Developing a Trustworthy AI culture in Scientific Research

Chiara Gallese

Abstract The development of a Trustworthy AI culture among researchers and students in the AI field is a major challenges for universities. Researchers are the main actors in stimulating AI technological progress, but they often lack of support, guidance and education regarding all the appropriate measures that should be taken in order to develop trustworthy AI systems. More often than not, ethical, privacy, and legal compliance is regarded as a mere bureaucratic burden that makes their work more difficult, as shown in the literature (Peloquin et al., 2020).

Scientific research benefits from some partial legal exemptions, but researchers involved in the development of AI systems still must comply with many applicable laws and regulations[1], such as GDPR, ePrivacy Directive, civil liability rules, contractual law, intellectual property law, and, soon, also new regulations such as the so called "AI Act". However, exemptions do not apply when the research output is commercially exploited, and that represents a challenge for technical universities strictly connected to industry. In fact, it is very common that researchers perform analysis on behalf of (or together with) other entities, such as companies, hospitals, NGOs, and Public Administration bodies. This means that for researchers it is often very difficult to understand where the line of the exemption applicability is drawn and how to perform the appropriate compliance when developing new AI tools.

It is important to note that the major legal issues are not triggered by a specific AI technique but instead by its concrete consequences on humans. Impact of AI systems on individual rights must be thoughtfully analyzed and explored from a legal and ethical point of view, because such systems can lead to negative consequences on citizens (for example, the lack of a sufficient explanation of their effects) especially when they require the use of personal data.

Eindhoven University of Technology · Carlo Cattaneo University - LIUC
E-mail: cgallese@liuc.it; c.g.gallese.nobile@tue.nl

[1] European Regulations are legal instruments that are binding and directly applicable in every Member State

Data protection, in fact, is one of the legal requirement that researchers probably encounter the most, because it regards not only student's and employees' data (so that senior researchers have to comply to relevant rules during the hiring process, during the mentoring of younger researchers, and during teaching activities), but also the data collection phase and secondary research use of data. The enacting of GDPR - which provided for harmonized rules in all Member States -, although introducing important mandatory principles such as transparency, right of explanation, privacy by design and by default, represented only a small step towards the realization of an European responsible and sustainable AI framework.

In April 2021, the European Commission published a proposal for a new Regulation laying down harmonised rules on AI. The proposal states that research should not amount to use AI systems in human-machine relations in a way that exposes natural persons to harm, and that research activities must be carried out in accordance with recognised ethical standards. It also lays down many requirements for putting into service AI systems, without providing for an explicit exemption for scientific research. These provisions, therefore, will be added to the existing laws and regulations to be followed by researchers, further complicating the regulatory framework in the academic field. It represents an additional step towards trustworthy AI, but it still doesn't ensure to fully reach this important goal.

According to the definition of the High-Level Expert Group on Artificial Intelligence (2019), trustworthy AI systems should have three components:

1. they should be lawful, complying with all applicable laws and regulations;
2. they should be ethical, ensuring adherence to ethical principles and values;
3. they should be robust, both from a technical and social perspective.

It has been noted that there are no proven methods to translate these conditions into practice (Markus et al., 2020; Mittelstadt, 2019), and literature shows how the complexity of stakeholders and the different expectations they have over the concept of transparency is a major challenge (Felzmann et al., 2019). This paper does not propose a method to develop trustworthy AI, but is aimed at identifying strategies and approaches to help researchers in developing trustworthy AI systems, by suggesting how to enhance the first two conditions (lawfulness and ethics).

The most important approach that could be adopted by universities is represented by mandatory training for students and researchers, in order to raise awareness on the importance of applying ethical and legal principles in their research: in fact, it is not possible for them to satisfy the first two conditions outlined above without being aware of the applicable laws, regulations and code of conducts. However, awareness is not important only for researchers whose aim is to develop trustworthy AI, but also for all AI researchers: the paradox is that it is not possible to understand when trustworthy AI is mandatory by law, when it is only recommended, and when it is not crucial, if an appropriate training is not provided to researchers. To identify and fill the knowledge gap within different universities and research institutes, a survey

should be conducted and administered to students, researchers and support staff.

One other fundamental measure is to provide - since the first stage of the research lifecycle - continuous support from research data stewards, privacy staff and Ethical Review Board members, who should contact students and researchers after the registration of each research project at a central level. The Netherlands can be a great example at this regards, due to its attention to data stewardship (Jetten et al., 2021). Providing support is crucial because the legal framework, especially regarding privacy, is very complex (Sartor and Lagioia, 2020).

An additional strategy could be also to promote interdisciplinary research groups, that would include researchers specialized in privacy, ethics and law into research projects related to the development of AI models; in particular, automated decision making systems that could be used by public administrations or by companies to refuse a service or a contract, and AI systems that require the use of sensitive data (such as medical imaging, mental health information, data about children or other vulnerable groups) need to be assessed by those expert, because otherwise their use may not be legal or could cause users to be exposed to legal claims. A literature review assessing approaches advocated in the trustworthy AI literature is needed to understand if the collaboration between AI researchers and Law researchers is increasing and to what extent the lawfulness condition of trustworthy AI is overlooked.

However, without raising awareness first, researchers in AI field would not always recognize the need for including experts from different disciplines and would most likely believe to be able to develop trustworthy AI models by themselves. In this context, joint master's degrees in AI & Law or AI & Ethics may not be the perfect solution because they would provide a false sense of self-sufficiency, while the complexity of the matter is challenging even for domain experts. It may be more efficient, instead, to provide appropriate education at a higher level (e.g., doctoral courses, specific training for post-docs and professors) while maintaining some specific courses in traditional degrees (for example, providing data protection essentials to students in STEM and legislation regarding AI to law school students).

Lastly, it is important to develop workflows, guidelines, tools and other resources in order to reduce the bureaucratic burden on researchers, without impairing the compliance process.

Keywords AI · European Law · Research · Ethics · Privacy

References

Felzmann H, Villaronga EF, Lutz C, Tamò-Larrieux A (2019) Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data & Society 6(1):2053951719860542

High-Level Expert Group on Artificial Intelligence (2019) Ethics Guidelines for Trustworthy Artificial Intelligence. European Commission

Jetten M, Grootveld M, Mordant A, Jansen M, Bloemers M, Miedema M, van Gelder CW (2021) Professionalising data stewardship in the Netherlands. Competences, training and education. Dutch roadmap towards national implementation of FAIR data stewardship. National Programme Open Science (NPOS)

Markus AF, Kors JA, Rijnbeek PR (2020) The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of Biomedical Informatics p 103655

Mittelstadt B (2019) Principles alone cannot guarantee ethical ai. Nature Machine Intelligence 1(11):501–507

Peloquin D, DiMaio M, Bierer B, Barnes M (2020) Disruptive and avoidable: Gdpr challenges to secondary research uses of data. European Journal of Human Genetics 28(6):697–705

Sartor G, Lagioia F (2020) The impact of the General Data Protection Regulation (GDPR) on artificial intelligence. European Parliament