For a 'good' governance of AI: the role of ethics and participatory guidelines*

Laura Palazzini
Professoressa ordinaria di Filosofia del diritto, Università LUMSA. Mail: palazzani@lumsa.it

1. What is governance

AI belongs to the so called 'emerging (bio)technologies' with specific characteristics: high level of complexity, high-speed development, new advanced technologies untested or lacking proof of evidence robust enough for safety and effectiveness, uncertainties concerning risks, an evolving scenario, which is unpredictable and/or unprecedented (difficult to foresee), possible long-term global or large impact, also for future generations, in different application areas. They are called frequently 'transformative' and 'disruptive', producing profound changes in the personal, medical, social, and/or economic environment that may offer significant potential benefits for individuals and society, but also challenges and risks.

Governance refers to (institutionalized) processes or practices whereby issues of common concern/challenging (as emerging technologies) should be managed, governed, and regulated. It is a framework between technoscience, ethics, and regulation: the goal is to establish and share principles/norms/rules, in order to 'govern' technologies in society. It entails the identification of: who governs (who has the institutional authority to evaluate, decide, and rule), what needs to governed (object), how it should be governed, who is the recipient of government, and when it is necessary to govern. Evaluation, regulation, decision-making, and policy-making (at local, national, international levels) are extremely difficult under such circumstances: difficult to plan and to envisage scenarios due to the exponential acceleration of technoscience. It is a problem we are facing now, with no definitive solution, but as a dynamic/living discussion.

After more than 50 years since bioethics was born, many 'traditional' topics of bioethics (beginning/end of life issues, cure/experimentation) have been debated and are still currently beeing debated and we found out that: ethical discussions are still there (ethical pluralism cannot reach in some very sensitive issues an ethical minimum consensus), regulations are heterogeneous (in some cases giving rise to the so-called 'bioethical tourism'), ambiguous (giving rise to different case-law), obsolete (they do not keep up with the pace of technological progress and always need to be updated, constantly lagging, the so-called 'law lag'). Only with regard to specific topics, as clinical trials or transplants, we have reached at least continental regulations (as in the EU), leaving space to ethical debate on applications and interpretations.

The need for a new way to 'govern' emerging technologies is rising, to avoid repeating past 'mistakes', to better manage the present and future of technoscientific development, considering its characteristics. In front of emerging scientific and technological development, on the one hand, there are expectations of great benefits/improvement, on the other hand, big challenges/concerns. Technophiles fear governance as an obstacle, limiting scientific and

* This publication is funded by the European Union - Next Generation EU, under the PRIN 2022 call, project "MEDICINE+AI, Law and Ethics for an Augmented and Human-Centered Medicine" (2022YB89EH) - CUP E53D23007020006.

Downloaded from www.biodiritto.org.
ISSN 2284-4503

technological freedom; technophobes hope that governance could hinder the bad use of technologies.

2. Which governance for AI: methodological requirements and approaches

We should reassess the discussion in order to improve governance. The discussion should be interdisciplinary and pluralistic but with innovations, above all in the AI field:

a) *enlarged interdisciplinarity* that should integrate other disciplines as informatics, engineering, cognitive sciences, huma sciences (psychology, sociology, anthropology, economy, political sciences), beyond ethics and law; interdisciplinarity should become *transdisciplinarity* engaged in a regular and permanent exchange of views, respecting the epistemological diversity of the different disciplines, but also being able to integrate in their own methodology and approach different inputs and views, learning to work together during each research phase (i.e. developers, computer scientists, engineers together with physicians, psychologists, social scientists, ethicists, lawyers);

b) *converging/convergent pluralism,* as a dialogical approach between tecnocentrism and human-centrism, not a negotiation procedure based on pragmatic agreement or compromise, but as a progressive hermeneutical and dialectical critical effort to grasp shared values/principles to find convergence;

c) *synthetic, dynamic and integrated ethics* reflection, in an ongoing process of designing, developing and deploying AI, in a constant effort to envisage scenarios (anticipatory ethics), in order to reflect before, during and not only after the technological application;

d) *timely and tailored reflection,* requiring constant and regular updates about scientific and technological novelty, with specific emphasis on specialisation in certain fields that rely on specific competencies and skills (as digital skills for informatics and AI);

e) *enlarged interaction*, considering the potential scenarios and possible impact of decisions, taking into account all the stakeholders, beyond local, national, regional, continental borders;

f) *tuned reflection* open to 'lay' discussion of non experts, engaging citizens in a participatory process (public engagement, alongside stakeholder engagement), in the decision-making process. When stakeholders see that their concerns and values are taken into account, they are more likely to support and trust the decision-making process: public acceptance and trust are key to the feasibility and sustainability of any potential intervention.

In the face of significant knowledge gaps, in the context of complex dynamic and accelerated systems, with difficulties in predicting with precision, there is an emerging need to develop decision tools that are sufficiently robust to manage deep uncertainty. It is possible to identify some 'approaches' on an ethical level:

a) *prudent approach* means not waiting for comprehensive scientific evidence (that may bring to inaction, as it could be impossible or, in any case, require to much time) and not enabling apriori

Downloaded from www.biodiritto.org.
ISSN 2284-4503

innovation (as liberalization could be dangerous, not knowing and forseeing in advance possible damages), but identifying the scientific requirements (tailored for each technology) or gathering the sufficient information in order to consider 'safety' or at least when 'safety' could be considered 'enough', in order to decide (in case of not enough safety) for a moratorium or a ban;

b) *critical approach* to reflect on pros and cons, in order to raise awareness both on a scientific and ethical level to build a more comprehensive understanding of the potential impacts and ethical considerations, leading to more informed and balanced decisions, that go beyond the technical feasability, addressing the ethical and societal, and/or ecological dimension;

c) *adaptive/flexible/reflexive approach*, with constant update and observation of developments, in order to readjust if necessary, and to have specific means and methods in place to be attentive to the emergent properties of the complex system responding to new information and changing conditions. Reflexivity involves ongoing prudent and critical evaluation of the governance processes and outcomes, to ensure they remain aligned with an ethical framework and societal values.

3. For a 'good governance' of AI: the role of ethics

The ethical principles in a human-centric perspective should be guiding principles supporting and grounding governance:

a) *meaningful human control:* humans need to keep control and oversight of what they design, program, and deploy; AI should provide 'support' to human decision-making, and cognitively 'assist' human decisions, but not 'substitute' or 'replace' them, in order to evoid the risk of decreasing human attention, with the possible consequence of reducing human skills (de-skilling), reducing responsibility (de-responsibilization), moving towards the technologization of choices that can impoverish and even cancel concrete interpersonal relationships (de-humanization);

b) *reliability and trustworthiness:* accurate checks should be made of both data (the accuracy of collection or the 'quality of data') and algorithms (the validation of software), to guarantee the beneficial (and not maleficent) introduction of various forms of AI;

c) *transparency and explainability:* given the impossibility – in case of indeterministic algorithms - in some circumstances also for software programmers and IT technicians to explain how the system has achieved certain results (the 'black box problem'), it is necessary to develop technology that can explain – as far as possible - every step of the decision or at least to inform users of the risk of opacity, in order for them to acquire critical awareness (above all in the applications to medicine);

d) *equality, fairness and non-discrimination:* the data collection,

Downloaded from www.biodiritto.org.
ISSN 2284-4503

classification, profiling, and clustering through algorithms may be discriminatory, based on exclusion criteria (e.g. age, gender, ethnicity, personal ability), intentionally or unintentionally. AI could worsen already existing inequalities, through possible biases in the design of the processing systems. This problem can be avoided by adopting a broad approach (known as the participatory approach), which could be representative of all individuals (design-for-all), regardless of any differences;

e) *data sharing:* the protection of privacy and confidentiality is often underlined as an obstacle to the development of AI, which is based on big data; the possibility/opportunity of "sharing" data arises, as a "social/common good", for the advancement of knowledge;

f) *public engagement and education:* it is also desirable to promote public debate on the developments and limits of AI so that all individuals can acquire the basics of "AI literacy", promoting active participation in social discussion, avoiding the "digital divide" and marginalization, stigmatization and exclusion of people. Education should be revisited, introducing humanities and ethics courses for engineers, computer scientists and data scientists, with a focus on ethics by design/in design, to ensure ethical awareness from the very beginning of the technological design, understanding the responsibilities involved in the development, deployment and use of AI;

g) *sustainability:* the sustainability of AI technologies should be ensured, both on a social and environmental level. It is necessary to take into account the importance of access for all human beings, considering the costs, and at the same time respect for the environment, including all living beings.

Ethics has progressively become an 'institutional practice'. The main expression of 'institutionalized moral reasoning' refers to ethics committees, established to produce reports, opinions and recommendations for governance on the ethical aspects of new technologies.

The interdisciplinary and pluralistic discussion carried out by those committees, in the framework of the interpretation and clarification of fundamental human rights, has contributed to identifying shared ethical principles, such as: the primacy of the human being and his/her dignity over the sole interest of science or society; the respect for integrity (safety) ; the duty of beneficence/non maleficence in interventions on human beings; the need to ensure autonomy as awareness of individuals; the need for equity of access and equal treatment of human beings.

The main documents are, at the European level: The High-Level Expert Group on Artificial Intelligence, in *Ethics Guidelines for Trustworthy AI,* 2019; European Group on Ethics in Science and New Technologies *Statement on Artificial Intelligence, Robotics and 'Autonomous Systems'*, 2018 and *Future of Work, Future of Society*, 2018; Council of Europe, Committee of experts on Human Rights, *Dimensions of Automated Data processing and Different Forms of Artificial Intelligence*. At an international level: Recommendation of the Council on Artificial Intelligence of the OECD, *Legal Instruments Artificial Intelligence* 2020; UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, November 24, 2021.

Downloaded from www.biodiritto.org.
ISSN 2284-4503

4. The ethics in the normative ecosystem: hard law and/or soft laws

The discussion on the choice of tools for governance depends on specific technology, fields of application, and the extent to which human rights might be engaged or challenged. If the challenges are relevant, consideration may be given to the opportunity of a legal regulation, as a translation of ethical principles into laws. There are good reasons to expect that loopholes will appear within existing regulations when it comes to emerging technologies.

The first choice is the decision to intervene or not intervene with regulation. Many countries are deciding not to intervene for several reasons: no need of regulation, as the impact and consequences of technology application are considered without relevant risks or as regulation might endanger scientific progress and freedom (of commerce and use). Abstain from intervention can often become dangerous, due to the intrusiveness of both technocracy and the market, alongside a 'technological far-west' (the so called techno-centrism).

Europe decided to intervene with a specific and detailed regulation on AI because of the risk of violation of fundamental human rights. The purpose of regulation is, at least, to clarify general principles and ban malevolent practices with unacceptable risks. In this sense AI Act is both a *hard and soft laws*: hard laws with binding rules for specific applications of AI considered dangerous for individuals and society (subliminal manipulation of ideas and behaviours, exploitation of vulnerable people, biometric identification remote in real time, inference of emotions in education and workplace, etc.); soft instruments refer to 'light', flexible, procedural laws, designed to define a minimum content, while leaving more or less broad scope for discretion when interpreting and implementing rules.

Soft instruments allow easier changes, adaptations, and reviewing, and do not require setting up a formally complete and timely regulatory framework, which may or may not eventually take place. It developed a methodology for the Ethical Impact Assessment of AI technologies based on rigorous scientific research, ethical discussion grounded in international human rights law, for its implementation in all stages of the AI system life cycle, and capacity-building materials to support countries' efforts to train government officials, policy-makers and other relevant AI actors, in order to plan appropriate risk prevention, mitigation and monitoring measures. Such assessments should be multidisciplinary, multistakeholder, multicultural, pluralistic, and inclusive (the EU focuses on the risks of surveillance, with social scoring and facial recognition; the manipulation of behavior and opinions as risks to be avoided).

At an international level emerges the need to ensure AI governance mechanisms with ethical oversight authorities, establishing, on an institutional level, independent public bodies responsible for control and providing a mechanism of continuous monitoring (e.g. UNESCO proposes a network of independent AI Ethics Officers to give support to this process at national, regional and international level); ensure possible mechanisms for monitoring and evaluation, such as an ethics commission, an AI ethics observatory, to evaluate their adherence to policy recommendations on ethical requirements.

The normative framework of AI is strictly connected to data governance (in the framework of Health Data Space), developing strategies that ensure the continual evaluation of the quality of training data for AI systems, including appropriate safeguards to protect the right to privacy and

Downloaded from www.biodiritto.org.
ISSN 2284-4503

promote open/sharing data, reviewing policies and regulatory frameworks, introducing mechanisms, such as open repositories for publicly funded or publicly held data and data trusts, to support the safe, fair, legal and ethical sharing of data; promote public and private efforts to create collaborative platforms to share quality data in trusted and secure data spaces; promote international collaboration (as opposed to power competition) on AI research and responsible innovation, engaging international organizations and research institutions, as well as transnational corporations, which can be a basis for the ethical use of AI systems by public and private entities, creating incentives for the public and private sectors to invest in this area (OECD, 2019); ensure that AI researchers are trained in research ethics and require them to include ethical considerations in their designs, products and publications (human oversight, safety and security, transparency and explainability, inclusiveness, sustainability); improve access to information and education, promote AI literacy to strengthen not only technological skills but also ethical critical thinking and the competencies needed to understand the use and implication of AI systems, in order to empower people and reduce digital divide and digital access inequalities; ensure broad participation, paying special attention to particularly vulnerable people or people in vulnerable situations.

5. The relevance of interdisciplinary and participatory guidelines in medicine

In the context of the complexity of existing or non-existing regulations, ethics plays an important role: that of filling gaps, integrating, and guiding interpretation where norms are not always complete, clear, or coherent. Ethics elaborates the general principles that constitute the framework for concrete applications. Within this framework, the role of guidelines or recommendations becomes evident: these are soft regulations that, within the broader horizon of legal norms and ethical principles, provide direction to professionals (in medicine, to healthcare practitioners) on how to make *good* use of AI. It is essential that these guidelines be *interdisciplinary* and *participatory*: interdisciplinary, meaning they arise from a dialogue involving the various disciplines implicated throughout the entire AI lifecycle, from design to application; participatory, meaning they are constructed not only *top-down* (drawing on ethical principles, fundamental rights, and the broader regulatory ecosystem), but also *bottom-up*, listening to all stakeholders — the voices of those who design, build, apply, and use these technologies. This includes the voices of computer scientists, doctors and healthcare professionals, patients, patient representatives, caregivers, in dialogue with ethicists, legal experts, and social scientists.

The participatory methodology involves, beyond literature review, empirical research, consisting of surveys aimed at all stakeholders using rigorous qualitative methods (e.g., validated questionnaires or those developed on interdisciplinary bases with statistically meaningful sampling), focus groups or structured/semi-structured interviews, public consultations, and so on. The goal of "participation" is to make the process of developing policy guidelines transparent and inclusive, taking into account not only the data emerging from literature (via expert analysis), but also the preferences, needs, and lived experiences of those directly involved in care. Such a methodology is important because it enables a better interpretation and application of general principles (which might otherwise remain abstract and detached from real circumstances), enhances adherence and acceptability by

Downloaded from www.biodiritto.org.
ISSN 2284-4503

professionals and patients, supports personalized care, and improves trust.

These guidelines should then become part of the training of those working in the field, as they promote the acquisition of AI literacy, a foundational competence to understand the proportional use of AI (balancing benefits and risks), with realistic expectations (avoiding both excessive techno-enthusiasm and equally excessive technophobic fears), and with awareness of risks and limitations. This contributes to the creation of an environment of critical trust, both on the part of physicians and of patients.

Downloaded from www.biodiritto.org.
ISSN 2284-4503