General AI and Transparency

Two critical points of the proposed EU AI Act

Federico L.G. Faroldi

i-lex

i-lex. Scienze Giuridiche, Scienze Cognitive e Intelligenza Artificiale
Rivista semestrale on-line: www.i-lex.it
Dicembre 2021
Fascicolo 2
ISSN 1825-1927

i-lex. Scienze Giuridiche, Scienze Cognitive e Intelligenza Artificiale
Rivista semestrale on-line: www.i-lex.it

GENERAL AI AND TRANSPARENCY
TWO CRITICAL POINTS OF THE EU AI ACT

FEDERICO L.G. FAROLDI*

Abstract. This paper puts forward two broad criticisms of the proposed EU AI Act. First, the proposal does not address the future developments of general intelligent systems and it is ill-equipped to deal with more general existing systems like GPT-3. Second, one suggestion found in the literature to meet the first criticism, i.e. to increase transparency, fails, and not only because the proposal itself fails to specify the transparency standards required, but because it is problematic to spell transparency out in a concrete, usable way.

Keywords: *General AI; Transparency; EU AI Act*

"*Once the machine thinking method had started, it would not take long to outstrip our feeble powers. At some stage therefore we should have to expect the machines to take control.*"
Alan Turing, 1951

1 Introduction

In recent years, with the extremely rapid progress in the field of artificial intelligence, the possibility of *singularity* has come to the attention of both specialists and the general public: machines that, having reached the stage of artificial general intelligence (AGI), or superintelligence, would give rise to an "explosive" moment that could

* Institute for Computer Science, University of Bern and Swiss National Science Foundation (CH); Center for Logic and Philosophy of Science, University of Ghent and Research Foundation Flanders, FWO (BE). federico.faroldi@ugent.be

radically and irreversibly change the planet and human civilization (cf. Kurzweil 2005, Bostrom 2014, Russell 2019).[1]

Already some time ago, Good introduced the ideas of superintelligence and singularity in the following way:

> *"Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind. Thus, the first ultra-intelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. … It is more probable than not that, within the twentieth century, an ultraintelligent machine will be built and that it will be the last invention that man need make."*

This raises many issues, but we focus, for the purposes of this paper, on the so-called *control problem*: how can we make sure that general intelligent, or so-called superintelligent, agents do not take control over us and the world?[2]

Section 2 argues that the proposed EU AI Act fails to address AI systems that do not have a concrete intended application. Therefore, (i) it fails to properly take into account the control problem due to potential general AI; (ii) even more modestly, fails to account for existing non-specific systems, like GPT-3.

Section 3 argues that one suggestion found in the literature to meet the first criticism, i.e., to increase transparency, fails, and not only because the proposal itself fails to specify the transparency standards required, but because it is problematic to spell transparency out in a concrete, usable way.

---

[1] In Faroldi 2020 and 2021a, I argued that such agents should be included in our normative practices, e.g. responsibility attribution.

[2] Assuming we want that, of course. It is (at least theoretically) imaginable, in fact, that humanity could want to defer and give control to a superintelligence.

2 General AI and the control problem in the EU AI Act.

There are several solutions that have been proposed to deal with the control problem. The most popular is alignment: superintelligent AI must be programmed to align with human values.[3]

The proposed EU AI Act does little to nothing to address the control problem, as I will argue in a moment.

This could be because the Commission does not want to overregulate a booming sector, and hinder future technological development.[4]

A more balanced approach could go in many ways, but it has to start from recognizing the problem in the first place. AI risk researchers are mostly in agreement that alignment has to be built in before a singularity is reached,[5] i.e. before it is too late for humans to gain back control.

However, the Commission writes that

> *"[t]he proposal sets a robust and flexible legal framework. […] it is comprehensive and future-proof in its fundamental regulatory choices (AI Act, Explanatory Memorandum)."*

This thought is probably based on the guiding idea that

> *"legal intervention is tailored to those concrete situations where there is a justified cause for concern*

[3] Cf. Russell 2019, Christian 2020. As is easy to imagine, this proposed solution raises perhaps more problems than it promises to solve, from the problem of identifying and formulating human values, to the problem of instilling them in the AI systems in question. A second solution proposes to limit the capabilities of a AGI by isolating it from the outside world. A third solution proposes instead to increase the capabilities of humans in various ways to be on par with superintelligent systems. Cf. Ngo 2020.

[4] "[the] proposal presents a balanced and proportionate horizontal regulatory approach to AI that is limited to the minimum necessary requirements to address the risks and problems linked to AI, without unduly constraining or hindering technological development or otherwise disproportionately increasing the cost of placing AI solutions on the market (AI Act, Explanatory Memorandum)".

[5] Cf Christian 2020, Ngo 2020.

> *or where such concern can reasonably be anticipated in the near future (ib.).”*

Given the current technological development, general AI cannot be anticipated to happen in the near future, i.e. within the next 5-10 years. But come what may, it does not matter, it seems, because:

> *“At the same time, the legal framework includes flexible mechanisms that enable it to be dynamically adapted as the technology evolves and new concerning situations emerge (ib).”*

So the proposed regulation will run after and follow, rather than shape, technological development, which seems to be a grave mistake when it comes to general AI and the control problem. The shared idea is that, when the singularity is reached, it will be too late to do anything about control.[6]

An AI system is considered high-risk if it is intended to be used in a specific way,[7] either listed in Art 6 or in Annex III.[8] Notable examples of high-risk systems include:

- AI systems intended to be used for the ‘real-time’ and ‘post’ remote biometric identification of natural persons;
- AI systems intended to be used for the purpose of determining access or assigning natural persons to educational and vocational training institutions;
- AI systems intended to be used for the purpose of assessing students in educational and vocational training institutions;
- AI systems intended to be used for recruitment or selection of natural persons; AI intended to be used for making decisions

[6] As an example, cf. Ord (2020) and Ngo (2020).

[7] “The classification of an AI system as high-risk is based on the intended purpose of the AI system, in line with existing product safety legislation. Therefore, the classification as high-risk does not only depend on the function performed by the AI system, but also on the specific purpose and modalities for which that system is used (EU AI Act).”

[8] “AI systems intended to be used as safety component of products that are subject to third party ex-ante conformity assessment; other stand-alone AI systems with mainly fundamental rights implications that are explicitly listed in Annex III [which] contains a limited number of AI systems whose risks have already materialised or are likely to materialise in the near future (EU AI Act).”

on promotion and termination of work-related contractual relationships;
- AI systems intended to be used by public authorities or on behalf of public authorities to evaluate the eligibility of natural persons for public assistance benefits and services; AI systems intended to be used to evaluate the creditworthiness of natural persons or establish their credit score;
- AI systems intended to be used to dispatch emergency first response services;
- AI systems intended to be used by law enforcement authorities for making individual risk assessments of natural persons in order to assess the risk of a natural person for crimes or the risk for potential victims of criminal offences;
- those with applications in migration, asylum and border control management;
- those with applications in the administration of justice and democratic processes.

We see that what matters in classifying an AI system as high-risk is its intended application in a field that is considered particularly relevant or worth extra care, and not at all the intrinsic characteristics of the AI system in question.

For instance, this proposed regulation could result in a low-level AI system used to categorize job application files in an alphabetical order as high-risk, and a superintelligent system with no concrete application or specific intended use, which might go on and influence millions of users in unforeseen, uncontrolled ways, as not high-risk.

This is problematic, because it seems to exclude currently existing systems that do not have a specific, concrete intended application, but are fairly general (without being instances of a general AI), such as GPT-3 and others.[9]

GPT-3 (short for Generative Pre-trained Transformer 3), created by Open AI, is a deep learning autoregressive language model that

[9] It is worth reminding here the definition of 'intended purpose' of Art. 3: "'intended purpose' means the use for which an AI system is intended by the provider, including the specific context and conditions of use, as specified in the information supplied by the provider in the instructions for use, promotional or sales materials and statements, as well as in the technical documentation".

generates human-like text. Originally released in mid-2020, GPT-3 was trained on almost half a trillion tokens, is capable of coding in CSS, JSX, Python, and has been exclusively licensed by Microsoft.

David Chalmers, a prominent philosopher, described it as "one of the most interesting and important AI systems ever produced",[10] but the MIT Technology Review says that its "comprehension of the world is often seriously off, which means you can never really trust what it says".[11] Other critics described it as unsafe, citing sexist, racist, and other biases. In one health-care simulation, GPT-3 advised a mental health patient to commit suicide.[12]

Regardless of this, GPT-3 is a system without an intended specific use (in the sense specified above): instead, it can be used in many, ways, unforeseen by its creators. Thus, it seems it cannot be classified as high-risk according to the EU proposal.

But suppose that it is used with a particular application in mind, which will not result in having it classified as a high-risk system.

This does not change the open nature of the system, which has now slipped regulatory measures.

That the Commission does not foresee AGI systems is also revealed by the provisions on human oversight, which require that a human:

> *"(d) be able to decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the high-risk AI system;*
>
> *(e) be able to intervene on the operation of the high-risk AI system or interrupt the system through a "stop" button or a similar procedure. (Art 14(4)).*"

These provisions do not take into account that an AGI could presumably (i) take control of its functioning by disabling and preventing any way to be switched off; (ii) make sure that human controllers are either fooled or convinced that everything is alright.

---

[10] https://dailynous.com/2020/07/30/philosophers-gpt-3/#chalmers

[11] https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion

[12] https://artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/

When it comes to (i), consider that even a non-malignant AGI, if built as many of the current AI systems, will have to maximize some reward function. Being switched off is definitely one way not to maximize the reward function, and therefore something to be avoided and prevented. Doing this with the full (supposed) power of an AGI is imaginably easy. When it comes to (ii), even without having reached a state of general intelligence, current algorithms are able to influence large numbers of people (e.g. through making fake news viral).

To these criticisms of the EU proposed regulations (and in particular of Art 14 in the present context), one can object that they are based on mere projections, or extrapolations, based on the capabilities of current AI systems, with no guarantee of what might or might not happen in the future. While this is true, one should also consider that the moment an AGI is created and is allowed to take control, there is a very high chance that it will not be possible to go back: once a singularity is reached, there won't be the possibility to switch off the systems in question. While there are many disanalogies, it is plausible to liken the singularity with nuclear annihilation: a point of no return. Seen in this light, it makes sense to legislate already now to safeguard and prevent this possibility, even if remote, rather than complain later, when too late.

But how to do that?

One proposal to address a similar concern suggests to widen Title IV, on transparency obligations, to apply across all AI applications regardless of specific purpose.[13] One way this could be achieved, according to the authors, is by requiring a "complete risk assessment of all an AI system's intended uses (and foreseeable misuses (ib.))".

However, this is also problematic: transparency is a notoriously complicated requirement, and the EU AI act proposal does not do anything to improve the situation. Let's see this aspect in turn.

3 Transparency in the EU AI Act.

[13] FLI position paper on the EU AI act, 2021 https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665546_en

The EU proposed AI Act imposes some transparency requirements to certain systems:

> *"Transparency obligations will apply for systems that (i) interact with humans, (ii) are used to detect emotions or determine association with (social) categories based on biometric data, or (iii) generate or manipulate content ('deep fakes')."*

While transparency is not at all a clear concept in the literature, it has to do with knowing how a system works, at least at a high level, to know how decisions are taken, for instance.[14] This seems also to be taken up in some existing legal sources.[15]

As a first point to notice, transparency obligations are not imposed on the same systems that are classified as high risk (see above).[16] Therefore, we can already exclude that more transparency is sufficient, *per se*, to solve the problem put forward in the previous section. If it were, in fact, it would be enough to require transparency to mitigate the potential problems of high-risk systems, without the need for further measures.

However, there's a further difficulty. What the EU Commission means with 'transparency' does not really seem aligned with the rest of the literature. The EU Commission seems to have in mind that transparency consists in the right to be told that one is interacting with an AI system:

> *"When persons interact with an AI system or their emotions or characteristics are recognised through automated means, people must be informed of that circumstance. If an AI system is used to generate or manipulate image, audio or video content that*

[14] This characterization is obviously very simple-minded.

[15] In Faroldi 2021b, I put forward critical remarks when it comes to transparency requirements, also with regard to the GDPR and a recent order of the Italian Corte di Cassazione. The GDPR prohibits automated decisions that have consequences for individuals, and establishes a right to meaningful information about the logic employed in these procedures. A recent Corte di Cassazione order (n. 14381/21, May 25, 2021) places emphasis on access to the elements an algorithm uses in a decision and the executive scheme in which the algorithm in question expresses itself.

[16] At least prima facie. It is possible that these two different fields of application end up covering exactly the same systems, although this is extremely unlikely.

> *appreciably resembles authentic content, there should be an obligation to disclose that the content is generated through automated means, subject to exceptions for legitimate purposes (law enforcement, freedom of expression). This allows persons to make informed choices or step back from a given situation."*

Fortunately, the EU AI Act also requires some form of control and "transparency", again for high-risk systems, in Art. 13, which is glossed in the following way:

> *"To address the opacity that may make certain AI systems incomprehensible to or too complex for natural persons, a certain degree of transparency should be required for high-risk AI systems. Users should be able to interpret the system output and use it appropriately. High-risk AI systems should therefore be accompanied by relevant documentation and instructions of use and include concise and clear information, including in relation to possible risks to fundamental rights and discrimination, where appropriate (EU AI Act, Preamble, 47)"*

The fact that Art. 14 requires human oversight, in such a way that the human in question:

> *"fully understand the capacities and limitations of the high-risk AI system and be able to duly monitor its operation, so that signs of anomalies, dysfunctions and unexpected performance can be detected and addressed as soon as possible (Art 14 (4)(a))"*

tells us that the certain degree of transparency (in the non-opacity sense) required for users is lower than that required for the controllers.

But is the transparency standard required for controllers even possible? Before answering this important question, it is necessary to check what it is meant with 'transparency' and its antonym, 'opacity', in the literature.

One can distinguish at least three senses of opacity, i.e. not transparency.

There is *intentional* opacity when the algorithm is not voluntarily made public in order to maintain a competitive or economic advantage. This type of opacity could be "solved" by simply publishing the algorithms in question - except for subsequent problems with competitiveness, etc.

There is *cognitive* opacity when the algorithm, while transparent, simply cannot be interpreted by most users due to their ignorance. In this case, moving to an open system, the opacity is not resolved ipso facto: an intermediate category of interpreters, explanators, or massive investments on the education of the population are needed.

There is *intrinsic or essential* opacity when it is the algorithms themselves, regardless of programmers and publication, that make decisions in a way that cannot be explained or interpreted by a human being, also regardless of cognitive capacity. This is the case with some machine learning techniques, and deep neural networks.[17]

It is of course this latter sense that is relevant, and it is often investigated under the label of 'interpretability'.

Lipton (2016) gives a systematic account of the relevant literature, and argues that the interpretability of models falls into two broad categories: that of transparency *stricto sensu* and that of *post hoc* explanations. In the first, what matters is understanding the mechanism by which the model works. In the second, what matters is extracting information from the models to clarify what exactly they learned.

Krishnan (2020) however, argues more radically that the concept of 'interpretability', is vague: thus**,** it is very hard to know whether a given technical solution is acceptable or not; and in any case often that of 'interpretability' is a means to achieve other ends (such as non-discrimination or justification). We should therefore not require algorithms to be interpretable, but focus the discussion on the true ends we want to achieve.

At a very high level, transparency should entail that we are in a position to know why a certain algorithm produced a certain output. The first 'why' we can distinguish is that of an explanation, which is of course a whole problem in itself. At a minimum, it should be

[17] For a similar tripartition, see Burrell (2016). There are also algorithms that are totally transparent because of their architecture, which makes them self-explanatory, such as linear regression, decision trees or rule-based systems.

informative about the causes and the processes that lead to that output, given that input, at an appropriate level of granularity.

The second 'why' we can distinguish is that of a justification, e.g. about the reasons in favor (or against) that particular output or set-up.

There is a definite sense in which this second 'why', that of justification, is more normative than the first.

The confusion between these two "why's" is already apparent in some technical proposals of explainability in machine learning.[18]

Given the multiple realizability and vagueness, it is unclear how these notions of transparency found in the technical literature can harmonize with what is or will be required in artificial intelligence legislation.[19]

4 Conclusion

In this paper I argued that there are (at least) two problematic points of the proposed EU AI Act: first, it is ill-equipped to deal with more general, already existing systems like GPT-3 and it is even in worse position to deal with possible strong or general artificial intelligent systems; second, I argued that one suggestion found in the literature, i.e. to increase transparency, fails, and not only because the proposal itself fails to specify the transparency standards required, but because it is problematic to lay them down in a concrete, usable way.

One recommendation that emerges from this short paper adheres to the principle that, when it comes to strong AI, the EU AI Act should lead and shape technological development, rather than just follow it, because once (if) a singularity is reached, it will be too late. The recommendation is to take into account already *now* the possibility of strong AI in the *future*. Much more work is needed in setting out what is to be done concretely, which is an active field of research in AI safety, but awareness and risk identification is a first step that can and should be already embedded in the EU AI Act.

[18] Cf. e.g. Juozapaitis et al (s.d.)., which I discuss at length in Faroldi 2021b.

[19] A couple of recent high-level approaches suggest to give more prominence to symbolic techniques in AI, especially when it concerns certain values (like transparency) in a legal context. Cf Billi (2021) and Faroldi (2021b).

Bibliography

Billi, M., A Symbolic Approach For Ensuring Fairness in AI. *i-lex* 14,1 (2021).

Burrell, J. 'How the machine 'thinks': Understanding opacity in machine-learning algorithms' (2016), *Big Data and Society* DOI: 10.1177/2053951715622512

Christian, B., *The Alignment Problem*, Norton, 2020.

Faroldi, F.L.G, *Responsabilità e ragione.* Satura editore, 2020.

Faroldi, F.L.G (2021a), Considerazioni filosofiche sullo statuto normativo di agenti artificiali superintelligenti, Revista Iustitia, 9, 2021.

Faroldi, F.L.G (2021b), Trasparenza dell'algoritmo e deep learning. Note logiche a margine della proposta di regolamento sull'Intelligenza Artificiale (Artificial Intelligence Act) della Commissione Europea e di un'ordinanza della Corte di Cassazione, Revista Iustitia, 10, 2021.

Good, I. J. ,1965, "Speculations Concerning the First Ultraintelligent Machine", in *Advances in Computers*, vol 6, Franz L. Alt and Morris Rubinoff, eds, pp31-88, 1965, Academic Press.

Goodman, B. and Flaxman, S., 'European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"' (2016) ICML Workshop on Human Interpretability in Machine Learning, arXiv:1606.08813 (v3); (2017) 38 AI Magazine 50.

Juozapaitis, Z. et al., Explainable reinforcement learning via reward decomposition. URL: http://web.engr.oregonstate.edu/~afern/papers/reward_decomposition__workshop_fin al.pdf.

Krishnan, M., Against Interpretability: a Critical Examination of the Interpretability Problem in Machine Learning, *Philosophy & Technology*, 33:487–502, 2020.

Lipton, Z. C., "The Mythos of Model Interpretability," https://arxiv.org/pdf/1606.03490.pdf, Jun. 2016.

Ngo, R., "AGI Safety from first principles", ms, 2020.

Numerico, T., *Big data e algoritmi.* Carocci, Roma, 2021.

Ord, T., *The Precipice*. Bloomsbury, 2020.

Russell, S., *Human Compatible*, Viking, 2019.

Russell, S. and Norvig, P., Artificial Intelligence: A Modern Approach, 4th Edition, Pearson, 2020.

Swapnil Nitin Shah, Addressing the interpretability problem for deep learning using many valued quantum logic, https://arxiv.org/pdf/2007.01819v1.pdf