observatorio cultura y comunicación
FUNDACIÓN alternativas

UAM Universidad Autónoma de Madrid

Commentaries on the proposal of the European Commission for harmonized rules on Artificial Intelligence

Lucía Ortiz de Zárate Alcarazo

Febrero 2022

FUNDACIÓN alternativas

ÍNDICE

1. On the definition of Artificial Intelligence

Artificial Intelligence (AI) is not only a broad field of study in which different scientific disciplines (engineering, linguistics, computer science, philosophy, psychology, etc.) come to term, but also a complex concept with diffuse limits and important implications. Heterogeneity behind AI can partially explain the difficulties to settle a uniquely shared definition for these set of emergent technologies. The role of the European Commission (EC) as a key international actor in the realm of "AI Governance" can help boost a common definition that can gather wide support and help unify the field of AI. This challenge comes with the recognition from the EC that a proper definition of AI needs to be "technology neutral" and "future proof" and the necessity of finding a definition that is "based on the key functional characteristics of the software" (EC, 2021, p.18). Additionally, the EC needs to overcome the demands of different stakeholders who request "a narrow, clear and precise definition for AI" (EC, 2021a, p.8).

In order to meet some of the challenges previously mentioned, in Article 3 (1) of its proposal for harmonized rules on AI, the EC has proposed the following **definition of AI** for its new regulatory framework (EC, 2021a, p.39):

"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".

In **Annex I** the techniques and approaches listed are:

a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;
b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;
c) Statistical approaches, Bayesian estimation, search and optimization methods.

Although the definition proposed by the EC is adequate for the goals and objectives set by the regulation proposal, we believe that there is still room for improvement in the following ways:

1) The socio-technical nature of AI needs to be more explicitly recognized in the definition set by the CE in order to accommodate the political, cultural and social limitations that the concept and the reality of AI entail. In this sense, we find that some definitions previously considered by the EC, such as the one proposed in 2019 by

the High-Level Expert Group (HLEG) on AI, reflect more clearly the human dimension of AI, and are a better fit to achieve the intended goal of technological neutrality.

2) The explicit inclusion of the methodologies and techniques used to develop AI systems can cause more harm than good. Although listing AI techniques can be useful to narrow down the number of AI technologies, it can become a hinderer to meet the goal of proposing a definition that is "future proof". The fast evolution of AI research and, thus, AI techniques makes more suitable to skip them from the central definition of AI. In this sense, it could be a better idea to focus on the functionalities and consequences of AI, that can overcome more easily fast-changing times, rather than on the techniques used
3) In the same vein, the list of AI techniques and methodologies could be included as an Appendix or a complement to the central definition, but never as a central element. This formulation has been suggested by the EC in previous documents (HLEG, 2019) and in the recent proposal by United Nations (UN) on the ethics of AI (United Nations, 2021).
4) As the definition of AI has a binding role for the proper compliance of the legislation, it would also be convenient to remark that it will be necessary to review the definition periodically, so that it does not become obsolete and, therefore, it does not generate legal gaps that can lead to violations of European citizens' rights.
5) The definition of AI needs to include the (possible) hardware component of these set of emergent technologies as many ethical and legal challenges emerge from this feature alone. This will be explained in more detail in section 5 of this document.

In line with these commentaries, we propose the following definition for AI. This definition is almost identical to the one proposed by the HLEG (2019), but some changes have been included.

"Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans, to operate with some aspects of autonomy, that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems may include several methods and techniques listed in Annex I[1]"

[1] Annex I would maintain the list of techniques included in the original proposal for AI regulation of the EC.

2. On Prohibited AI practices

2.1. Facial recognition

After careful consideration we think that facial recognition applications and techniques used for real-time identification should be listed as prohibited practices within the European Union (EU). Although the definition, included in Article 3, for biometric data mentions the collection of facial images, it would be convenient to address this technique separately from others due to its particular characteristics and intended use. We propose that facial recognition can only be used under the same conditions consider for biometric data which appear in Article 5. In that scenario, facial recognition would be listed in Annex III along biometric identification and categorization of natural persons.

As in the case of the other prohibited AI practices, facial recognition used in unjustified scenarios directly violates the rights and freedoms of EU citizens at least in the following ways:

1) Privacy violations that emerge from the compilation, analysis, and use of personal information.
2) Risk of massive surveillance.
3) Restrictions on human autonomy that come from constant surveillance.
4) Bias: Different studies have shown that data used to train facial recognition algorithms is often biased in such a way that harms most vulnerable groups (Lohr, 2018). Due to an overrepresentation of white men in the data used to train the algorithms, they tend to identify people from these groups with much more precision and accuracy than people who hold different physical characteristics. As it is the case for other AI systems, gender bias is present in facial recognition algorithms that usually have more difficulties identifying white women, black women (and other racialized women) than men. In the same vein, it is also important to pay especially attention to the case of transsexual women or people who do not belong to groups with hegemonic representation (men wearing makeup, women with facial hair, etc.). These people, who are common targets of hate crimes and are in great need for protection, could find themselves frequently in situation where facial recognition is legal, but paradoxically are the ones that are the less likely to be correctly identified by AI systems due to their low presence in databases.

Thus, apart from being illegal, bias in facial recognition systems is harmful for two practical reasons:

a) If it was the case that a person being sought for posing an imminent danger to other people (potential criminal) is not a white man, his/her identification would be more difficult and, therefore, there would be less possibilities for capturing him/her. This could occur if no matches were found or in the case of a false identification (that would affect innocent people, usually the vulnerable ones).
b) If facial recognition is used to look for a missing person or potential victim of a crime (another of the assumptions contemplated by the regulation), it will be more difficult to find a person who is not a white man. This is especially problematic since women, black people, transgender people, and other discriminated groups are the most frequent victims of crimes.

2.2. Autonomous weapons

We proposed that the used of autonomous weapons is listed as a prohibited practice and, thus, included in Article 5, Title II, of the final AI regulation proposal. It is surprising that this practice is not included, neither as prohibited nor as high-risk, in the regulation proposal. As it is the case for other AI practices, we do not find any exception that could make the use of autonomous weapons legal due to the harms it could cause to large portions of population.

3. ON HIGH-RISK AI SYSTEMS

High-risk AI systems are listed in Chapter 1 of Tittle III, and Annex III. The proposal stablishes that any AI system that accomplishes the two following criteria must be considered as high-risk:

1) AI systems intended to be used as safety component of products that are subject to third party ex-ante conformity assessment;
2) Other stand-alone AI systems with mainly fundamental rights implications that are explicitly listed in Annex III

Although Annex III presents an adequate list of potential harmful practices of AI for citizens, we find that, at least two more applications should be included in the list in order to protect UE citizens' fundamental rights: AI systems used for healthcare and fully autonomous vehicles (from now on referred as autonomous vehicles).

3.1. Health

Health is considered one of the most promising areas for AI application (EC, 2021b). However, experts have been alerting for several years about the ethical problems involve in the use of this kind of systems in health care (Bartoletti, 2019). Among the different areas in which AI can be used, health is probably one of the most sensitive due to its centrality in everyone's lives. The nuclear role that health plays in our lives is mainly due to the fact that all of us, and our loved ones, are going to be affected by it on several occasions through our live times. Additionally, these are moments of great vulnerability (physical and psychological) for patients and their companions. Thus, there is a general demand for excellency in healthcare and health professionals.

For these reasons, besides the potential benefits that AI systems can bring to the health and medical realm (improvement of medical diagnoses, medication and treatment prescription, treatment supervision, etc.), most people feel reluctant about it, and prefer to be treated by human doctors and traditional technologies (ONSTI, 2021). Although some unwillingness can be explained due to citizens' unfamiliarity with AI systems, we firmly believe that these applications must be included in the high-risk AI systems list. To make this point clearer we present the following list of risks that AI systems in healthcare can present for citizens:

1) The main reason for classifying AI applications as high-risk is that any failure of the system can cause significant and irreparable damage to people's lives.
2) AI applications in healthcare are not still able to meet accountability requirements. To date, the CE has not established solid criteria that enables accountable processes if any failure or malpractice occurred during the use of these systems. The regulation on AI would be an excellent opportunity to set clear and solid criteria to take accountable AI systems in healthcare to term.
3) Strong privacy issues and dilemmas emerge with the use of these systems. Large amounts of data and personal information are needed in order to make optimal medical diagnoses or any other type of medical evaluation through AI systems. Any failure, error or misuse of this data could not only have fatal consequences for patients' health and lives, but it could also represent a direct and immediate violation of their privacy and fundamental rights.
4) Harmful biases are also present in AI systems conceived for medical uses. Nowadays, most studies and research on this field is based on white western men. This makes diagnosing diseases to anyone who falls outside of this group more difficult as results are less accurate. In this sense, it is especially significant the presence of gender bias in health. This phenomenon makes that AI applications becomes less effective when patients to be treated are women and diseases to be diagnosis only occur in female bodies (menstruation, pregnancy, sexual and reproductive health, etc.). This and other

genuinely female medical conditions are less likely to be detected and treated successfully due to presence of still unresolved biases.

We would like to propose that until points (2) and (4) are not sufficiently solved most AI systems used in healthcare, medical facilities, or medical interventions (excluding chatbots, assistance robots, etc.) are listed as PROHIBITED. As points (1) and (3) have no definitive solution, only measures to reduce the harm can be approved, we propose that once points (2) and (4) are solved these AI systems are listed as HIGH-RISK. Thus, and following the format of the proposal, Annex III should include a point similar to the following:

Health:

a) AI systems intended to be used in hospital or medical facilities to: diagnose any disease, prescribe medication or any treatment intended to heal, evaluate health conditions or perform medical interventions.

3.2. Autonomous vehicles

We think that autonomous vehicles should be regulated by the European legislation and the AI regulation. We proposed that these AI systems are listed as high-risk due to the potential harms they can cause, not only to users, but also to not users (pedestrians).

In this case, autonomous vehicles use should be conditioned to the same requirements (Chapter 2) that the rest of high-risk AI systems listed in Annex III.

4. ON REQUIREMENTS FOR HIGH-RISK AI SYSTEMS

Despite a substantial list of requirements for high-risk AI systems, providers and user is included in Chapter II (Articles from 8 to 15) and Chapter II (Articles from 16 to 29) of the regulatory proposal made by the EC (2021), we would like to call the attention of the committee on the following points that can improve the regulation framework for AI in the UE:

1) The first requirement for any high-risk AI system should be that, previously to the systems implementation, responsibilities if something goes wrong are properly settled. In the current proposal the accountability process is not clearly explained and could cause harmful legal gaps for citizens and AI users. In this sense, the list of mechanisms included in these two chapter, but especially in Article 17 on the quality management system, should be expanded and explained in more detail.

2) Article 13 tackles transparency and information requirements for users. In this regard the proposal contemplates the need for transparent information to users of high-risk AI applications. Here it is important to focus on an important issue that is has been widely overlook in the proposal, AI explainability. In the case of any AI system, but especially in the case of high-risk systems, it is not enough to fulfill the principle of transparency, but also the principle of explainability.

   The proposal contemplates that the people who supervise AI systems have a sufficient understanding of the different characteristics and features of AI. However, it is important to note, especially in the case of high-risk AI applications in the public sector, that in order to ensure accountability, citizens need to be able to fully understand how the AI system that is being used works and why it takes certain decisions and no others. This is a complex challenge since the sociodemographic reality of European citizens is very diverse. In this sense, the adoption of ethical AI systems entails the challenge of making these systems understandable for everyone (regardless of his/her social status, educational level, etc.). The existence of digital gender, age, race, and other, gaps make especially important to ensure that the communications between algorithms and people are adapted to the reality of the last. If this does not happen, the digital divide will grow larger, and the digitization process will create more unequal societies.

3) Article 14 establishes that high-risk systems need human supervision to guarantee their correct use. Although human supervision is important, it is worthwhile noting that this requirement does not guarantee that the AI is used correctly, and different studies how shown that this measure is less effective that what it was believed at first (Busuioc, 2021) as these people are also subject to bias. Although the regulatory proposal acknowledges this fact, we think that it needs to be better reflected in the regulation and that it should be articulated in terms of supervision teams (instead of one person supervising).

   "Automation bias" is commonly present in people who have to take decisions based on recommendations or information that comes from a computer/algorithm/intelligent system. When people suffer from "automation bias" they tend to rely non-critically on the system's recommendations. Thus, these people follow the proposed results and conclusions by default and, perpetuate the bias or mistakes made by the machine. As in the case of other biases, ending it is a complex task, but significant efforts can be made to reduce its effects. In this sense, we propose that:

a) Mechanism to ensure that AI supervisors know the existence of "automation bias" are properly established.
b) The regulation binds that AI supervision is carry out by diverse groups of more than one person for any high-risk AI system through gender (and other) quotas. Beyond the biases that arise in human-technology relationships, we all have biases inherent to our social and cultural condition that are practically impossible to overcome. In this sense, the only way to reduce these biases to a minimum is to gather in teams that can represent social plurality and diversity.

5. On MODERATE/LOW-RISK SYSTEM

The regulation proposal for AI does not formally address moderate/low-risk systems because, as suggested by the EC, they do not pose significant risks to the fundamental rights of European citizens. For these applications, the EC proposes the development of codes of conduct as mentioned in Article 69. Although it is true that not all AI systems present the same level of risk (in terms of violations of fundamental rights), it is necessary to rethink the scope of the regulatory framework to include, at least, some aspects of moderate/low-risk AI systems.

Although moderate/low-risk systems do not involve a direct violation of the rights of EU citizens, they can harm them indirectly. In this sense, we find convenient to incorporate some considerations regarding these systems in the legislation to ensure that they are subjected to several of the same mechanisms (external audits, data representativeness, transparency, technical documentation, etc.) as high-risk AI systems. To exemplify the potential harms of moderate/low-risk AI systems we have listed in this document some frequent applications that can be harmful for women and violate their rights as citizens. The following examples do not meet the principles of justice and equality held by the EU.

1) **Chatbots:** Different studies show that most chatbots have female names and/or voices, as well as other attributes traditionally related to women (West et al., 2019). Although the assistance roles for which most chatbots are used are worthy and necessary tasks, it is a reality that they are normally associated with subordination, help and/or care roles. In line with other studies, we think that the systematic attribution of female characteristics to these AI systems perpetuates harmful associations for women since they reinforce damaging stereotypes. These practices could be regulated for the future law, at least for chatbots used in the public sector, so equity in the distribution of roles represented by AI systems can be guarantee. This measure would result in a similar proportion between "female" and "masculine" chatbots or even in the implementation of gender-neutral chatbots.

2) **Robots:** Women rights can also be violated through a certain representation of the physical characteristics of robots. This fact is not contemplated in the regulation proposal but should be listed in the final regulation. Although the use of robots is yet not very common, the number is expected to increase significantly in the coming years (EC, 2021c). To date, most robots with humanoid appearance have hold very specific physical features that respond to, commonly harmful, gender stereotypes. These robots usually present features commonly associated white wester women with slim bodies (sometimes even sexualized).

   As in the case of chatbots, robots that imitate human appearance are mainly oriented to interact with people in situations of assistance, help and care. In these circumstances, human resemblance is sought to increase and facilitate people's confidence on robots. The selection of the previously mentioned characteristics for robots perpetuates the belief that women should occupy certain roles of assistance and care, and, in addition, it is assumed that these physical characteristics are more welcoming and pleasant than others. These representation of women through robots reinforces a bias message about what the normal appearance of women should be like; perpetuates harmful stereotype; and makes invisible the reality of other types of bodies that are present in our societies. For these reasons, the European regulation should contemplate the need to legislate on robots' perpetuation of gender stereotypes and other well-proven harmful representations of women.

3) **Sexual labor:** Closely relate to the previous example, we think that the UE should regulate, and in this case prohibited, the use of robot for sexual labor. Although European legislation is not homogeneous in terms of prostitution, as well as other types of tasks classified as sex work or labor, the regulation on AI should be able to address the issue for the case of robots which use is intended for this type of purposes. This point must be approached from two different perspectives:

   a) The review and consequent application of the European legislation on prostitution to find a joint way to deal with the issue.
   b) The development of intelligent robots (AI) rights.

6. ON EXPANDING ETHICAL NORMAL

Although the AI regulation proposal attempts to legally shield several of the ethical principles previously proposed by the CE and its HLEG (justice, non-harm, explainability and human autonomy), there are different points where there is still room for improvement:

1) The regulation does not adequately address the principle of explicability. As previously mentioned, transparency of data and algorithms is not enough to guarantee an ethical use of high-risk AI systems, or any other type of AI system. In order to ensure accountability (which constitutes one of the fundamental pillars of modern liberal democracies) it is not only essential that expert users understand how AI works, but also citizens whose lives are going to be affected by the decisions made and/or influenced by AI systems. This implies a commitment from the EU and all the organizations that comprise it to improve communications and relations between new technologies and people.
2) It is necessary to broaden the ethical reflection on the lack of diversity in the field of AI and how this reflects on an insufficient shielding of the rights of the most vulnerable groups in the European regulation on AI.
3) It would be especially important to propose certain measures to assure fairness and AI justice. As many people with low education levels, precarious jobs, and subjected to conditions of social exclusion are going to lose their jobs, it is urgent that the EU develops plans to proportionate these people sufficient and dignified live conditions, so they do not renege from the digitalization.
4) AI justice research should explore the relationship between AI systems and the elderly. In addition to the gender digital divide, there also exists the generational digital divide that mainly affects older people who do not feel part of the digitization process. This is due to lack of knowledge about AI and their own self-perception as incapable of understanding the way new technologies work. It would be interesting to explore the possibility to make specific plans for the elderly in this matter.
5) The principle of no harm is not well covered by the regulation as it does not include explicit prohibitions and limitations of autonomous weapons, autonomous vehicles and sex robots.

7. ON MECHANISMS TO GUARANTEE AI GOVERNANCE

There are certain mechanisms that can help ensure that the AI governance approach properly works:

Prevention mechanisms

Prevention mechanisms are those that are aimed at avoiding, *ex ante,* AI misuses and failure. Within this set of mechanisms the most effective one is the implementation of AI regulation (like the one the EU is working on) in order to avoid that certain practices and uses of AI take place. The classification of AI applications based on the risks they pose to human rights, as well as the establishment of the requirements that they must comply with, is a fundamental step in this direction.

It is also important to develop and implement a solid ethical framework (as the one proposed by the UE) that is known by citizens, so they are aware of their rights against AI as European citizens.

Supervision mechanisms

Supervision mechanisms are those aimed at ensuring the correct functioning of the AI once it is already implemented. In this regard, there are different and complementary mechanism that can be introduced[2]:

1) Periodic data audits.
2) Implementation of diverse and interdisciplinary teams to supervise high-risk AI systems.
3) The creation of European and national agencies that monitor AI systems functioning in the European territory. This could include the elaboration of:

    a) AI Maps
    b) Reports on the use of AI in the private sector
    c) Reports on the use of AI in the public sector
    d) Surveys of citizen satisfaction in relation to AI services
    e) 5) Studies to identify how to improve the relationship between AI and people

[2] Some of these mechanisms are already contemplated in the regulation but we have introduced small to them.

Compensation mechanisms

Compensation mechanisms are those that star operating when, accidentally, or intentionally, a failure and/or error occurs during the use of an AI system causing harm to one or more people, so they have the right to ask the relevant authorities for accountability measures. In this sense, the main compensation mechanism is the existence of accountability processes that enable compensation to those affected and establishes sanctions (penal or administrative) for those set accountable.

In order to guarantee accountability, it is necessary that the regulation on AI requires that all AI systems used in European territory are transparent and explainable. Transparency allows citizens to have the necessary information to evaluate AI systems, while explainability should guarantee that not only the relevant information is available, but also that it is understandable for everyone.

8. RESUME OF KEY POINTS

It is essential to keep in mind, as it has been pointed out at the beginning of this document, that several aspects of the regulation may become obsolete in a few years due to the rapid changes experience in AI and other emerging technologies. This requires that the regulation on AI (mainly the ethical standards, databases, and techniques for AI, is periodically subjected to review.

We now present a resume of this document through some key points:

1) Autonomous weapons and sex robots should be listed as PROHIBITED AI practices.
2) Autonomous vehicles and AI systems for healthcare should be listed as HIGH-RISK AI systems.
3) Further analysis and studies on how AI affect women are needed in order to present a fair and ethical AI regulation.
4) Include some moderate/low-risk AI systems' characteristics in the regulation as an addition to code of conduct.
5) Analyze the necessity of developing and introducing robots' rights.
6) Include the presence of hardware in the definition of AI since many possible discriminations and violations of rights (mainly of women, but also of other groups) are associated to this feature of AI systems.
7) Exclude AI techniques and methods from the central part of its definition and focus on its functionalities.
8) Need to expand the ethical framework of AI in different areas, but especially in terms of gender equality.
9) Reinforce accountability mechanisms.

10) Include the constant changing reality of AI, regarding techniques and methods used for its development, and data, in the regulation. Regardless of specific political affiliations, it is important to recognize that the European society is increasingly diverse in relation to gender, sex, religion, race, ethnicity, etc. Therefore, an ethical use of AI that is in accordance with European values needs to represent these realities in the data and the design of algorithms.

References

Bartoletti, I. (2019, June). AI in healthcare: Ethical and privacy challenges. In Conference on Artificial Intelligence in Medicine in Europe (pp. 7-10). Springer, Cham.

Busuioc, M. (2021). Accountable Artificial Intelligence: Holding Algorithms to Account. Public Administration Review, 81(5), 825–836. https://doi.org/10.1111/PUAR.13293/FORMAT/PDF

European Commission. (2021a). Laying down harmonized rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

European Commission. (2021b). Study on eHealth, Interoperability of Health Data and Artificial Intelligence for Health and Care in the European Union. https://euagenda.eu/publications/study-on-ehealth-interoperability-of-health-data-and-artificial-intelligence-for-health-and-care-in-the-european-union

European Commision- (2021c). 2030 Digital Compass: the European way for the Digital Decade https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A52021DC0118

HLEG. (2019a). A definition of Artificial Intelligence: main capabilities and scientific disciplines. https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-maincapabilities-and-scientific-disciplines

Lohr, S. (2018). Facial recognition is accurate, if you're a white guy. *New York Times*, *9*(8), 283.

ONSTI. (2021). Report on AI applications. https://www.ontsi.es/es/publicaciones/Estudio-aplicacion-inteligencia-artificial

United Nations (2021). First Draft of the recommendation on the ethics of artificial intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000373434

West, M., Kraut, R., y Ei Chew, H. (2019). I’d blush if I could. Closing gender divides in digital skills through education. [ltima visita 17.03.2021]. https://en.unesco.org/Id-blush-if-I-could