Philosophy & Technology (2023) 36:87
https://doi.org/10.1007/s13347-023-00690-z

EDITOR LETTER

On the Brussels-Washington Consensus About the Legal Definition of Artificial Intelligence

Luciano Floridi[1,2]

Accepted: 4 December 2023 / Published online: 27 December 2023

We are so used to the political cacophony of partisan disagreements and misunderstandings that sometimes we forget to cherish evidence of progress. This is true even regarding Artificial Intelligence (AI), a topic that should attract more facts than fiction, and hence more evidence-based policies. The ethical and legal debate about *why* (here piles of science fiction mingle with serious problems) and *how* (from more competition to better innovation and protection of human rights) AI should be regulated is internationally intense, and lively on both sides of the Atlantic (Floridi, 2023). Not even the EU and the US can agree on a single text (or definition, as we shall see), let alone the rest of the world. Indeed, there are plenty of disagreements even within the EU.[1] Looking at the headlines (mass media complain a lot but are often part of the problem of disinformation), it may seem that the most one can achieve are scaremongering warnings, pious recommendations, and empty good intentions, a sort of climate change debate *déjà vu*. And yet, there has been some valuable progress. Some corners of the world are still considering how to nudge producers and users of AI to behave properly, but Brussels and Washington are moving forward in terms of legislation, while plenty of legal developments are on their way. With some hard-acquired and carefully protected optimism, one may speak of a Brussels-Washington consensus emerging. Let me clarify.

According to the Artificial Intelligence Index Report 2023 (see Fig. 1): "An AI Index analysis of the legislative records of 127 countries shows that the number of bills containing "artificial intelligence" that were passed into law grew from just 1 in 2016 to 37 in 2022. An analysis of the parliamentary records on AI in 81 countries likewise shows that mentions of AI in global legislative proceedings have increased nearly 6.5 times since 2016."[2]

[1] See for example: https://www.euractiv.com/section/artificial-intelligence/news/eus-ai-act-negotiations-hit-the-brakes-over-foundation-models/

[2] For more information see https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report-2023_CHAPTER_6-1.pdf

✉ Luciano Floridi
luciano.floridi@yale.edu

[1] Digital Ethics Center, Yale University, 85 Trumbull St., New Haven, CT 06511, USA

[2] Department of Legal Studies, University of Bologna, Via Zamboni, 27, 40126 Bologna, Italy

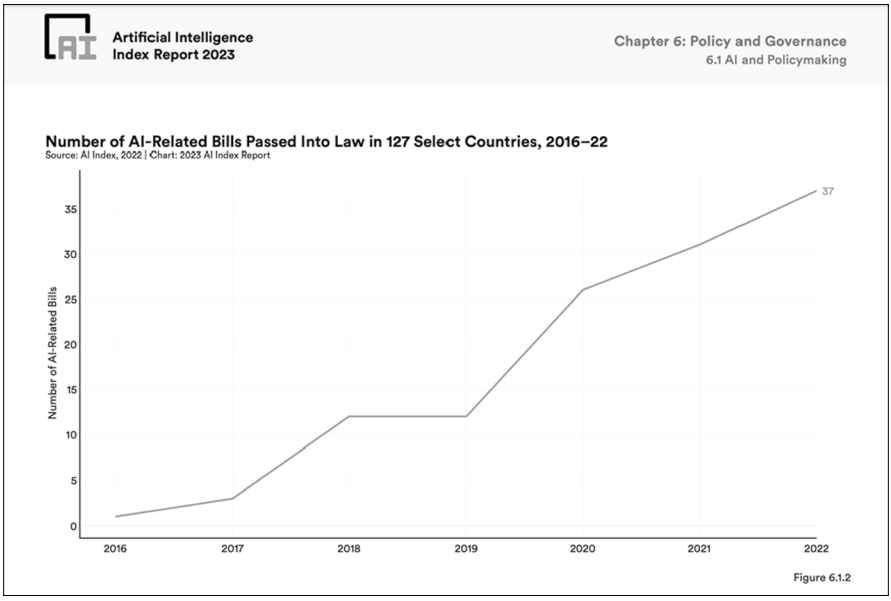

Fig. 1 AI-related bills passed into law (2016–2022), source: Artificial Intelligence Index Report 2023

Gentle invitations to do the right thing are being replaced by enforceable requests of compliance. Admittedly, it is still unclear *when*, but there is no doubt about *whether* (Floridi, 2021) the AI industry will be regulated like other sectors.

Of all these initiatives, the two most influential and well-known are, of course, the European *AI Act* and President Biden's *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* (hereafter the Executive Order). What the regulatory frameworks will be in each case, once the dust settles, is a matter of negotiation and specific implementation,[3] and plenty of speculations not worth considering. So, a comparative, in-depth analysis of the two texts would be a fascinating exercise. However, I am very happy to leave this for another occasion, or to someone else, because it is also complicated, and doing it properly could be fun, if at all, only for the readers. Instead, in this short article, I would like to focus on one crucial feature that the two documents share, which seems to have gone unnoticed. It is a feature of great significance, and evidence of the kind of slow progress that sometimes we fail to appreciate but should cherish. Both documents offer a *legal definition* of what they mean by AI, that is, not a *scientific definition* in terms of necessary and sufficient conditions, but an explicit statement about what technology they are addressing and regulating. They do not agree yet because the AI Act is still being discussed. But the Executive Order's definition agrees with the old AI Act definition (see below; it is the one proposed by the Commission), builds on

[3] See for example https://www.politico.com/news/2023/11/02/senate-ai-bill-biden-executive-order-00124893

Table 1 The definitions of AI in the original version of the AI Act and the Executive Order

AI Act, Article 3, Definitions Commission Proposal (CP)	*Executive Order, Sec. 3. Definitions*
For the purpose of this Regulation, the following definitions apply: (1) 'artificial intelligence system' (AI system) means *software* that is developed with one or more of the techniques and approaches listed in Annex I and can, *for a given set of human-defined objectives*, generate outputs such as *content*, predictions, recommendations, or decisions influencing the *environments* they interact with. (emphasis added)	(b) The term "artificial intelligence" or "AI" has the meaning set forth in 15 U.S.C. 9401(3): a *machine-based system* that can, *for a given set of human-defined objectives*, make predictions, recommendations, or decisions influencing *real or virtual environments*. Artificial intelligence systems use machine- and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action. (emphasis added)

it, and, like anyone who has learned a good lesson, does a slightly better job, yet only "slightly" because it has a significant omission (it fails to refer to "content', more on this in a moment). Meanwhile, the AI Act definition has changed twice, each time with increasing confusion, as we shall see presently. So, the quiet, yet remarkable novelty is that Brussels and Washington essentially, even if not entirely or ultimately, agree on what does and does not count as AI and hence on the scope of the regulatory frameworks they propose. This consensus is not good news for any AIpocalyptic and Singularitarian (followers of the Singularity) journalists, scientists, futurologists, intellectuals, and other clickbaiters who are chasing fame and headlines by warning that AI is some kind of Alien Intelligence that may come to dominate our lives and treat us like its pets. Existential risks can be left to Hollywood movies.

Let us have a look by starting from the original definition proposed by the EU Commission. Table 1 contains a synopsis of the two definitions[4,5] side by side.

Four aspects (emphasized in the text) of the two definitions are worth some comments.

First, strictly speaking (and "strictly" is how the law tends to speak), the AI Act in the Commission Proposal (hereafter CP) concerns only software, not hardware. The Executive Order is more inclusive, and more verbose, as the expression "machine-based system" is plausibly meant to capture both hardware (machine) and software (system). Back to the CP, appliances, gadgets, robots, or wearables, for example, will be subject to the legislation only as far as they run on software that is described in Appendix 1, which essentially covers everything: "machine learning approaches […], logic- and knowledge-based approaches […], and statistical approaches […]".[6] The problem is not Appendix 1, which is necessarily and sufficiently inclusive, but

[4] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206

[5] https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/?utm_source=link

[6] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206

the disincentive that the software-only definition may cause. For example, fridges, dishwashers, washing machines and even vehicles may need to remain on the safe side of "artificial stupidity" to avoid having to comply with the AI Act (CP version). A scenario becomes plausible in which companies start dumbing down ("de-AI-ing") or at least stop smartening up their products in order not to be subject to the AI Act. The problem of an innovation disincentive or rather *premium* – the premium terminology helps one understand that compliance overheads may be counterbalanced by the need to compete, so that, for example, AI will be included to sell more AI-powered cars than those which are not despite the extra burdens – is old in the philosophy of technology, and not new in the AI debate, since it already emerged when the now defunct (and never really fruitful) debate on a robotax developed around 2017.[7] The solution is to ensure that the AI Act applies without hurting, hence the debate about the levels of risk, and how risk is modelled (Novelli et al., 2023). But what kind of trade-off is reached, and where the threshold is placed – so that some goods do qualify as being subject to the AI Act (so that one can sell a fridge with AI "inside", for example) but do not generate disincentivising compliance-related costs because the "AI inside" is not high risk – is a debate worth of some of the subtlest philosophical minds.

The second aspect concerns the crucial phrase "for a given set of human-defined objectives". It occurs identically in both definitions,[8] which recognize the entirely and only human nature of any end, goal, or objective pursued using AI. This means that if something goes wrong, if a mistake is made, if there is any bias, if there are cases of discrimination, in short, and more abstractly, if any misuse of AI occurs, then one must "cherche l'humanité" behind the technology. The rhetoric of what AI does or wants is silly at best, and an intentional distraction at worst, meant to deflect attention away from individuals' and organisations' causal, moral, and legal responsibilities. This is the kernel of what I like to call the Brussels-Washington consensus about the nature of AI understood as a technology designed, developed, and deployed by people who ultimately are to be praised, if anything goes well, or blamed, ethically and legally, if anything goes wrong. Anything else is sci-fi, and the EU-US regulators are not taking it seriously, or at least so it seems in the Commission Proposal. Let me add two final remarks to contextualize the phrase "for a given set of human-defined objectives".

Conceptually, the phrase also occurs in textbooks about AI, but with a significantly different meaning. In the classic textbook (Russell & Norvig, 2021), for example, the phrase does not refer to how AI works – that is, how it is designed, guided, and constrained by human-oriented objectives – but to how its behaviour is evaluated externally, that is, how AI performs with respect to expectations (objectives) that are human-defined. The latter interpretation, which is not what the two documents endorse, leaves the possibility of scenarios where AI outperforms any human-defined objectives, which is understood here only as a benchmark.

[7] https://en.wikipedia.org/wiki/Robot_tax

[8] In the AI Act it has occurred since earliest versions: https://artificialintelligenceact.eu/wp-content/uploads/2022/05/AIA-COM-Proposal-21-April-21.pdf

Table 2 The definition of AI in the ISO/IEC 23894

ISO/IEC 23894
3.1.4 artificial intelligence system
AI system engineered system that generates outputs such as content, forecasts, recommendations or decisions *for a given set of human-defined objectives* (emphasis added)
Note 1 to entry: The engineered system can use various techniques and approaches related to artificial intelligence (3.1.3) to develop a model (3.1.23) to represent data, knowledge (3.1.21), processes, etc. which can be used to conduct tasks (3.1.35)
Note 2 to entry: AI systems are designed to operate with varying levels of automation (3.1.7)

Historically, "for a given set of human-defined objectives" already occurred in earlier versions of the *International Standard ISO/IEC 23894 on Information Technology — Artificial intelligence — Artificial intelligence concepts and terminology*. Table 2 shows the official version published in 2022, but the project started in 2018, and drafts were circulated and debated since then.[9]

It is plausible that the AI Act and the Executive Order ultimately owe their approach to ISO/IEC 23894, at least indirectly (see below the discussion of the National AI Initiative Act of 2020). There is more. Notice the presence of "content" in Table 2. This is the third aspect of the two definitions that is worth emphasizing. Re-read the two definitions in Table 1, and you will notice that the AI Act CP, like the ISO/IEC 23894, carefully places *content* (animations, images, music, sounds, photographs, texts, videos, voices, etc.) as the first of the kind of outputs that qualify AI. Yet the Executive Order does not even mention it. This is astonishing. Any debate about education, the job market, the entertainment industry, the future of mass media, copyright, Intellectual Property, fair use, fake news or deepfakes, phishing, disinformation, political debates, manipulation of public opinion, propaganda, and so forth, requires an essential acknowledgement of the critical role played by the automation or "AIfication" of content production. Indeed, this is one of the most challenging aspects of the AI revolution. Somewhat incoherently, given the emphasis on home security, for example, the Executive Order does not include this crucial aspect, nor does it have any safety "such as" clause that we see present in the AI Act CP and the ISO/IEC 23894. Strictly speaking (once again), according to the Executive Order, AI concerns only "predictions, recommendations, or decisions". In this sense, a lot of generative AI, for example, is not covered. Why such omission? A plausible explanation, barring conspiracy theories, lobbying strategies, and conceptual mistakes (and this is a lot of barring, I know), is linked to the fact that the Executive Order explicitly adopts the whole definition, including the phrase "for a given set of human-defined objectives", from the National AI Initiative Act of 2020 (NAIIA), which became law on January 1, 2021 (see Table 3).

As you can see, the only (irrelevant) difference is that the NAIIA§9401 is more structured and less discursive than the Executive Order. Now, the NAIIA§9401

[9] https://www.iso.org/standard/74296.html

Table 3 The definition of AI in the NAIIA

NAIIA§9401. Definitions
(3) Artificial intelligence
The term "artificial intelligence" means a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. Artificial intelligence systems use machine and human-based inputs to—
(A) perceive real and virtual environments;
(B) abstract such perceptions into models through analysis in an automated manner; and
(C) use model inference to formulate options for information or action

is, understandably and justifiably, much more defence- and security-oriented than the ISO/IEC 23894 or the AI Act, and this might have contributed to creating such a blind spot in the Executive Order about the crucial role of AI in content generation. Oversimplifying, predictions, recommendations, and decisions are all that matter in situations of competition, security, and conflict, not content, which, therefore, drops off the radar (pun intended). Of course, there may be many other reasons, but whatever the explanation, this is a mistake that should be rectified in the future.

We come to the fourth and last feature of the definitions that I wish to discuss here: their reference to environments. In this case, the Executive Order is more careful and explicit, once again following verbatim NAIIA§9401, which refers to "real or virtual environments". Yet both definitions agree that AI influences the spaces we inhabit, no matter whether analogue or digital. I will not comment at length on the choice of words – as if the virtual were not real (more on this later) – or the granularity of the statement. Perhaps, it is helpful to make sure that virtual environments are covered explicitly (again, more on this presently) and, when security and defence contexts are the primary concern, being clear that what you are talking about also includes cyberspace and, hence, cyberwar, is vital. Furthermore, I was told that the EU introduced the distinction "physical or virtual" (see Table 5, EP Mandate definition) to have a backdoor for a potential extension of the AI Act to the metaverse. Whatever the reasons behind the distinction, what matters is that the two documents should (and to a reasonable extent do, if one reads the whole texts) take their own definitions seriously. AI is a force for positive and negative changes when it comes to *all* environments, and it should be regulated accordingly. Any "human-centric"-only rhetoric smacks of old-fashioned modernity. Not because it is wrong, but because it is not right enough. AI must be at the service of not only all humanity but also the whole environment – any environment – or we risk forgetting not just its social costs but its environmental impact as well. AI can be a great force for good, but it must be used as such, not wasted to fuel more consumerism while further damaging the environment. So, the *human-defined objectives* mentioned by both definitions should not be merely consumer- and citizen-oriented. The objectives must be *socially preferable* and *at least* ecologically sustainable.

End of the four considerations. The time has come to summarise the *initial* Brussels-Washington consensus about what counts as AI for legal purposes. Both sides

Table 4 The definitions of AI in the OECD AI Principles 2018 and 2023

OECD AI Principles 2018 (OECD18)	OECD AI Principles 2023 (OECD23)
An AI system is a machine-based system that can, *for a given set of human-defined objectives*, make predictions, recommendations or decisions influencing real or virtual environments. It does so by using machine and/or human-based inputs to: i) perceive real and/or virtual environments; ii) abstract such perceptions into models through analysis in an automated manner (e.g. with ML, or manually); and iii) use model inference to formulate options for information or action. AI systems are designed to operate with varying levels of autonomy. (emphasis added)	An AI system is a machine-based system that, for *explicit or implicit objectives*, infers, from the input it receives, how to generate outputs such as predictions, *content*, recommendations, or decisions that can influence *physical or virtual environments*. Different AI systems vary in their levels of autonomy and adaptiveness after deployment. (emphasis added)

of the Atlantic agree that AI is an artefact (software or machine-based system) that can, *for a given set of human-defined objectives*, generate outputs such as predictions, recommendations, or decisions, influencing any kind of environment. They should both stress the importance of content.

The next question is whether this consensus is going to be universally accepted as a starting point. Don't hold your breath, for the disappointing answer is, at best, not yet.

The OECD (see Table 4) recently published its revised definition of AI (OECD23) and, surprisingly, dropped the clause "*human-defined"* that correctly qualified its previous definition (OECD18). It is a mistake, but I hope, to use a sports metaphor, not a forced one (lobbying).

At the same time, OECD23 improves on OECD18 in two respects. It speaks of "physical or virtual" (I suggest reading the "or" inclusively, as and/or, like the Latin/logic *vel*) environments, which we have seen is better than "real and virtual". And it does include a significant reference to "content", even if its occurrence in the text, after "predictions" and before "recommendations", is odd and looks like an afterthought. Unfortunately, it now fudges the point about "objectives", adding a distinction that classically makes no difference: "explicit or implicit". One is left wondering what this may mean (a polite, British way of saying that it probably makes no sense). In this fundamental respect, the previous definition in OECD18 was much preferable. You can still find it in other documents by the OECD, such as *Artificial Intelligence in Society* (2019).[10] For the absence of the phrase "for a given set of human-defined objectives" opens the door to potential sci-fi scenarios, with AI systems having a mind of their own and selfish objectives as well. All this is problematic because there is now a lack of coherence between the Brussel-Washington initial consensus and the OECD regarding what counts as AI for ethical and legal purposes. And this incoherence matters because definitions are not just wonderful entertainment for philosophers, but the places where clear and exact boundaries are

[10] https://www.oecd-ilibrary.org/sites/8b303b6f-en/index.html?itemId=/content/component/8b303b6f-en

Table 5 The definition of AI in the three versions of the AI Act

	Commission Proposal	EP Mandate	Council Mandate	Draft Agreement
127	For the purpose of this Regulation, the following definitions apply:	For the purpose of this Regulation, the following definitions apply:	For the purpose of this Regulation, the following definitions apply:	
128	(1) 'artificial intelligence system' (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with;	(1) 'artificial intelligence system' (AI system) means ~~software~~**a machine-based system** that is ~~developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined~~**designed to operate with varying levels of autonomy and that can, for explicit or implicit** objectives, generate outputs such as ~~content,~~ predictions, recommendations, or decisions**, that influence physical or virtual environments** ~~influencing the environments they interact with~~;	(1) 'artificial intelligence system' (AI system) means ~~software~~**a system** that is ~~developed with one or more of the techniques and approaches listed in Annex I and can, for~~**designed to operate with elements of autonomy and that, based on machine and/or human-provided data and inputs, infers how to achieve** a given set of ~~human-defined objectives, generate~~**objectives using machine learning and/or logic- and knowledge based approaches, and produces system-generated** outputs such as content **(generative AI systems)**, predictions, recommendations, or decisions, influencing the environments ~~they interact with~~**with which the AI system interacts**;	

precisely drawn for the scope, applicability, and enforcement of regulations and recommendations. Unfortunately, things got worse recently. The definition in the Commission's AI Act proposal has gone through two revisions, each of which has made mincemeat of a good starting point. Table 5 is messy enough to convey the point even visually. [11]

The EP Mandate (EPM) version rightly drops "software" in favour of "machine-based system", which the Council Mandate (CM) correctly reduces to "system". So far, so good. Both drop the reference to Appendix I, and this is also a simplification that may be welcome. But now EPM introduces "explicit or implicit" objectives that we saw are unclear, to say the least. Luckily, the CM version drops this change and simply refers to "objectives". This is good. Unfortunately, CM indicates that "a system … produces system-generated outputs". This is unassailable – what else could a system generate? – but also useless. More nonsense is added in terms of "elements of autonomy" (too vague to be informative), and "infers how to achieve" (this is poorly written, confusing, and conceptually wrong). The good news is that the "environments" are no longer specified as virtual or not, which is more in line with the digital revolution and a twenty-first-century culture of "onlife" experience that no longer distinguishes between online and offline, analogue and digital environments. And "content" is duly kept in its significant position. So, there is hope for the final agreement and the Brussels-Washington consensus to prevail. And this leads me to the last point I wish to make by conclusion.

The temptation to synthesize the previous definitions into one is too strong, and I shall not resist it. So Table 6 offers a suggestion. I have kept the AI Act's structure, style and level of abstraction as proposed by the Commission and the Executive Act. Still, the definitions we have seen above do not refer to *learning*, which

[11] https://artificialintelligenceact.eu/wp-content/uploads/2023/08/AI-Mandates-20-June-2023.pdf

Table 6 A revised definition of AI

Artificial Intelligence (AI) refers to an engineered system that can, for a given set of human-defined objectives, generate outputs – such as content, predictions, recommendations, or decisions – learn from data, improve its own behaviour, and influence people and environments

is a fundamental feature that discriminates new forms of AI from other artefacts, that is, its ability to be trained on past data and improve its performance based on its own output, to put it simply. And all of them seem to forget that the influence exercised by AI is not just on any environment but also on people. So, I have taken the liberty of adding the two specifications. Relying on the same approach shared by the Brussels-Washington consensus, the outcome seems to be a further improvement that avoids the problems highlighted above:

As I remarked above, this is not a scientific definition but a legal one that could work to set the scope of the AI Act (also in connection with the other pieces of legislation that make up the EU regulatory architecture about digital technologies). Who knows, it may even help in reaching a final agreement and a Brussels-Washington consensus, at least about what the law is discussing and regulating. But I offer it with no illusion about its potential success.

Acknowledgements Many thanks to Emmie Hine, Jessica Morley, Claudio Novelli, and Renée Sirbu for our conversations and their insightful comments on previous versions of this article. They improved it significantly, much to my recurrent (I wrote this before) embarrassment and relief.

References

Floridi, L. (2021). The end of an era: from self-regulation to hard law for the digital industry. *Philosophy & Technology, 34*(4), 619–622.

Floridi, L. (2023). *The ethics of artificial intelligence - principles, challenges, and opportunities*. Oxford University Press.

Novelli, Claudio, Federico Casolari, Antonino Rotolo, Mariarosaria Taddeo, and Luciano Floridi. 2023. "Taking AI risks seriously: a new assessment model for the AI Act." *AI & Society*, online.

Russell, Stuart J., and Peter Norvig. 2021. *Artificial intelligence: A Modern Approach*. Fourth edition ed. Hoboken: Pearson.