Essay Six

Digital Worlds[1]

"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 ..."[2]

T. Maudlin (2019)

"…. we will start to see the most powerful forms of AI emerge when simulated AI agents are able to talk to each other as part of proper communities."[3]

A. Clark (2019)

Perhaps we might to leave to the experts the important general contexts here of contemporary Catholic Social Teaching,[4] the nature of artificial intelligence (AI) today,[5] and on AI and the pertinent distinctions between wants and needs – all mentioned in the conference Invitation Letter.[6] By contrast, I will try to assemble merely some reminders about how English-language speakers use some key conference expressions. This will help elucidate four of the several important elements the organizers have detailed for our considerations. These are, first, what AI is and its challenges, then AI and the present blurring of "the distinction between virtual reality

and 'real reality,'" third, AI and ethical responsibility, and finally AI and "a new and particularly pernicious kind of vulnerability."[7]

I begin by trying to make more explicit just what we are talking about when we talk about AI. I then describe several main differences between virtual reality in AI and the real itself. Next I offer some remarks on AI and ethical responsibility, and I conclude with several further remarks on AI, ethical responsibility and vulnerability.

1. Digital Technologies Rapidly Advancing[8]

> "... *the rapid advance of the digital technologies* at the beginning of our present century," write the organizers, "presents unprecedented challenges."[9]

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 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."[10] Ten years later came another main advance in AI, the inaugural work in 2010 on machine learning.[11] Every year since, advances have continued, especially with respect to AI and deep neural networks (DNNs).[12]

Thus, AI today uses greatly developed digital systems as artificial neural networks.[13] These networks are "computer systems inspired by the way that neurons interconnect in the brain."[14] Further, 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."[15] Here, however, we are not worried about issues concerning digital

systems generally; we are concerned with AI in particular. But just what is AI anyway?

In everyday British English (BE), the expression "artificial intelligence" denotes "the capacity of a machine to simulate or surpass intelligent human behavior." And in everyday 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."[16]

Note that the idea of intelligence appears in both current usages but in different forms.[17] In the case of BE, AI is roughly defined with respect to a machine that is able "to simulate . . . intelligent human behavior." By contrast, in AE, AI is roughly defined with respect a machine able "to perform . . . activities normally thought to require intelligence." Simulating intelligent human behavior however is not identical with performing activities normally thought to require intelligence. Evidently, these common uses of the expression AI do not identify what exactly is to be simulated, whether some actual human intelligent behavior such as expressing sympathy in words and gestures or some any merely virtual human activity like playing a game. That is, just what the expressions "intelligent" and "intelligence" denote here remains vague.[18]

Besides these lexicographical indications, some online technical 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."[19] Here we find again the notion of intelligence. But unlike previously, we have as well a partial list of the kinds of intelligent activities that computers are supposed to be capable of performing.

Further, Wikipedia, drawing on several standard works, defines AI in late November 2019 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. Leading AI textbooks define the field as the study of "intelligent agents": any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals.[20]

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".[21] 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.

Thus, many definitions of AI draw on the widespread assumption that the intelligence at issue in artificial intelligence is human intelligence. And the basic uncritical idea is that AI aims to simulate the very same thing as embodied human intelligence itself.

This problem was already recognized in 1956 at Dartmouth College in the USA when specialists hesitated on which of two options they 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" without those ambiguities. The winner in those discussions was the name "artificial intelligence", even though many conceded at the end that the name "augmented intelligence" was

a more accurate description of what they were doing. Yesterday's specialists had disagreed.

Today's specialists themselves still do not agree on exactly what AI is. However, one useful and now standard approach, however, is to define AI in terms of its goals.[22] On this approach AI is understood as a field that aims at building systems whose goal, along one dimension, is matching human performance or some ideal rationality, and, along another dimension, constructing systems that reason or simply act. In tabular form then AI looks something like this.

[**Defining AI in Terms of Possible Goals**] [23]

	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 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.[24] 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.[25] In their Preface, the authors define AI as follows:

> "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."[26]

This professional description, however, leaves out a basic distinction between what AI was before 2010 and what it has become today: the distinction between symbolic AI and hybrid AI. Symbolic AI uses hard-coded rules based on techniques derived from deductive reasoning to recognize patterns in discrete objects and their interrelations. By contrast, hybrid AI mixes symbolic AI with reinforced learning gained by training in 3D environments to enable hybrid systems to recognize not just three-dimensional objects but also many other things besides objects only. These systems embodied in some advanced robots are able to pick out what matters; they "manipulate the world and create their own data through their own actions."[27]

But we need to simplify. Let us say then that when we talk here about AI we are talking about advanced hybrid computer systems developed over the last ten years that are either based on human reasoning[28] or human rationality,[29] and advanced hybrid computer systems that are aimed at either matching human thinking/reasoning or matching human or rational acting. Thus AI today either tries to simulate human reasoning or both human reasoning and human acting.

With this particular idea of AI in mind, how then is it that we may properly say, with the conference organizers, that AI today – that is, hybrid AI – "presents unprecedented challenges?"

On reflection, it seems that there are probably many good reasons why hybrid AI presents unprecedented challenges,[30] especially for philosophical ethics.[31] Let me mention just three of these reasons, a particularly important one in the dramatic changes affecting democracies in Trump's America, Johnson's UK, and Ukraine's eastern provinces, the continually expanding uses of AI "to skew perceptions of how others in the community will vote – which can alter the outcomes of elections." [32]

Recall that in so-called democratic elections, "the pattern of network connections influences what voters believe about others' vot-

ing intentions. This influence matters because people shift their own perspectives and voting strategies in response. . . ." People of course form their political opinions in many ways. One major way is the way some members of a group share their information by contact through social media. What some recent game-theoretical analyses demonstrate, however, is that otherwise unbiased networks "can be rewired in ways that lead some individuals to reach misleading conclusions about community preferences. And, ultimately, these misperceptions can even sway the course of an election."[33]

Central to the manipulations that generate such increasingly unfair elections is the work of sophisticated hybrid AI systems with deep learning capacities.[34] Evidence of such manipulation has accumulated in the US, the UK, and the EU.

To appreciate the central role of hybrid AI, consider briefly just how such manipulations of social media occur in elections. "Online social networks are highly dynamic systems," two experts write recently, "that change as a result of numerous feedbacks between people and machines. Algorithms suggest connections; people respond; and the algorithms adapt to the responses. Together these interactions and processes alter what information people see and how they view the world. . . ." [35]

A group of seven correspondents recently in *Nature* provide us with a second, shorter example of how AI "presents unprecedented challenges" in 2020. "Studying AI agents [hybrid AI systems like some advanced robots] as if they are animate," they write soberly, "moves responsibility for the behavior of machines from their designers, thereby undermining efforts to establish professional codes for AI practitioners."[36]

And an equally short example gives us, for now, a final example. Consider then that however complex contemporary defensive elements are in even hybrid AI systems, these systems can always be

expertly hacked, for the weaknesses of even the most current hybrid AI systems are known. "A hacker could use these weaknesses to hijack an online [hybrid] AI-based system," one expert writes recently, "so that it runs the invader's own algorithms."[37] After Brexit and the Trump election, we all realize what that can mean for the future.

2. 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. It is not only that we are able to digitize reality and construct its digital representations, but 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." [38]

Although this statement from the conference Invitation Letter includes several important claims, we may focus on the assertion 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." Two things here now need our attention. First, just as with the expression "AI," we need to confirm that we all apprehend the key expressions, "the real" and "the virtual," in the same main senses. Second, we need to understand how, if at all, digitizing reality, informational overload, and the constant dissemination of 'facts' and 'fake facts' could convince many people that the real and the virtual are interchangeable.[39]

Many people today think of the virtual exclusively in terms of virtual reality. One popular source for the ongoing discussion of just what virtual reality is suggests that "we now have three kinds of reality – normal reality. . . , virtual reality, and augmented reality."[40]

"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. "Normal reality" is taken to be the world of everyday life.

Another important popular source reports that virtual reality "is a computer-simulated environment simulating physical presence in real or imagined worlds. . . an experience that can be similar to or completely different from the real world."[41] Besides 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."[42]

Our main concerns, however, are neither with the nature of normal, augmented, or virtual reality, nor with experiencing computer-simulated environments. Rather, our concerns are with understanding what the nature of "the virtual" itself might be when the virtual is carefully contrasted with the real.

Begin then not with ordinary everyday understanding but this time with our ordinary philosophical understandings of the two key expressions.[43] Thus, in BE, ordinary philosophical understandings, "the real" denotes whatever is existing "in fact and not merely in appearance, thought, or language. . . ." And "the virtual" denotes what exists essentially, "actually, or by strict definition." In particular, in computing verbiage, the virtual is something "not physically existing but made by software to appear to [be existing] from the point of view of the program or the user. . . ."

That is, in today's philosophical parlance, the expression "the real" denotes whatever is "existing objectively in the world regardless of subjectivity or conventions in thought or language," while the expression "the virtual" denotes whatever exists only "in the mind . . . though not in actual fact, form or name." In computing, the virtual

is what "is created, simulated, or carried on by means of a computer or computer network. . . ."

What induces confusion in the use of the cardinal expression "the real," I think, are at least two matters. The first are the very different uses of "the real" in physics and especially in quantum physics. And the second is the use of two very closely related expressions in connection with "the real."

As for the first source of confusion, trying initially to understand how physicists use the expression "the real" in physics and quantum physics seems to be relatively straightforward. The physics dictionary tells us that the "the real" is what exists in a "directly observable" state. By contrast what is not directly observable may be said to be what exists merely as a construction "that enables the phenomenon to be explained in terms of quantum mechanics."[44]

Many philosophers are deeply involved not just with physics but also with ongoing debates in the philosophy of science about scientific realism. And, although some of us try to follow some of these matters the best we can, quite frankly I simply lack the appropriate knowledge in advanced mathematics to be able properly to summarize such matters here. I have supplied some references to several excellent recent books however that may be of help to others.[45]

A citation from one of the internationally distinguished scientists, however, may give us an initial sense about how confusing current talk about "the real" remains. Thus, "It has become almost de rigueur in the quantum foundations literature," the American physicist Tim Maudlin writes in his 2019 book, "to misuse the terms 'realist,' 'realistic,' 'antirealist,' and 'antirealistic.' . . . 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. . . . 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." [46]

But if this is the case for even the best of our scientific theories today, we are left confused about what these theories are describing. They are finally no more than the objects of personal attitudes, rather than any so-called real objects or states of affairs objectively existing in the world independently of our minds and our languages.

With regards to a second cause of current confusion, consider the key expressions "the actual" and "the true." If someone says, the *AHDE* reports, "she showed real sympathy for my predicament," then the implication is one of "authenticity, genuineness, or factuality." When Thoreau writes however about "rocks, trees . . . the actual world," he means not any merely potential or possible world but "the existing world." When Bertrand Russell recommends that "it is undesirable to believe a proposition when there is no ground whatsoever for supposing it true," the implication is one of "consistency with fact, reality, or actuality."[47]

We need to observe, however, 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 Thoreau's use of ""the actual" cited above denotes, we said, "the existing world," that is, the presently existing world.

Contrast for example the main senses of the expression "actual" in the sentence, "He had no illusions about himself as [an] actual. . . soldier," and in the sentence, "Husbands were chosen as much on eventual as actual salary." In the first sentence "actual" denotes the person described as having no illusions about himself as he is existing at the time of the description, whereas in the second sentence "actual" denotes a present salary in strict contrast with any future salary.[48]

In more particular uses, for example in modal logic or the logic that examines necessity and possibility, the expression "the actual"

is used with respect to the world. Thus, the actual world is "the world as it is" as contrasted with the possible world, the world as "it might have been."[49] In our contexts, this usage suggests that "the actual" is what we also call "the real." When linked with the lexicographical remarks above we may then take "the actual" as denoting "what, presently, really is the case."

Note that some other philosophical uses of "the virtual" today go back to the twentieth-century French philosopher, Gilles Deleuze (1925-1995) who "used the term virtual," one standard reference reports, "to refer to an aspect of reality that is ideal, but nonetheless real. An example of this is the meaning, or sense, of a proposition that is not a material aspect of that proposition (whether written or spoken) but is nonetheless an attribute of that proposition."

In the history of philosophy, Duns Scotus (c.1266-1308) understood the virtual as something existing "as if" it were real. In the twentieth century, the German philosopher Hans Vaihinger (1852-1933) even developed a philosophy of "as-if" called fictionalism, a critical reflection on the use of fictions known to be false but useful for coping with some problems where true solutions seem apparently impossible to achieve. Hence some ordinary uses of the virtual denote what is not the case in fact but nonetheless is virtually the case.

The differences and similarities here among the three expressions, "the actual," "the real," and "the true" account partly for the relatively recent uses today of apparently redundant phrases such as "true facts" and "real facts." In ordinary AE usage, facts themselves are understood as things that are known to exist or to have existed. So the expression "real facts" may seem redundant. Since many argue that facts can be nothing other than true, the expression "true facts" may also seem redundant. Given the prevalence and the persuasive powers of "fake news" today, however, these apparent redundancies remain quite useful for emphasizing both the nature of fake news

and its alternatives. To simplify once again, we may accordingly take here the key expression "the real" to denote what actually exists presently, independently of our thinking or saying so. We may take the other key expression "the virtual" to denote what actually exists presently in our minds or sentences only.[50]

Now we are in a position to ask more clearly how digitizing reality, informational overload, and the constant dissemination of 'facts' and 'fake facts' could convince many people that the virtual and the real are interchangeable. How then could one come to believe that what actually exists presently in our minds or sentences be interchangeable with what exists presently independently of our minds or sentences? To answer such a general question, I think we need to focus on a particular domain, for example, that of AI and ethical responsibility.

3. AI and Ethical Responsibility

> "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." [51]

One might at first want to argue that AI as we have narrowed its definition has nothing to do with ethical responsibility. The users of AI may very well have special ethical responsibilities with respect to AI, but AI itself is ethically neutral.

This common approach, the alleged moral neutrality of AI, does not, however, resist close critical examination,[52] for AI essentially depends on persons who develop its programs, however complex. AI developers cannot cast the responsibility for errors (usually called by the euphemism, "miscalculations") on the complexity of some

AI programs or on the lack of detailed enough specifications by their clients. When some AI developers reject ethical responsibility, "they fail to recognize that, in the process of developing software, they are not just instantiating specifications and implementing programs, but they are additionally providing a service to society."[53]

Drawing on some earlier work,[54] three AI researchers suggest what some ethicists may consider a useful but initial distinction only. Thus we may think of some AI developers exhibiting negative responsibility, that is, producing "correct artifacts without considering the potential effects and influences of the artifacts in society. By contrast, positive responsibility considers the consequences that the developed machine may have among users."[55] Negative responsibility may protect many AI developers from legal liability, but it does not deny their ethical responsibility, particularly with respect to their clients, users, fellow professionals, and the public.

Notably, these dimensions are 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 formulates no fewer than eight separate principles regarding the ethical behaviors of AI professionals.

Much more generally, in June 2019 the leaders of the twenty 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.[56] 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."[57] Also, in August 2019, the G7

leaders formally launched the International Panel on Artificial Intelligence (IPAI), including a call for research projects that include a large place for the ethical dimensions of AI.[58]

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."[59] "Computational artifacts should fulfill moral values together with common functional requirements," several AI professionals recently write.[60]

"Beside correctness, reliability, and safety," these professionals continue, "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.

The complexities here cluster especially around just how the various moral and ethical values at issue are to be understood. Some rather distinctive philosophical work has tried to contribute to the rather fundamental theme of the interactions of persons with one another through machine interfaces like the Internet.[61] In this phenomenological philosophical domain, interaction between humans and humans and between humans and 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."[62] Among those categories are communities and moral and ethical concepts themselves.

Although the phenomenological literature on community and ethical concepts such as ethical responsibility is large, three related although different approaches may be roughly sketched as follows:[63]

	Artefact / 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, they tend 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.

	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 its design and in its 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.

	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, however, remains too much in the background: the issue of ethical responsibilities of AI developers with respect to the unprecedented vulnerability of their artifacts. I turn briefly to this issue now in my last section.

4. AI, the Ethical, and Vulnerability

> ". . . [D]igital reality," the organizers write, "has introduced *a new and particularly pernicious kind of vulnerability* that prevents us from detecting how, through the power of invisible digital algorithms, our thought and decision-making processes are influenced. . . ."[64]

What is this new kind of vulnerability? We have become vulnerable by having "our thought and decision processes influenced through our ignorance of having been manipulated and our "genuine needs" not being met, because "wants and needs have become confused and illegitimate wants have been transformed into needs." Is this really new? Let us suppose the ways in which this manipulation and transformations operate are new because of the apparent omnipresence of digital, in particular, the newest self-learning technologies.

Generally speaking, a person's vulnerability is his or her liability to be "physically or emotionally hurt."[65] According to this British usage, vulnerability is "the state or quality" of a person to likely 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).[66] In phenomenological philosophy, some important work on ethics and vulnerability goes back to the early work of the Danish philosopher, Peter Kemp (1937-2018).[67]

The main senses of the expression "vulnerability" for our own concerns here with AI, ethical responsibility, and vulnerability, have to do with the susceptibility of persons to ethically harmful or injurious attacks. What is ethically injurious is what substantively undermines a person's normal capacities to act in accordance with their most well-considered ethical values.

For example, most reflective persons think that preserving their own privacy and that of those who are close to them personally and professionally is a basic ethical value for their relations, both with themselves and others. Privacy in this ethical sense, the sense of having and acting with the fundamental freedom from secret or un-

wanted disturbance or intrusion, is not just a legal but a specifically moral value that must be continually respected in persons' daily actions.[68] This is so because privacy is intrinsically linked to issues concerning personhood and self-identity.

Let me highlight here but three specific ways only of just how hybrid AI trades on the vulnerability of many persons. Besides hybrid AI's capacities to invade the most intimate corners of individuals' privacy,[69] a second exploitation of persons' lives is hybrid AI's capacities to introduce major bias into how many persons are treated.[70] The third is the use of facial recognition techniques to track people's movements without their knowledge or consent.[71]

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 differ among cultures and individuals. When something is private to a *person*, it usually means that something is inherently special or sensitive to them."[72]

The vulnerability of privacy that may be violated here by hybrid AI systems is preeminently the general right all persons hold to determine for themselves just what they are willing to share with others.

This kind of vulnerability is importantly different from what persons may undergo unknowingly, as in many public health systems, when AI algorithms are applied to their health records. 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."[73]

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. Appropriate codes for developing minimal biases in AI algorithms have yet to win effective consensus. Persons' vulnerability to hybrid AI systems might thus be called the vulnerability to algorithmic biases.[74]

A third kind of vulnerability to today's advanced AI systems is persons' vulnerability to unwanted identification through AI facial recognition systems. "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."[75]

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 use of hybrid AI facial recognition systems. Some uses, for example in demonstrable security contexts, are unobjectionable. But many other uses, for example in tracking individual students participating in legally authorised demonstrations against some government university or government policies, seem clear violations of persons' vulnerability.

In short, while people exhibit many different kinds of vulnerability, either with respect to diseases or to recurring natural disasters or to increasing climate change, different kinds of vulnerability, especially with respect to the misuse today of hybrid AI systems, raise particularly acute issues about ethics and social justice.[76]

Envoi: A Culture for AI?

If we have on hand then reasonably good definitions of what AI looks like now, just how far will future developments of AI take us?

At least one major element in assessing such future developments, however, is already reasonably clear. For not only is AI continually developing, the field of ethics itself is also continually developing.[77]

Philosopher Andy Clarke's response to such a question merits noting. He believes that unlike human development that includes cultural systems, "there is nothing similar for AI systems. Their development," he continues, "will take off when something similar to culture exists for them – some way for them to create the conditions under which they can learn. My best guess," he concludes, "would be that we will start to see the most powerful forms of AI emerge when simulated AI agents are able to talk to each other as part of proper communities." [78] Will they also, some may ask, be ethical entities?

Endnotes for Essay Six

1 This is a revised version of an invited paper first presented in shorter form at the second international conference on Integral Human Development in the Digital Age on the theme, Informational Overload, AI, and Responsibility, held at The Ukrainian Catholic University in Lviv from 26-28 February 2020. I thank V. Turchynovskyy for his generous invitation and participating scholars and students for their questions and comments.

2 A. Clark, *The Philosophy of Physics: Quantum Theory* (Princeton: PUP, 2019), p. xii.

3 M. Segal, "Interview with A. Clark on 'A Philosopher's View of Robots'," *Nature,* 571 (25 July 2019), p. 18; cited hereafter as "Clark 2019").

4 The context of fruitful multidisciplinary dialogue on matters of public relevance with special reference to "Integral Human Development" is "Catholic Social Teaching (CST) wherein 'Integral Human Development' is employed as a key concept for facing the complex challenges and opportunities involved in dignified human development and social progress." Please note that, in accordance with the contexts specified in the Letter of Invitation, the particular contexts of this presentation are to be found in the May 2019 publication of the *Conseil Permanent de la Conférence des évêques de France, Qu'est-ce que l'homme pour que tu penses à lui? Eléments d'anthropologie catholique* (Paris: Bayard, 2019), *passim.*

5 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/.

6 See for example, G. Brock and D. Miller, "Needs in Moral and Political Philosophy," in: *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition), ed. E. N. Zalta (ed.), https://plato.stanford.edu/archives/sum2019/entries/needs/.

7 Conference Letter of Invitation 18 July 2019: cited hereafter as "*Conference Letter* 2019."

8 In the text angle brackets, < >, enclose text excluded from the oral presentation.

9 *Conference Letter 2019*, my italics.

10 Clark 2019. See M. Segal, *Surfing Uncertainty: Prediction, Action, and the Embodied Mind* (Oxford: OUP, 2016).

11 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]."

12 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).

13 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).

14 *Conference Letter 2019*.

15 "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).

16 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*.

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

18 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).

[19] 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."

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

[21] 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.

[22] 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 enough to account for the full gamut of intelligence capability . . . [Russell] ignores the strain of twentieth-century thinking whose holistic contextual understanding of reasoning 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).

[23] 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.

[24] Cf. L. Drew, "Agency and the Algorithm," *Nature,* 571 (245 July 2019), S 19- S 21.

[25] See *AIMA.*

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

[27] Heaven 2019, p. 165. Thus, 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).

[28] 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/.

[29] 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/.

[30] Cf. for many carefully detailed and recent examples 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.

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

[32] For details 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 favor [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]).

[33] Bergstrom and Bak-Coleman 2019, p. 40.

[34] See another massive leading textbook, I. Goodfellow, Y. Benglo, and A. Courville, *Deep Learning* (Cambridge, MA: MIT Press, 2016).

[35] Bergstrom and Bak-Coleman, p. 41.

[36] Moss *et al.,* p. 176. In their letter to *Nature* the authors refer to *The Social Construction of Technological Systems,* ed. W. E. Bijker *et al.* (Cambridge, MA: MIT, 2012).

[37] Heaven 2019, p. 164.

[38] *Ibid.*, my italics.

[39] With all the advances in AI today and more to come, perhaps a short distraction here is not 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).

[40] See "Virtual Reality" on Google; accessed 12 November 2019.

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

[42] *Ibid.*

[43] Again, for current BE I am using here the *SOED*, and for current AE I am using here the *AHDE*.

[44] "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."

[45] See for example 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.

[46] Maudlin 2019, pp. xi-xii; Maudlin's italics. Cf. his comments on the 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.

[47] 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.*

[48] Examples and adaptations from the *SOED.*

[49] S. Blackburn, *The Oxford Dictionary of Philosophy,* 3rd ed. (Oxford: OUP, 2016).

[50] See the magisterial philosophical discussion in 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).

[51] *Conference Letter 2019.*

[52] 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/.

[53] *Ibid.*

[54] 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).

[55] Turner, Angius, and Primiero 2019.

[56] 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.

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

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

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

[60] Turner, Angius, and Primiero 2019.

[61] 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/.

[62] *Ibid.*

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

[64] *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."

[65] See however several fresh philosophical reflections on the surprisingly rich notion of the expression "vulnerability" understood mainly not just in exclusively negative but also in positive senses. Cf. for example J. N. Chung's critical views in her "Review of M. D. K. Ing *The Vulnerability of Integrity in Early Confucian Thought* [NY: Oxford UP, 2017]," *Mind,* 129 (2020), 299-307.

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

[67] P. Kemp, *Théorie de l'Engagement,* 2 vols. (Paris: Le Seuil, 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.

[68] 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).

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

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

[71] 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 Jack Ma and Elon Musk the fonder of Tesla, and the urgency of AI developers insisting on respecting ethical and social principles.

[72] Wikipedia; accessed 16 November 2019.

[73] 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."

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

[75] Wikipedia; accessed 16 November 2019. For the history and the technology see the rest of this quite extensive article.

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

[77] See for example S. Franklin, "Ethical Research," *Nature,* 574 (31 October 2019), pp. 627-630.

[78] Clark 2019, p. 18.