsibility?

Such a question however requires more sustained investigation than what we can undertake here. Note then a third and more fundamental issue.

3. The Virtual and the Real

"The digital reality empowered by AI control and management of big data," the organizers write further, "has become so powerful that *the distinction between virtual reality and 'real reality'* is blurred.... we are now able to convince masses of people, through informational overload and the constant dissemination of 'facts' and 'fake facts,' that what is real is virtual and that what is virtual is real."[58]

Although these statements include several complex claims, perhaps we may focus here on just one simple claim. The claim is that some uses of AI are sufficient to "convince masses of people... that what is real is virtual and that what is virtual is real."

But if some persons believe falsely that what is real is virtual and that what is virtual is real, what then exactly is the difference between the real and the virtual? Someone might reply: if the real may properly be taken as what exists independently of our thinking and of our using language, then the virtual may be taken properly

[57] See H. Ledford, "Millions Affected by Racial Bias in Health-Care Algorithm," *Nature,* 574 (31 October 2019), 608–609.

[58] Information Letter, my italics.

as what exists only dependently on our thinking and on our using language.[59]

Two things here need attention. First, just as earlier with the expression "AI," so too now we need to inquire whether we are all understanding the key expressions, "the real" and "the virtual," in the same main senses. And, second, we also need to understand how informational overload and the constant dissemination of 'facts' and 'fake facts' could ever convince persons that the real and the virtual are interchangeable.[60]

Many people today think of the virtual exclusively in terms of virtual reality. "Virtual reality" in this ordinary view is a "fully synthetic world," something persons experience when wearing virtual reality headsets. By contrast, "augmented reality" is what persons experience when 3D graphics are overlaid onto the world we experience in everyday life.[61] "Normal reality" is taken to be the world of everyday life.[62]

Other people think today of virtual reality as "a computer-simulated environment simulating physical presence in real or imagined

[59] For the central ontological distinction here between the dependent and the independent see T. E. Tahko and E. J. Lowe, "Ontological Dependence," in *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), ed. E. N. Zalta, URL, <https://plato.stanford.edu/archives/win2016/entries/dependence-ontological/>.

[60] With all the advances in AI today and more to come, perhaps a short distraction here may not be entirely out of order. Consider then some of the resonances in the lovely title of a new French novel by G. Naij, *Ce matin maman a été téléchargée* (Paris: Buchet-Castel, 2019).

[61] "Traditional technologies for virtual reality (VR) and augmented reality (AR) create human experience through visual and auditory stimuli that replicate sensations associated with the physical world. The most widespread VR and AR systems use head-mounted displays, accelerometers and loudspeakers as the basis for three-dimensional, computer-generated environments that can exist in isolation as overlays on actual scenery." X. Yu *et al.,* "Skin-integrated Wireless Haptic Interfaces for Virtual and Augmented Reality," *Nature,* 575 (21 November 2019), p. 473.

[62] *Wikipedia*; accessed 12 November 2019.

worlds… an experience that can be similar to or completely different from the real world."[63] For besides providing virtual reality headsets, some virtual reality systems use "multi-projected environments to generate realistic images, sounds and other sensations that simulate a user's physical presence in a virtual environment."[64]

But our main concerns here are neither with the nature of normal realities, nor augmented one, nor virtual realities, nor with experiencing computer-simulated environments. Rather, our concerns are with understanding better what the nature of "the virtual" itself might be when the virtual is carefully contrasted with the real while all too often beings confused with the real. How does such confusion arise? Such confusion arises from at least two puzzling matters.

One puzzling matter is the variety of very different uses of the expression "the real" in contemporary physics. And a second puzzling matter is the variety of current uses of an expression that are closely related to the expression "the real." This expression is "the actual."

As for the first source of confusion about the real and the virtual, understanding how physicists use the expression "the real" in physics and quantum physics may seem relatively straightforward. We consult a dictionary. For example, the latest edition of the Oxford physics dictionary tells us that the "the real" is what exists in a "directly observable" state.[65] By contrast, what is not real is what is not directly observable. Thus, most physicists substitute for the ontological distinction between the independent and the dependent the empirical

[63] *Wikipedia*; accessed 12 November 2019.

[64] *Ibid.*

[65] "Virtual State" in *A Dictionary of Physics,* ed. R. Rennie, 7th ed. (Oxford: OUP, 2015). The article reads in its entirety: "Virtual State. The state of the virtual particles that are exchanged between two interacting charged particles. These particles, called photons, are not in the real state, i. e. directly observable; they are constructs to enable the phenomenon to be explained in terms of quantum mechanics."

distinction between the observable and the unobservable. Accordingly, "the real" may be said to exist as more than a construction, while "the virtual" exists merely as a construction, something "that enables the phenomenon to be explained in terms of quantum mechanics."

Still, much scientific talk itself about "the real," as this dictionary entry suggests, remains confusing. Thus, "it has become almost de rigueur in the quantum foundations literature," the American physicist and philosopher of science Tim Maudlin writes in 2019, "to misuse the terms "realist," "realistic," "antirealist," and "antirealistic." [But]… in the proper meaning of the term, *physical theories* are neither realist nor antirealist…. It is *a person's attitude toward a physical theory* that is either realist or antirealist….[66]

Yet if even some of the best scientific theories today are misusing the expression "the real," then it's no wonder that many persons can be left confused. Many people can become confused, that is, about whether persons endorsing such scientific understandings of the world are trying to describe no more than the objects of personal attitudes, or rather are trying to describe states of affairs existing objectively in the world regardless of our minds and our languages.

As for a second cause of confusion about the central sense of the expression "the real," consider the closely related expression "the actual." When someone writes about "rocks, trees… the actual world," as the American Transcendentalist Henry David Thoreau (1817–1862)

[66] The scientific realist maintains that in at least some cases, we have good evidential reasons to accept theories as true, or approximately true, or on-the-road-to-truth. The scientific antirealist denies this" (Maudlin 2019, pp. xi–xii; Maudlin's italics). Cf. his comments on some physists' many-worlds views in his "Review of P. Lewis, *Quantum Ontology: A Guide to the Metaphysics of Quantum Mechanics," Inference: International Review of Science,* 3 (23 November 2017), Issue 3 (online). See also A. Becker, *What Is Real? The Unfinished Quest for the Meaning of Quantum Physics* (London: John Murray, 2018), and the two excellent books of T. Maudlin, *Philosophy of Physics: Space and Time* and *Philosophy of Physics: Quantum Theory,* both published by Princeton UP respectively in 2012 and in 2019. Each book has extensive bibliographies.

did so memorably in his classic work *Walden, or Life in the Woods* (1854), he or she is writing not about merely any potential or possible world but about "the existing world."[67] We need to note then that using the expression "the actual" requires distinguishing between something that exists or has existed in fact and something that exists only in the present. That's why, as in our example, Thoreau's use of the expression ”the actual," in the expression the *American Heritage Dictionary* cites, denotes "the existing world," that is, the presently existing world. When linked to such authoritative lexicographical observations we may then take the expression "the actual" here as mainly denoting "what, presently, is the case."

This sense is however importantly different from the main sense of "the real." For unlike its close cousin "the actual," the expression "the real" is not normally restricted to what is existing in the present. "The real" is more inclusive. That is, the expression "the real" may also denote the future, and, on some views, the past as well.

But besides "the actual," what about "the real's" other close cousin, "the virtual"? In the medieval period, the Scots philosopher Duns Scotus (c.1266–1308) understood the virtual as something existing "as if" it were real. This usage of the expression "the virtual" returned early in the twentieth century in the German philosopher Hans Vaihinger's "philosophy of as-if."[68] We might then argue that some uses of the expression "the virtual" denote what is in fact not the case. Instead, they denote what merely could be taken to be the case.[69]

[67] See the note on synonyms for "real" in the *AHDE.* Note that the rubric "Synonyms" is dropped in the more recent 5th edition of the *AHDE.*

[68] See Hans Vahinger's 1911 book, *Die Philosophie des Als Ob.* See also T. Nagel, "As If: Review of K. A. Appiah's *As If Idealization and Ideals*," *The New York Review of Books,* 5 April 2018, pp. 36–38.

[69] Cf. R. Turner, N. Angius, and G. Primiero, "The Philosophy of Computer Science," *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition), ed. E. N. Zalta (ed.), https://plato.stanford.edu/archives/spr2019/entries/computer-science/.

Accordingly, we ourselves may take the expression "the real" to denote here what actually exists presently, what exists independently of our thinking or saying so. And we may take the other key expression "the virtual" to denote what does not actually exist but what could be taken to exist.[70]

These kinds of observations may help explain just how even some sophisticated discussions of hybrid AI today can bring many persons to confuse the real with the virtual, often with quite serious ethical consequences.[71] But now still another question emerges:

> Q3. Exactly how do confusions about the difference between "the real" and "the virtual" lead directly to confusions between what is ethical and what is not?

4. AI and Ethical Responsibility

A final citation from the Invitation Letter reads: Sometimes "our genuine needs are not being met because our illegitimate wants have been transformed into needs through digital deception. This confusion of wants and needs in turn changes the very meaning of what it means to be human."[72]

By way of comment, one might at first decide to argue that, to the contrary, AI has nothing to do with ethical responsibility. The developers

[70] See G.-G. Granger, *Le probable, le possible et le virtuel: Essai sur le rôle du non-actuel dans la pensée objective* (Paris: Editions Odile Jacob, 1995).

[71] On the increasing efforts to require ethical considerations in research on AI today, especially on the consequences of biased algorithms on the increase of social harm to vulnerable persons, see E. Gibney, "The Battle to Embed Ethics in AI Research," *Nature,* 577 (30 January 2020), 609.

[72] Invitation Letter. Cf. J. Nida-Rümelin, "Digital Humanism," *Max Planck Research,* February 2020, pp. 10–15. See his book, co-authored with N. Weidenfeld, *Digitale Humanismus* (Munich: Piper Verlag, 2018). I thank J. Casanova for this reference.

and users of AI may very well have special ethical responsibilities. But just like any other technology, AI itself is ethically neutral.[73]

But alleging that AI is itself ethically neutral does not resist critical examination.[74] For in the end AI depends on persons who develop its programs. So AI developers cannot, as many do, just blame the responsibility for errors (usually, and revealingly, called by the euphemism, "miscalculations") on the complexity of some AI programs. Nor can they blame them on their clients' failures to having provided detailed enough specifications.

Much more generally, in June 2019 the leaders of the 20 largest economies in the world, the G20, issued the G20 AI Principles. Despite their trade and especially AI rivalries, both the US and China signed the statement.[75] In June 2019 also China's National New Generation of Artificial Intelligence Governance Committee published its list of ethical principles supposed to be governing those working in AI development. The principles, which resembled those issued in Europe by the OECD the preceding month, included "harmony, fairness and justice, respect for privacy, safety, transparency, accountability, and collaboration.[76]

Moreover, in August 2019, the G7 leaders formally launched the International Panel on Artificial Intelligence (IPAI), including a call for research projects that significantly incorporate the ethical

[73] J. Ladd, "Computers and Moral Responsibility: A Framework for Ethical Analysis," in *The Information Web: Ethical and Social Implications of Computer Networking,* ed. C. C. Gould (Boulder, Colorado: Westview, 1988).

[74] R. Turner, N. Angius, and G. Primiero, "The Philosophy of Computer Science," *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition), ed. E. N. Zalta (ed.), https://plato.stanford.edu/archives/spr2019/entries/computer-science/.

[75] See the Editorial in *Nature,* 572 (22 August 2019), p. 415. See also S. Kaufman, "*La Bataille de l'intelligence artificielle," Le Monde,* 14 November 2019), p. 30.

[76] S. O'Meara, "China's Ambitious Quest to Lead the World in AI by 2030," *Nature,* 572 (22 August 2019), p. 428.

dimensions of AI.[77] This call was echoed in the Second World AI Conference held in Shangai from 29–31 August 2019. All of this work in common took place against the backgrounds of the very important December 2017 Montréal Declaration for the Responsible Development of AI.[78]

However, groups everywhere are still working on the problem of demonstrating "transparency in how algorithms make decisions. [And, at present,] there are no agreed standards for this."[79] For, as several AI professionals write recently, "computational artifacts should fulfill moral values together with common functional requirements."[80]

Many of the complexities here cluster around just how the various moral and ethical values at issue are to be understood.[81] Some rather distinctive philosophical work has tried to contribute to the basic theme of the interactions of persons with one another through machine interfaces like the Internet.[82] In this philosophical domain, interaction between humans and humans and between humans and

[77] Again, the Editorial in *Nature,* 572 (22 August 2019), p. 415 noted above.

[78] See latest versions available online.

[79] O'Meara 2019, p. 428.

[80] Turner, Angius, and Primiero 2019. "Beside correctness, reliability, and safety," the citation continues, "computing systems should instantiate moral values including justice, autonomy, liberty, trust, privacy, security, friendship, freedom, comfort, and equality. For instance, a system not satisfying equality is a biased program, that is, an artifact that "*systematically* and *unfairly discriminates* against certain individuals or groups of individuals in favor of others. [But although mostly] everybody would agree that computing artifacts should satisfy those moral values," just how such values are to be reconciled with functional requirements in software development remains both complex and controversial.

[81] See for example R. Juste, "Four Ethical Priorities for Neurotechnologies and AI," *Nature,* 551 (9 November 2017), 159–163.

[82] See for example much of the work cited in L. Introna, "Phenomenological Approaches to Ethics and Information Technology," *Stanford Encyclopedia of Philosophy* (Fall 2017 Edition), ed. E. N. Zalta, https://plato.stanford.edu/archives/fall2017/entries/ethics-it-phenomenology/.

machines are often known as "the phenomenon of the virtualisation of interaction."

"Most of our current thinking about ethics," one of the main researchers in this phenomenological field has observed, "implies a certain sense of community based on reciprocal moral obligations that are largely secured through situated, embodied practices and institutions that are often overlapping and mutually inclusive. If these practices and institutions become virtualized, then it would seem that we need to reconsider some of our most fundamental human categories."[83] Among those categories are communities and moral and ethical concepts themselves.[84]

Although the phenomenological literature on community and ethical concepts such as ethical responsibility is large, three related although different approaches may be roughly tabulated as follows.[85]

	[I.] Artifact / tool Approach
View of technology / society relationship	Technologies are tools that society draws upon to do certain things it would not otherwise be able to do. When tools become incorporated in practices it tends to have a more or less determinable impact on those practices.
Approach to ethical implications of technology	The task of ethics is to analyze the impact of technology on practices by applying existing or new moral theories to construct guidelines or policies that will 'correct' the injustices or infringements of rights caused by the implementation and use of the particular technology.

[83] *Ibid.*

[84] See for example, T. Garcia, *Nous* (Paris: Grasset, 2016) and his interview, N. Sarthou-Lajus, "*Ce Qui Fait Communauté : Entretien avec Tristan Garcia»*, *Etudes,* N° 4265 (November 2019), pp. 57–66.

[85] This rough sketch is Introna's in his Stanford Encyclopedia entry noted above.

	[II] Social Constructivist Approach
View of technology / society relationship	Technology and society co-construct each other from the start. There is an ongoing interplay between the social practices and the technological artifacts (both in their design and in their use). This ongoing interplay means that technological artifacts and human practices become embedded in a multiplicity of ways that are mostly not determinable in any significant way.
Approach to ethical implications of technology	The task of ethics if to be actively involved in disclosing the assumptions, values and interests being 'built into' the design, implementation and use of the technology. The task of ethics is not to prescribe policies or corrective action as such but to continue to open the 'black box' for scrutiny and ethical consideration and deliberation.

	[III] Phenomenological Approach
View of technology / society relationship	Technology and society co-constitute each other from the start. They are each other's condition of possibility to be. Technology is not the artifact alone it is also the technological attitude or disposition that made the artifact appear as meaningful and necessary in the first instance. However, once in existence artifacts and the disposition that made them meaningful also discloses the world beyond the mere presence of the artifacts.
Approach to ethical implications of technology	The task of ethics is ontological disclosure. To open up and reveal the conditions of possibility that make particular technologies show up as meaningful and necessary (and others not). It seeks to interrogate these constitutive conditions (beliefs, assumptions, attitudes, moods, practices, discourses, etc.), so as to… question the fundamental constitutive sources of our ongoing being-with technology.

Each of these three current approaches to views of society and technology on the one hand and to the ethical implications of technology on the other clearly has much to offer future critical reflection. One central issue here, however, remains too much in the background. And that is the issue of the ethical responsibilities of AI developers today with respect to the unprecedented vulnerability of their artifacts.

When some AI developers reject ethical responsibility, "they [themselves] fail to recognize," as three philosophers argued in 2019, "that in the process of developing software, they are not just instantiating specifications and implementing programs. [The developers] are additionally providing a service to society."[86]

These philosophers also suggested what some ethicists may consider a useful distinction between negative and positive ethical responsibility. Thus, we may think of some AI developers exhibiting negative responsibility. They do so when they develop "correct artifacts without considering the potential effects and influences of the artifacts in society." By contrast, we may think of other AI developers showing positive responsibility. They do so when they develop

[86] Turner, Angius, and Primiero 2019. "Beside correctness, reliability, and safety," the citation continues, "computing systems should instantiate moral values including justice, autonomy, liberty, trust, privacy, security, friendship, freedom, comfort, and equality. For instance, a system not satisfying equality is a biased program, that is, an artifact that "*systematically* and *unfairly discriminates* against certain individuals or groups of individuals in favor of others. [But although mostly] everybody would agree that computing artifacts should satisfy those moral values," just how such values are to be reconciled with functional requirements in software development remains both complex and controversial.

Ibid., my italics. See also in the same place, "Often times we make the very decisions that are not good for us, unaware of the fact that we have been manipulated. At other times, our genuine needs are not being met because our illegitimate wants have been transformed into needs through digital deception. *This confusion of wants and needs* in turn changes the very meaning of what it means to be human."

artifacts only after considering critically "the consequences the developed machine may have among users."[87]

We need to grant that negative responsibility may protect AI developers from legal liability. But negative responsibility does not keep them from bearing ethical responsibility, particularly with respect to their clients, users, fellow professionals, and the public. These ethical dimensions are even important enough for many engineering companies and associations to have articulated software ethical codes such as the "Software Engineering Code of Ethics and Professional Practice." This ethical code, for example, formulates no fewer than eight separate principles regarding the ethical behaviors of AI professionals.[88]

However, groups everywhere are still working on the problem of demonstrating "transparency in how algorithms make decisions. [And, at present,] there are no agreed standards for this."[89] Many complexities cluster around just how the various moral and ethical values at issue are to be understood.[90] And some ongoing philosophical work has tried to contribute to the fundamental theme of the interactions of persons with one another through machine interfaces like the Internet.[91]

[87] Turner, Angius, and Primiero 2019.

[88] Note however the critical cautions of B. Mittelstadt, "Principles Alone Cannot Guarantee Ethical AI," *Nature Machine Intelligence,* 1 (4 November 2019), 501–507.

[89] S. O'Meara, "China's Ambitious Quest to Lead the World in AI by 2030," *Nature,* 572 (22 August 2019), p. 428.

[90] See especially A. Jobin, M. Ienca and E. Vayena, "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence,* 1 (2 September 2019), 389–399, https://doi.org/10.1038/s42256-019-0088-2, with excellent further reading section.

[91] See for example much of the work cited in L. Introna, "Phenomenological Approaches to Ethics and Information Technology," *Stanford Encyclopedia of Philosophy* (Fall 2017 Edition), ed. E. N. Zalta, https://plato.stanford.edu/archives/fall2017/entries/ethics-it-phenomenology/.

Thus, the precise relationships between AI today and ethical responsibility are yet to be plotted. Perhaps then a final question for further critical discussion now comes into view:

> Q4. Do today's AI developers (perhaps like some researchers in ethics at UKU) enjoy a certain sense of community based on reciprocal moral obligations and ethical solicitations?

If so, then these very obligations are largely secured "through situated, embodied practices and institutions that are often overlapping and mutually inclusive."[92] But now I must conclude.

Envoi

Regardless of critical issues still remaining here, we nonetheless have on hand several reasonably good reminders of what AI looks like today. And we also have on hand some observations on AI and vulnerability, on the virtual and the real, as well as on AI and ethical responsibility. Perhaps one general issue for ongoing critical discussion of today's hybrid AI is just how to develop further the basic distinctions between human intelligence and machine intelligence. In concluding, then, a final question, a fundamental question, might go something like this:

> Q5. Must sufficiently critical discussions of hybrid AI today confront the utterly basic differences between intelligence and comprehension, between the intelligence of the brain and that of the mind, between – may we say here? – the intelligence of the spirit and the wisdom of the heart?

[92] *Ibid.*

The responses if not the answers I must for now leave up to others. But may I quite simply ask that, if long ago Ezekiel was right in the epigraph at the beginning of this paper, then, in truth, does responding finally to such dark questions require a new spirit – and a new heart?