Nicola Palladino

A digital constitutionalism framework for Ai

(doi: 10.53227/113109)

Rivista di Digital Politics (ISSN 2785-0072)
Fascicolo 3, settembre-dicembre 2023

Ente di afferenza:
()

Nicola Palladino

A digital constitutionalism framework for Ai: security and fundamental rights in the Ai Act[1]

A DIGITAL CONSTITUTIONALISM FRAMEWORK FOR AI: SECURITY AND FUNDAMENTAL RIGHTS IN THE AI ACT

Ai is increasingly crucial in everyday life and social relations, raising both expectations and concerns. There is a growing consensus regarding the need to establish a trustworthy and human-centric framework to unlock the full potential of this technology. As a result, we are witnessing a proliferation of initiatives aimed at creating ethical codes for Ai development. However, many studies highlight concerns about a «principle-to-practice» gap, noting that Ai developers and deployers often struggle to ensure the effectiveness and enforcement of the principles they adhere to. This article seeks to bridge the gap by combining the approaches of Societal constitutionalism and Science and technology studies. It aims to provide a digital constitutionalism framework for Ai ethical governance, advancing the discussion on how to incorporate ethical and human rights standards into the socio-technical design of Ai systems. By analyzing the case of the Artificial intelligence act, the article illustrates the roles and responsibilities of various actors in translating fundamental rights into technical and organizational arrangements. It emphasizes the significance and concerns surrounding a hybrid constitutionalization process.

KEYWORDS *Artificial Intelligence, Digital Constitutionalism, Artificial Intelligence Act, Societal Constitutionalism, European Union.*

[1] This article is the extended version of a short paper presented at the International Conference on AI for People in November 2023 and to be published in the related conference proceeding (Proceedings of the 2nd International Conference on AI for People: Democratizing AI, Springer, Forthcoming).

Nicola Palladino, Università degli studi di Salerno - Via Giovanni Paolo II, 132 - 84084 Fisciano, email: npalladino@unisa.it, orcid: 0000-0001-5472-5814.

ISSN 2785-0072

1. Introduction: The challenge of a human-centric and trustworthy Ai

The term Artificial intelligence (Ai) refers to a wide range of technologies, including expert systems, machine learning, natural language processing, artificial neural networks, computer vision, and knowledge representation. According to the Organisation for economic co-operation and development (Oecd), Ai could be defined as «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»[2]. As the European Commission High-level expert group on Ai specified, Ai systems do that «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 Hleg 2019a). In doing so, Ai systems can make predictions, recommendations, and decisions with a certain degree of autonomy, adapting to new cases and changing virtual or physical environments, based on an unprecedented number of parameters. While these Ai characteristics could be extremely useful to address the complexity of the modern world, augment human capability, and empower people, they also raise a series of concerns (Bollier 2018; Renda 2019). Due to the huge amount of data necessary to train Ai systems, they pose questions in terms of privacy and data protection. Autonomy and adaptability can result in challenges related to the opaqueness of output and the scale and remediability of potential mistakes. Moreover, the large adoption of Ai applications also raises questions related to manipulation and disinformation, job market and workplace rights, and energy and natural resources consumption (Brownsword *et al.* 2017; Turner 2018).

In the past few years, governments, the private sector, civil society, and the technical community reached the awareness that the full potential of this technology is attainable only by building a trustworthy and human-centric framework. In this view, Ai systems must be aligned with societal values and governed through accountable arrangements to avoid both misuse of Ai applications capable of harming people and underuse because of a lack of public acceptance (Floridi *et al.* 2018).

Not by chance, in the past few years, we have witnessed a flourishing of initiatives setting ethical codes and good governance principles for Ai development that usually converge around a common set of guiding principles,

[2] See: https://oecd.ai/en/wonk/ai-system-definition-update.

including respect for human autonomy, prevention of harm, fairness, privacy, transparency, and explainability (Fjeld *et al.* 2020; Jobin *et al.* 2019; Whittaker *et al.* 2018). However, practitioners are still struggling to translate governance principles into operational routines, which is commonly referred to as the «principles-to-practices» gap (Mittelstadt 2019; Schiff *et al.* 2021a). While some authors talk about «ethics washing» practices put in place by companies to delay or soften state regulation (Greene *et al.* 2019; van Dijk and Casiraghi 2020), there are many reasons to consider factors related to Ai systems' socio-technical productive process as the cause of the gap (Hallensleben *et al.* 2020; Schiff *et al.* 2021b; Shneiderman 2020). Ai systems, indeed, are giving rise to a novel governance layer in which, consciously or less, social values and regulatory mechanisms are embedded into digital architectures, while they are also influenced by factors such as division of labor, organizational culture, operational routine, governance arrangements, and the broader regulatory environment (Palladino 2023b). This complexity may result in a lack of awareness by developers of the social implications of their job, different interpretations of the same principles, functional separation and lack of communication between more technical or social-oriented components in the process, as well as unclear accountability mechanisms and attribution of responsibilities (Schiff *et al.* 2021b).

Initiatives currently on the table do not provide developers and deployers with sufficient guidance on how to implement principles for trustworthy and human-centric Ai within concrete contexts. Ai ethics initiatives are, for the most part, a collection of high-level abstract principles and do not provide enough assistance to contextualize principles in concrete situations, solve conflicts and trade-offs between them, and deal with the functional separation among technical and non-technical actors in the Ai development (Morley *et al.* 2021). Moreover, they lack enforceable «legal and professional accountability mechanisms» to counter the organizational pressures that developers face to constrain the time and costs of their projects and realize new profitable Ai applications (Mittelstadt 2019; Rakova *et al.* 2021).

Combining societal constitutionalism (Teubner 2012) and science and technologies studies approaches (Musiani *et al.* 2016) this article will contribute to close the gap between principles and practices by outlining a digital constitutionalism framework for a trustworthy and human-centric Ai.

As we will explore in greater detail in Section 2, a digital constitutionalism framework allows for the reframing of the concerns raised by Ai ethics in terms of a constitutionalization process. The constitutionalization of the digital world, with Ai representing the most advanced point in the contemporary digitalization of societies, refers to the process of establishing rules that unleash the potential of the digital social subsystem while institutionalizing

limitation mechanisms that set the boundaries of legitimate operations for digital systems and preserve the integrity and autonomy of individuals and other social spheres.

By relying on fundamental rights conceived as counter-institutions against the exploitative tendencies of digitalization, a digital constitutionalism approach can make the most of already existing institutional settings to deal with implementation, interpretation, and enforcement issues. Furthermore, a digital constitutionalism framework for Ai is also crucial «to maintain the character of our political communities as constitutional democratic orders» especially considering how «Ai systems increasingly shape our collective and individual environments, entitlements, access to, or exclusion from, opportunities, and resources» (Yeung 2020, 81).

However, digital constitutionalism should not be understood solely in legal terms; rather, it indicates the interplay between the social processes of technical design and norms creation (Santaniello *et al.* 2018). For this reason, it also implies a «hybrid» process in which various actors such as public authorities, international organizations, trade unions, and professional communities contribute to developing proper legal, technical, and organizational standards to embed fundamental rights within digital architectures. A digital constitutionalism framework then contributes to clarifying the roles and responsibilities of different actors, improving accountability. It also enhances coordination between technical and non-technical actors, ensuring the necessary integration between the matter's computer science and social science aspects.

Section 3 will apply the digital constitutionalism framework specifically to the Ai field. In doing so, it will identify a series of rule sets (comprising coding rules, security rules, inclusionary rights, and exclusionary rights) essential for the constitutionalization of this domain. Furthermore, the section outlines the different roles and tools employed by the diverse stakeholders engaged in this process, paying particular attention to the capabilities of the technical community to anticipate and address the social and political implications of technical specifications.

Section 4 discusses the European approach as a concrete example of a digital constitutionalism framework in the Ai sector. The section delves into the Artificial intelligence act (Aia), illustrating how it encompasses all the rule sets previously identified and how it outlines a hybrid process of governance in which Ai providers and standard-setting organizations are entrusted to identify the most proper solutions to the requirements for trustworthy and human rights-based Ai established by the public authority. However, the article also warns about the potential capture by private interests within standard-setting organizations and the risk of technological solutionism.

2. A digital constitutionalism perspective

As noted, «international human rights standards offer the most promising basis for developing a coherent and universally recognized set of standards that can be applied to meet many (albeit not all) of the normative concerns currently falling under the rubric of Ai Ethics» (Yeung *et al.* 2020, 80). International human rights standards offer several advantages in developing a trustworthy and human-centric framework. In many cases, human rights norms are already internationally recognized and supported by national and international institutions. Also, they rely on a system of reflexive evaluations capable of solving tensions and conflicts between competing concerns.

However, traditional constitutional instruments are undermined by the transnational, mostly private, and, above all, infrastructural nature of Ai systems power abuses.

As emphasized by Grimm (2016), the weakening of state authority due to transnational modes of governance is a significant threat to constitutionalism. The conventional understanding of constitutional order necessitates the «concentration and monopoly of public power that allows comprehensive regulation» within a specific territory and the recognition of a political community, serving as the constituent power, instituting self-imposed constraints on the exercise of public power (Santaniello *et al.* 2018).

The constitutionalization of international law (Wet 2006) has been put forward as a solution to the challenges brought about by transnational modes of governance. This can be seen as a form of «compensatory constitutionalism» (Peters 2006) that supplements and addresses the voids created at the domestic level, with norms of international law that might acquire a quasi-constitutional character (Gardbaum 2009).

Although theories concerning the constitutionalization of international law identify intriguing trends and remedies to counter the decline of nation-state authority, they offer minimal insight into preserving constitutional guarantees and fundamental rights within transnational private regimes. International human rights law falls short of constitutionalizing the international order in the digital domain. Given its emphasis on nation-states, it doesn't have a direct impact on major technology companies. Furthermore, its articulation of general principles is not well-suited for governing a complex socio-technical ecosystem like Ai.

Gunther Teubner's theory of societal constitutionalism can offer a more robust conceptual framework. Drawing from Luhmann's theory of social systems (2010) and the subsequent advancements by Sciulli (1992) and Thor-

nhill (2011), the German scholar initiates his analysis from the dynamics of social differentiation.

In this view, the more a social subsystem achieves autonomy, the more it establishes 'its own systemic logic based on a specific means of communication' that facilitates meaningful interaction within the subsystem (like money in the economic subsystem and law within the legal subsystem). As the activities of a subsystem gain significance for the broader social system, they give rise to what Teubner terms «expansionist» and «totalizing» tendencies (2011; 2012). This implies that the subsystem can impose its rationale on other social spheres to reproduce itself. Teubner calls 'autonomous matrix' (2011, 209) this anonymous power process capable of jeopardizing the autonomy and integrity of individuals and communities.

According to this perspective, the rise of the Internet and digital technologies in our societies can be conceived as a process of autonomization of an emerging digital subsystem. In the wake of Lessig (2009), we can identify in the «code» the communicative means of the digital subsystem, meaning by this not some programming language, but rather the socio-technical architecture which, by combining software, hardware, and human components, makes the interaction between different social actors in the digital world possible, shaping their experience and disciplining their behavior. While the code constitutes the means of communication of the digital subsystem, digitization or datafication (George 2020) can be interpreted as its logic. The latter, therefore, consists of an incessant process of conversion of social reality into digital information in order to be further processed and elaborated to extract further information with added value.

In this view, the process of constitutionalization arises with the establishment of both constitutive and limitative rules, allowing, on the one hand, to free the «potential of highly specialized dynamics» of the subsystem, and on the other hand, to institutionalize self-limitation mechanisms that preserve the integrity and autonomy of individuals and other social spheres (Teubner, 2004, 12). These set of rules also include fundamental rights understood as «social and legal counter-institutions» (Teubner 2011, 210) that embody the resistance of «flesh-and-blood human beings against the structural violence of the matrix» (*ibidem,* 213). Fundamental rights carry out both an inclusive and exclusionary function to counteract the expansionistic nature of digitization logic. Inclusionary rights ensures people with the access to the specific 'means of communication' of the subsystem and then to the possibility to take part in the definition of the foundational norms of the subsystem. In so doing people can exert control over the digitalization process to make sure that the development of digital technologies would be socially beneficial and serve hu-

man flourishing, rather than the contrary. Exclusionary rights define what are the boundaries of legitimated subsystem operations, preventing or sanctioning misuses of digital technologies that may put in danger the autonomy and integrity of individuals and communities.

Worth noting, in this view to be truly effective, fundamental rights must be translated into the specific means of communication of the sub-system and inscribed into its logic. Talking about digitalization, this means that fundamental rights need to be incorporated into the socio-technical architecture of digital technologies, including programming, algorithms, internal policies, and operational routines (Palladino 2021).

According to Teubner's approach, constitutionalisation is inherently a hybrid process, for two main order of reasons. First, the introduction of self-limitation mechanisms in the subsystem is the result of the external pressure exercised by individuals, communities, and other social spheres. Second, transposing fundamental rights into the logic and the code of the digital subsystem is a complex task that requires the involvement of a wide range of actors with different roles and responsibilities. These considerations should make clear that the constitutionalization of the digital world could not occur through private self-regulation or on state regulation alone.

3. Digital constitutionalism and Ai

It could be said that Ai technologies represent the most advanced point in this deployment of digitalization logic (George 2020; Van Dijck 2014) in which the external environment (the society) is exploited to extract digital data to be processed in order to elaborate added-value data commonly used to profile and classify people, predict their behavior, and make decisions based on this (Zuboff 2019). Ai technical choices and specifications have an immediate governance value, and the impact on people and society must be carefully considered at all the stages of the Ai life cycle.

This extraction-processing-elaboration process deploys its logic irrespective of its consequences on human life and then is likely to produce negative externalities endangering the integrity of people, institutions, and communities (Latzer and Just, 2020; Brownsword *et al.* 2017; Turner 2018).

In this context, Ai could be considered one of the most intriguing cases for applying a digital constitutionalism framework. But what does this mean in more precise terms?

As mentioned earlier, the constitutionalization of a subsystem occurs when frictions with other social spheres lead to the development of constitu-

tive and restrictive rules, along with fundamental rights that serve as a form of human resistance against the exploitative tendencies of the subsystem's autonomous matrix. Thus, we can identify different types of rules necessary for the constitutionalization of the digital subsystem, and then also of Ai systems, by combining constitutive or limitative functions with the safeguard of the autonomy and integrity of the digital subsystem (autonomous matrix) on the one hand, and of individuals and community (human rights) on the other hand (see Figure 1).

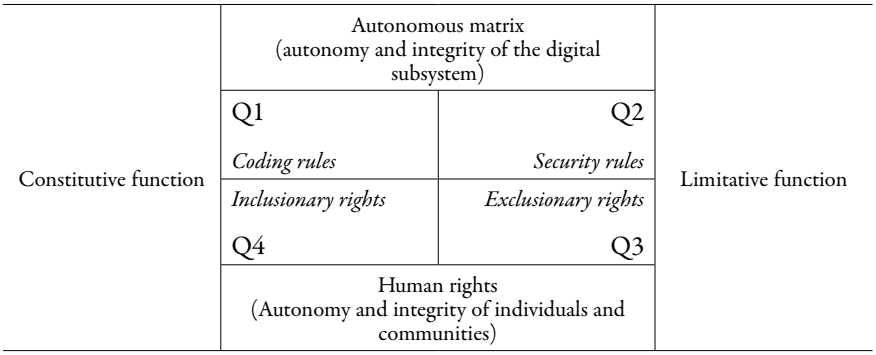

Fig. 1 Constitutionalization rules.
Source: Author's elaboration.

On the upper-left quadrant (Q1), the combination of the constitutive function pole with the autonomous matrix pole identifies the area of the Coding rules.

The code, we said, is the specific 'mean of communication' that made interaction in the digital world possible and shapes the specific potential of digitalization. In our case, it defines the sociotechnical infrastructure that allows the development and functioning of Ai systems. Coding rules encompass the set of prescriptions, possibilities, and constraints associated with programming languages, statistical methods, software and hardware specifications, as well as operational procedures. These rules are essential for the functionality and effectiveness of Ai systems. They ensure that Ai systems can effectively process and interpret data, make decisions, and deliver the desired outcomes in a consistent and reliable manner.

Even though we are discussing coding rules independently here, it's important to note that the other sets of rules within this framework also need to be translated into coding rules to be effective.

At the intersection between the autonomous matrix and the limitative function we can find the Security rules quadrant (Q2). These norms govern

the behavior of actors in order to safeguard the boundaries and integrity of Ai systems. Rules in this quadrant aim to prevent or sanction attacks, damages or malfunctioning of Ai systems, thus ensuring their stability, reliability, and resilience. For example, they might encompass access control measures, encryption protocols, data authentication procedures, and intrusion detection mechanisms, along with legal sanctions for both threat actors or organizations unable to ensure system security.

By combining limitative functions and human rights, we arrive at the domain of Exclusionary rights (Q3). These rules are designed to safeguard the autonomy and integrity of individuals, persons, and institutions, protecting them from the potentially exploitative tendencies of Ai systems by establishing the limits of their legitimate operations. These rules may refer to areas such as privacy, fairness, and safety. For example they may include mechanisms like data anonymization, data minimization, differential privacy, strict access controls, bias prevention and detection, testing and validation procedures, or safeguards against automated decision-making and automation bias.

Finally, the intersection between human rights and constitutive function gives rise to the area of Inclusionary rights (Q4). These rights aim at ensuring social shaping of and people's control over Ai development to make sure that it is socially beneficial and human-centric in the sense to place human needs, values, and wellbeing at the forefront of Ai design. Inclusionary rights include norms related to the transparency of Ai systems, accountability of Ai providers, and stakeholders' involvement at various stages of the Ai lifecycle. Inclusionary rights may include documentation provisions about the algorithms, data sources, and decision-making processes within Ai systems, explainability and traceability measures, auditability procedures, impact and risk assessment of Ai systems, as well as human oversight arrangements.

Moreover, stakeholders involvement is deemed crucial in order to ensure that the mathematical formalization of ethical and social issues and their solutions correspond to actual social views and needs.

However, broadly speaking, a human-centric and trustworthy approach to Ai requires the collaboration of different kind of actors with different roles and responsibilities (see Table 1), according to the hybrid nature of the process of digital constitutionalization.

Tab. 1 Different Actors involved in the Constitutionalization process.

Subject	Roles	Instruments
States	Define requirements and establish external accountability system	Regulation, law
Technical community	Develop, test and improve socio-technical standard to comply with requirements	Standards, guidelines, certification
Organizations	Management: define system requirements according to regulatory compliance, stakeholders expectations and business need; Establish internal accountability system (definition of role and responsibility); internal oversight; provide external accountability; technical team, developer, deployer: implement most proper technical specifications	Quality management system, risk management system, ethical committee, code of conduct, technical documentation, technical arrangements
International organization	Harmonize norms at the international level	Guidelines, declarations, recommendations, resolutions, treaties
Civil society and public opinion	Advocacy and watchdog function	Press campaign, vote, petitions, boycott, consumption habits strike, legal action

Source: Author's elaboration.

Most of the ethical choices that need to be made while developing and deploying Ai systems have an inherently political nature since they entail what kind of values we want to embed within this technology, their hierarchy, and the kind of changes we want to see in our society. For this reason, States and other public decision-makers are the most qualified subjects to determine which values and requirements must be met by Ai systems being endowed with the necessary authority and democratic legitimacy.

International organizations can play a pivotal role in ensuring that rules established by states do not differ too much, thus preventing hindrances to the transnational supply chain in Ai production or favoring a cherry-picking attitude and unfair competition. They can frame an issue around a common set of normative claims and influence national policymakers. Recent examples, such as the Oecd Recommendation on Ai or Unesco's recommendations on the Ethics of artificial intelligence, whose wording found its way into the G7 statement and the Eu Aia, are good illustrations of this influence.

Technical communities, with their expertise, are not only the best placed but also have the sole ability to translate requirements stemming from state law or transnational multistakeholder fora into operational standards. Better than others they can understand the effects and impact of certain technical specifi-

cations and are able to identify the socio-technical arrangements and architectures needed to prevent fundamental rights abuses and the endangerment of people and communities (Liddicoat and Doria 2012; Palladino 2021).

They also play a crucial role in the assessment of the compliance with established requirements for Ai systems.

Through the development of technical standards, guidelines, and certification schemes, technical communities can make a fundamental contribution both in the implementation and in promoting convergence towards a common normative framework (Lewis *et al.* 2021).

Media and public opinion can play a significant role too. To the extent that they manage to politicize the implications of certain technical choices (Santaniello *et al.* 2016), they can not only prompt public intervention and exert pressure on the company by leveraging its reputation but also transform fundamental rights into elements of economic rationality, making features such as privacy protection, encryption, fairness determinants of user/consumer behavior.

Of course, while all the previous actors could provide requirements, guidelines, or rules, most of the efforts in implementing these inputs into workable and operational arrangements fall back on the organizations that develop, deploy, and manage Ai technologies and applications.

To simplify, within organizations, we can distinguish between the responsibilities of management, which includes Ceos, senior, and middle management, and the technical staff, which encompasses team leaders, developers, and deployers. In short, the former is entrusted with defining Ai system requirements in accordance with regulatory compliance, stakeholders' expectations, and business needs. They also have to establish internal and external accountability mechanisms and oversee the entire Ai system's lifecycle. The latter should identify and implement the most appropriate technical specifications based on the established system requirements while taking into account stakeholders' perspectives.

Finally, it is worth noting the crucial role carried out by the concept of «trustworthy Ai» or «trustworthiness» in inscribing fundamental rights into the digitalization logic.

According to the International organization for standardization (Iso) definition, trustworthiness refers to the capability of a technological system to satisfy stakeholder expectations, mostly in terms of safety, reliability, and efficiency (Iso/Iec 2020). Trustworthiness is usually considered a key factor for the acceptance and adoption of a new technology, and to prevent rejection, loss of opportunities, and investments, as already occurred, for example, with nuclear power or genetically modified organisms in some contexts. Tru-

stworthiness is even more relevant in the case of Ai systems, which are based on the collection and processing of behavioral data, and mostly used to make decisions that can impact human beings.

In this view, fundamental rights could be considered as a necessary component to create digital technology that could be trusted and avoid rejection of the digitalization process. By imposing some limitations on the type of operations that Ai applications can perform and on the modalities in which data are collected and processed, Ai developers can create systems that respect stakeholder expectations and allow the digitalization logic to progress without endangering individuals and other social spheres.

4. The Artificial intelligence act: A concrete example of constitutionalization of Ai?

For several years now, European institutions have been paying significant attention to the field of Ai. They've been churning out a series of documents that seem to pave the way for one of the most intriguing examples of societal constitutionalism applied to the digital realm. This is particularly notable in terms of two key aspects of this process: embedding fundamental rights within the design of Ai systems, and the hybrid nature of the process itself. Moreover, fundamental rights are quite explicitly conceived as a mean to attain «trustworthy Ai», thereby ensuring Ai uptake across society.

Indeed, «building trust around the development and use of Ai» is a pillar of the European Commission strategy «Artificial intelligence for Europe» (Eu Com 2018), to be achieved by ensuring «an appropriate ethical and legal framework based on the Union's values and in line with the Charter of Fundamental Rights of the Eu» (*ibidem,* 3), a concept further developed in the communication, titled not by chance, «Building trust in human-centric artificial intelligence» (Eu Com 2019).

To this purpose, the Commission planned to leverage on a broader regulatory framework, including safety and product liability legislation, cybersecurity rules, and the General data protection regulation (Gdpr), to be adapted and integrated with Ai-specific initiatives.

In this regard, a key point was the drafting of the «Ethics guidelines for trustworthy Ai» (Ai Hleg 2019b), in which Ai trustworthiness is related to the compliance with a series of principles (respect for human autonomy, prevention of harm, fairness, explicability) and requirements (human agency and oversight; robustness and safety, privacy, transparency, diversity, environmental and societal well-being, accountability) grounded in fundamental rights, to

be implemented through a mix of technical and non-technical methods (for example explanation, testing and validation methods; quality of service indicators; standardization, certification).

The hybrid nature of the process becomes evident, as the ethics guidelines have been formulated by a multistakeholder High-level expert group on artificial intelligence established by the Commission. This group was comprised of 52 members, predominantly scholars, researchers, and developers situated in universities, public administrations, private sectors, and third sectors, often already involved in other significant initiatives in the field, such as the Ethically aligned design by Ieee or the Asilomar Ai principles (Palladino 2021).

Furthermore, the European Ai alliance, a broad and open multi-stakeholder platform boasting over 2700 members, was established to provide extensive input for the Ai High-level expert group's efforts and undergo a piloting phase to test the guidelines.

Drawing upon feedback from the piloting phase, which highlighted that «while a number of the requirements are already reflected in existing legal or regulatory regimes, those pertaining to transparency, traceability, and human oversight are not specifically covered under current legislation» (Eu Com 2020, 9), the «White paper on artificial intelligence» called for an Ai-specific regulatory framework in order to establish an «ecosystem of trust». The framework should ensure compliance with Eu rules, including the rules protecting fundamental rights and consumers' rights, thereby providing citizens with the confidence to embrace Ai applications and organizations with the legal certainty to innovate using Ai.

These objectives are being implemented by the Aia draft, which is currently in its final stage of approval[3].

According to the Commission's explanatory memorandum (Eu Com 2021), through the establishment of proportional requirements and responsibilities for all participants in the value chain, the proposal aims to advance and safeguard several rights protected by the European Charter, including: the right to human dignity (Art. 1), the right to privacy and personal data protection (Art. 7 and 8), the right to be free from discrimination (Art. 21), and to gender equality (Article 23), the upholding of freedom of expression (Art. 11) and assembly (Art. 12), the assurance of the right to a fair trial and an impartial judge, the presumption of innocence, and the rights of the defense (Art. 47 and 48); the rights of workers to equitable and just working conditions (Art.

[3] At the time of the final review of this article, a political agreement has been reached among the Commission, the Council, and the Parliament concerning the ultimate version of the Aia. Technical teams are currently dedicated to transposing the political agreement into the Act's text, before the final votes by both the Council and the Parliament.

31), consumer rights (Art. 38), the rights of minors (Art. 24), and the rights of individuals with disabilities (Art. 26).

A closer inspection of the requirements and mechanisms through which the Aia draft seeks to implement fundamental rights, reveals that it encompasses all the kinds of rules we deemed necessary for the constitutionalization process.

First of all, the Aia draft provides a series of exclusionary rights, consisting of the prohibition of certain types of applications deemed incompatible with the values of the Union and human rights (applications with manipulative or discriminatory purposes, real-time biometric recognition systems). Moreover, the draft sets out a series of specific requirements that high-risk Ai applications are required to fulfill mandatorily in order to prevent or mitigate their potential negative impact on European citizens and society. These demanded criteria significantly impact the architectures and the governance of these technologies. They include the implementation of risk management systems, quality control, data governance, and mechanisms for human oversight.

Furthermore, the Aia draft also provides for inclusionary rights. Transparency obligations are introduced that not only stipulate a range of information to be communicated to end users but also necessitate the establishment of automatic recording systems for system operations (logs) and other accessible technical documentation. These measures aim to make the choices made by the machine «explainable» and thus enable scrutiny of Ai systems developed by Big tech companies. Also most of the Aia's human oversight measures could be included into this category, inasmuch as they allow human operators to monitor and intervene in the functioning of Ai systems. Taken as a whole, these norms may ensure that Ai systems remain socially beneficial and «human-centric», preventing the exploitation of human dignity and autonomy to merely serve the digitalization logic.

Another set of rules concerned with the cybersecurity and robustness of Ai systems corresponds to what we defined as «security rules». These norms aim to ensure the Ai system's resiliency as regards «errors, faults or inconsistencies that may occur within the system or the environment in which the system operates, in particular due to their interaction with natural persons or other systems» and the «attempts by unauthorized third parties to alter their use or performance by exploiting the system vulnerabilities» (Eu Com 2021, art.15). In so doing they define the boundaries the range of authorized operation that can impact on Ai systems coming from outside, in order to safeguard the integrity and autonomy of the system itself.

But above all, the Aia seems to depict a hybrid constitutionalization process, through which constitutive and limitative rules are translated into code and embedded into the socio-technical structure of Ai systems. The

draft stipulates that the requirements set forth by the regulation are to be met through specific technical solutions developed in alignment with international standards and acquired scientific knowledge, rather than being strictly legislated: «the precise technical solutions to achieve compliance with those requirements may be provided by standards or by other technical specifications or otherwise be developed in accordance with general engineering or scientific knowledge at the discretion of the provider of the Ai system. This flexibility is particularly important, because it allows providers of Ai systems to choose the way to meet their requirements, taking into account the state-of-the-art and technological and scientific progress in this field» (Eu Com 2021, 13).

Indeed, in May 2023, the European Commission published a standardization request addressed to the European standardization organizations (Cen, Celec, Etsi) to develop a series of Harmonised standards to comply with the Aia's requirements.

Harmonised standards are standards specifically designed by a recognized European standards organisation to support Eu legislation, following a request from the European Commission.

They are published in the Official journal of the european union (Ojeu) and adhering to them carries a «presumption of conformity» with the essential requirements.

The Eu strategy offers several advantages. It prevents technical specifications prescribed by law from becoming rapidly outdated due to the pace of technological development.

Above all, it involves the people concerned with day-to-day Ai system development, deployment, and management in translating legislative requirements reflecting human-rights concerns into workable coding and organizational routines capable of structuring and setting the boundaries of Ai systems' operations.

This move is deemed crucial for the affirmation of a digital constitutionalism perspective, insomuch the technical community is best placed to translate ethical concerns and political claims into «code» and operative arrangements.

However, this hybrid approach to the governance of Ai also entails risks and challenges. One of the most problematic aspects of Aia draft indeed concerns the effectiveness of the established requirements.

Requirements apply only to a limited number of Ai applications, those classified as high-risk, and mostly through self-assessment processes and a preventive check before market entry is foreseen in only an extremely limited number of cases.

Secondly, the regulation permits the release of applications deemed high-risk while awaiting the development of standards that are meant to ope-

rationalize the requirements. This allows the use of these applications prior to the establishment of clear mechanisms to limit potential threats to individuals' fundamental rights. However, even when standards will be developed, it must be acknowledged that these are untested procedures, and it cannot be assumed that we will be able to develop procedural and technical arrangements capable of mitigating the risks associated with certain types of Ai applications to an acceptable level.

Thirdly, standard-setting organizations are essentially private non-profit entities and may be susceptible to capture by special interests. The influence of private companies within these organizations could lead to the softening of certain provisions or the oversight of particular aspects.

Besides that, they are technical communities whose mindset might lead to adopt a technological solutionism approach, reducing complex ethical matters into mathematical representations (Lee and Floridi 2021; Morley *et al.* 2021; Palladino 2023a) that may not necessarily align with the actual needs and concerns of individuals and society.

This leads us to consider that the hybrid nature of digital constitutionalization processes also conceals a darker side, which is the mutual interdependency of different institutional logics. This results in the integration of private schemes and concerns that could undermine mechanisms of social control and accountability.

Further threats to the protection of fundamental rights arise from the exemptions granted by EU institutions from the regulatory requirements for military, defense, and national security purposes, especially in the context of migration and border control.

For this reason, as outlined in the previous paragraph, it is deemed crucial that other actors, such as the media and public opinion, also take part in the digital constitutionalization process, ensuring that the implications of certain technical choices are discussed in the public debate.

5. Conclusions

The policy documents produced by the European Union institutions in deploying their own Ai strategy provide a compelling example of how the essential elements we have identified for the effective constituttionalization of the digital subsystem can be concretely formalized and implemented.

Firstly, there is the centrality of fundamental rights, understood as counter-institutions capable of countering the exploitative nature of the digitization logic in society.

Secondly, European Ai policies involve the implementation of fundamental rights by defining requirements and control systems that must be incorporated into the architectures and design of Ai systems. At the same time, however, the operationalization of fundamental rights is not conceived as an exclusive prerogative of public authorities. On the contrary, it is achieved through the involvement of a variety of actors, both in the development of legislative production and in the design of the most appropriate solutions to translate the legal requirements into specific technical terms. This latter aspect reflects the inherently hybrid nature of processes of constitution-making in the digital sphere.

This article also warned about potential risks stemming from the European strategy. The constitutionalization of Ai and the digital system could be undermined by special interests capture, the limitedness of pre-market assessments, the lack of well-established standards and an overreliance on purely techno-solutionist approaches.

These considerations suggest that further research is needed on how to translate rights, principles and requirements into operational standards, as well as on the reliability of the methods and procedures indicated in those standards, in order to close the principles-to-practice gap. Furthermore, researchers and developers should pay attention to including the perspective of the people affected by Ai applications in Ai design and building governance mechanisms, ensuring that adopted technical solutions actually correspond to societal needs.

Recalling that agreed-upon and well-established standards to deal with Ai technologies' social and political implications do not exist yet, policymakers should carefully consider, while actively fostering and investing in Ai research, the suspension of commercialization for at least the most concerning high-risk Ai applications.

The hybrid constitutionalization of Ai governance could be further strengthened through the establishment of a dedicated regulatory authority, multistakeholder in nature, and possessing the necessary expertise. This authority would be tasked with overseeing the effective compliance of solutions proposed by standard-setting organizations, as well as those implemented by public and private organizations involved in the development and deployment of Ai with the statutory requirements for trustworthy and fundamental rights based Ai.

However, the most significant driving force in this process would be a broad awareness of the social and political implications and impacts of the technical specifications of Ai applications. This awareness would, on one side, encourage public opinion to politicize the values and political goals to

be incorporated into Ai socio-technical architecture and, on the other side, make individuals involved in the design, deployment, and management of Ai systems more accountable for the consequences of their work. To this purpose, the dialogue between different disciplines and professions should be encouraged, and new professional roles and curricula integrating computer and social sciences should be created. Furthermore, digital literacy should be widely spread in educational programs at all levels. This initiative ensures that individuals are equipped with the necessary tools to critically engage with, evaluate, and contribute to the ongoing advancements in Ai technology. By incorporating digital literacy into educational curricula, society can ensure that people that is not only technologically skilled but also well-versed in considering the social and political implications inherent in Ai applications.

6. Acknowledgements

The authors has received funding from the MIUR the Italian Minister of University and Research under the project "Cybersecurity (As A) Public Policy The Institutionalization Of Platform And Network Security In The Eu And Italy" based at the University of Salerno (Prot. 2020X5LAK7); and the European Union's Horizon 2020 Research and Innovation Programme under the HUMAN+COFUND Marie Skłodowska-Curie grant agreement No. 945447.

References

Ai Hleg (2019), *A Definition Of Ai: Main Capabilities And Disciplines*, European Commission High Level Expert Group on Ai.

Ai Hleg (2019), *Ethics Guidelines For Trustworthy Ai, European Commission High Level Expert Group on Ai*, accessed 24 May 2023 at https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Bollier, D. (2018), *Artificial Intelligence, The Great Disruptor*, Washington D.C., The Aspen Institute.

Brownsword, R., Scotford, E. and Yeung, K. (2017), (eds.) *The Oxford Handbook of Law, Regulation and Technology*, First edition, New York, Oxford University Press.

Eu Com (2018), *Artificial Intelligence for Europe*, European Commission.

Eu Com (2019), *Building Trust in Human-Centric Artificial Intelligence*, European Commission.

Eu Com (2020), *White Paper on Artificial Intelligence*, European Commission.

Eu Com (2021), *Proposal for a regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence* (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, Com (2021) 206 final, European Commission.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., Srikumar, M. and Berkam Center (2020), *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for Ai*, in «SSRN Electronic Journal». doi: 10.2139/ssrn.3518482.

Floridi, L., Cowls, J, Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P. and Vayena, E. (2018), *Ai4People — An Ethical Framework for a Good Ai Society: Opportunities, Risks, Principles, and Recommendations*, in «Minds and Machines», 28(4), pp. 689-707.

Gardbaum, S. (2009), *Human Rights and International Constitutionalism*, in in J. L. Dunoff and J.P. Trachtman (eds), *Ruling the World?: Constitutionalism, International Law, and Global Governance*, Cambridge, Cambridge University Press.

George, É. (2020), *Digitalization of Society and Socio-Political Issues 2: Digital, Information, and Research*, Hoboken, John Wiley & Sons.

Greene, D., Hoffmann, A. L. and Stark, L. (2019), *Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning*, https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1261&context=hicss-52.

Grimm, D. (2016), *Constitutionalism: Past, Present, and Future*, Oxford, Oxford University Press.

Hallensleben, S., Hustedt, C., Fetic, L., Fleischer, T., Grünke, P., Hagendorff, T., Hauer, M., Hauscke, A., Heesen, J., Herrmann, M., Hillerbrand, R., Hubig, C., Kaminski, A., Krafft, T., Loh, W., Otto, P. and Puntschuh, M. (2020), *From Principles to Practice. An Interdisciplinary Framework to Operationalise Ai Ethics*. doi: 10.13140/RG.2.2.31757.97764.

Iso/Iec (2020), *Overview of Trustworthiness in Artificial Intelligence*.

Jobin, A., Ienca, M. and Vayena, E. (2019), *The Global Landscape of Ai Ethics Guidelines*, in «Nature Machine Intelligence», 1(9), pp. 389–99.

Latzer, M. and Just, N. (2020), *Governance by and of Algorithms on the Internet: Impact and Consequences*, in «Oxford Research Encyclopedia of Communication», Oxford, Oxford University Press.

Lee, M. S. A. and Floridi, L. (2021), *Algorithmic Fairness in Mortgage Lending: from Absolute Conditions to Relational Trade-offs*, in «Minds and Machines», 31(1), pp. 165-91.

Lessig, L. (2009), *Code 2.0*, New York, Basic Books.

Lewis, D., Filip, D. and Pandit, H. J. (2021), *An Ontology for Standardising Trustworthy Ai*, in Hessami, A. G. and Shaw, P. (eds), *Factoring Ethics in Technology, Policy Making, Regulation and Ai*, in «IntechOpen». doi: 10.5772/intechopen.97478.

Liddicoat, J. and Doria, A. (2012), *Human Rights and Internet Protocols: Comparing Processes and Principles*, accessed 1 January 2023 at www.internetsociety.org.

Luhmann, N. (2010), *Potere e complessità sociale*, Milano, Il Saggiatore.

Mittelstadt, B. (2019), *Principles Alone Cannot Guarantee Ethical Ai*, in «Nature Machine Intelligence», 1(11), pp. 501–7.

Morley, J., Morton, C. Karpathakis, K., Taddeo, M. and Floridi, L. (2021), *Towards a Framework for Evaluating the Safety, Acceptability and Efficacy of Ai Systems for Health: An Initial Synthesis*, in SSRN. doi: 10.2139/ssrn.3826358.

Musiani, F., Cogburn, D. L., DeNardis, L. and Levinson, N. S. (2016), *The Turn to Infrastructure in Internet Governance*, New York, Palgrave MacMillan.

Palladino, N. (2021), *The Role of Epistemic Communities in the «Constitutionalization» of Internet Governance: The Example of the European Commission High-Level Expert Group on Artificial Intelligence*, in «Telecommunications Policy», 45(6), 102149.

Palladino, N. (2023a), *A "Biased" Emerging Governance Regime for Artificial Intelligence? How Ai Ethics Get Skewed Moving from Principles to Practices*, in «Telecommunications Policy», 47(5); doi: : 10.1016/j.telpol.2022.102479.

Palladino, N. (2023b), *The Blind Watcher: Accountability mechanisms in the Artificial Intelligence Act*, in «The Quest for Ai Sovereignty, Transparency and Accountability», Rio de Janeiro: Data and Artificial Intelligence Governance Coalition of the United Nations Internet Governance Forum, pp. 145-60.

Peters, A. (2006), *Compensatory Constitutionalism: The Function and Potential of Fundamental International Norms and Structures*, in «Leiden Journal of International Law», 19 (3), pp. 579–610.

Rakova, B., Yang, J., Cramer, H. and Chowdhury, R. (2021), *Where Responsible Ai meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices*, in «Proceedings of the Acm on Human-Computer Interaction», 5(1), pp. 1-23.

Renda, A. (2019), *Artificial Intelligence: Ethics, Governance and Policy Challenges: Report of a Ceps Task Force*, Brussels, Centre for European Policy Studies.

Santaniello, M., De Blasio, E., Palladino, N., Selva, D., De Nictolis, E. and Perna, S. (2016), *Mapping the Debate on Internet Constitution in the Networked Public Sphere*, in «Comunicazione politica», 17(3), pp. 327-54.

Santaniello, M., Palladino, N., Catone, M. C. and Diana, P. (2018), *The Language Of Digital Constitutionalism and the Role of National Parliaments*, in «International Communication Gazette», 80 (4), pp. 320-336.

Schiff, D., Borenstein, J., Biddle, J. and Laas, K. (2021), *Ai Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection*, in «Ieee Transactions on Technology and Society», 2(1), pp. 31-42.

Schiff, D., Rakova, B., Ayesh, A., Fanti, A. and Lennon, M. (2021), *Explaining the Principles to Practices Gap in Ai*, in «Ieee Technology and Society Magazine», 40(2), pp. 81-94.

Sciulli, D. (1992), *Theory of Societal Constitutionalism: Foundations of a Non-Marxist Critical Theory*, Cambridge, Cambridge University Press.

Shneiderman, B. (2020), *Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered Ai Systems*, in «Acm Transactions on Interactive Intelligent Systems», 10(4), pp. 1-31.

Teubner, G. (2004), *Societal Constitutionalism: Alternatives to State-Centred Constitutional Theory?*, in «Transnational Governance and Constitutionalism», Hart Publishing.

Teubner, G. (2011), *Transnational fundamental rights: Horizontal effect?*, in «Netherlands Journal of Legal Philosophy», 40(3), pp. 191-215.

Teubner, G. (2012), *Constitutional Fragments: Societal Constitutionalism and Globalization*, Oxford, Oxford University Press.

Thornhill, C. (2011), *A Sociology of Constitutions: Constitutions and State Legitimacy in Historical-Sociological Perspective*, Cambridge, Cambridge University Press.

Turner, J. (2018), *Robot Rules: Regulating Artificial Intelligence*, London, Palgrave MacMillan.

Van Dijck, J. (2014), *Datafication, Dataism and Dataveillance: Big Data between Scientific Paradigm and Ideology*, in «Surveillance & Society», 12(2), pp. 197-208.

Van Dijk, N. and Casiraghi, S. (2020), *The "Ethification" Of Privacy And Data Protection Law In The European Union. The Case Of Artificial Intelligence*, Bruxelles Privacy Lab Working Papers, 6(22).

Wet, E. D. (2006), *The International Constitutional Order*, in «International & Comparative Law Quarterly», 55(1), pp. 51-76.

Whittaker, M., Crawford, K. and Dobbe, R. (2018), *Ai Now Report 2018*, Ai Now Institute, accessed 24 May 2023 at https://ainowinstitute.org/publication/ai-now-2018-report-2.

Yeung, K., Howes, A. and Pogrebna, G. (2020), *Ai Governance by Human Rights-Centered Design, Deliberation, and Oversight*, in F. Pasquale and S. Das (eds.), *The Oxford Handbook of Ethics of Ai*, New York, Oxford University Press.

Zuboff, S. (2019), *The Age of Surveillance Capitalism*, New York, Public Affairs.