AI

MDPI

Essay

AI and Regulations

Paul Dumouchel

Département de Philosophie, Université du Québec à Montréal, 455 Boulevard René Lévesque, Montréal, QC H2L 4Y2, Canada; dumouchel.paul@uqam.ca

Abstract: This essay argues that the popular misrepresentation of the nature of AI has important consequences concerning how we view the need for regulations. Considering AI as something that exists in itself, rather than as a set of cognitive technologies whose characteristics—physical, cognitive, and systemic—are quite different from ours (and that, at times, differ widely among the technologies) leads to inefficient approaches to regulation. This paper aims at helping the practitioners of responsible AI to address the way in which the technical aspects of the tools they are developing and promoting directly have important social and political consequences.

Keywords: intelligence; AI; regulations; data-driven agents; ethics; politics; moratorium; joint cognitive systems

1. Introduction

The issue of regulating AI has recently gained urgency with the appearance of ChatGPT and the proposed moratorium on research in AI. Should AI be regulated? How? To what extent? And by whom? A major difficulty when trying to answer questions posed in this way is that, in a sense, there is no such thing as AI. Artificial intelligence conceived as a single reality, as a unitary object, does not exist. There is no single, unique underlying power present in the different manifestations of what we refer to as AI. Artificial intelligence is not a mysterious, undefinable, ambivalent entity that threatens us with extinction and promises miraculous progress.

What does in fact exist are different research agendas and technologies, machines, apps, algorithms, programs, and scientific instruments: numerous devices and heterogeneous systems that fulfill myriads of different purposes, from the most futile to the most essential. "AI" does not refer to an immaterial entity that exists in itself, daily gaining in force and extension; "AI", as the term is commonly used, points towards a large number of physical objects, including sensors, optic fibers, data centers, screens, computers, Wi-Fi transmitters, microchips, underground cables as well as their assemblage in commonly encountered devices and gadgets, like smart phones, personal computers and all the connected objects of the internet of things (IoT). All this makes AI "real" for end users. "AI" also refers to immaterial entities, programs, algorithms, data, and lines of code, all of which are "written stuff".

The thesis I wish to defend is that this popular misunderstanding of the nature of artificial intelligence has important consequences concerning how it should be regulated. Viewing AI as something that exists in itself, rather than as a set of cognitive technologies the characteristics of which—physical, cognitive, and systemic—are quite different from ours (and that, at times, differ widely among the technologies) leads to inefficient approaches to regulation. It limits our ability to anticipate the consequences of AI's foreseeable developments and social diffusion. It undermines attempts to protect ourselves from the social and political dangers it presents. It limits our ability to anticipate the consequences of AI's foreseeable developments and social diffusion. It undermines attempts to protect ourselves from the social and political dangers it presents.

check for updates

Citation: Dumouchel, P. AI and Regulations. *AI* **2023**, *4*, 1023–1035. https://doi.org/10.3390/ai4040052

Academic Editors: Pablo Rivas and Gissella Bejarano

Received: 24 October 2023
Revised: 25 November 2023
Accepted: 27 November 2023
Published: 29 November 2023

Following this short introduction, Section 2 rapidly retraces the history of the idea that intelligence is essentially a quality, one that artificial systems share with humans, and that this resemblance between artificial and natural cognitive systems trumps their differences. Section 3 reviews two approaches to the dangers of AI that reflect this illusion in different ways. Section 4 turns to how we should understand regulation, including its main goal and purpose. Section 5 analyzes three central characteristics of artificial cognitive systems, which Section 6 compares with the corresponding characteristics of natural cognitive systems. Finally, Section 7 draws some conclusions regarding how we address the question of regulating AI.

2. Artificial Intelligence: The Origin of a Misunderstanding

The conjecture that motivated early AI research not only assumed that intelligence (or thought) is a form of computation but also that there is no difference between human thought and the performance of an artificial cognitive system. Both were to be seen as exemplifying the same mysterious quality: intelligence. The 1956 meeting that launched artificial intelligence as a research program was based on the assumption "that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" [1]. The claim that there is no difference between human thought and the performance of an artificial cognitive system thus rested on the possibility of creating a machine able to duplicate the results of diverse intellectual operations.

Notwithstanding the unrealistic character of the conjecture that every aspect of any feature of intelligence can be so precisely described that a machine can be made to simulate it, and, for many years, the subsequent failure of that project to yield substantial technical results—what is commonly known as AI winter [2]—the claim about the identity of human and artificial intelligence was widely received. Any success at mechanically duplicating any intellectual activity was taken as a proof of the truth of the conjecture. This unfounded conclusion rested on two then (and still) popular reductions. The first is that the brain is a computer, or sufficiently like a computer, for thought to be identified with computation. The second is that, as a cognitive agent, a human person can be identified to his or her brain [3]. Given that human and artificial intelligence were viewed as substantially the same, people assumed that they could be compared and measured on a single scale, giving rise to the dream and fear of the day when AI would surpass human intelligence.

We have made cranes and backhoes to do some things better than we can and also to accomplish tasks that we cannot. Similarly, we make artificial cognitive systems to perform better and more rapidly some operations that we can carry out, and others that we cannot. We should be neither surprised nor alarmed by the fact that they often perform better than us. That is precisely why we made them. If computers were as slow and as prone to make mistakes as we are when they calculate, we would not use them. To understand the advantages and dangers to which interacting with artificial cognitive systems may lead, we need to focus on their particular characteristics as cognitive agents, to focus on the different ways they perform their cognitive tasks, not stand in awe of our own creations.

The claim that humans are like any other animal is in part similarly motivated by ideological reasons. However, no one believes that there are no differences between a human pancreas and a genetically modified bacteria that produces insulin, or that the differences between the two organisms pale in view of the fact that all aspects of this peptide hormone have been so precisely described that it can be artificially produced. On the contrary! Biology is a discipline that inquires into the differences between forms of life, natural and modified, in relation to the continuity that exists among them. These differences are also at the heart of the regulations that we put in place to create a protective framework for our interactions with other organisms.

As Luisa Damiano and I argued elsewhere, there are reasons to believe that the cognitive domain is heterogeneous rather than homogenous. There are many different forms of cognitive systems and of cognition, and this diversity is revealed by, among other

things, the types of mistakes that different cognitive systems make. For example, both humans and computers make mistakes, but they do not make the same types of mistakes. When calculating, humans commonly make mistakes due to inattention. In consequence, we have evolved techniques to correct such mistakes, preferably, but not always, before they invalidate the final result. Because we calculate slowly, this corrective procedure is often successful. Computers do not make such mistakes and do not need the discipline of reviewing the operations they have already performed, but, as we will see, they make different types of mistakes.

3. False Starts: A Moratorium on AI Research and Ethics

The fact that AI corresponds to many different types of systems and material objects constitutes a major difficulty when trying to characterize it in relation to regulation. Further, these different systems intervene in highly different environments—social (here, "social" is understood to mean instances where AI (like apps on smart phones, search engines, personal assistants, or ChatGPT) interact with the general public, but that is only one of the environments where artificial cognitive systems are prevalent), industrial, administrative, regulatory, security, scientific, etc. Finally, there is often no identity between a system as it is technically defined and the various applications to which it may lead and with which the lay public is confronted. These cannot always be identified as the same individual. However, the term "AI" is commonly used to designate both types of objects. On one hand, for example, there is machine learning and large language models (LLMs), and on the other hand, the applications of machine learning, LLMs, and other techniques in facial recognition, language tutoring, personal assistants, and products like ChatGPT. "Artificial intelligence" is a blanket term that encompasses both fundamental research and development and specific applications. What should be the main target of regulation: fundamental research or its various applications?

In the first case, regulating fundamental research on AI seems difficult if not impossible because a heterogeneous collection of technical innovations and research endeavors does not constitute a proper object for regulation, which inevitably necessitates specifying the characteristics common to the objects or practices that are to be regulated. This is the difficulty which, I believe, underlies the recently proposed moratorium on research in AI. As long as we think of AI as a single entity, and of "intelligence" as the only thing that the different smart machines that populate our world have in common, trying to control it directly appears like a good solution.

A moratorium, however, is not a form of regulation. As a form of control, it is more like a double bind, a contradictory injunction that is doomed to fail [4]. On one hand, a moratorium is more or less the equivalent of saying: "Stop! Let's wait and see. Let's hold back further development". On the other hand, because it is a temporary suspension, it assumes that the development of AI will at some point start again. All that the moratorium aims to do is to slow down the growth of AI until we have adapted sufficiently to the inevitable societal changes that it will bring about. It is doomed to fail because it does not and cannot specify the conditions that need to be met for it to be lifted. It is also doomed to fail because the idea of a moratorium on research in AI is simply too vague. Does it target all research in AI including the development of new mathematical tools, or only, say, machine learning or LLM? This imprecision in the formulation of the moratorium reflects the illusion that AI essentially is only one and the same thing.

Not only is the call for a moratorium on fundamental research not a form of regulation, it is its very opposite. That call is made to the industry to voluntarily slow down the pace of AI development. It is a bit like saying to the industry "Please regulate yourself." Why is that call made? Is it because we think that we cannot do it, that only the industry and only specialists have the necessary knowledge and expertise to control AI? The proposed moratorium, if accepted, would then appear as the proof that the developers of AI are responsible social actors who are willing and able to take a step back (It is therefore not surprising that the call for a moratorium was supported by major actors in the field: not

only by well-known researchers like Yoshua Bengio and Stuart Russell, but also by Elon Musk, Steve Wozniak, Emad Mostaque, and the CEOs of other companies that produce artificially intelligent devices.) It is not clear whether we should take their word for it.

The second branch of the alternative, namely, that applications are what should be regulated, also faces serious difficulties as long as we think that what essentially makes these technologies different is that they are "intelligent". This suggests that their activity should be constrained, disciplined, and evaluated in the same way as that of other intelligent agents, for example, humans. This is the assumption that underlies the plan to control intelligent machines through ethics. It generally takes one of two forms.

The first consists of the numerous warnings concerning the ethical consequences of interacting with artificial agents rather than with other humans. For example, the consequences that these interactions may have on an individual's ability to think critically, to develop a sense of self [5], or a sense of justice [6], and more generally on the different virtues that one should or needs to acquire [7]. The second is the promotion of the development of artificial moral agents (AMAs), which would be artificial agents on which we impose ethical norms that they are programmed to respect in different circumstances and situations [8–11].

The first proposal suffers from the fault that this virtuous strategy and its associated criticisms are not essentially geared towards the particular characteristics of artificial cognitive systems. Apart from the specific threats to privacy due to the characteristics of artificial agents that will be analyzed later, the dangers that these critics identify, though generally real, are not limited to the deployment of artificial cognitive systems. Endangering individuals' ability to think critically and making them unable to face unforeseen adversity or to give meaning to their own life are not new phenomena. They are not peculiar dangers characteristic of AI. The same allegations have been made of television, and before that, of newspapers, theatre, literature in general, and, in ancient Greece, Plato argued that those dangers were inherent to writing as opposed to the spoken word! Many of these fears are not ill-founded, but they correspond to recurrent difficulties of social life, rather than reflect the specific consequences of the introduction of artificial cognitive systems.

The second proposal, in which artificial moral agents would be created, faces different difficulties. It overlooks central differences between artificial agents and human actors, such as the distinguishing fact that moral rules and human laws can be challenged and resisted. We are, to some extent, bound by human rules and laws, but we are not forced to obey them, and this is related to the type of cognitive systems that we are. That human laws and morals can be transgressed and challenged constitutes a fundamental aspect of social life [12]. Our capacity to question laws and rules places their acceptance and rejection in a space of public discussion, a space of dissensus, as Rancières [13] says. Coded instructions, in contrast, cannot be resisted. Failure to follow them is, for a system, a malfunction. It is not an action. AMAs are artificial agents, but they are not moral, no matter how complex and autonomous we may wish them to be. Giving AMAs more "moral code" would transform the particular moral views of the individuals who design and control such machines into inescapable regularities of proper functioning.

Furthermore, ethics is not a form of regulation. Regulations and legal provisions do not aim directly at making agents moral, nor is their goal to define what is a good life, or a moral life (In the sense that persons subject to the same system of laws and regulation can have very different conceptions of what is moral and of what is a good life.) They also do not require agents to be moral (i.e., to never act immorally). Regulations seek to avoid certain consequences and to obtain (or at least to encourage) certain outcomes. Think of air traffic regulations or regulations concerning the use and sale of medications. Concentrating on the ethical qualities of artificial agents, either indirectly, such as by targeting whether they enhance or endanger the virtues of those who interact with them, or directly, by making them act morally, is to treat them if they were "intelligent", in the sense in which we use the term in relation to humans (This confusion is also what brings some to wonder if we should give artificial agents rights, or if they can be considered responsible.) It is, to use

an image, to treat them either as bad companions that we would like our children to avoid, or as children in whom we seek to inculcate moral rules. However, they are only machines. A purely ethical approach to them is not only doomed to fail, but it also circumvents and undermines the fundamental purposes of regulation, which are to protect us from dangers related to the use of different objects and to facilitate socially beneficial results.

4. On the Use and Purpose of Regulations

In any domain, regulations usually intervene when we cross the threshold of application, that is, when the "object" leaves the laboratory to enter the social or natural world. Research on viruses provides a good illustration of this point. We are essentially interested in whether or not the little creatures can escape the lab. Regulators' interest in the research itself is commanded by the consequences of a leak. Drafting adequate regulation is guided by the nature of the object and technologies involved in view of the consequences of an accidental leak or of a voluntary introduction of the virus into the environment. Regulators are generally not interested in the research as such, but only indirectly in its uncontrolled consequences. For example, in discussions concerning the regulation of what is called "gain of function" research on viruses [14], the debate focuses on the dangers resulting from the fact that a consequence of such research is that the viruses could become more virulent or contagious or more easily jump the "species barrier". Drafting good regulations focuses on characteristics of viruses in general, how different research involving them can modify some of these characteristics, what the consequences could be, and how any ensuing risks should be managed.

In the case of AI, it should be the same. Reflection on the need for regulation and the type of regulation necessary should focus on the characteristics of artificial cognitive systems that determine the central aspects of their possible applications. The fact that they are in some way "intelligent" is a too vague and ill-defined aspect to be useful here. Furthermore, it easily leads to imagining inscrutable probable effects, unfathomable benefits, and unknowable menaces, all of which are beyond the scope of what regulations can address.

The objection that, in this case, research cannot be separated from application, and that, here, there is no inside and outside of the laboratory, does not really stand. Even if working on large language models inevitably leads, at some point, to creating a functioning device, like, for example, ChatGPT, making it accessible to anyone for just about any purpose is a different issue. Similarly, research on new medications, for example, is by definition always applied; however, it is when the medication is to be administered to persons that regulations become central. *Mutatis mutandis*, a similar approach to the problem, should be adopted here. What is fundamental for regulation is not which new products or technologies are developed, but how they are introduced into our social world, at what price, and with what consequences. How do we protect the public from the dangers and downsides they occasion while optimizing the advantages they provide? (Regulations are always, to some extent, political, because it is usually not the same groups who reap the advantages and who suffer the disadvantages. Therefore, regulations also inevitably reflect power relationships between groups).

It is the interface between the object and the public that is at the heart of any regulation. As Lucy Suchman [15] argued many years ago, humans and artificial cognitive systems do not share the same cognitive domains. Not only are there many things that a human may know and that an artificial cognitive system does not, and vice versa, but there also always are types of information that are inaccessible to either one or the other. It is those differences that should be at the heart of our reflection on regulation. As Woods and Hollnagel [16] point out, in order for humans and artificial cognitive agents to successfully accomplish a joint task, to avoid failures and accidents as much as possible, human operators must always be aware of what the machine is doing, of why it is doing it, and of what it is likely to do next. It is fundamental for operators to know the basic characteristics of the machine's agency and cognitive domain.

The examples of joint cognitive systems used by Woods and Hollnagel come mostly from working environments in industries or hospitals. These are environments where the task to be accomplished together is (usually) relatively well-defined and delimits to an extent the cognitive abilities of the machine. Universal Turing machines, as their name indicates, are much more versatile; they can be put to numerous different uses. Nonetheless, we always interact with them through some more or less specific program that allows us to accomplish some, but not all, joint cognitive tasks. In every case, the material or immaterial characteristics of the machine (and generally both), as well as our own characteristics, delimit the interactions and joint cognitive activity that are possible. In consequence, whether we interact with a search engine, a word processing program, or any other app or cognitive agent, there is a sense in which, together, we form a joint cognitive system.

Joint cognitive systems are engaged in doing something together. They pursue a common objective to which both the natural and the artificial agents contribute. Their respective contributions are different, and their different agencies are central to the success of their endeavor. Any joint enterprise requires a more or less complex division of labor, which is determined in part by what each of the agents can and cannot do. To regulate interactions between humans and AI, it is therefore important to inquire into how the agency of artificial cognitive agents differs from that of natural cognitive agents. Because there is a vast array of different tasks that may be accomplished by humans and different artificial cognitive systems working together, we should focus on characteristics that are common to all or most artificial agents.

5. Intelligent Agents and AI

All artificial cognitive systems that correspond to what we call AI contain or comprise an "agent". A classic definition in AI states that "(a)n agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators". This definition from Russell and Norvig ([17], pp. 31) is somewhat misleading because it claims that the agent "perceives" its environment through sensors and "acts" upon it through actuators, and this suggests that the artificial intelligent agent is in the world as is a person who sees with her eyes and moves an object with her hand. However, the notion of an agent in AI is much more abstract than this definition suggests, given that an agent is completely specified by its agent function, the mathematical function that maps precept sequences into actions. Such an agent, therefore, is essentially a mathematical object. It is not a physical thing that interacts directly with the world in which we live. The dynamic environment in which it can act autonomously and to which it responds intelligently is a digital one. For such an agent, to "perceive" is to receive data input and to "act" is to produce a data output. Sensors and actuators, where they exist, intervene down the line, so to speak, having been calibrated to yield the right type of data or to transform the output data into the correct event in the world: printed numbers on a screen, a change in the trajectory of a plane, or in the dosage of drug delivered to a patient.

As a different classical definition puts it: "Autonomous agents are computational systems that inhabit some complex dynamic environment, sense and act autonomously in this environment" [18] What is this environment? It is a model of the world, one that has been made in view of whatever it is we would like the intelligent agent to do. Russell & Norvig ([17], pp. 34) also give the following table (Table 1) of the types of percepts (data) to which different agents are exposed: a medical diagnosis system receives information concerning patients' symptoms and conditions, a system that analyzes satellite images receives pixels of different colors and intensity, a refinery controller receives temperature and pressure readings, and so on.

In each case, the data both determines and reflects the model of the world which the agent "inhabits", which it "senses", and to which it responds more or less intelligently and autonomously.

Table 1. Types of precept to which different artificial agents are exposed according to Russell and Norvig.

Agent	Precept	Action	Goal	Environment
Medical diagnosis system	Symptoms, findings, patients' answers	Questions, tests, treatments	Healthy patients, minimize cost	Patient, hospital
Satellite image analysis system	Pixels of varying intensity and color	Print a categorization of scene	Correct categorization	Images from orbiting satellites
Part picking robot	Pixels of various intensity	Pick up parts and sorts into bins	Place parts in correct bins	Conveyor belt with parts
Refinery controller	Temperature readings	Open and close valves, adjust temperature	Maximize purity, yield, safety	Refinery
Interactive English tutor	Typed words	Print exercise, suggestions, corrections	Maximize students' score on tests	Set of students

That model only contains data of certain types, coded in a certain way. It constitutes, by definition, an imperfect, partial, and incomplete representation of the world. Incomplete and imperfect, first in the sense that the model is designed in view of the task that we want the agent to accomplish. Much information and types of information that are, or appear, irrelevant for that purpose are left out of the model. Big Data notwithstanding, processing a lot of information is expensive in both computer and monetary resources, and there are pressures to make the model simpler and more manageable. It follows that the model is imperfect and partial in another sense, namely, relative to the specific task for which the agent is designed (This corresponds to an experience that anyone who has interacted with automated public or private services has probably made. In many cases, whether the appalling stupidity and incompetence of the intelligent agent is by design or by accident remains an open question.) It is true that complex models that reflect ongoing changes in the world can be made, and autonomous agents that adapt in innovative ways to the changes can be designed. The models nonetheless remain limited. There are always aspects of the world that are not modelized, either because they were not anticipated or anticipated not to be relevant or judged too improbable to be worth the effort. Finally, there are necessarily aspects of the world—modelized or not—that are assumed never to change.

Artificial agents are data-driven agents who "inhabit" a model of the world (Talking of agents in this context always trades on two different meanings of the term. A mathematical function that maps inputs onto outputs is not, properly speaking, an agent, therefore, when we attribute to such mathematical agents anthropological characteristics, like "perceiving" or "inhabiting" a model of the world, it should be remembered that this use is always metaphorical). No matter how rich and complex the model and the data they process, they are "literal agents" that are exclusively ruled by the agent function that determines them. Literal agents can also be defined as pure agents in the sense that an artificial agent cannot do anything else than what it does: recognize faces, make predictions about future trading on the stock market, identify cancerous cells in data from fMRI, process job applications, etc. Artificial cognitive systems can be highly "intelligent" in the sense that they can perform extremely well whatever task they have been designed to do, but they are completely bound by that task, and cannot take distance from it. This is true of supervised and unsupervised machine learning as well. Once a system has learned to do something, it cannot be made to learn something different. It cannot expand its knowledge outside of that original domain.

The recent ChatGPT, which is much more versatile than other artificial agents, including previous LLMs, may seem to be able to escape those two difficulties: the prison of its model and literality. It can pretty much answer any question about anything. Yes, but can it do anything else? Asked if it can emulate a plane autopilot, ChatGPT answered "I cannot emulate any other physical or software system. If you need information about any such system, I would be happy to…". ChatGTP is considered by some as an example of General Artificial Intelligence (GAI); however, it cannot do what a GPS does or even what the humble app that tells you when the next bus will arrive does. It cannot emulate any other physical or software system. Asked if it can accept data in any form other than written language, it answered that it cannot accept data otherwise than as written text

(ChatGPT was asked these questions on 7 September 2023. The situation has apparently changed since and the more recent version can also accept images.) All that ChatGPT can do is talk, or rather talk, talk, talk! Just as when we say of someone that "all he or she can do is talk", similarly, ChatGPT has neither interest nor access to the truth or falsity or consequences of what it "says". That it is a literal agent captive of its model of the world is further illustrated by two other aspects of its behavior.

First, bias and offensive content. OpenAI hired hundreds of Nigerian workers to label harmful content to train the agent to recognize such content when exposed to it and to avoid producing biased offensive statements [19] (This illustrates clearly that the social influence of this new technology sometimes begins before the machine is actually fully developed.) There have nonetheless been numerous reports of bias on its part. Interestingly, ChatGPT has, for example, been known to refuse to tell jokes about women or persons from India, but not about English males. It has also been accused of being biased in favor of Biden, Democrats, and left leaning figures in general, and against Trump and Republicans. This aspect of its performance illustrates first that there is no general rule against bias. Why Englishmen but not Indians? Presumably because the system has to be trained to recognize and avoid some statements relative to specific classes of people: Jews, Black persons, women, Muslims, etc. Second, it indicates that what people find offensive not only differs from group to group, but also in time: what was offensive yesterday is not today and what is offensive today may very well not be tomorrow. Bias is not something which is simply *in* the data; it happens in the world.

Second, what are referred to as "hallucinations". This refers to when the system responds in an erroneous manner or reports events that did not happen, lists references that do not exist, or quotes invented sources or legal decisions, while claiming that all of this is true and real. In short, "hallucinations" correspond to when ChatGPT turns into a false news and alternative facts engine. One evident difficulty here is that there is no way of knowing how big the problem is. How often does it happen? It is impossible to check in all cases if its claims are true or correct, not only because its data base is so large that no one person or even community of scholars would be able to supervise all its answers, but also because ChatGPT can interact with thousands of persons simultaneously (it has already interacted with millions) and it does not necessarily give the same answer to the same question to everyone, or the same answer to the same question twice to the same person.

The only way an artificial agent has of being informed of its mistakes is whatever feedback we have decided to give it, that is to say, all it has is what we have modelized. If we have not anticipated the issues that may cause it to fail, then it does not have that information. To put it in a more metaphorical way, an artificial agent is not interested in what takes place in the world, or where it takes place, beyond what we have made it interested in it. For example, a face recognition system may know with extreme precision the distance that separates the camera from the face on which it focuses, but it knows nothing of where it is and very little of anything else that is happening there. Data-driven agents are prisoners of the digital world in which they live and they have no way to escape. They may be able to adapt to that world better and faster than we do to those aspects of ours that correspond to the data they receive, but of the aspects and events that have not been modelized, they know nothing.

Not only is the artificial agent unaware of its mistakes, but many times there is a sense in which the agent did not make any mistake. Often, when its action leads to what in our world is a mistake, it simply follows the complex rule that governs its behavior. The "mistake" did not happen in the relation between the data input and output: it took place in the world or somewhere in the material system which allows it to act in the world. Such mistakes exist in relation to the social or technical function that the agent is assigned to fulfil in view of the expectations and goals of the persons with whom it interacts, or in view of whatever effect we want to it produce. However, for the agent itself, there is no such thing as a mistake. That is why, as O'Neil [20] clearly shows, in the absence of

adequate human feedback, many mistakes will become entrenched in the agent's behavior through reinforcement.

Data-driven literal agents who inhabit a digital environment are also nowhere and everywhere. They are nowhere in particular because a GPS, a weather app, or a money converter can be accessed from anywhere there is an internet connection. That is also why they are everywhere. These are two aspects of their absence from the world. Unlike a person or an animal (or any physical object for that matter), there is no place where artificial agents are to be found to the exclusion of any other place. Artificial agents are also nowhere in a different sense. Where is the agent in a GPS? Or should we ask where is the agent *of* a GPS or any other app? The fact is that there is no possible answer to such questions. Whatever it is that responds to your queries when you search for a restaurant or an itinerary is neither in my phone nor yours, neither in his car nor her tablet. The agent is not anywhere in the world. Related to its absence of location is the fact that the same agent can in principle interact simultaneously with large numbers of individuals who are all in different places. In relation to all, it is the same agent, but it interacts, or more precisely, it appears to interact, with each one individually.

We never encounter the agent itself for it ultimately is a mathematical function. We are only exposed to some consequences of its functioning, and there are many other consequences of its actions or decisions of which we are not and cannot be directly aware. In consequence, such artificial agents are, to us, invisible and ubiquitous. They inhabit a digital model of the world. As literal agents, they are unaware and uninterested in the consequence of their actions in the world. As pure agents, they are unable to escape the role they have been assigned.

6. Human Actors and Artificial Agents

Some basic attributes of human individuals and artificial intelligent agents are in many ways perfect opposites. To begin, a person, unlike an artificial agent, is always somewhere in particular and visible. Unless one takes measures to dissimulate oneself, one's presence is inevitably known to others who are present. That is to say, when humans are present, they are not only somewhere in particular and visible, but their presence is also reciprocal, unless measures have been taken to hide.

For artificial agents, on the contrary, as we have just seen, the default rule is that they are invisible and that special measures need to be taken in order to make the agent's "presence" and action known. For example, a message appears on the screen of your computer asking you to allow some app to have access to your contacts and images or track your position. Even then, independently of what you may have answered, you are unsure of how many or which agents are mining your data and what information they are collecting. The action of a data-driven agent remains unknown, unless we are specially informed or if its consequences are immediately perceivable, as when appears on a screen a text that answers our query, or whenever the agent reacts directly to our demands.

An important aspect of a natural agent's physical presence in the world is that the agent can act directly only where it is present, which constitutes a limited domain of potential publicity (The issue of privacy is inextricably linked to this difference between the "natural" public dimension of natural agents and the invisibility and anonymity of digital agents.) Without my resorting to special indirect means, my action remains limited to the here and now, where it is exposed to the scrutiny of others and their reactions. This is not the case of the actions of ubiquitous artificial agents that are not in the world. They are invisible and not sensitive to the reactions of those who experience the consequences of their actions except insofar as we make them so.

Natural agents are, in a sense, inevitably public. Data-driven artificial agents, on the contrary, are known only (and only partially known) by the consequences of their actions. We have no way of being aware of their presence and action whenever their action's consequences cannot be directly perceived or cannot clearly be attributed to the intervention of the agent. Because persons are physical objects located in a particular place,

they are interested in all that occurs in the world that may influence their situation. A person may find an event or situation boring or be unaware of its importance, yet the event or situation will become interesting if it has an incidence on the person's situation. What happens in the vicinity of where a natural person is and the consequences of that person's action are especially relevant because they risk affecting him or her immediately.

On the contrary, the algorithm that reads your bank card and password, and commands the opening of the slot where bills appear is not interested in whether the money was picked up by the legitimate cardholder or a thief, or in the fact that a short in the mechanical device that pushes the bills out caused a fire. It is domain-bound and absent from the world where such things as muggings and fires take place. The artificial agent is perfectly indifferent to all that happens in the world, including the consequences of its actions, apart from what we have made it able to take into account, even if those consequences directly affect its ability to fulfill its task. There is one last important difference between humans and artificial agents, a difference that is indirectly illustrated by the vast literature in economic theory concerning the agent/principal relationship.

The principal, in economics, is the person that an agent represents, and the agent is a person who exercises, by delegation, the function of the principal or more generally acts in the principal's name. Central to this economic literature is the danger that the agent may use his/her position to promote his/her own interest, rather than that of the principal, and the means available to the latter to protect himself or herself. This danger arises from three conditions. The first is that the agent has interests of his or her own. The second is that the agent has access to information which the principal ignores. The third is simply that the agent's action inevitably escapes to some extent the control of the principal. Beyond the economic literature's obsession with conflicts of interest, what the agent/principal problem clearly illustrates is that human actors can move in and out of a role. They can stop being an agent, not only when it suits their interest, but simply, say, after 5:00 p.m. They can take on different roles at different times. Unlike pure literal agents, humans can take distance from whatever role or function they have been assigned or have chosen to adopt for a while (This is related to what was said earlier, that, unlike artificial agents, humans can disobey laws and moral rules. These are not to them like coded instructions with which they have to comply.) An artificial agent, because it is completely defined by its agent function, *is* that role or function, and nothing else.

Unlike a human actor, an artificial agent has no interests of its own. It cannot take any distance from its role or function, and it ignores any consequence of its actions that it has not been designed specifically to be made aware of. In the case of an artificial agent, the principal can determine a priori the extent of information to which the agent has access. All the reasons why an agent might not follow the principal's instructions perfectly have been removed. Artificial agents provide the perfect solution to the principal/agent problem as it is commonly framed. Furthermore, an artificial agent is able to interact simultaneously with thousands of persons, flawlessly representing the interest of its principal. Where, previously, hundreds of human agents were needed. Individuals that all had their own interest, who may have hesitated to act as instructed, perhaps because of some consequences of their action (as Atkins [10] writes page 117 concerning artificial lethal agents "we do not want the agent to derive its own beliefs concerning the moral implications of the use of lethal force, but rather to apply those that have been previously derived by humanity as prescribed in the laws of war and rules of engagement"), it is now possible for there to be only one pure literal agent.

7. The Need for Regulation: A Political Analysis

It is often claimed that financial reasons are what motivate the drive to replace people doing numerous clerical jobs and services with intelligent artificial agents. This may the case, but the evident result of such automatization is that it augments and entrenches the power of those who control the machines (Those who control the AI systems are not always those who resort to them to automatize the jobs and service. For example, across the world,

many government agencies outsource to private companies the task of automatizing some of their services, and, in the process, abandon their right of inspection and control over those aspects of their services.) The creation, development, and implementation of AI agents are, at this point, already dominated worldwide by a few large corporate actors. The main dangers of the large-scale diffusion and application of AI are not ethical, nor do they concern such metaphysical questions as whether artificial agents should have rights or whether AI, upon becoming conscious, will try to exterminate us. The dangers are political. The most evident consequence of the growing importance of AI agents in all walks of life, and main driving force behind this transformation is that they change the power relationships between the different social actors and groups.

Artificial cognitive systems and associated information and communication technologies have been central in the unprecedented concentration of wealth and power that we are witnessing. This transformation is not, however, the result of some form of technological determinism. Nor does it constitute an inescapable destiny. Rather, it is the consequence of how artificial cognitive agents are deployed in most sectors of our social environment. More precisely, it is the consequence of the fact that they are deployed in such a way that they deprive those who interact with them of the ability to respond to their actions otherwise than through very limited channels.

AI, artificial agents, and cognitive systems can, and often are, used in very different ways in various professions. Especially, as I argued elsewhere, in natural sciences, the meaning and value of data automatically collected or processed by artificial agents, the results at which the AI arrives, and the way the data are interpreted are not judged in the final instance by the machines themselves. "In all cases in the natural sciences, the locus of interpretation of the data, remains within the discipline itself. Both what constitutes data and how it should be interpreted are under the collective jurisdiction of specialists of the domain and that final authority is recognized by governments, funding agencies and the general public" [21].

The main reason why scientific and professional communities retain the power to accept or reject the conclusions reached by artificial cognitive systems is because scientists and artificial cognitive agents are essentially trying to do something together. They are engaged in an enterprise with an objective that is always beyond whatever the machine can do. Their collaboration results in many different joint cognitive systems. In scientific research, AI and artificial cognitive agents constitute scientific instruments, and, like all scientific instruments, they both incorporate and produce knowledge [22]. These extraordinarily powerful tools contain hypotheses and a model of the world that are known by those who employ them. Scientists are well aware of the differences between their own way of knowing the world and the cognitive domains of the machines they use. They also have a goal and expectations about how what they are searching for should look. It is to these, their objectives and their expectations, that they refer to determine if the artificial system's results are sound, when and where they incorporate noise, or are as they should be, or reveal an interesting anomaly [23].

In enterprises, administrations, and services, artificial agents are, on the contrary, most commonly thought of as replacing human employees. AI is commonly advertised (sold) as a labor cost-reducing option that will provide better service to consumers. Experience suggests that this is not always what happens. The limited cognitive domain which artificial agents inhabit and their lack of interest in the consequences of their actions qualify them poorly to be deciders in the last instance. Their ubiquity and the fact that they are literal agents entails that their decisions can have consequences affecting numerous persons and that they remain unaware and unconcerned by those consequences. It therefore seems unwise to let them decide in the last instance. Finally, the fact that they and their actions are invisible deprives the persons who interact with them of the ability to react directly and efficiently to their actions.

When artificial cognitive systems are built and deployed as part of joint cognitive systems, they are viewed as "collaborators", as co-laborers in a task that is beyond the

function that defines them as an agent. That task, whether it be a service to which certain people are entitled or safely flying a large plane to its destination, is something that only human operators can understand and human operators are the only ones able to judge its success.

Consider, for example, safely flying a large plane with hundreds of passengers to its destination on time. It is a very complex and difficult process, something that no human agent can accomplish by him or herself alone. Only through the collaboration and coordination of many different persons and systems—pilots, co-pilots, automatic pilots, air traffic controllers, radar, satellites, airport authorities, ground personnel inside the airport and on the tarmac, etc.—can it be achieved. While this process is going on, no one has absolute authority, no one is in complete control. Sometimes it is the pilot, sometimes the air traffic controller, sometimes one or another automation system, sometimes the airport authorities. When all goes well, it seems that everyone's task can be reduced to an ordered list of procedures, that coordination among these procedures can more or less be reduced to a question of timing and that each participant's role is equivalent to an agent function. However, things do not always go well and when what happens is different from what is expected, less rigid relations between the agents involved become necessary. As became evident in the two tragic Boeing 737 Max accidents which happened when the flight control system malfunctioned. They happened because the pilots could not correct the planes' trajectory, for the flight control system was designed to function independently and to overrule their commands. In fact, more subtle interactions between participants are often necessary simply to realize that things are not going well when the problem is not as sudden and evident as when a plane is nose diving.

Rather than conceiving of the relation between AI and humans as one of replacement and of comparison on a unique scale, where one or the other is superior, we should think of their interrelations in terms of joint cognitive systems relative to specific goals and purposes. Such a change entails, however, more than simply seeing them in a different light because artificial agents are ultimately as we construe and make them. Building artificial cognitive systems while keeping in mind the objective that they are to be parts of joint cognitive systems entails giving operators and those who interact with the artificial agents more power and authority over what they are doing together, not less, as has mostly been the case in the massive introduction of artificial agents in the social world over the last twenty years. This is where the fundamental problem lies, and it is political, not ethical, not metaphysical.

Funding: This research received no external funding.

Institutional Review Board Statement: Does not apply; the research does not involve any animals or humans.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. McCarthy, J.; Minski, M.L.; Rochester, N. A Proposal for the Darthmouth Summer Research Project on Artificial Intelligence. 1955. Available online: https://raysolomonoff.com/dartmouth/boxa/dart564props.pdf (accessed on 24 October 2023).
2. Winter, A.I. Available online: https://en.wikipedia.org/wiki/AI_winter (accessed on 24 October 2023).
3. Vidal, F.; Ortega, F.A. *Being Brains: Making the Cerebral Subject*; Fordham University Press: New York, NY, USA, 2017.
4. Bateson, G. *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*; University of Chicago Press: Chicago, IL, USA, 1972.
5. Cohen, J.E. The emergent limbic media system. In *Life and the Law in the Era of Data Driven Agency*; Hildebrandt, M., O'Hara, K., Eds.; Edward Elgar: Cheltenham, UK, 2020; pp. 60–79.
6. Delacroix, S.; Veale, M. Smart technologies and our sense of self: Going beyond epistemic counter-profiling. In *Life and the Law in the Era of Data Driven Agency*; Hildebrandt, M., O'Hara, K., Eds.; Edward Elgar: Cheltenham, UK, 2020; pp. 80–99.

7. Wynsberghe, A. Designing Robots for Care: Care-Centered Value Sensitive Design. *Sci. Eng. Ethics* **2013**, *19*, 407–433. [CrossRef] [PubMed]
8. Arkins, R. *Governing Lethal Behaviour in Autonomous Robots*; Chapman & Hall/CRC: New York, NY, USA, 2009.
9. Boyles, R.J. Philosophical signposts for artificial moral agent frameworks. *Suri* **2017**, *6*, 92–109. Available online: http://suri.pap73.org/issue9/Boyles_SURI_2017.pdf (accessed on 24 October 2023).
10. Scheutz, M. The Case for Explicit Ethical Agents. *AI Mag.* **2017**, *38*, 57–64. [CrossRef]
11. Wallach, W.; Allen, C. *Moral Machines Teaching Robots Right from Wrong*; Oxford University Press: Oxford, UK, 2008.
12. Hildebrandt, M. *Smart Technologies and the End(s) of Law*; Edward Elgar: Cheltenham, UK, 2015.
13. Rancières, J. *Dissensus: On Politics and Aesthetics*; Continuum: London, UK, 2010.
14. Sanders, D.A. Research on viruses is essential but can never be risk-free. *Times High. Educ.* **2021**. Available online: https://www.timeshighereducation.com/features/research-viruses-essential-can-never-be-risk-free (accessed on 24 October 2023).
15. Suchman, L. *Human-Machine Reconfiguration: Plans and Situated Actions*; Cambridge University Press: Cambridge, UK, 2006.
16. Woods, D.; Hollnagel, E. *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering*; Taylor & Francis: New York, NY, USA, 2006.
17. Russell, S.; Norvig, P. *Artificial Intelligence a Modern Approach*; Prentice-Hall: New York, NY, USA, 1995.
18. Burgin, M.; Dodig-Crnkovic, G. A Systematic Approach to Artificial Agents. *arXiv* **2009**, arXiv:0902.3513. Available online: https://arxiv.org/abs/0902.3513 (accessed on 24 October 2023).
19. Exclusive: OpenAI Used Kenyan Workers on Less than 2$ Per Hour to Make ChatGPT Less Toxic in Time magazine. Available online: https://time.com/6247678/openai-chatgpt-kenya-workers/ (accessed on 24 October 2023).
20. O'Neil, C. *Weapons of Math Destruction*; Crown Books: New York, NY, USA, 2016.
21. Dumouchel, P. Data agency and knowledge. In *Life and the Law in the Era of Data-Driven Agency*; Hildebrant, M., O'Hara, K., Eds.; Edward Elgar: Cheltenham, UK, 2020; p. 52.
22. Baird, D. *Thing Knowledge A Philosophy of Scientific Instruments*; University of California Press: Berkeley, CA, USA, 2004.
23. Cognitive Neuro-Science at the Crossroad. Nature 2022. Available online: https://www.nature.com/articles/d41586-022-02283-w (accessed on 24 October 2023).