Information Systems Frontiers (2023) 25:71–102
https://doi.org/10.1007/s10796-022-10300-6

From the "rush to ethics" to the "race for governance" in Artificial Intelligence

Vasiliki Koniakou[1]

Accepted: 30 May 2022 / Published online: 28 June 2022

Abstract
This paper engages with the emerging field of Artificial Intelligence (AI) governance wishing to contribute to the relevant literature from three angles grounded in international human rights law, Law and Technology, Science and Technology Studies (STS) and theories of technology. Focusing on the shift from ethics to governance, it offers a bird-eye overview of the developments in AI governance, focusing on the comparison between ethical principles and binding rules for the governance of AI, and critically reviewing the latest regulatory developments. Secondly, focusing on the role of human rights, it takes the argument that human rights offer a more robust and effective framework a step further, arguing for the necessity to extend human rights obligations to also directly apply to private actors in the context of AI governance. Finally, it offers insights for AI governance borrowing from the Internet Governance history and the broader technology governance field.

Keywords Artificial Intelligence (AI) · Algorithms · Human Rights · AI Ethics · AI governance · Science and Technology Studies (STS) · Internet Governance (IG)

1 Introduction

During the last two decades, Artificial Intelligence (AI) and various kinds of algorithms have rapidly become integral for numerous sectors and industries, from recommendation services, online content moderation, and advertising to the provision of healthcare, and policy-making (Aizenberg & van den Hoven, 2020; Cath, 2018; Gerards, 2019). Bearing the promise of swift, rational, objective, and efficient decision-making, they are often employed to inform or make critical decisions in a constantly growing number of central and socially consequential domains, including but not limited to the justice system (Giovanola & Tiribelli, 2022; Yeung et al., 2019). From this angle, enabling data-driven, automated decision-making, advanced reasoning and processing features, AI and algorithms are considered to offer new opportunities for individuals, and the society at large, to improve and augment their capabilities and wellbeing (Floridi et al., 2018; Smuha, 2021b; Tegmark, 2017). They are also expected to contribute to global productivity (Agrawal et al., 2019), the achievement of sustainable development goals (Pedemonte, 2020; Vinuesa et al., 2020), as well as broader environmental objectives, from green technologies (Elshafei & Negm, 2017; Mishra et al., 2021) to climate change (Cowls et al., 2021).

However, apart from the benefits they offer, we have gradually come to realise that AI systems[1] and algorithms also pose a wide range of risks (Gerards, 2019; Radu, 2021; Taeihagh, 2021) and ethical challenges (Stahl, 2021). Instances of discrimination and bias (Binns, 2017; Borgesius, 2018; Lambrecht & Tucker, 2018), online disinformation and opinion manipulation (Allen & Massolo, 2020; Cadwalladr, 2020), private censorship (Gillespie, 2014, 2018) pervasive monitoring (Feldstein, 2019a; Kambatla et al., 2014), as well as adverse job market effects (Agrawal et al., 2019; Vochozka et al., 2018), have raised serious concerns. For example, algorithmic processes and automated decision-making may reinforce and widen social inequalities, (Chander & Pasquale, 2016; O'Neil, 2016; Risse, 2018), as the same key functionalities that lead to more accurate and informed decisions may also perpetuate bias and discrimination (Floridi et al., 2018; Miller, 2020a, b; Murray et al.,

✉ Vasiliki Koniakou
koniakou@eltrun.gr

[1] Department of Management Science & Technology, ELTRUN Research Center, Information Systems Technology Laboratory (ISTLab), Athens University of Economics and Business, Evelpidon 47A, 11362 Athens, Greece

[1] The terms AI and AI systems are used interchangeably, under the definition offered in Section 2.

2020), either due to fragmented or non-representative databases or because algorithms tend to reproduce the prejudices already existing in our societies (Miller, 2020a, b). Additionally, AI may have a considerable impact on how individuals deliberate and act, affecting our autonomy (Laitinen & Sahlgren, 2021; Vesnic-Alujevic et al., 2020), or even undermining self-determination (Danaher, 2018). This way, AI systems and algorithms may also foster repression, and authoritarian practices (Feldstein, 2019b). Consequently, it is widely recognised that if such technologies "are poorly designed, developed or misused, they can be highly disruptive to both individuals and society" (Fukuda-Parr & Gibbons, 2021).

Simultaneously, as AI systems are increasingly becoming embedded in several social contexts, they also become highly relevant for human rights[2] (Aizenberg & van den Hoven, 2020; Gerards, 2019). For instance, algorithms are implemented to make or inform critical decisions that define individuals' suitability, or entitlement to life-affecting opportunities and/or benefits (McGregor et al., 2019; Yeung et al., 2019). They are used to screen applicants' CVs and make recommendations for study or job openings. They are also employed to assess individuals' income and credit scores to determine their access to the credit system, or their eligibility for state subsidies. Such decisions affect or even interfere with a wide array of human rights, such as freedom from discrimination, right to work, and right to education (Latonero, 2018; Raso et al., 2018a). Additionally, algorithms used for online content moderation have been frequently accused of private censorship, or opinion-shaping (Borgesius, 2018; Gillespie, 2020; Gorwa et al., 2020), while massive biometrical surveillance and facial recognition algorithms threaten our privacy in an unprecedented way (Smith & Miller, 2021). Ultimately, AI and algorithms have been also introduced to the justice system, giving rise to concerns that either in the form of Law Tech (Kennedy, 2021), or as prediction and risk assessment mechanisms, they may affect or interfere with the right to equality before the law, fair trial, freedom from arbitrary arrest, detention, and exile, or even with the rights to liberty, and personal security (Asaro, 2019b; Završnik, 2020).

The accelerating pervasiveness and ubiquity of AI systems, combined with the growing public concern about their ethical, social, and human rights implications have brought to the forefront questions related to the steering of such technologies towards socially beneficial ends (Floridi et al., 2018; Stahl et al., 2021a; b; Taddeo & Floridi, 2018). Initially, the main response to concerns about the negative aspects of AI and algorithms has been a turn to ethics (Smuha, 2021a), or as Boddington describes it, "a rush to produce codes of ethics for AI, as well as detailed technical standards" (Boddington, 2020). Governments, intergovernmental organisations, public actors, civil society groups, and private sector stakeholders turned to value-based norms in search of "ethical AI" or "responsible AI" (Fukuda-Parr & Gibbons, 2021; Smuha, 2021a), to ensure the ethical and societally beneficial design, development, and deployment of AI (AI-DDD). Currently, it seems that we are at a turning point, as governments and intergovernmental organisations have taken several initiatives toward developing specific regulatory instruments. Essentially, the focus on steering AI and algorithms toward ethical and societally beneficial ends is shifted from primarily relying on ethical codes to gradually introducing binding legislation (Maas, 2022; Smuha, 2021b; Taeihagh, 2021). However, although several modes of governance have been suggested (Almeida et al., 2021; Smuha, 2021b), AI governance remains a relatively underdeveloped field of academic research and policy practice (Cihon et al., 2020; Taeihagh, 2021).

In this context, human rights are highly relevant and have been frequently discussed. The impacts and implications of AI, algorithms and automated decision-making have emerged as important areas of human rights concern during the last decade. Moreover, human rights have been proposed as a source of ethical standards (Gerards, 2019; Yeung et al., 2019) or design principles (Aizenberg & van den Hoven, 2020; Umbrello, 2022). They have been also suggested as offering a better alternative than ethics in terms of an accountability framework (McGregor et al., 2019). Additionally, some researchers argue that a more direct application of human rights law can provide clarity and guidance in identifying potential solutions to the AI challenges (Stahl et al., 2021a, b). From a similar point of view, human rights have been suggested as governance principles, to "underlie, guide, and fortify" an AI governance model (Smuha, 2021a). Most notably, they have been identified by the High-Level Expert Group on AI (AI HLEG), as offering "the most promising foundations for identifying abstract ethical principles and values, which can be operationalised in the context of AI" (AI HLEG, 2019b). Similarly, the United Nations (UN) has highlighted the role of the Universal Declaration of Human Rights (UDHR) in offering a basis for AI principles (Hogenhout, 2021).

Considering the turn to governance in conjunction with the growing relevance of human rights in the AI-DDD and AI governance discourse, this paper engages with the question "how should we regulate AI?" from a Law and Technology and Science and Technology Studies (STS) point of view. Grounded in international human rights law theory,

[2] The term "human rights" is used here as a synonym to "fundamental rights" as this is the more common and familiar locution employed in international law. For the source of the "human rights gap" as a concept see Zalnieriute, M., & Milan, S. (2019). Internet Architecture and Human Rights: Beyond the Human Rights Gap. Policy and Internet, 11(1), 6–15.

and business and human rights scholarship, the study seeks to supplement the emerging AI governance literature from three angles. First, it focuses on the shift from steering AI through ethics and means of soft regulation to AI governance through particular national and intergovernmental binding instruments. In that context, it offers a bird-eye overview of the developments in AI governance, focusing on the comparison between ethical principles and binding rules for the governance of AI, and critically reviewing the latest regulatory developments.

Secondly, turning to the role of human rights in steering AI, through this paper I wish to take a step further the argument that human rights offer a more robust and effective framework to ensure the AI-DDD for the benefit of society. More specifically, building upon human rights and business discourse, I argue that human rights may offer more than aspirational and normative guidance in AI-DDD as well as in AI governance, if they are employed as concrete legal obligations that directly apply to both public and private actors. In my view, in an increasingly AI-mediated world the direct application of human rights obligations to private actors in terms of AI governance or through a new human rights treaty for the private sector and AI, constitutes a critical step to adequately protect human rights. Moreover, it is a meaningful way to ensure that such technologies will contribute to the flourishing of society (Gibbons, 2021; Stahl et al., 2021a, b). This aspect of direct human rights application in AI governance is not yet addressed in the relevant literature. This way, I offer a new perspective to AI governance and human rights research, providing arguments from the direct horizontality discourse and the business and human rights field.

Finally, looking beyond human rights, I offer insights to the emerging AI governance scholarship building upon the rich tradition of theories of technology, technology governance and the Internet Governance (IG) history. I start from the observation that whereas the field of AI raises very deep and broad philosophical questions, the ethical and governance questions are not necessarily new (Niederman & Baker, 2021), nor exclusively inherent to AI (Stahl et al., 2021a, b). Additionally, the discussion regarding the steering of new and disruptive technologies towards ethical and societally beneficial ends is part of a broader discourse on the relationship between technology and society (Benedek et al., 2017; Bucchi, 2009; Strobel & Tillberg-Webb, 2009) that is almost as old as human history (Black & Murray, 2019). From this angle, I argue for the necessity to examine what is at stake in AI governance and seek insights by studying the governance trajectory of other major disruptive technologies, such as the Internet.

The rest of the paper is structured as follows: In Section 2 I discuss the definition of the key terms and concepts. In Section 3, I focus on the turn from ethics to governance. I argue that although ethics are vital to steer AI-DDD to the benefit of society they are not sufficient to ensure it. Moreover, regardless of the positive steps in AI governance adequate human rights protection is not necessarily ensured yet. In Section 4, I aim to advance the discourse regarding the role of human rights, suggesting the direct application of human rights obligations to private actors in the context of AI governance. Before offering my closing remarks (Section 6), in Section 5 I ask the question of what AI governance can learn from IG and the broader field of technology governance, addressing the relevance of IG as a source of insights, and three critical points of consideration building upon the IG experience.

2 Key terms and concepts

Given that conceptual clarity and terminological consistency are yet to be achieved in AI literature (Collins et al., 2021; Larsson, 2020; Surden, 2020), it is essential to start by clarifying the way the key terms will be used. Moreover, as will be further discussed in Sections 3 and 5, the way AI systems are defined is critical for policymaking (Bhatnagar et al., 2018; Collins et al., 2021). Additionally, the field of international human rights may be significantly complex. The focus here is on the UN human rights system, which comprises the UN human rights principles along with the institutional mechanism to encourage and monitor compliance by the states (Buergenthal, 2006).

2.1 Artificial Intelligence

Almost like any other term that is shared across several different disciplines, receiving also a fair share of public and media attention, AI is riddled with multiple interpretations (Scherer, 2016), covering different aspects (Haenlein & Kaplan, 2019), functions, and functionalities, ranging from the capabilities of a smartphone to those of self-operating vehicles, or robots (Ertel, 2017; Risse, 2018). Through their systematic literature review of the field, Collins et al. (2021) identified a large variety of different definitions. According to their findings, apart from the lack of cohesion, a noteworthy observation is that most definitions tend to focus on what AI systems are capable of, instead of what AI actually is. The term can be traced back to 1956 when AI was described by John McCarthy as "the science and engineering of making intelligent machines"(McCarthy, 2018). More recent definitions describe it as "the ability of a machine to perform cognitive functions that we associate with human minds, such as perceiving, reasoning, learning, interacting with the environment, problem-solving, decision-making, and even demonstrating creativity"(Rai et al., 2019) or the process that "enables the machine to exhibit human intelligence,

including the ability to perceive, reason, learn, and interact, etc."(Russel & Norvig, 2020).

Outside academia, the EU Commission defined it as "a collection of technologies that combine data, algorithms and computing power" (European Commission, 2020), while the HLEG described it as "software (and possibly also hardware) systems designed by humans 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 can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions" (AI HLEG, 2019). Using human cognition as a point of reference, and highlighting the wide range of forms and applications, in his report to the EU Parliament, Szczepański, (2019) defined it as the "term used to describe machines performing human-like cognitive processes such as learning, understanding, reasoning and interacting. It can take many forms, including technical infrastructure (i.e. algorithms), a part of a (production) process, or an end-user product."

The common thread amongst the majority of the descriptions and proposed definitions is the increasing capacity of machines to perform specific cognitive functions, roles and tasks currently performed by humans (Dwivedi et al., 2021). Thus, a key component of the definition of AI is the element of intelligence. Within the AI discourse intelligence is a somehow elusive term (Bryson & Theodorou, 2019; Ray, 2021). Although researchers suggest that in the context of AI intelligence should not be understood in terms of human intelligence (Bryson & Theodorou, 2019; Collins et al., 2021), the mere reference to intelligence almost automatically brings to mind human-related and human-premised associations, while it is commonly closely connected with notions such as sentience, sensibility, consciousness, self-awareness, or intentionality (Ertel, 2017). While such an anthropocentric approach is an integral part of human cognition, it may be still rather misleading or even unsuitable (Sætra, 2021). As Russel and Norvig (2020) note, even if a certain degree of anthropomorphism is not only expected but also descriptively helpful, the emphasis on "human-like" intelligence may be deceiving. Stemming from this observation, Bryson argues that adopting a simple definition of intelligence is of at most importance for developing the appropriate governance mechanisms (Bryson, 2020).

Yet, AI is not a single nor a stand-alone technology but "a deeply technical family of cognitive technologies" (Kuziemski & Misuraca, 2020), which include various techniques, subfields and applications (Gasser & Almeida, 2017), from machine learning, and natural language processing, to robotics and systems of super-intelligence (Raso et al., 2018a, b; Stahl et al., 2021a, b). In turn, the term "artificial intelligence" is a collective noun, an "umbrella term" employed to describe a cluster of technologies and applications (Dubber et al., 2020; Latonero, 2018), linked to and embedded in other technologies (Stahl et al., 2021a, b). The relevance of a variety of different technologies, processes, and procedures makes it difficult to clearly delineate artificial intelligence in a single way and across all contexts, particularly as often distinguishing between them is considerably hard (Stahl, 2021). Moreover, due to the impressive pace at which formerly cutting-edge innovations become mundane, "losing the privilege of being categorized as AI", it is not always clear which technologies can be labelled as AI (Raso et al., 2018a).

In this paper, AI is perceived and studied as a time and space contextualised, enhanced computation process, which is intentionally designed, based on predefined rules and data input (Bryson, 2020). Using the term I denote the various forms of software (and their physical carrier whenever relevant) designed to perform such enhanced computational functions, including problem-solving, pattern recognition, analysis, recommendation and decision-making (Yavar 2018). Studying the ethical and societally beneficial AI-DDD, I embrace AI as a general-purpose technology (Dafoe, 2018; Trajtenberg, 2018), the disruptive effects of which may be "as transformative" as the industrial revolution (Gruetzemacher & Whittlestone, 2022). I focus equally on the externalities of AI systems and algorithmic procedures (i.e. their impact and effects, including unintended outcomes), and the impactful role of the actors involved in their design, development, and deployment, wishing to highlight the centrality of their agency.

2.2 Human Rights

The UN has been in charge of initiating the drafting of the first major international instrument containing a specific set of rights reserved for all human beings, in response to the atrocities of the Second World War (Kanalan, 2014—see also the preamble to the UDHR). Accepted on December the 10th 1948, the UDHR constitutes a foundational text in the history of human rights and combined with the two "twin" covenants, the International Covenant on Economic, Social and Cultural Rights (ICESCR), and the International Covenant on Civil and Political Rights (ICCPR) they constitute the core of the "UN human rights system." The rights enshrined in these treaties, serving as the basic moral and legal entitlements of every human being, have a dual function, both as legal requirements under international law, and as norms encapsulating and reflecting moral, ethical, and social values (Bilchitz, 2016a; Tasioulas, 2013). Today, human rights are deeply rooted in contemporary politics and

law, recognised in political practice and legal institutions globally (Etinson, 2018). In this context, the UN human rights system constitutes the key benchmark of international human rights protection, as most of the 193 Member-States of the UN have ratified at least one major human rights treaty, including the UNDHR.[3]

Thus, even though not uncontroversial (Hopgood, 2018), nor equally applied universally (Tharoor, 2000), human rights represent a rare sum of principles and norms that are widely shared and institutionalised globally. Under the 'tripartite typology of human rights obligations,' the subjects of the international human rights obligations ought to 'respect, protect and fulfil' human rights (Asbjørn, 1987). In short, "respecting" human rights entails the obligation to refrain from taking any action that would infringe upon the enjoyment of these rights; "Protecting" human rights refers to the duty to prevent violations of human rights, via taking concrete measures; while "fulfilling" human rights relates with the duty to facilitate the realisation of and enjoyment of human rights[4] (Alston & Quinn, 2017). Yet, who is the subject of these obligations? The answer to this question is closely related to a crucial and increasingly debated characteristic of international human rights that is also a key point of this article.

Human rights are vertical in nature. The term "vertical" implies that the state, placed on a higher field than the individual, is the obligation-holder, while the individual is the right-holder (Lane, 2018a). The rationality behind the vertical application, which is closely related to the historical trajectory of human rights (Witte, 2009), reflects the view that the state is the key perpetrator of individuals' rights and freedoms, particularly given the far greater power it possesses. Building upon the power asymmetry between the state and the individuals, human rights are intended to serve as a shield, protecting people from the power of the state (Dawn & Fedtke, 2008). The "vertical nature/effect" is also the result of the international law fundamental principles. Based on international law, for an actor to be directly bound by international human rights it should be recognised as a subject of international law (Alston, 2005; Clapham, 2006; Kanalan, 2014). Given that the states constitute the original subjects and primary actors of international law (Bilchitz, 2016a; Kampourakis, 2019; Lane, 2018b),[5] only the states are directly bound by international human rights law and treaties.

Due to the vertical application of international human rights law, and the so-called "state-centric model"(de Aragão & Roland, 2017), private law relationships are broadly considered to be immune from direct human rights effects (Cherednychenko, 2007). Therefore, the protection of human rights between actors that lay on the same level, in other words between private individuals or entities, depends on domestic legislation and the extent to which the states have translated human rights into their national legal order, establishing the necessary framework and the structures required to ensure they are sufficiently protected. The degree to which this model is adequate and effective in the contemporary world is contested and will be addressed in Section 4. For now, it would suffice to note that the increasing power asymmetries between individuals and corporations, most prominently Transnational Corporations (TNCs), and the shifted balance of power to negatively affect human rights, combined with the failure of states to foster meaningful mechanisms to protect human rights at a national level, safeguarding access to meaningful remedy and redress, has intensified the discussions over a paradigm shift in international human rights law (Kampourakis, 2019; Zamfir, 2018).

3 Steering AI: From Ethics to Governance

3.1 The rush to ethics

3.1.1 Ethical principles and guidelines as the early response to AI challenges

As mentioned in the Introduction, the early response to the growing recognition that AI and algorithms may also have adverse impacts has been a rush to develop and promote a wide range of value-based norms, ethical codes, and declarations (Boddington, 2020; Radu, 2021; Smuha, 2021a). Even though Floridi and Cowls (2019) rightfully observe that the ethical debate is almost as old as the emergence of AI as a field of research, there has been a remarkable proliferation of ethical principles related to AI since 2016 (Winfield et al., 2019), as the application of AI and algorithms drastically increased during the mid-2010s (Bryson, 2019). From that moment on, harnessing the potential of AI while mitigating, or at least balancing its negative effects and harmful consequences became a pressing priority, centred around the need to make AI more "ethical". Thus, AI Ethics came under the limelight, since almost all the major stakeholders eagerly engaged in an unofficial competition to develop and publish their own set of ethical norms, soon as ethical concerns related to AI gained momentum. Governments and intergovernmental organisations formed ad hoc expert committees, tasked to offer policy recommendations. Simultaneously, ethical

[3] According to the UN there are 195 sovereign states in the world, 193 of which are members of the UN.

[4] See Maastricht Principle No. 6, as cited in the Office of the United Nations High Commissioner for Human Rights (n 35) 15.

[5] Codified in Codified in Vienna Convention on the Law of Treaties, Jan. 27, 1980, 1155 U.N.T.S. 331.

guidelines have been developed by several private actors, companies, research entities, think tanks, and policy bodies (Jobin et al., 2019).

Hence, the term AI Ethics practically refers to the field of moral principles, ethical guidelines, codes, frameworks and declarations, intended to inform, guide and secure the ethical AI-DDD across several different sectors (Muller, 2020a, b; Whittlestone et al., 2019). At a regional level, the EU appointed the HLEG (AI HLEG, 2019a, b), to gather expert input from diverse stakeholders groups to produce guidelines for the ethical use of artificial intelligence, emphasising the key role of the EU Charter of fundamental rights for informing and guiding AI development. Similarly, the CoE established in February 2018 a Task Force on AI within the European Commission for the Efficiency of Justice (CEPEJ) to "lead the drafting of guidelines for the ethical use of algorithms within justice systems, including predictive justice", and an Expert Committee on human rights dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT). Outside Europe, the UN established the Ad Hoc Expert Group on the Ethics of AI (Hogenhout, 2021), while the Organisation for Economic Co-operation and Development (OECD) appointed the expert group on AI in Society (OECD, 2019). Each of these expert groups produced a distinct set of principles and ethical guidelines for the steering of AI and algorithms towards ethical, and by extension, societally beneficial ends.

At a national level, more than thirty countries, such as the UK, France, Germany, China, Japan, Canada, Finland, and the United Arab Emirates (UAE) have started drafting or even implementing national AI strategies (Dutton, 2018), centred around notions such as "ethical implementation", "good and trustworthy AI" (Jobin et al., 2019; UK Government, 2021). Additionally, professional bodies, civil society organisations, and various think-tanks developed ethical principles and guidelines, aimed at guiding practitioners, and shaping AI-DDD to the benefit of individuals and the society at large. The Future of Life Institute developed "The Asilomar AI Principles" in January 2017. At the same time, the Association for Computing Machinery US Public Policy Council (USACM) published "The Statement on algorithmic transparency and accountability", while in March 2017 the Institute of Electrical and Electronics Engineers (IEEE) provided "The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems." Slightly later the same year, the University of Montreal published "The Montreal Declaration for Responsible AI." Several non-state actors, and most prominently a handful of major technology companies, have also issued their own ethical declarations, principles, and ethical codes, expressing their commitment to the ethical and responsible development and deployment AI and algorithms, either individually (see for example Facebook, 2020; IBM, 2020; Microsoft, 2017; Pichay, 2018; Pinjušić, 2022) or jointly, (Partnership on AI, 2020).

Reviewing the reaction to concerns over the negative implications of AI, it seems that the need for some basic rules to guide AI-DDD was widely recognised, while ethics have largely dominated the discussions (Radu, 2021). Currently, the Algorithm Watch's AI Ethics Global Inventory includes more than 170 guidelines (Algorithm Watch, 2020), as regardless of the policy initiative, ethical codes remain the most elaborated response to the challenges of AI and algorithms. But are ethics the most appropriate way to ensure the ethical sound and socially beneficial AI-DDD? To answer this question one needs to consider the function of ethical norms as modalities of technology governance, as well as the merits and drawbacks of AI Ethics in specific.

3.1.2 The promises and limitations of ethical codes and guidelines

Ongoing historical and sociological research on the role of ethics and values-based norms in technology development has demonstrated that ethical codes and moral principles are valuable for informing and shaping the research and development of technologies in a responsible way (Basart & Serra, 2013; Doorn, 2012). They raise awareness and contribute to the ethical education of professionals while setting the key criteria against which unethical behaviour may be censured (Stark & Hoffmann, 2019). They are also vital, underpinning design and standardization, becoming translated into technical requirements, informing engineering studies and technical decisions (Floridi, 2016; Lloyd, 2009), and supplementing or substituting formal regulation (Hildebrandt, 2017). Moreover, ethical principles assist in summarising a variety of complex ethical issues and challenging moral questions into a few central elements, which in turn can be clearly understood, and reflected upon by people from diverse fields and different backgrounds. This way, they facilitate the development of a common ground, which "can form a basis for more formal commitments in professional ethics, internationally agreed standards, and regulation" (Whittlestone et al., 2019). By the same token, the softer approach they offer may buy valuable time for research, political, social, and legal inquiry to catch up with the developments (Larsson, 2020).

Particularly in the context of AI, the proliferation of various ethical guidelines and principles reframed the meaning and metrics of progress in terms of AI-DDD. It essentially shifted the attention from a purely technical assessment, focused on performance criteria, to a definition of progress that is also premised on the ethical and social aspects of AI (Scantamburlo et al., 2020). This way, ethics-informed approaches complemented the conceptualisation of progress in AI and significantly contributed to making the ethical and

socially beneficial design, development and deployment the benchmark of progress in AI, and a crucial objective in AI-DDD. In turn, the fact that progress in such a transformative and all-purpose technology (Dafoe, 2018; Gruetzemacher & Whittlestone, 2022; Taeihagh et al., 2021) is defined in terms of ethics instead of technocratic criteria introduces a human-centric element that is valuable to ensure that AI will be used for the benefit of society at large, and the flourishment of humanity.

Furthermore, although steering, regulation and governance are commonly associated with the normative functions of law (Hildebrandt, 2018), the law does not enjoy a monopoly in governing human behaviour. In fact, legislation is only one of the governance modalities, namely the factors that shape and affect human actions. Social norms, including principles of ethics and morality, the market, as well as design characteristics equally shape human conduct (Lessig, 1999). In daily life, for example, we abstain from acting in certain ways not necessarily because it is prohibited by law, but because specific types of behaviour may entail ethical or social disgrace, disrepute, or isolation. From this angle, social norms and ethical codes are a significant and impactful way to affect and channel human behaviour. This has led some researchers to suggest that soft-law approaches, ethical principles, and guidelines offer a better solution for AI governance (Floridi, 2018; Floridi & Cowls, 2019; Taddeo & Floridi, 2018).

However, although ethical requirements affect our behaviour, while they often overlap with legal rules and already existing legislation, they are not binding (Horner, 2003). Practically the extent to which social or ethical norms affect our actions and shape our behaviour is contingent on the consequences and the deterrent mechanisms, namely on how we evaluate the negative impacts in comparison to whatever benefit we may gain. Hence, the normative power of social, more and ethical norms depends considerably on the impact of the attached consequences in each context (Hagendorff, 2020). In turn, without established and meaningful mechanisms providing concrete incentives for principles to become practise the normative effect of ethics is rather limited (Whittlestone et al., 2019). Based on this observation, it is suggested that ethics do not necessarily lead to an actionable chain of steps that can effectively establish the much-needed set of rules in AI-DDD (Saslow & Lorenz, 2019). As Hagendorff (2020) remarks in his acute but insightful criticism of the dominance of deontological AI ethics, "the enforcement of ethical principles may involve reputational losses […], or restrictions on memberships in certain professional bodies, yet all together, these mechanisms are rather weak and pose no imminent threat." Similarly, Black and Murray (2019) stress that "if we are to seek to control the way corporates and governments use AI, then ethics cannot substitute for law or other forms of formal regulation."

Apart from problematic enforcement, the limitations of ethics become clearer if we examine their normative function, applicability, and effects, as well as the motives behind them. The broader domain of AI-DDD is largely dominated by a handful of American, Chinese and some EU companies, such as Google, Amazon, Facebook, Apple, IBM, ATOS, Microsoft, Baidu, Alibaba and Tencent which have been all remarkably eager to develop and adopt ethical guidelines, in most cases before government and intergovernmental organisations engaged with AI Ethics, (see for example the initiatives undertaken individually and jointly by Google, Facebook, IBM, Apple, Atos etc. mentioned in Section 3.1). The impressive variety of principles and declarations, in conjunction with the commonly vague language in which ethical codes are drafted, has led researchers to suggest that there is a lack of conceptual clarity and concrete direction, which significantly hampers their practical impact (Asaro, 2019a; Yeung et al., 2019). Even though there is a common core, as all declarations and guidelines include a set of shared themes, such as fairness, privacy, accountability, safety and security, transparency and explainability, non-discrimination and human oversight, and the key discoursive tools are relatively shared, they are rarely concretely defined (Fjeld et al., 2020; Scantamburlo et al., 2020), and there can be overwhelming differences in how the principles are interpreted and materialised.

Critics remark that the majority of AI principles "are too broad to be action-guidings" (Vesnic-Alujevic et al., 2020; Whittlestone et al., 2019), adding that the vague way principles and guidelines are drafted, combined with the close connection between ethical guideless and morality makes them open to various interpretations and contingent on cultural differences, undermining their effectiveness and universality (Asaro, 2019b; Whittlestone et al., 2019; Yeung et al., 2019). For instance, the UN Special Rapporteur Philip Alston argues that framing guidelines as "ethics" renders them meaningless, hindering their normative impact, noting that "as long as you are focused on ethics, it's mine against yours." Similarly, Binns notes that whereas everyone may agree on the centrality of fairness, what it entails exactly, as well as how it is to be achieved may vary considerably among different individuals or groups (Binns, 2017), while different stakeholders may value fairness differently, especially in comparison to other values and objectives. The cultural differences may also add to the difficulty in translating such abstract values into concrete measures that can be easily assessed regarding their effectiveness and their objective capacity to serve as guidelines for practitioners. Thus, the fact that there is no consensus between the major players in the field leaves developers and designers with little guidance and the wider public with no clear view of what "ethical AI" means in practice, which is particularly problematic

considering that AI is largely globalised and transborder (Saslow & Lorenz, 2019).

Moreover, codes on the ethical AI-DDD tend to predominantly focus on particular ethical issues, features, functions or consequences (Larsson, 2019; Stahl et al., 2021a, b; Vesnic-Alujevic et al., 2020). Reviewing the major ethical issues in emerging technologies, Stahl et al. suggested that they can be broadly divided into two main categories, namely issues with individual impacts, such as privacy, autonomy, treatment of humans, identity and security, and issues with broader societal impacts, such as digital divides, collective human identity and the good life, responsibility, surveillance, and cultural differences (Stahl et al., 2017). Building upon this classification, Vesnic-Alujevic et al. (2020) concluded in a similar categorisation, observing that "ethical debates about AI mainly focus on individual rights only" while the dimension of societal challenges and implications is often overlooked or marginally addressed. Thus, although the development and deployment of AI may be consequential and impactful to a diverse set of fundamental questions, the core principles and guidelines for AI Ethics tend to focus on a limited number of concerns (Fukuda-Parr & Gibbons, 2021; Vesnic-Alujevic et al., 2020). Similarly, the "noise" created by the big technology companies around their own set of rules, may marginalise the voice of citizens of civil society groups, which have significantly fewer resources to support and equally distribute their views of ethical AI (Saslow & Lorenz, 2019).

Some scholars have noted that apart from the acknowledgement of the ethical challenges of AI and algorithms, these private-driven initiatives may also conceal further objectives (Bietti, 2019; Slee, 2020; Yeung et al., 2019). Firstly, through developing their own set of principles, particularly the leading technology companies, may seek to set the narrative of ethical AI on their own terms. Defining what the ethical development and employment of AI systems and algorithms entail in practice, such companies not only promote their own interpretation of "ethical and responsible AI" but also establish the criteria for assessing AI-DDD. In turn, shaping how ethical AI is perceived in a way that serves their interests and reflects their priorities may mitigate the risk of reputational cost. Similarly, this way private actors may reserve for themselves a privileged position, establishing themselves as pioneers in the field (Slee, 2020).

Additionally, such self-declared commitments may be practically simply proclamatory, invoked merely for 'ethics washing' (Bietti, 2019; Metzinger, 2019; Muller, 2020a, b), promoted to avoid scrutiny, criticism or even direct regulation (Black & Murray, 2019; Rességuier & Rodrigues, 2020), or used for branding, or reasons related to corporate social responsibility (CSR) (Wettstein, 2012). For instance, Asaro contends that the primary purpose of ethical declarations and self-developed and adopted ethical rules by large corporations is to prevent the introduction of legally binding obligations and "foster a brand image for the company as socially benevolent and trustworthy"(Asaro, 2019b). Similarly, Bietti observes that AI Ethics are currently "weaponized in support of deregulation, self-regulation or hands-off governance" (Bietti, 2019) The emphasis on industry-led ethical codes and commitments may lead to the assumption that the authority and responsibility to steer and control such technologies "can be devolved from state authorities and democratic institutions upon the respective sectors of science or industry"(Hagendorff, 2020). From this angle self-promulgated and adopted ethical codes are invoked to avoid the introduction of binding legal rules, introducing a framework of self-regulation instead of concrete, specific and binding rules (Asaro, 2019b; Daly et al., 2020; Wagner, 2018). This way they serve as a shield from direct regulation (Wagner, 2018), and a vehicle to introduce and establish self-governance.

This challenges not only the actual normative impact of ethical guidelines but also the premise and intentions behind these voluntarily adopted codes. Going this argument a step further, Hao (2019) observes that it is doubtful whether such declarations produce tangible and auditable outcomes in terms of AI-DDD, while Black and Murray (2019) suggest that there are empirical and normative reasons against the reliance on such soft forms of governance for AI.

Yet, even if such commitments genuinely stem from the best of intentions, it is still fair to doubt whether industry actors can actually form adequate norms for the ethical development and deployment of such technologies, and be trusted to enforce and police adherence to them. Particularly in cases in which their monetary interests are at stake, the implementation of ethical requirements without supervision is at least questionable (Hagendorff, 2020). Furthermore, considering that engineers and developers typically lack systematic ethical training (Bednar et al., 2019; Martin et al., 2021), and they are not actively encouraged to reflect upon the ethical aspects of their work (Bairaktarova & Woodcock, 2017; Hagendorff, 2020; Slee, 2020), especially in corporate environments (Lloyd, 2009; Troxell & Troxell, 2017), it is questionable whether, in absence of binding rules and requirements of accountability and transparency, the ethical commitments will indeed guide the development of AI and algorithms. Furthermore, considering that not rarely does AI-DDD entails balancing risks and benefits, or conflicting rights and interests, it is questionable whether private actors without proper guidance are capable of successfully engaging with such delicate and complex tasks.

Finally, turning to their actual normative impact, the absence of specific enforcement tools, and the reliance on self-commitment and reputational costs put their effectiveness and normative impact in question. Looking at the individual level, a survey focusing on the effects in

decision-making making processes of the ethical code developed by the ACM, found that the impact, in absence of other incentives, was rather trivial (McNamara et al., 2018). Moreover, considering that a handful of giant technology companies, including Amazon, Google, IBM, and Microsoft, have the leading role in AI, algorithms, and machine learning (Maguire, 2021; Nemitz, 2018), reputational costs can be arguably relatively easily counterbalanced through the substantial resources available for public relations and other corrective actions that may significantly reduce any damage to the company image. Thus, as the reputational costs seem to be rather insufficient, or at least negligible, ethical principles and guidelines have a rather limited normative capacity to meaningfully ensure the ethical and socially beneficial AI-DDD. From this angle, Rességuier and Rodrigues (2020) argue that while promising, ethical codes in AI are also equally problematic, as not only their effectiveness is yet to be demonstrated but also "they are particularly prone to manipulation, especially by industry."

Thus, the current codes and ethical guidelines can guide AI-DDD only partially. They do add to the awareness around the ethical implications of AI systems and algorithms, but remain silent or at least abstract on the specific role of the companies in avoiding, mitigating and remedying these implications, and set no accountability and redress mechanisms. Moreover, overly relying on ethics, without other structures and governance benchmarks may ultimately reduce them to a mere checklist, turning fundamental values into a box you simply need to click to be on the safe side (Hagendorff, 2020). Yet, the salience of AI Ethics arguably reflects the recognition that algorithms and AI are not simply "another utility that needs to be regulated once it is mature"(Floridi et al., 2018) turning the question of how these principles can be translated into practice through governance (Winfield & Jirotka, 2018) both immediate and demanding.

3.2 The race to governance

3.2.1 The need for governance and different approaches to AI governance

Stemming from the limitations of ethics and the shortcomings of soft-law instruments as means of governance, several scholars have emphatically stressed the urgency to turn from soft to hard-law solutions, and from steering models premised on ethics to governance models based on law and binding obligations (Black & Murray, 2019; Bryson, 2020; Nemitz, 2018). Arguing for the need for regulation they highlighted the transformative power of AI, its relevance for human rights, and the risks it poses on an individual and social level, but also the disruptive effects it may have on social structures (Cihon et al., 2020; Crawford, 2021; LaGrandeur, 2021). On the policy level, the wide range of ethically important and societally impactful implications of AI systems and algorithms, as well as the adverse effects of such technologies on an individual and social level, has led governments and intergovernmental organisations to progressively shift their attention from the promulgation of ethical codes and principles to specific legal instruments (see for example the "Proposal for a Regulation laying down harmonised rules on Artificial Intelligence", by the EU Commission). This shift of attention combined with a sense of urgency has created what Smuha describes as a race for AI governance since national, regional, international, and supranational organisations are in the process of considering and assessing "the desirability and necessity of new or revised regulatory measures"(Smuha, 2021b).

Looking to mitigate the risks, while enhancing trust and legal certainty, international public and private stakeholders have engaged in a competition to draft and promote AI governance models (Radu, 2021; Taeihagh, 2021). Hence, the need and urgency to regulate AI, placing the design, development and deployment of AI systems and algorithms within a specific regulatory environment, seem anymore undeniable (Almeida et al., 2021; Gasser & Almeida, 2017).

However, how to regulate AI remains an open question, as disruptive technologies tend to challenge traditional governance paradigms (Cath, 2018; Maas, 2022). Thus, whereas literature provides with a large number of proposed governance models, most of which are still largely premised on ethics or remain primarily expressed in the language of ethics (Black & Murray, 2019; Whittlestone et al., 2019), we do not have yet "a functional model that is able to encompass all areas of knowledge that are necessary to deal with the required complexity" (Almeida et al., 2021). Regulatory proposals for AI governance vary from suggesting building upon existing norms and instruments (Scherer, 2016), and public international law (Kunz and Héigeartaigh, 2020; Yeung et al., 2019), to the establishment of completely new, specialised institutions (Kemp et al., 2019), and from centralised to international alternatives (Cihon et al., 2020, 2021; Erdélyi & Goldsmith, 2018). At the same time, governments initiatives, at both the national and international level, for the time being, seem to be either "technology-centric", focusing mostly on individual AI applications, or "law-centric", focusing on the effects of AI applications on specific legal fields (Maas, 2022), principally having a risk-based approach (Pery et al., 2021; Scantamburlo et al., 2020).

"Technology-centric" approaches engage with specific applications of AI technology, singling out particular use-cases that regulation should focus on, such as autonomous cars, drones, robotics etc. However, Maas (2022) stresses that such an approach "emphasizes visceral edge cases, and is therefore easily lured into regulating edge-case challenges

or misuses of the technology (e.g. the use of DeepFakes for political propaganda) at a cost of addressing far more common but less visceral use cases." Additionally, it leads to patchwork regulation and fragmentary regulatory responses, as it promotes an ad-hoc, problem-solving orientation (Liu & Maas, 2021). Reversely, the law-centric approach does not focus on the individual application, but on the relevant legal doctrine, exploring how AI applications may change or challenge the scope or assumptions of existing legislation (Crootof & Ard, 2020; Petit, 2017). Instead of starting from the technology and its applications here the point of departure is the legal system. The problem with this approach is that it leads to the segmentation of the regulatory responses, as it ties the regulatory reaction to specific legal doctrines, such as privacy, contract law, consumer protection etc., while the effects of a specific AI application may be relevant to more than one domain of law. Furthermore, focusing only on the law, it may neglect other means of regulation, such as design and standardisation. Therefore, Maas (2022) remarks that both approaches represent somehow "siloed policy responses".

Nevertheless, AI and algorithms are not developed, nor deployed in a vacuum. Some of the challenges posed by AI systems and algorithms are subject to already existing legislation (Black & Murray, 2019; Cannarsa, 2021). For example, consumer protection law, anti-discrimination legislation, as well as privacy and data protection rules already apply in several AI applications. Yet, it is still essential to review the regulatory framework and critically evaluate it considering also the unpredictable outcomes of AI and taking steps towards the establishment of specific governance structures when necessary. Simultaneously, in sectors lacking regulation, it is urgent to identify and assess the risks and potential harms, considering also the sometimes unpredictable outcomes of AI systems (Reed, 2018). There is also a growing volume of literature that argues for the need to develop new means of governance and regulatory instruments, as the existing structures cannot successfully meet the challenges and respond to the issues AI raises (Stahl et al., 2021a, b; Taeihagh, 2021). However, the governance of such all-purpose (Dafoe, 2018; Trajtenberg, 2018), transformative (Gruetzemacher & Whittlestone, 2022), widely disruptive, highly complicated, and still emerging technology is far from a simple task (Radu, 2021; Smuha, 2021b).

3.2.2 Designing a model for AI governance

Challenges in governing AI arise from various sources. They range from the choice of the most suitable and appropriate approach; the proper combination of modalities of governance; the identification, and engagement of all the relevant stakeholders, to the very definition and conceptualisation of AI and algorithmic processes (Taeihagh, 2021). Moreover, questions regarding how we should regulate "a changing technology with changing uses, in a changing world" (Maas, 2022) as well as what exactly is that we seek to regulate (Black & Murray, 2019; Maas, 2022) represent crucial aspects of AI governance that we need to consider. Simultaneously, it is essential for legislation to strike a balance between protecting society from potential harms and allowing AI technologies to develop and advance for the benefit of society. As Floridi et al. (2018) remark, "avoiding the misuse and underuse of these technologies" is both challenging and critical in ensuring that AI will serve society.

Now, if we wanted to put the key challenges for AI governance in some sort of order, from a legal and regulatory point of view, the very first challenge for AI governance would be the lack of conceptual clarity. As in every other domain and particularly so in the field of technology governance, it is of paramount significance to have proper knowledge of what is to be regulated (Kooiman, 2003; Larsson, 2013a, 2020; Reed, 2018). Thus, framing AI is an integral part of formulating adequate governance structures and responses (Perry & Uuk, 2019), given that "the definition is in itself a form of conceptual control" with significant impacts on the governance discourse (Larsson, 2013a, b, 2020). However, as mentioned in Section 2, AI is riddled with multiple interpretations and competing definitions (Almeida et al., 2021; Haenlein & Kaplan, 2019), while, beyond the shared points Collins et al. (2021) have found between the various descriptions and conceptualisations, there is no consensus (Stahl et al., 2021a, b).

Whereas the existing definitions may be sufficient to offer an idea of the broader scope of AI and the issues at stake, allowing us to discuss AI governance, and explore the priorities, they are arguably not specific and detailed enough to allow the application of governance structures (Stahl et al., 2021a, b). Moreover, without a common framework and an agreed-upon starting point, it is hard for policy-makers to determine what aspects of AI and algorithms applications are desirable and which are not (Bhatnagar et al., 2018; Larsson, 2020), to take the appropriate regulatory measures. Additionally, the fact that AI is not a single technology, but a collection of technologies and applications (Gasser & Almeida, 2017; Latonero, 2018; Raso et al., 2018a), the relevance of a variety of processes, procedures, and components, the nonlinear way in which algorithms and machine learning work (Robbins, 2019), combined with the rapid pace at which the technologies that are considered to be part of AI change and become replaced by others (Raso et al., 2018a) further obscure the picture, making governance particularly challenging (Radu, 2021).

Furthermore, to promote a governance model that will foster "AI for human flourishing" (Stahl et al., 2021a, b), it is significant to also critically examine the black-box approach, the innate opacity and inherent unexplainability, as well as

the unpredictable nature of algorithms. Such features, often repeated with a sense of truism as necessary components of the AI definition, may diminish the accountability of AI designers, owners and operators, and reduce the contestability of their decisions (Edwards & Veale, 2017; Hildebrandt, 2016), rendering the justification of the outcome impossible or unnecessary, even in cases of damaging, unfair or discriminatory results (Bayamlioglu, 2018). Unless we find a meaningful way to address these issues and challenge their premise and entailments, effective governance through law may remain particularly difficult (Leenes et al., 2017; Santoni de Sio & Mecacci, 2021). To that end, Bryson (2020), Bryson & Theodorou (2019) suggests that policy-making should be premised on a human-centric approach, based on an understanding that embraces technologies as end-product of design, choice and intentionality, which can be transparent, documented and explainable, at least to the extent necessary for accountability reasons, if so mandated by the law.

Beyond the difficulties in framing AI, a subsequent challenge involves engaging all the key stakeholders, balancing the asymmetries between them, and deciding upon the most suitable and appropriate governance model. For a long time AI research has been far away from the interests periphery of governments and the public (Smuha, 2021a, b), as technology companies have been in charge of AI development so far (Jang, 2017). Thus, the field became largely dominated by private entities, which had at their disposal not only state-of-the-art equipment and ample proprietary data but also leading researchers and sufficient discretion. This way, the field was principally industry-driven and self-governed, primarily through ethical codes and declarations. However, lately, there is an ever-increasing interest on behalf of governments, intergovernmental and supranational organisations, non-governmental organisations, research institutions, civil society groups, and the public at large (Radu, 2021; Taeihagh, 2021). Currently, the AI governance field constitutes a global arena, in which multiple stakeholders from various fields, with divergent resources, interests, motives, and familiarity with the topic, compete for power and authority to influence AI governance (Butcher & Beridze, 2019; Dafoe, 2018). Moreover, the "race for AI regulation" (Smuha, 2021b) has politicized the area (Radu, 2021) inducing competition among the stakeholders that try to steer the governance quest to their benefit, or reserve for themselves a leading position.

In this context, the traditionally leading role of the state is significantly challenged as private entities, and particularly a handful of technology companies leading the field, enjoy considerable informational and resource advantages compared to national governments. They arguably have enhanced familiarity with the field (Guihot et al., 2017; Taeihagh et al., 2021), and it is questionable whether they will be willing to share their insights, while most probably this will be a quid pro quo. Additionally, as the key role of private entities is hardly in question, governments and intergovernmental organisations need to remain cautious of the risks involved in over-delegating power and authority to private hands, considering technology companies can be notoriously difficult to control and supervise (Chenou & Radu, 2019). Building on this observation, Nemitz (2018) warns that in the context of technology governance IG has set a rather dangerous culture of "lawlessness and irresponsibility," permitting extensive discretion, accumulation of substantive power, and ultimately allowing a handful of private technology companies to become the de facto governors (Suzor, 2019). Thus, adequately engaging private actors, delegating them the tasks that they are better equipped to fulfil while ensuring that they will not abuse their power, authority, and privileged position entails designing and deciding upon the most appropriate governance model, as well as introducing the necessary checks and balances to private and public power.

From a similar point of view, creating the appropriate model for governance and assigning roles to the different actors and stakeholders is also rather tricky, and equally crucial for the success of the AI governance regime. There have been several suggestions aimed to solve the governance mode puzzle, proposing innovative governance approaches, such as decentralized multistakeholder models, and hybrid or adaptive forms of governance (Brundage et al., 2018; Dafoe, 2018; Radu, 2021; Smuha, 2021b), as well as different approaches towards governance (Cihon et al., 2020, 2021; Kemp et al., 2019), aimed at combining the competences and de-facto governance power of the relevant stakeholders.

In this context, it is necessary to ensure the balance between the competing governments, as well as the adequate representation of smaller countries and the developing world. So far, the race for AI governance is led by the countries hosting some of the key technology companies with a leading role in AI development, such as the US, China or Japan, and governments along with intergovernmental organisations that seek to proactively develop thorough and prospective governance models to secure a leading role in the AI future, as the EU and several European governments have been doing since the mid-2010s. The rest of the world, and particularly the less advanced countries, struggle to set their objectives and priorities, to have a place on the table (Radu, 2021). Given that AI governance is essentially the quest of setting the foundations for the future of a disruptive technology that is expected to change the world, it is arguably of at most importance to secure that all countries will have a place on the table, and ensure that they will have a voice in the negotiations. Consequently, one of the most significant questions relates to defining and prescribing the role of non-state actors, as well as balancing the power,

information, and resource asymmetries between private and public actors, as well as between the countries (Nemitz, 2018; Taeihagh, 2021; Taeihagh et al., 2021).

Finally, the appropriate combination of modalities of governance and the specific role of law in AI governance remain open issues (Almeida et al., 2021; Cath, 2018), bearing significant consequences for the overall form of governance as well as for the stakeholders involved (Smuha, 2021b). Law has a rather famously complicated relationship with technology. Its competence and suitability in terms of technology governance have been multiply contested, and the context of AI and algorithms is no exception. As has happened previously, in the context of Internet regulation, the capacity of law to serve as an effective vehicle of governance for AI, is debated. Yet, we need to remind ourselves that the law has successfully sustained numerous "revolutionary innovations" adapting and remaining relevant. Nonetheless, there are still some noteworthy challenges. For example, the pace of technological advances is a fairly obvious one (Larsson, 2020; Perry & Uuk, 2019). Additionally, the reference to a variety of different technologies, often hard to be told apart, can be a further challenge. Moreover, the choice of rules, the balance between over-regulating and regulatory vacuum, as well as the different domains of law that are relevant and applicable requires carefully assessing the existing instruments, the necessity for intervention, the impacts, and potential spill-over effects without losing sight of coherence (Smuha, 2021b).

3.2.3 The steps we have taken and the path ahead

Although the focus on ethics has somehow overshadowed the translation of principles into concrete regulation, and the concretization of guidelines into specific binding requirements (Radu, 2021), the scenery is rapidly changing, as governments and intergovernmental organisations are increasingly moving towards the introduction of specific and binding rules to govern AI. In this context, when discussing AI governance, a reference to the GDPR seems an inevitable commonplace, both because the Regulation is one of the most visible and well-known pieces of legislation relevant to AI (Almeida et al., 2021; Stahl et al., 2021a, b), and because it is widely embraced and celebrated as efficiently and effectively tackling several of the critical issues, particularly related to privacy, transparency, explainability, and documentation of algorithmic processes and automated decision-making procedures.

Although not explicitly referring to AI systems and algorithms, a set of specific GDPR provisions affect not only the collection and processing of personal data by AI and algorithms but also the design and deployment of AI and algorithms, as well as algorithm-based decisions (see for example Articles 4, 6, 9, 22, 25, 35). For example, regarding the opacity of algorithms and the problematic accountability of automated decision-making, the Regulation introduces the principles of transparency and explainability, and an enhanced accountability model jointly with the requirement for detailed documentation (Article 30). Moreover, although not necessarily purported to solve all the challenges of AI, data protection laws offer suitable responses to several AI issues, as the right to privacy crucially relates to a number of other rights and freedoms, such as equality and non-discrimination (Hildebrandt, 2019), which are relevant in terms of AI.

However, the GDPR does not cover all of the negative or challenging AI ethical and societal implications (Busacca & Monaca, 2020; for an interesting discussion of AI and the GDPR see also Mitrou, 2019), while privacy and data protection are only part of the AI-related concerns. Sometimes the scope of critical provisions for AI is too narrow (see for example Article 22), while the Regulation offers limited guidance on how to achieve a balance between the obligations and requirements of the GDPR and the objective of promoting AI research and applications that respect these obligations (Sartor & Lagioia, 2020). Explainability remains challenging and obscure, also due to the relatively vague way in which the requirement to provide explanations is phrased in the Regulation (Hamon et al., 2021, 2022). Issues of discrimination are only partially addressed, as Article 9 on special categories of personal data does not include "categories of colour, language, membership of a national minority, property" which may also lead to discriminatory outcomes through AI and algorithms (Ufert, 2020). Additionally, critics have underlined that "paying high fines instead of complying with the GDPR could be a preferable path for major digital companies" which in turn limits the actual effectiveness of the provisions (Vesnic-Alujevic et al., 2020). Finally, even though the extraterritorial effect significantly expands the reach of the Regulation, the GDPR is hardly a global instrument. Furthermore,

Another noteworthy instrument is the Proposed EU Regulation of AI, the Artificial Intelligence Act (AIA). The Act aims to foster an "ecosystem of trust that should give citizens the confidence to take up AI applications and give companies and public organisations the legal certainty to innovate using AI." Notably, it constitutes the first-ever attempt to enact a horizontal regulation of AI, through an instrument that is specifically intended to govern AI, signifying the decisive step from soft to hard law. It is also indicative of the EU strategy to become a pioneer in AI governance by introducing a framework premised on the EU values and the key EU regulatory principles. The proposed legal framework focuses on the specific utilisation of AI systems, having a risk-based approach, and will also enjoy extraterritorial jurisdiction, (Pery et al., 2021). It applies to all providers, assigns responsibility to users, importers, distributors, and

operators, and seeks to ensure compliance with fines that go well beyond those of the GDPR. In Article 3, AI systems are defined as "software that is developed with [specific] techniques and approaches [listed in Annex 1] 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." As noted for the GDPR, the Proposed Act does not constitute a global solution, yet if adopted it will arguably serve as a blueprint for similar instruments.

Yet, the AIA Proposal currently is far from ratification, while several of the key points of the proposed legislation have attracted criticism and are open to debate. For example, Smuha et al., (2021). remarks that the Act fails to accurately recognise the risks and harms associated with different kinds of AI and future AI applications. It is suggested that for "trustworthy AI" it is essential to establish a mechanism that will allow the Commission to expand the list of prohibited AI systems and propose banning existing manipulative AI systems such as DeepFakes, social scoring and some biometric identification systems. They also stress that in many cases the proposal does not provide sufficient protection for fundamental rights, nor effective redress mechanisms or a meaningful framework for the enforcement of legal rights and duties, while public participation is not adequately protected and promoted (Smuha et al., 2021). Similarly, Ebers et al., (2021) while celebrating the innovative elements of the proposed Act, also emphasise the absence of effective enforcement mechanisms, criticising the self-enforcement structure for raising concerns of under-regulation. They argue that without external oversight, and meaningful ways to ensure access to remedy to the affected parties the risk-based approach does not adequately protect individuals' rights (Ebers et al., 2021). From the same angle, Veale and Borgesius (2021) highlight that obligations on AI systems users, may fail to protect individuals given that the draft Act does not provide a mechanism for complaint or judicial redress available to them.

Reviewing the most recent regulatory initiatives, it becomes apparent that AI governance constitutes a pressing priority. Nevertheless, although positive steps have been taken there is still a long way ahead. Hopefully, the growing public attention will lead to wider and deeper discussions about the most appropriate responses to AI challenges. Yet considering the ever-increasing penetration of AI systems and algorithms in contemporary society, and their relevance to human rights, a human-centric, rights-based approach ought to underpin the governance initiatives and any regulatory instruments. To that end, additional research and further negotiation are also necessary to ensure greater inclusivity and diversity, fair participation, and meaningful representation of all the views, as well as to explore the role key actors should have, considering their place in the AI governance ecosystem along with their interests and agenda (de Almeida et al., 2021; Larsson, 2020; Perry & Uuk, 2019). Yet, broad public debate and democratic deliberation are still lagging behind technological development and policy-making in the context of AI governance (Vesnic-Alujevic et al., 2020).

4 Human Rights and AI Governance

4.1 Human rights in AI Ethics and the AI governance discourse

Human rights are highly relevant within the AI Ethics and AI governance discourse from multiple angles. First, AI and algorithms have emerged as a key area of human rights concern during the last decade, as their adverse effects on human rights became increasingly apparent and alarming (Bachelet, 2021; Fukuda-Parr & Gibbons, 2021). As noted in the introduction, AI systems and algorithms are routinely employed in ways particularly relevant and commonly impactful for human rights (Latonero, 2018; Risse, 2018; Yeung et al., 2019). The ubiquitous role of AI in our daily lives across public and private contexts could adversely impact the rights and freedoms of citizens all over the world on a scale and in ways not always clearly foreseeable (Saslow & Lorenz, 2019). Whereas considering the negative implications of AI systems for human rights privacy, data protection, and discrimination, are often discussed, McGregor et al. (2019) stress that there is also a variety of human rights issues that are less apparent and studied, while the bias and discrimination that are repeatedly reinforced by AI systems may lead to further adverse impacts for human rights.

For example, the wide employment of Amazon's face recognition technology "Rekognition" by US law enforcement and immigration services, as well as by private companies in search for employees, has created a number of human rights-related controversies, as it tended to falsely match the images of women with darker skin colour with those of arrested people to a disproportionate degree (Godfrey, 2020). The bias of the system discriminated against these women affecting their access to work, or most importantly, their rights to life, liberty and security. Similarly, automated credit scoring may affect employment and housing rights, or the rights to work and access to education, in ways that are not always obvious ex-ante. Moreover, "the increasing use of algorithms to inform decisions on access to social security potentially impacts a range of social rights" (McGregor et al., 2019) including family life, as algorithmic bias in identifying children at risk may have devastating effects on already vulnerable families. Additionally, as Rachovitsa and Johann (2022) remark, the employment of AI systems in terms of digital welfare state initiatives often falls short of meeting basic requirements of legality.

Human rights are also highly relevant in the AI discourse as in the quest for ethical and societally beneficial AI-DDD they are commonly invoked, either as guidelines for AI Ethics or as principles for AI governance (Fukuda-Parr and Gibbons, 2021; Muller, 2020a, b; Smuha, 2021a). They are mentioned in most of the ethical principles and guidelines developed by national and intergovernmental organisations and research groups. For example, both the EU and the UN have identified human rights as forming "the most promising foundation for identifying abstract ethical principles and values" and central in the effort to ensure the development and employment of AI for the benefit of society (AI HLEG, 2019b; Hogenhout, 2021). Similarly, the CoE has emphasised the vital need to safeguard human rights along with their relevance in informing and shaping AI Ethics (Mijatović, 2018). Moreover, the Toronto Declaration clearly states that it builds upon "the relevant and well-established framework of international human rights law and standards." Respect for human rights is the also first principle of the IEEE ethical framework for AI (The IEEE Global Initiative, 2017) and several other AI Ethics codes developed by research groups and think tanks (Algorithm Watch, 2020).

Human rights are also commonly mentioned in the AI ethical guidelines and principles of several private entities and technology companies working on AI-DDD (Asaro, 2019a). For example, human rights are invoked among the guiding principles within Facebook's Five Pillars of Responsible AI (Pesenti, 2021). Human rights and the UNDHR are also explicitly noted in the Microsoft Global Human Rights Statement (Microsoft, 2020). However, as Alston (2019) remarks the"token references" to human rights, and the self-proclaimed commitment to respect human rights as a stand-alone principle in private-sector AI Ethics codes, are commonly ornamental. The codes rarely provide a comprehensive list of rights that individuals may invoke against the company, nor a redress system in case of violations. Access to remedy is implied and not safeguarded, while external auditing or any kind of human rights monitoring is rarely mentioned. This is not necessarily surprising, as, in terms of such codes, human rights are not perceived with the sense of legal rights, but merely as ethical principles (Hagendorff, 2020).

Building on this observation, a number of researchers and human rights advocates have suggested building "ethical" and "responsible" AI on the basis of human rights, essentially premising AI governance on human rights instead of ethics (Saslow & Lorenz, 2019; Smuha et al., 2021; Yeung et al., 2019). It is argued that human rights can both establish and reaffirm the human-centric nature AI-DDD ought to have, but also introduce actionable standards and binding rules, complementing and expanding upon ethics (Saslow & Lorenz, 2019). Developing AI governance models and rules with human rights standards as a premise, while holding AI designers, developers and operators accountable to protect individuals' fundamental rights and freedoms may effectively address and overcome many of the limitations of ethics (Pizzi et al., 2020; Saslow & Lorenz, 2019; Yeung et al., 2019). Human rights provide a deeper and more thorough framework to analyse the overall effect of algorithmic decision-making, determine harm and address accountability (McGregor et al., 2019). Moreover, anchoring AI governance to the international human rights law can offer a more robust, comprehensive and widespread framework for AI governance (Cath, 2018; Smuha, 2021a; Yeung et al., 2019), providing "aspirational and normative guidance to uphold human dignity and the inherent worth of every individual, regardless of country or jurisdiction" (Latonero, 2018). Furthermore, as Stahl et al. (2021a, b) suggest "it seems plausible that a more direct application of human rights legislation to AI can provide some clarity on related issues and point the way to possible solutions."

Due to their dual nature as legal and ethical entitlements, human rights are indeed both relevant and suitable to be the foundations for controlling and steering AI and algorithms. Unlike the multiplicity of ethical principles and self-adopted guidelines, the support of the UN human rights system is substantial on a global scale (Risse, 2018). Serving as the basic moral entitlements of every human being, they are deeply rooted in contemporary politics and law, recognised in political practice and legal institutions globally (Etinson, 2018). Thus, contrary to ethics, human rights are universal, offering a common set of principles that can be applied globally (Smuha, 2021a). Considering the global reach of a variety of AI systems and algorithms, along with the calls for a governance system of international nature, this is a considerable benefit, as human rights provide a globally legitimate and comprehensive framework. Of course, the human rights system is not flawless. It has several limitations, however, it evolves over time, reacting to the developments and the challenges in society. The UN has established a rigorous and robust system of Special Rapporteurs and Obunspeople who have identified and set in monition a variety of initiatives aimed at improving the level of human rights protection globally, as well as responding to the challenges posed by digital technologies (Fukuda-Parr & Gibbons, 2021; Pizzi et al., 2020).

Additionally, although relatively vague, the international human rights law system is far clearer and more specific than ethical guidelines, particularly when it comes to rights and harms. Contrary to the multiple ways in which crucial ethical principles may be defined in terms of ethical guidelines, it clearly sets out obligations and entitlements, and harms, articulating specific enforceable duties and providing support and guidance related to what needs to be avoided, and how it ought to be mitigated and remedied (McGregor et al., 2019). The analysis through the lenses of the international

human rights law system goes well beyond abstract notions such as "bias," "privacy" or "fairness" focusing on the impact of specific choices, actions, and activities on human rights in a relatively measurable manner. Similarly, it offers means, ways and principles to weigh and balance rights, providing a framework for resolving tensions between conflicting rights and clearly establishing in a non-arbitrary way when and how particular rights may be restricted, including also redress and contestation mechanisms (McGregor et al., 2019; Yeung et al., 2019). This is valuable, especially in grey zones of AI employment and cases in which risks and benefits need to be weighted. It also provides well-established and thorough means of classifying and labelling harm through the establishment of a comprehensive and elaborated set of internationally agreed-upon substantive and procedural rights.

The international human rights system involves also a rich theoretical background and analytical lenses, combined with ample discursive tools (Latonero, 2018; Risse, 2018). From this angle, Van Veen and Cath stress that "human rights, as a language and legal framework, is itself a source of power because human rights carry significant moral legitimacy and the reputational cost of being perceived as a human rights violator can be very high" (van Veen & Cath, 2018). Moreover, in vast contrast to self-commitments, the international human rights law system includes a well-established institutional framework comprised of dedicated monitoring bodies and agencies, along with built-in accountability, advocacy and redress mechanisms, aimed to ensure compliance with human rights principles globally. Thus, whereas ethical guidelines that cannot be invoked in court, nor is it possible to monitor compliance to them, the international human rights regime is a legal framework that is binding, and adherence to its rules can be monitored.

There is also a steadily growing branch of literature that suggests the translation of human rights into technical requirements, essentially proposing the hardwiring of human rights into technical standards, both in terms of AI and more broadly in digital technologies at large (Krishnamurthy, 2019; Yeung et al., 2019; Zalnieriute & Milan, 2019). Whereas promising, this ambitious suggestion that builds upon the strong normativity of code (Hildebrandt, 2008), and the normative effects of standardisation and technical decision-making, is challenging and may be difficult to be fully implemented soon (Mueller & Badiei, 2019). Although the embodiment of law and human rights into design has numerous significant merits (Hildebrandt, 2011a, 2015), as we have seen for example regarding privacy and data protection, it may also lead to rather inflexible results that may even have reverse effects in practice (Hildebrandt, 2011b; Koops & Leenes, 2014). Moreover, due to the strong legalism of code (Diver, 2021), it is argued that governance through design, code and standardisation may adversely impact human dignity (Brownsword, 2017, 2019).

Evidently, human rights have moved from the periphery to the core of the AI governance discourse (Latonero, 2018; Raso et al., 2018b). Although expressed in different ways, there is a broad consensus over the centrality and relevance of human rights in the broader AI discourse and their particular role in AI Ethics and AI governance (Fukuda-Parr & Gibbons, 2021).[6] From anchoring AI ethics and the broader AI governance discourse to human rights on a theoretical level to translating human rights into technical requirements and design principles, the international human rights system is widely perceived as a source of appropriate and adequate responses to the AI governance dilemmas from multiple aspects. This view is also embraced here, in the sense that international human rights law offers a concrete, specific and comprehensive set of almost universally accepted and institutionalised rights, specific obligations, and expectations. It offers means to balance and restrict risks, ways to assess, mitigate and remedy the harms, combined with thick analytical and discoursive lenses, as well as guidance, compliance, redress, and monitoring mechanisms.

Nonetheless, I do not intend to suggest that human rights are flawless, or without limitations, nor that an AI governance model premised on human rights will be an all-covering solution that would magically resolve all AI-related challenges. What I argue, instead is that human rights offer a promising framework for AI governance, which is, at least in comparison, substantially better than the mere reliance on soft-law and self-adopted ethical commitments (McGregor et al., 2019; Yeung et al., 2019). However, how such a model can be actually materialised in a way that will practically offer effective and meaningful solutions constitutes a challenge on its own, given the state-centric model of human rights, and the highly privatised characteristics of the AI industry.

4.2 Vertical vs. Horizontal Application, the UNGPS and the Treaty for Business and Human Rights

International human rights law is traditionally interpreted and applied as created by and for nation-states (Brownlie, 2019; Ziemele, 2009). Thus, although there are other spheres of international law, such as humanitarian and criminal law, in which private actors can be held directly responsible for violations of the norms they embody, any obligations arising from human rights treaties are legally binding only on

[6] On the growing literature suggesting human rights as a source of principles for human rights see also (Latonero, 2018; Leslie et al., 2022; McGregor et al., 2019; Risse, 2018; Taddeo & Floridi, 2018).

states (Kampourakis, 2019).[7] This state-centric model also referred to as the vertical effect of human rights, entails that private actors, from natural persons to legal entities, are not directly bound to human rights. Hence, under the status quo, the states by entering into human rights treaties undertake international obligations regarding human rights, and then based on their duty to protect human rights, they are required to establish a legal framework and the necessary regulatory structures that introduce obligations for private actors concerning human rights within its domestic legal system (Lane, 2018a) The state "is the sole "originator" of obligations for third parties and the main "enforcer" of those obligations. Consequently, individuals have no human rights claims against each other outside the terms of the framework set up and established by the state itself (Bilchitz, 2016a), which practically means that in case of violation whether there will be a remedy largely depends on the ability and the willingness of the state to fulfil its obligations (George & Laplante, 2017; Ramasastry, 2015).

Yet, because of divergent political ideologies, or financial strategies, the governments may be more or less willing or even able to hold companies accountable for human rights through domestic legislation (Ramasastry, 2015). For example, limited liability, intended to encourage entrepreneurship, innovation, and to attract investments, may significantly limit individuals' access to meaningful remedies for human rights infringements by private actors (Rights & Look, 2021). Hence, the failure (or unwillingness) of the state to introduce human rights duties in private law leaves the affected individuals without protection and access to remedy, as they cannot invoke their rights against the private perpetrator (Ramasastry, 2015). Moreover, it is possible that whereas the state has taken every reasonable action to protect human rights through domestic law, the private entity may infringe upon them in a less obvious or unexpected manner (Bilchitz, 2016a).

For example, based on the vertical effect of human rights whether a private company can dismiss an employee or restrict her right to manifest her religion depends on whether the state has introduced specific protections in the national legal order, by enacting the necessary provisions. In the absence of direct obligation, the employee can only invoke her human rights vertically, against the state, for failing to fulfil its duty to protect her human right through domestic legislation. While this example is a relatively obvious and clear-cut case, it would be arguably more difficult for the employee to say that she was not hired due to her religion if her CV was reviewed by a biased algorithm. It is trickier to see the relevance of an algorithm doing credit background checks with access to education. Moreover, it may be hard for an Internet user to realise that the news feed she reads through was filtered and indexed in a certain way by an algorithm that may be interfering with her freedom of expression (Gillespie, 2020). In fact, the technologically-mediated exercise of human rights within a hybrid public–private context significantly obscures the possibility to identify a human rights violation and pinpoint the perpetrator (Klonick, 2018; MSI-NET, 2017). However, even if they manage to identify both the infringement and the perpetrator, whether the violation will cease, and whether the victim will have access to meaningful remedy depends on how human rights have been translated into domestic law.

Although there is rich literature on the shortcomings of the indirect application of human rights to private actors, to make my point more clear, I will briefly address some of the main reasons why the state-centric model is problematic. Firstly, adequate human rights protections established in domestic law may find significant opposition from private interests. Considering their centrality to the economy, as well as the significant lobbying power some private entities have, major companies and TNCs may oppose, delay, or tone down such legislation (Hamdani & Ruffing, 2017). Furthermore, domestic law applies only within a specific jurisdiction, while in the globalised economy we are currently living jurisdictionality may be a significant challenge to enforce national human rights protections (de Aragão & Roland, 2017). Jurisdictional doctrines such as *forum non conveniens* may pose serious difficulties in even commencing legal proceedings against a corporation across borders (Rights & Look, 2021). Additionally, seeking remedy via legal action against the state for failing to fulfil its human rights obligations may be particularly challenging, as it can be hard to prove, while the overall court procedures are in many countries a lengthy and costly process that not all citizens can afford (Deva, 2017; MacChi, 2018). Consequently, it seems that for a number of reasons, domestic provisions seem to have limited effect, and often appear to be inadequate to provide meaningful protection and access to remedy (Kanalan, 2014).

The relationship between private actors and international human rights law has been a subject of intense political and scholarly debate for over four decades (Rights & Look, 2021; Zalnieriute, 2019; Zamfir, 2018). Apart from the limitations of the state-centric model, the very rationale behind the vertical application of human rights is increasingly brought to question. The growing realisation that the power of private actors to impact human rights is gradually surpassing that of states (George & Laplante, 2017), along with the observation that in contemporary society, private actors rapidly become ever more involved in various human rights-relevant sectors, has brought up with urgency the question of how far human rights should extend. For

[7] International Covenant on Civil and Political Rights art. 2, Dec. 16, 1966, 999 U.N.T.S. 171.

example, major technology companies, and online information service providers, such as Google, Facebook, or Twitter may interfere with the freedom of expression of millions of people on a daily basis in ways virtually impossible for any state (Balkin, 2018; Sander, 2019). The indubitable power of private actors to negatively affect human rights combined with the realisation that non-state actors are increasingly in a position to affect individuals' human rights in ways more impactful and consequential than the states (de Aragão & Roland, 2017; Lane, 2018c), has brought to the forefront the change in the global balance of power between state and non-state actors, challenging the traditional state vs. individual asymmetry view.

The calls to extend human rights obligations to non-state actors, while finding new ways to expand the reach of human rights into the private sphere have become louder in the last two decades. Researchers from the business and human rights field have emphatically stressed the accountability gap created by the status quo, highlighting the limitation of soft-law mechanisms, such as various codes and CSR (Blitt, 2012; Kampourakis, 2019; Wettstein, 2012). Simultaneously, IG scholars have emphatically highlighted the adverse effects of the human rights gap in IG, suggesting that steps should be promptly taken to include private actors in the international human rights law framework. (Zalnieriute, 2019). In this context, the UN has taken several initiatives aimed at improving the status quo, nevertheless, the most widely accepted and endorsed benchmark of the corporate human rights discourse are the UN "Guiding Principles on Business and Human Rights" (UNGPS) (Ruggie, 2011). The principles, which are premised on the obligations of the states to respect, protect, and fulfil human rights, build upon the recognition that the central role of business enterprises in contemporary society necessitates adequately engaging them in the respect, protection, and promotion of human rights. UNGPs introduce the notion of "human rights responsibility" that is attached to business enterprises "regardless of their size, sector, location, ownership, and structure." They also suggest a three-step approach, namely "a policy commitment" to human rights, "a human rights due diligence process" intended to identify, prevent, mitigate, and plan for remedies, and a process to redress and remedy any adverse human rights impacts.

Although the principles constitute a positive step, they have been strongly criticised for not introducing specific binding requirements, and not establishing a concrete implementation mechanism. The non-binding character of UNGPs is said to be in vast contrast with the sacred role of human rights, while the fact that the principles are optional significantly undermines their actual impact (Hazenberg, 2016; Pillay & Curiae, 2014). Moreover, as Zamfir (2018) notes "according to a 2017 study for the European Parliament, although much progress has been achieved in implementing the UNGPS (for example, the OECD Guidelines have been aligned to the UNGPs and new tools have been developed), human rights abuses by corporations persist." Thus, the limited effect of the UNGPs combined with the growing awareness of the impactful way in which private actors can negatively affect human rights has fuelled discussions on finding new ways to address the vacuum.

The urgency to take the next step, and impose direct horizontality of human rights, making international human rights law directly binding to non-state actors, and more specifically to corporations, has been suggested by several scholars (Deva, 2017; George & Laplante, 2017; Pillay & Curiae, 2014; Wettstein, 2012). Some even argue that based on human rights theory, as well as the foundations, the *raison d'être* of human rights, such a shift would not necessarily constitute a drastic departure (Bilchitz, 2017; Kampourakis, 2019). From this angle, apart from the practical benefits of this change, in terms of access to remedy for those affected, and enhanced protection of human rights between actors on the same level, this paradigm shift would rectify the profound asymmetry between businesses' rights and obligations (Bilchitz, 2016b; Zalnieriute, 2019). It would also reaffirm the significance and relevance of human rights in contemporary society, across all different fields and contexts of human conduct.

In practice, this shift can take two forms, either by extending the already existing human rights obligations to private actors or by drafting a new international human rights treaty that would also apply to non-state actors. This issue is anymore firmly placed on the agenda of international law-making, while it seems that if this plan is to flourish in the near future, a new treaty will probably be the way forward, as in June 2014 the UN Human Rights Council established an intergovernmental working group (IGWG) tasked to elaborate an international legally binding instrument on Transnational Corporations and Other Business Enterprises concerning human rights. The Working Group presented a Zero Draft of a Treaty and an Optional Protocol in July 2018 and the Third Revised Draft in August 2021. Even though direct horizontality remains highly controversial, these developments may be highly relevant for the future of AI governance.

4.3 Direct application of human rights in AI governance

Based on the two previous sub-sections human rights are both relevant and normatively appropriate to serve as the foundation of AI governance. The UN human rights system may provide an invaluable organisation framework, offering a set of widely recognised rules, concrete rights and obligations, combined with a well-established and elaborated institutional and theoretical framework that

can contribute to AI governance becoming human-centric, rights-premised, as well as focused and actionable. Hence, arguably human rights are critical to ensure that AI will be developed and employed for societally beneficial ends and the flourishment of humankind, perceived as binding rules and not mere signposts. Yet, finding the proper way to meaningfully and effectively inject them into AI governance, as well as in the entire lifecycle of AI-DDD and algorithmic design, employment and decision-making is a crucial and challenging task. Moreover, considering that the AI industry is highly privatised, the vertical effect of human rights and the state-centric model in the application of human rights obligations may limit their effectiveness (Rachovitsa & Johann, 2022).

For instance, as previously discussed, the translation of human rights obligations into domestic law may not always be adequate, while the ambiguity of AI systems and algorithms and their ubiquitous presence in our lives may significantly obscure the possibility of identifying an infringement and invoking the relevant provision. Moreover, building upon the previous discussion, jurisdictional obstacles may limit the effectiveness of domestic law in regulating international corporations, and the ability of courts to enforce their verdicts outside their jurisdiction. Simultaneously, big technology companies, such as the ones leading the AI industry, have significant resources and leverage mechanisms that may allow them to shape or at least affect domestic law in their favour.

Stemming from these observations, I argue for the direct application of human rights to public and private actors in AI governance through a concrete treaty focusing on human rights in the context of AI-DDD. In my view, the enactment of such an instrument is critical to ensure that AI will be developed in an ethical and societally beneficial way, in line with human rights. This approach may offer "a powerful additional tool in our armour to regulate more effectively the activities of difficult regulatory targets" (Deva, 2017) within AI design, development, and deployment, supplementing AI governance while enhancing the level of respect and protection for human rights, independently of the states and their willingness or objective ability to hold private actors accountable for human rights violations. My suggestion, which builds upon the Business and Human Rights treaty and the international human rights and businesses discourse, is by no means a fully developed nor a standalone solution and will be probably complementary to other governance initiatives. To support my case, I will refer to the key role of the private sector in the AI industry, the adverse effects of regulatory competition between states and the arbitrage private companies may have in shaping AI governance in a way that does not safeguard respect to human rights, and finally the promise of not only respecting but also promoting human rights through AI.

As we have already discussed, the majority of algorithms and AI systems are privately designed, owned, and operated, given that private technology companies enjoy the leading position in AI research and innovation (Jang, 2017; Popkin et al., 2020). Most notably, from the screening of CVs to credit background checks, and from recommendation systems to online content moderation, private companies run the show. Even in cases in which algorithms or AI systems are used by public actors, in terms of border control, healthcare provision, or the justice system, the AI and algorithms employed are privately designed and developed, while their deployment is commonly outsourced to a private entity (Levy et al., 2021; Pizzi et al., 2020). Moreover, as Saslow and Lorenz (2019) observe "AI development in the private sector has followed the quick and dirty paradigm that Facebook popularized: 'move fast and break things.'" This rush to innovate and develop products and services is at odds with meaningfully and in detail reviewing the human rights implications of a new product/service, exploring and mitigating the risks.

Consequently, considering the limitations of ethical guidelines, having the means to directly impose human rights obligations upon private actors as well is key to ensuring that AI-DDD will respect human rights at each and every stage of AI-DDD. Moreover, in case of interference with human rights and adverse effects, it will be possible to effectively cease the interference, hold those responsible accountable and get access to meaningful remedies (McGregor et al., 2019), without depending on domestic regulation, and overcoming jurisdictional obstacles, regardless of whether it will be used by a private actor or public authorities. Reversely, the state-centric regime based on our current experience may significantly challenge access to remedy and redress (Pizzi et al., 2020).

Moreover, the introduction of directly binding human rights provisions is necessary to restrain the influence big technology companies may exercise on domestic legislation and to revoke in practice the regulatory competition between states, which, aiming to attract leading AI companies and investments in AI research and innovation, tend to significantly lower the threshold of human rights protection in the context of AI governance. The fact that the private actors will be held directly responsible under international human rights law will essentially diminish their regulatory leverage, allowing the more efficient and effective application of human rights in the context of AI, not only internationally but also at a national level. This development will arguably rectify, or at least supplement existing and proposed regulatory instruments, given that it has been observed that human rights do not tend to be central in national AI strategies (Fjeld et al., 2020; Fukuda-Parr & Gibbons, 2021; Latonero, 2018). For instance, as noted earlier even the Proposed AI Act by the EU leaves a lot to be desired in terms of the

adequate protection of fundamental rights and freedoms, as noted earlier (Smuha et al., 2021).

Reversely, the direct application of human rights on private actors irrespective of domestic legislation will also significantly reduce the other side of the regulatory arbitrage that the current situation permits. Namely, if human rights apply directly to public and private actors then the states will have no means to put pressure on the private companies in the AI industry, minimising the dangerous liaisons the current system allows. Furthermore, considering countries with a leading role in AI, or investing to gain a privileged position in the race for AI, such as China, or Saudi Arabia, are not bound by the UN human rights system, the direct application of human rights on the private entities may allow holding the companies directly responsible, even if they are located in a non-signatory country. To that end, the way the GDPR defines its scope can serve as a blueprint.

Looking beyond their function as negative rights, the direct application of human rights obligations to private actors will be valuable to ensure the ethical and socially beneficial design, development and deployment of AI systems and algorithms, not only in terms of respecting and protecting human rights in the context of AI-DDD, but also with the prospect of promoting human rights through these technologies for the flourishing of humanity. By introducing direct obligations, human rights will not be an afterthought anymore, but rather a guiding principle that will inform the design, development and deployment of AI systems and algorithms, not premised on the goodwill of the private actors, but on their binding duties. It will be also vital in reinforcing human rights respect and protection in other fields, in which AI systems are employed. For instance, the employment of algorithms and AI systems for content moderation on the Internet has created extensive freedom of expression challenges (Balkin, 2017; Gillespie, 2020; Tirosh, 2017), while their vast employment for news indexing and the creation of personalised recommendations and targeted advertising may adversely affect autonomy and self-determination (Laitinen & Sahlgren, 2021). Holding those designing and deploying such algorithms accountable for human rights in the context of AI governance, bears the promise to improve the situation, without necessarily introducing new regulatory instruments on the Internet.

Furthermore, a treaty can offer substantial guidelines to the technology companies, as well as the individual developers and designers as to what is expected from them and how to develop AI that would benefit humanity. The process of drafting it, although arguably arduous and time-consuming, provided that it will bring together researchers, academics and scientists, along with policy-makers, human rights experts, civil society groups and representatives of the industry, can contribute to further exploring key concepts that are highly relevant and heavily contested in AI governance discourse, such as explainability, transparency, accountability, and algorithmic fairness. Moreover, such a process is critical to fostering international cooperation in the governance of a transborder technology with potentially global impacts.

At the end of the day, although it may seem radical at first, my suggestion is not unheard of. On the contrary, it builds upon the long and thick discourse in the business and human rights field, jointly with the remarks offered by human rights and technology advocates. More than a position, it is a call for action, since, whereas we all seem to agree that technologies that penetrate the "lifeworld" producing consequential impacts that shape and affect individuals' options and choices, rights, and freedoms, should not be left outside the human rights discourse and system (Mylly, 2009), the regulatory initiatives still fall short to introduce a concrete framework, which specific human rights protections that can be equally invoked against both public and private actors.

5 AI governance beyond human rights

Engaging with the main question underpinning this paper, namely "how should we regulate AI?" so far, I focused on the role of human rights in promoting the ethically sound and societally beneficial AI-DDD. Slightly shifting my focus, in this section, I discuss what insights AI governance discourse can derive from the broader field of technology governance, and more specifically from IG. AI, and AI governance in particular, emphatically poses a variety of ethical, policy and regulatory questions, which although urgent, are in fact hardly new. Ethical dilemmas and moral questions focused on the effects and the impact of technology (Doppelt, 2001; Feenberg, 1994; Grimes & Feenberg, 2013), as well as concerns regarding the legal, ethical and social externalities of specific technologies, and the objective to steer technological progress towards ethical and societally beneficial ends, are as old as technology and society (Strobel & Tillberg-Webb, 2009). For example, the objective of taming a new and disruptive technology is an integral part of technology governance since the dawn of the field (Zimmerman, 1995). Moreover, some of the key questions regarding the governance of AI are inherent in the field of technology governance and have been thoroughly discussed in terms by several researchers, from various angles (see for example Feenberg, 1994; Winner, 1977; Hess, 2015).

Hence, since AI governance, both as an academic field and as a public policy objective, is still unfolding, it may be useful to seek insights into the trajectory of how governance modes and mechanisms emerged in other fields of technology. In that context, the history of IG may be particularly relevant and insightful, shedding light on less apparent

challenges, assisting policymakers to set the priorities and objectives based on educated decisions, and avoiding mistakes or omissions that led to adverse policy and/or regulatory outcomes. Building on these observations, I wish to highlight the role IG may have as an example for AI governance, and discuss three critical points from technology governance that may be valuable for AI governance, building upon the IG experience.

5.1 IG and AI Governance

Even though algorithms and AI systems currently pose one of the most significant regulatory challenges in terms of technology governance in the twenty-first century, IG has certainly been, and to an extent still remains (Radu, 2019), the most challenging technology governance "experiment" of the twentieth century. The field loosely defined as "IG" is a constantly expanding area of policy competition, interdisciplinary research, public debate, and ideological quarrel over how and by whom the Internet ought to be governed (DeNardis, 2014). As the research field that led to the emergence of AI was still developing, the Internet has been already awarded the label of "the most significant disruptive innovation" (Greenstein, 2015; Lyytinen & Rose, 2003), and the quest for developing and establishing the most appropriate governance model for the Internet started during the late 1990s.[8] Until then, for the almost thirty years that the Internet was been developing inside research institutions and universities, any tasks of coordination and oversight were largely trusted upon the engineers and developers working on its expansion, establishing a form of unofficial, ad-hoc governance that was gradually institutionalised (Castells, 2003; DeNardis, 2014; Radu, 2019).

In the early trajectory of the Internet, the lack of direct regulation and the absence of governmental interference in the form of active steering created the assumption that the Internet did not require or allow any form of external control, which soon became particularly popular (Murray, 2007). Moreover, cyber-libertarianism, essentially a technology premised version of libertarianism, has nourished the belief that industry-driven self-governance offers a better alternative to technology governance than state-initiated regulation through law (Johnson & Post, 1996; Perritt, 1997; Reidenberg, 1997), popularising private governance in the field of technology. Thus, the Internet advanced for a long time under a market-driven, self-governance model, as after it was commercialised, private entities got the leading role, leading to the excessive privatisation of several aspects of its governance (Belli et al., 2017; Jørgensen & Zuleta, 2020). After several shifts and twists, today the Internet is governed through a multi-stakeholder model by national governments, civil society, and a variety of private entities and non-governmental institutions.

Black and Murray (2019) have already highlighted the similarities and relevance between AI and IG, while Nemitz (2018) has noted that the influence of IG on the AI governance discourse is relatively apparent, while he tried to draw attention to the negative implications the example set by IG may have on private power in terms of AI governance. Apart from sharing the title "disruptive innovation", having both attracted a fair share of public attention and media coverage, the Internet and AI also share a wide array of similarities in their history. For example, they both started as state-funded research, developing away from the public eye for a considerable time. The launch of Sputnik 1 by the Soviet Union (USSR) on 4 October 1957 is credited as one of the main incentives behind the development of the Internet (Curran et al., 2016). The original network, which eventually became the Internet, was developed out of the military research project under the US Department of Defense (DoD), subsidized by the Advanced Research Projects Agency (ARPA), with the assistance of Rand, MIT, the University of California and the British National Physical Laboratory (NPL) (Radu, 2019). Almost at the same time, AI was being established as a research field at Dartmouth College workshop, during a period in which research into machine intelligence, neuronic functions and robotics was receiving significant funding from the American, British and Japanese governments (Radu, 2021). Hence, they both started as stated funded research.

Although following different paths and different paces, they both advanced largely free from regulatory burdens for the largest part of their early years, while after state funding was withdrawn, they both became dependent on the private sector (Greenstein 2015; Radu, 2019, 2021). The leading role of the private sector and the commercial application of the technology seems to be critical for their success, which also brought to the forefront their negative aspect, calling for governance (Mueller, 2013; Reed, 2004; Smuha, 2021a). In the case of the Internet, it was the dissemination of sexually explicit content, defamation instances, and the sharing of copyrighted material that attracted the attention of the legislator (Zittrain, 2006). On the other hand, in the context of AI, it was the ethical challenges of its implementation, combined with instances of algorithmic bias and discrimination across different sectors combined with extensive surveillance and datafication of marginalised communities that emphatically brought the need for steering to the table (Radu, 2021; Taeihagh, 2021).

[8] For more details see for example Mueller, M. (2002). *Ruling the Root: IG and the Taming of Cyberspace*. MIT Press, Zittrain, J. L. (2008). *The Future of the Internet and How to Stop It The Harvard community has made this article openly Please share how this access benefits you. Your story*. 10.1086/261502.

In the meantime, the absence of state funding and oversight allowed private actors, and most prominently technology companies, to govern the respective fields, essentially vesting de facto governing power and authority to private hands (Bietti, 2019; Black & Murray, 2019; Zittrain, 2006). In both cases, the private actors used their power and privileged position to set some at least basic governing structures, reserving for themselves considerable discretion in setting the means, the rules, the objectives and the priorities for governance, framing, or at least influencing the narrative. For the Internet it was through the proliferation of the cyber-libertarian narrative and various forms of Terms of Use (ToS) that the self-governance model was established, while in AI, it was through the rush to ethics and the self-promoted ethical codes that self-regulation has been so far promoted. Yet it is fair to admit that at least in the case of AI, the governments realised relatively earlier both the risks inherent in AI systems and algorithmic decision-making, and the opportunities in leading the table of AI governance and took initiatives to address the risks and challenges of AI technologies. However, is it soon enough to avoid "AI Libertarianism" (Black & Murray, 2019) and prevent the establishment of private custodians of human rights in the context of AI-DDD, and an elite of industry governors that will enjoy a privileged position in AI governance with weak or meaningless restrictions?

Libertarianism is a common thread running across several technologies, from the Internet to blockchain (Zamani, 2022). In the context of IG, the radical line of cyber-libertarianism was equating state governance and law with hierarchical structures, cumbersome bureaucratic processes, and slowness portrayed as the opposite of the bottom-up, open, inclusive private governance. Such narratives were particularly prominent and influential during the nascent days of the Internet and the early stages of IG, leading to assumptions that ascribed governing power and authority to private actors, leading to governing arrangements that hampered the application of law online, favouring private power instead (Chenou, 2014). Black and Murray (2019) appear rather pessimistic, arguing that "if one were to predict the outcome of this based upon our experience of the internet regulation case study, it does not make for happy reading." Nevertheless, I dare to be slightly more positive, given that the experience of IG can serve as a valuable source of insights, that may inform our choices and regulatory decisions, assisting in avoiding choices that led down to unhappy paths.

Additionally, the advent of AI governance coincides with a rather momentous point in the history of IG (Suzor, 2019). Large scale regulatory reforms initiated by several governments and intergovernmental organisations, seek to introduce meaningful restrictions on arbitrary private power, and effective protective measures for individuals' rights and freedom (Redeker et al., 2018). This movement in IG has fuelled broad discussions about human rights and human dignity in terms of technology governance, as well as concerning the necessity for adequate restriction to private power, highlighting the limitations of self-governance and market rules for governing general-purpose, disruptive and human rights-relevant technologies. Hence, as AI governance is advancing within this discursive, research and policy momentum, we may avoid repeating the past.

5.2 What IG may teach to AI governance

5.2.1 Between overregulating and underregulating: Exceptionalism and the limits of existing law

Wu (2010) observes that in terms of technology regulation we tend to become obsessed with "the newest new," perceiving each and every new technological development as necessarily disruptive, revolutionary, transformative, and exceptional. Based on this understanding, we seek equally new, transformative, and exceptional means to govern it, in a pendular movement from pessimism, technophobia, and fear, to techno-utopianism, excessive enthusiasm, and recklessness. Such new means may in turn lead to over-regulation or under-regulation, depending on whether new rules and structures and considered necessary or the reduction of regulatory control is deemed more appropriate. For instance, in the context of IG, strong exceptionalism, combined with the view that the Internet was so disruptive and novel has led to regulatory failures, such as the deregulation and the excessive privatisation of several of its key aspects (Chenou, 2014), or the excessive, and/or hastily-introduced regulation in others (Goldman, 2009; Tushnet, 2015).

Whereas there is no question that AI, algorithmic and smart technologies, in general, will have a major impact on society, introducing enormous and far-ranging transformations (Gruetzemacher & Whittlestone, 2022), some of the key questions regarding their governance are, as already noted, hardly new. Thus, as we strive to determine the means, modes, and principles to govern AI, we need to resist the temptation of exceptionalism and consider what we can learn from the experience of the governance of other technologies. Hence, to avoid over-regulation, hastily introduced, or ill-premised governance structures, as well as to prevent losing valuable time to establish the much-needed regulations, we need to avoid exaggerating not only the risks and opportunities these technologies bear but also the element of novelty and exceptionality. In this context, considering the adverse effects that AI may have on the job market, for instance, the introduction of specific legislation that focuses only on AI and access to work may be a more exceptional way of handling this challenge than finding meaningful ways to ensure the protection of the right to work extending or

expanding upon existing frameworks or finding meaningful ways to ensure the application of already existing rights.

5.2.2 Definitions, conceptualisations, governing images, and governance

IG "is neither a homogenous object of governance nor of study" (Radu, 2019). From this angle, IG shares with AI the lack of conceptual and terminological clarity, as well as the relevance of multiple technologies, techniques and processes, since none of them is a monolithic artefact. Drake (2004) has suggested that the lack of consensus regarding the conceptual core of IG is also reflected in the lack of consensus "about which issues, and institutions are and should be involved in what manner". This may be also true in the case of AI, particularly since the variety of technologies and applications relevant may trigger distinct bodies of law and regulation (Stahl et al., 2021a, b). In turn, a specific definition may be more favourable than others. However, the definition and specific framing of AI are very significant and impactful from a governance perspective. As Miller (2012) explains, policymaking is a power struggle that involves meaning capturing as well as the dominance of one narrative over the others.

This process is closely related to the formation of governing images, which in turn "have an important, even decisive, influence on the unfolding of governing processes," as they serve as key points of departure "for the selection of governing instruments and taking governing action" (Kooiman, 2003). From that aspect, they also have a normative dimension, since they influence or even form the understanding of what and how it is to be governed in a consequential manner. Similarly, Ezrahi argues that there is an inherent "fictional layer" in every governance model, that has also particular performative dimensions. He claims that there is some kind of a collective, unconscious imagination, an "imaginary", that produces ever-changing images of what constitutes legitimate power and authority, which in turn compete for enactment and institutionalisation in the political arena (Ezrahi, 2012). For instance, in the context of IG, the way the Internet was perceived under the influence of the "cyberspace" metaphor and the cyber-libertarian narrative had broad and far-reaching ramifications regarding the attribution of power and authority, as well as in the priorities and the means of governance (Wyatt, 2004; Larsson, 2013c).

Consequently, considering the vital role of conceptualisations (Larsson, 2013a) it is critical in the context of AI governance to address the conceptualisation and framing of AI as an integral step toward developing governance structure. It would be valuable to demystify AI systems and algorithms, focusing on how the element of intelligence is to be perceived, and promoting the adoption of comprehensive yet simple definitions (Bryson, 2020), that can be widely shared and become the common ground for governance. Hence, apart from critically assessing the suggested definitions in academia and policy, we need to choose the definition as well as our discursive tools carefully, remaining mindful of their entailments, connotations, and impacts on governance. In the same context, it is essential to thoroughly review the surrounding narratives and the way they shape or affect the priorities and means of governance, along with the implications they may have regarding power and authority in governing AI.

5.2.3 Democratic control, human rights premise, and the significance of human rights-trained and empowered designers/developers

Finally, learning from the IG history and experience, and deriving useful insights from the current momentum, at which IG is sought to be injected with meaningful human rights protections and democratic governance principles (Suzor, 2019), it is important to place AI governance within a framework of democratic scrutiny, conscious democratic control (Almeida et al., 2021; Vesnic-Alujevic et al., 2020), and human rights protections (Leslie et al., 2022; Yeung et al., 2019). Whereas in modern constitutional democracies such a request sounds self-evident or presumed already satisfied, in fact, technology governance is commonly a non-democratic procedure (Dotson, 2015; Feenberg, 1994). Decisions about technology development and deployment are largely taken behind closed doors, either by private entities or through technocratic bodies, that premise their decisions on non-democratic, and non-political criteria. In the context of the Internet, most of the critical policy issues were commonly framed or perceived as purely technical ones. Thus, issues closely related to access to the Internet or privacy, such as the availability of Internet Protocol (IP) addresses, or the encryption of information transmitted through different Internet protocols, received little public attention and political debate, regardless their public policy implications. It was only recently that Internet policy and relevant legislation got into the spotlight in political debates and public discourse, as was the case of Net Neutrality in terms of the US elections (Gibson, 2017), becoming issues of democratic debate and deliberation.

In the context of AI governance, whereas the rapid deployment and adoption of AI-based products and services across various sectors, including public administration and the justice system, is increasingly promoted, "there is a notable lack of prior scrutiny, democratic oversight and public debate" (Rachovitsa & Johann, 2022). The rapid pace of technological change and the complexity of the technologies may make the democratic process and procedures seem rather outdated, and irrelevant or unsuitable to steer such a disruptive technology, while the promise of progress may

make democratic control seem unnecessary. Nevertheless, considering the wide variety and the range of adverse effects AI systems and algorithms may have, and their potential impact, creating the necessary mechanisms to ensure not only that policy and decision-making about such impactful and consequential technologies will reflect and adequately represent the views of the citizens, but also that critical decision will be publicly discussed, adequately debated and scrutinised and that any decision-making process will be transparent and open to challenge, is of at most importance. Hence, it is crucial to premise AI governance upon a democratic framework and prevent the decoupling of its governance from democratic control and decision-making.

Beyond democratic control, returning to human rights and perceiving them not merely as binding legal rights, but as a set of foundational principle that should inform and shape the design, development and deployment of new technologies, in AI governance we have the opportunity to premise AI-DDD directly on a new, more focused and actionable approach to human rights. What the IG history taught us is that in designing and governing such all-purpose, transformative technologies with impactful consequences, human rights cannot be an after-thought. On the contrary, a human rights-premised, human-centric approach should lead and underlie all design, development, and deployment stages of a new technology. Respect for human rights, as well as avoiding interference with fundamental rights and freedoms should be a core design principle, guiding technical decision-making, design and standardisation. From this angle, it is not enough to premise AI governance on human rights at a policy level only. Human rights-centred approach should also underpin design, standardisation, and technical decision-making in terms of AI-DDD (Article19, 2019),

In turn, this last remark highlights the need for human rights training as a core element of the designers' and developers' curriculum, and the necessity to develop adequate practical tools that will empower and encourage designers and developers in making not only ethically sound, but also human rights-informed decisions. In the context of the Internet and IG, since the mid-2010s there is a remarkable effort devoted to the development of practical tools that would allow practitioners to make ethical and human rights-informed decisions while engaging in design and standardisation.[9] Similarly, perceiving AI and algorithms as malleable, human creations that are critical for human rights, highlights the necessity for designers and developers to have not only adequate ethical training and tools to reflect upon the different ethical aspects of their designs (Minkkinen et al., 2021; Papagiannidis et al., 2021), but also sufficient training, support, and encouragement to reflect upon the human rights implications and engage in human rights due diligence in the course of their tasks.

To that end, it is significant to take the necessary steps to inject their training with human rights modules and develop adequate tools to support them in design, development, and deployment procedures, as well as the necessary mechanisms to encourage and empower them to make decisions that not simply respect human rights, but hopefully also use the affordances of AI technology to promote them. From this angle, a human rights-premised model for AI governance requires the active participation and the meaningful engagement of not only the policymakers, the legal community, human rights experts, sociologists and ethicists, but also of the technical and scientific community developing and designing AI systems and algorithms.

6 Concluding Thoughts

AI no longer resides in science fiction movies and books. From recommendation systems, and virtual assistants, to self-driving vehicles and algorithmically informed court decisions, various applications of algorithmic systems, and processes of divergent automation and intelligence, are already widely employed. The proliferation of AI in the last two decades has been a catalyst for automation and efficiency in several domains but has also had a wide range of harmful consequences, and unanticipated adverse effects, including algorithmic bias, and problematic legal certainty due to questions regarding transparency, accountability and liability. It also enabled enhanced surveillance on an unprecedented scale and led to various cases in which AI systems and algorithms had an adverse impact on human rights, such as freedom of expression and privacy. Although concerns of various sorts and validity, combined with ethical dilemmas, political discussion, and policy debate constitute an integral part of the trajectory of every new technology (Muller, 2020a, b; Wu, 2010) ethical considerations and governance questions about AI and algorithmic systems have attracted significant attention, which continuously grows as the technologies in question become more ubiquitous.

In the context of amplifying the benefits and mitigating the risks, AI governance is an emerging field of regulatory and normative theorisation, experimentation, and debate, over the means and modes of controlling this sum of new and disruptive technologies, that is increasingly attracting public, academic, and political attention on a global scale. In this paper, I argued that to steer these new and powerful forces of disruption toward the benefit of society it is necessary to shift our attention from ethical principles and

[9] See for example RFC 8280, (Internet Research Task Force (IRTF), RFC 8280, by. N. Ten Oever and C, Cathy, Research into Human Rights Protocol Considerations, October 2017, available at https://www.rfc-editor.org/rfc/rfc8280.txt) and the human rights check lists developed by the.

guidelines to specific governance mechanisms and structures. Whereas AI Ethics have been integral in identifying several of the key risks and challenges in the field of AI systems and algorithmic decision-making, governance and hard-law instruments are necessary to ensure that AI will be designed, developed and deployed for the benefit of society (Black & Murray, 2019; Fukuda-Parr & Gibbons, 2021; Yeung et al., 2019).

Although there is a wide consensus that international human rights law provides a valuable organising framework of concrete rights and binding obligations (Rachovitsa & Johann, 2022), as well as an elaborated institutional structure (Yeung et al., 2019) to support and monitor adherence, it remains unclear how international human rights, as they currently stand, can effectively address the challenges posed by the use of AI systems and algorithms (Rachovitsa & Johann, 2022). Taking up this challenge, while focusing on the relevance and the role human rights ought to have in AI governance, I argued for the necessity and normative suitability of employing human rights not merely as ethical principles or governing guidelines, but as binding obligations, applicable to every private and public actor in the context of AI governance, through a treaty for human rights in AI that will equally apply to both public and private actors. Although not unproblematic, nor uncontroversial, my argument aims to supplement the existing literature offering insights and argumentation from the direct horizontality and the business and human rights discourse.

Looking beyond human rights and considering the wide array of challenges in AI governance, I suggest that IG may offer several valuable insights that will allow AI governance to be a sum of informed decisions, allowing us to avoid regulatory failures and create a better governance model. Building upon the experience of IG, AI governance ought to be premised on a model that will balance between over-regulation and under-regulation, regulatory/governance innovations and trust to traditional, well-established rules that can adequately respond to the challenges posed by AI. Similarly, whatever form this model ultimately takes, it should be premised on clear and specific definitions, concrete conceptualisations, and expectations about AI, that will inform and shape the governance mechanisms. Moreover, considering its power and effects, while learning form the IG past, it is significant to ensure that AI governance will be premised on conscious democratic decision-making, human rights-premised design, and human rights-aware designers and developers.

Nonetheless, nothing in this paper is to be read as a fully developed suggested solution. My remarks, apart from inputs to the AI governance discourse, are points of departure aimed to encourage academic discussion about the available paths that will enable human rights to meaningfully fulfil their role in AI governance, as well as deeper and more thorough research into the relationship between AI and IG.

Declarations

Funding and/or Conflicts of interests/Competing interests The author has no relevant financial or non-financial interests, nor any competing interests to declare that are relevant to the content of this article.

References

Agrawal, A., Gans, J. & Goldfarb, A. (2019). Artificial Intelligence, Automation, and Work. In *The Economics of Artificial Intelligence* (pp. 197–236). University of Chicago Press. https://doi.org/10.7208/chicago/9780226613475.003.0008

Aizenberg, E. & van den Hoven, J. (2020). Designing for human rights in AI. *Big Data and Society*, *7*(2). https://doi.org/10.1177/2053951720949566

Algorithm Watch. (2020). *AI Ethics Guidelines Global Inventory*. Algorithm Watch. https://inventory.algorithmwatch.org/?sfid=172&_sf_s=urban&sort_order=_sfm_i_date+desc+alpha%0Ahttps://inventory.algorithmwatch.org/%0Ahttps://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/

Allen, J. R., & Massolo, G. (2020). *AI in the Age of Cyber-Disorder | ISPI*. https://www.ispionline.it/en/pubblicazione/ai-age-cyber-disorder-28309

Alston, P. (2005). *Non-State Actors and Human Rights*. Oxford University Press.

Alston, P. Report of the Special rapporteur on extreme poverty and human rights, United Nations UN Doc. A/74/493 (2019). https://doi.org/10.2139/ssrn.2534341

Alston, P., & Quinn, G. (2017). The nature and scope of states parties' obligations under the international covenant on economic, social and cultural rights. In *Economic, Social and Cultural Rights* (pp. 3–76). Routledge. https://doi.org/10.4324/9781315257044-2

Aragão, D. M., & Roland, M. C. (2017). The need for a treaty: Expectations on counter-hegemony and the role of civil society. In *Building a Treaty on Business and Human Rights* (pp. 131–153). Cambridge University Press. https://doi.org/10.1017/9781108183031.007

Article19. (2019). *Governance with teeth: How human rights can strengthen FAT and ethics initiatives on artificial intelligence -*. https://www.article19.org/resources/governance-with-teeth-how-human-rights-can-strengthen-fat-and-ethics-initiatives-on-artificial-intelligence/

Asaro, P. M. (2019a). *A Review of Private Sector AI Principles: A Report Prepared for UNIDIR*.

Asaro, P. M. (2019b). AI ethics in predictive policing: From models of threat to an ethics of care. *IEEE Technology and Society Magazine, 38*(2), 40–53. https://doi.org/10.1109/MTS.2019.2915154

Bachelet. (2021). Urgent action needed over artificial intelligence risks to human rights | | UN News. In *UN News* (pp. 5–9). https://news.un.org/en/story/2021/09/1099972

Bairaktarova, D., & Woodcock, A. (2017). Engineering Student's Ethical Awareness and Behavior: A New Motivational Model. *Science and Engineering Ethics, 23*(4), 1129–1157. https://doi.org/10.1007/s11948-016-9814-x

Balkin, J. M. (2017). Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation. *SSRN Electronic Journal*, 1149–1210. https://doi.org/10.2139/ssrn.3038939

Balkin, J. M. (2018). Free speech is a triangle. In *Columbia Law Review*.

Basart, J. M., & Serra, M. (2013). Engineering Ethics Beyond Engineers' Ethics. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-011-9293-z

Bayamlioglu, E. (2018). Contesting Automated Decisions. *European Data Protection Law Review*, *4*(4), 433–446. https://doi.org/10.21552/edpl/2018/4/6

Bednar, K., Spiekermann, S., & Langheinrich, M. (2019). Engineering Privacy by Design: Are engineers ready to live up to the challenge? *Information Society, 35*(3), 122–142. https://doi.org/10.1080/01972243.2019.1583296

Belli, L., Francisco, P. A., & Zingales, N. (2017). Law of the land or law of the platform? Beware of the privatisation of regulation and police. In *Platform regulations: How platforms are regulated and how to they regulate us*, (Issue December, pp. 41–64).

Benedek, W., Kettemann, M. C., & Senges, M. (2017). The Humanization of IG: A Roadmap Towards a Comprehensive Global (Human) Rights Architecture for the Internet. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2798361

Bhatnagar, S., Alexandrova, A., Avin, S., Cave, S., Cheke, L., Crosby, M., Feyereisl, J., Halina, M., Loe, B. S., Ó hÉigeartaigh, S., Martínez-Plumed, F., Price, H., Shevlin, H., Weller, A., Winfield, A., & Hernández-Orallo, J. (2018). Mapping intelligence: Requirements and possibilities. In *Studies in applied philosophy, epistemology and rational ethics* (Vol. 44, pp. 117–135). https://doi.org/10.1007/978-3-319-96448-5_13

Bietti, E. (2019). *From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy*. https://papers.ssrn.com/abstract=3513182

Bilchitz, D. (2016a). Corporations and the limits of state-based models for protecting fundamental rights in international law. *Indiana Journal of Global Legal Studies, 23*(1), 143–170. https://doi.org/10.2979/indjglolegstu.23.1.143

Bilchitz, D. (2016b). The Necessity for a Business and Human Rights Treaty. *Business and Human Rights Journal, 1*(2), 203–227. https://doi.org/10.1017/bhj.2016.13

Bilchitz, D. (2017). Corporate Obligations and a Treaty on Business and Human Rights. In S. Deva, & D. Bilchitz (Eds.), *Building a treaty on business and human rights: Context and contours* (pp. 185–215). Cambridge University Press. https://doi.org/10.1017/9781108183031

Binns, R. (2017). Fairness in Machine Learning: Lessons from Political Philosophy. *Conference on Fairness, Accountability, and. Transparency,* 1–11. http://arxiv.org/abs/1712.03586

Black, J. & Murray, A. (2019). Regulating AI and machine learning: setting the regulatory agenda (complementar). *European Journal of Law and Technology*, *10*(3), 1–17. http://eprints.lse.ac.uk/102953/4/722_3282_1_PB.pdf

Blitt, R. C. (2012). Beyond Ruggie's Guiding Principles on Business and Human Rights: Charting an Embracive Approach to Corporate Human Rights Compliance. *SSRN Electronic Journal*, *48*(1). https://doi.org/10.2139/ssrn.1907778

Boddington, P. (2020). Normative modes: Codes and standards. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford handbook of ethics of AI* (Issue July, pp. 123–140). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.7

Borgesius, F. Z. (2018). Discrimination, artificial intelligence, and algorithmic decision-making. In *Council of Europe*. https://www.ftc.gov/news-events/blogs/business-blog/2020/04/using-artificial-intelligence-algorithms

Brownlie, I. (2019). Principles of Public International Law. In *Verfassung in Recht und Übersee* (Vol. 14, Issue 1). Oxford University Press.

Brownsword, R. (2017). From Erewhon to AlphaGo: For the Sake of Human Dignity, Should We Destroy the Machines? *Law, Innovation and Technology, 9*(1), 117–153.

Brownsword, R. (2019). *Law Technology And Society: Re-Imagining The Regulatory Environment*. Routledge.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., … Amodei, D. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. ArXiv. http://arxiv.org/abs/1802.07228

Bryson, J. (2019). The past decade and the future of AI's impact on society. In *Towards a new enlightenment? A transcendent decade* (Vol. 11, pp. 1–34). https://www.bbvaopenmind.com/wp-

Bryson, J. (2020). The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation. In *The Oxford Handbook of Ethics of AI*. www.oxfordhandbooks.com

Bryson, J. J. & Theodorou, A. (2019). How Society Can Maintain Human-Centric Artificial Intelligence. In M. Toivonen & E. Saari (Eds.), *Human-Centered Digitalization and Services* (pp. 305–323). Springer. https://doi.org/10.1007/978-981-13-7725-9_16

Bucchi, M. (2009). Beyond technocracy: Science, politics and citizens. In *Beyond Technocracy: Science, Politics and Citizens*. https://doi.org/10.1007/978-0-387-89522-2

Buergenthal, T. (2006). The Evolving Human Rights System. *The American Journal of International Law, 100*(4), 783–807.

Busacca, A., & Monaca, M. A. (2020). Processing of Personal Data and AI: GDPR Guarantees and Limits (Between Individual Data and BIG DATA). *Studies in Systems, Decision and Control, 288*, 51–64. https://doi.org/10.1007/978-3-030-45340-4_6

Butcher, J., & Beridze, I. (2019). What is the State of Artificial Intelligence Governance Globally? *RUSI Journal, 164*(5–6), 88–96. https://doi.org/10.1080/03071847.2019.1694260

Cadwalladr, C. (2020). *Fresh Cambridge Analytica leaks'shows global manipulation is out of control*. The Guardian. https://www.theguardian.com/uk-news/2020/jan/04/cambridge-analytica-data-leak-global-election-manipulation?CMP=Share_AndroidApp_Slack

Cannarsa, M. (2021). Ethics Guidelines for Trustworthy AI. In *The Cambridge handbook of lawyering in the digital age* (pp. 283–297). Cambridge University Press. https://doi.org/10.1017/9781108936040.022

Castells, M. (2003). *The Internet Galaxy Reflections on the Internet*. Oxford University Press.

Cath, C. (2018). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2133), 20180080. https://doi.org/10.1098/rsta.2018.0080

Chander, A. & Pasquale, F. (2016). The Racist Algorithm? *Michigan Law Review*, *498*. http://ssrn.com/abstract=2795203

Chenou, J. M. (2014). From Cyber-Libertarianism to Neoliberalism: Internet Exceptionalism, Multi-stakeholderism, and the Institutionalisation of IG in the 1990s. *Globalizations, 11*(2), 205–223. https://doi.org/10.1080/14747731.2014.887387

Chenou, J. M., & Radu, R. (2019). The "Right to Be Forgotten": Negotiating Public and Private Ordering in the European Union. *Business and Society, 58*(1), 74–102. https://doi.org/10.1177/0007650317717720

Cherednychenko, O. O. (2007). Fundamental Rights and Private Law: A Relationship of Subordination or Complementarity? *Utrecht Law Review, 3*(2), 1–25.

Cihon, P., Maas, M. M., & Kemp, L. (2020). Fragmentation and the Future: Investigating Architectures for International AI

Governance. *Global Policy, 11*(5), 545–556. https://doi.org/10.1111/1758-5899.12890
Cihon, P., Maas, M. M., & Kemp, L. (2021). Should Artificial Intelligence Governance be Centralised?: Design Lessons from History. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3761636
Clapham, A. (2006). *Human Rights Obligations of Non-State Actors*. Oxford University Press.
Collins, C., Dennehy, D., Conboy, K. & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management*, *60*. https://doi.org/10.1016/j.ijinfomgt.2021.102383
Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2021). The AI Gambit — Leveraging Artificial Intelligence to Combat Climate Change: Opportunities, Challenges, and Recommendations. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3804983
Crawford, K. (2021). Time to regulate AI that interprets human emotions. *Nature, 592*(7853), 167. https://doi.org/10.1038/d41586-021-00868-5
Crootof, R. & Ard, B. (2020). Structuring Techlaw. *Harvard Journal of Law & Technology*, *34*(2). https://doi.org/10.2139/ssrn.3664124
Curran, J., Fenton, N., & Freedman, D. (2016). Misunderstanding the internet. In *Misunderstanding the Internet* (Routledge).
Dafoe, A. (2018). *AI Governance: A Research Agenda*. www.fhi.ox.ac.uk/govaiagenda
Daly, A., Hagendorff, T., Li, H., Mann, M., Marda, V., Wagner, B. & Wang, W. W. (2020). AI, Governance and Ethics: Global Perspectives. In *SSRN Electronic Journal* (Issue June). https://doi.org/10.2139/ssrn.3684406
Danaher, J. (2018). Toward an Ethics of AI Assistants: An Initial Framework. *Philosophy and Technology, 31*(4), 629–653. https://doi.org/10.1007/s13347-018-0317-3
Dawn, O. & Fedtke, J. (2008). Human Rights and the Private Sphere. *UCL Human Rights Review*.
de Almeida, P. G. R., dos Santos, C. D., & Farias, J. S. (2021). Artificial Intelligence Regulation: A framework for governance. *Ethics and Information Technology, 23*(3), 505–525. https://doi.org/10.1007/s10676-021-09593-z
de Witte, B. (2009). The crumbling public/private divide: Horizontality in European anti-discrimination law. *Citizenship Studies*. https://doi.org/10.1080/13621020903174670
DeNardis, L. (2014). *The Global War for IG*. Yale University Press.
Deva, S. (2017). Conclusion: Connecting the dots: How to capitalize on the current high tide for a business and human rights treaty. In *Building a treaty on business and human rights* (pp. 472–494). Cambridge University Press. https://doi.org/10.1017/9781108183031.019
Diver, L. (2021). Interpreting the Rule (s) of Code : Performance , Performativity , and Production. *MIT Computational Law Report*.
Doorn, N. (2012). Responsibility Ascriptions in Technology Development and Engineering: Three Perspectives. *Science and Engineering Ethics, 18*(1), 69–90. https://doi.org/10.1007/s11948-009-9189-3
Doppelt, G. (2001). What sort of ethics does technology require? *Journal of Ethics*. https://doi.org/10.1023/A:1011956206973
Dotson, T. (2015). Technological Determinism and Permissionless Innovation as Technocratic Governing Mentalities: Psychocultural Barriers to the Democratization of Technology. *Engaging Science, Technology, and Society*, *1*, 98–120. https://doi.org/10.17351/ests2015.009
Drake, W. (2004). Reframing IG discourse: Fifteen baseline propositions. *IG: Toward a Grand Collaboration*.
Dubber, M. D., Pasquale, F., & Das, S. (2020). *The Oxford Handbook of Ethics of AI*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.001.0001
Dutton, T. (2018). An Overview of National AI Strategies. In *Medium*. https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd
Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., … Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*. https://doi.org/10.1016/j.ijinfomgt.2019.08.002
Ebers, M., Hoch, V. R. S., Rosenkranz, F., Ruschemeier, H. & Steinrötter, B. (2021). The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS). *J*, *4*(4), 589–603. https://doi.org/10.3390/j4040043
Edwards, L. & Veale, M. (2017). Slave to the Algorithm? Why a "right to an explanation" is probably not the remedy you are looking for. *Duke Law & Technology Review*, *16*(18). https://doi.org/10.31228/osf.io/97upg
Elshafei, G., & Negm, A. (2017). AI Technologies in Green Architecture Field: Statistical Comparative Analysis. *Procedia Engineering, 181*, 480–488. https://doi.org/10.1016/j.proeng.2017.02.419
Erdélyi, O. J. & Goldsmith, J. (2018). Regulating Artificial Intelligence Proposal for a Global Solution. *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 95–101. https://doi.org/10.1145/3278721.3278731
Ertel, W. (2017). *Introduction to Artificial Intelligence*. Springer International Publishing. https://doi.org/10.1007/978-3-319-58487-4
Etinson, A. (2018). *Human Rights*. Oxford University Press.
European Commission. (2020). White paper on artificial intelligence: A European approach to excellence and trust. In *European Commission*. https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
Ezrahi, Y. (2012). Imagined democracies: Necessary political fictions. In *Imagined democracies: Necessary political fictions*. Cambridge University Press. https://doi.org/10.1017/CBO9781139198769
Facebook. (2020). *Facebook's five pillars of Responsible AI*. Retrieved October 27, 2021, from https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai/
Feenberg, A. (1994). The technocracy thesis revisited: On the critique of power. *Inquiry (united Kingdom), 37*(1), 85–102. https://doi.org/10.1080/00201749408602341
Feldstein, S. (2019a). *The Global Expansion of AI Surveillance*. https://carnegieendowment.org/files/WP-Feldstein-AISurveillance_final1.pdf
Feldstein, S. (2019b). The Road to Digital Unfreedom: How Artificial Intelligence is Reshaping Repression. *Journal of Democracy, 30*(1), 40–52.
Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3518482
Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083). https://doi.org/10.1098/rsta.2016.0112

Floridi, L. (2018). Soft Ethics and the Governance of the Digital. *Philosophy and Technology*, *31*(1). https://doi.org/10.1007/s13347-018-0303-9

Floridi, L. & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608F92.8CD550D1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities. *Risks, Principles, and Recommendations., 28*, 689–707. https://doi.org/10.1007/s11023-018-9482-5

Fukuda-Parr, S., & Gibbons, E. (2021). Emerging Consensus on 'Ethical AI': Human Rights Critique of Stakeholder Guidelines. *Global Policy, 12*(S6), 32–44. https://doi.org/10.1111/1758-5899.12965

Gasser, U., & Almeida, V. A. F. (2017). A Layered Model for AI Governance. *IEEE Internet Computing, 21*(6), 58–62. https://doi.org/10.1109/MIC.2017.4180835

George, E. R., & Laplante, L. J. (2017). Access to remedy: Treaty talks and the terms of a new accountability accord. In *Building a treaty on business and human rights* (pp. 377–407). Cambridge University Press. https://doi.org/10.1017/9781108183031.016

Gerards, J. (2019). The fundamental rights challenges of algorithms. *Netherlands Quarterly of Human Rights, 37*(3), 205–209. https://doi.org/10.1177/0924051919861773

Gibbons, E. D. (2021). Toward a More Equal World: The Human Rights Approach to Extending the Benefits of Artificial Intelligence. *IEEE Technology and Society Magazine*. https://doi.org/10.1109/MTS.2021.3056295

Gibson, G. (2017). Net neutrality repeal gives Democrats fresh way to reach millennials | Reuters. *Reuters Online*. https://www.reuters.com/article/us-usa-internet-election-analysis-idUSKBN1E901R

Gillespie, T. (2014). The Relevance of Algorithms. In *Media Technologies* (pp. 167–194). The MIT Press. https://doi.org/10.7551/mitpress/9780262525374.003.0009

Gillespie, T. (2018). Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media. In *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data and Society*, *7*(2). https://doi.org/10.1177/2053951720943234

Giovanola, B. & Tiribelli, S. (2022). Weapons of moral construction? On the value of fairness in algorithmic decision-making. *Ethics and Information TechnologyEthics and Information Technology*, *24*(3). https://doi.org/10.1007/s10676-022-09622-5

Godfrey, C. (2020). Legislating Big Tech: The Effects Amazon Rekognition Technology Has on Privacy Rights. *Intellectual Property and Technology Law Journal*, *25*. https://heinonline.org/HOL/Page?handle=hein.journals/iprop25&id=175&div=&collection=

Goldman, E. (2009). *The Third Wave of Internet Exceptionalism*. *497*, 1–3. http://blog.ericgoldman.org/archives/2009/03/the_third_wave.htm

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data and Society*. https://doi.org/10.1177/2053951719897945

Greenstein, S. M. (2015). How the Internet became commercial: innovation, privatization, and the birth of a new network. *In The Kauffman foundation series on innovation and entrepreneurship*.

Grimes, S., & Feenberg, A. (2013). Critical theory of technology. In S. Price, C. Jewitt, & B. Brown (Eds.), *The SAGE handbook of digital technology research*. SAGE Publications Ltd.

Gruetzemacher, R. & Whittlestone, J. (2022). The transformative potential of artificial intelligence. *Futures*, *135*. https://doi.org/10.1016/j.futures.2021.102884

Guihot, M., Matthew, A. & Suzor, N. P. (2017). Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence. *Vanderbilt Journal of Entertainment & Technology Law*, *20*(2), 385. https://doi.org/10.31228/osf.io/5at2f

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review, 61*(4), 5–14. https://doi.org/10.1177/0008125619864925

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines, 30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hamdani, K. & Ruffing, L. (2017). Lessons from the UN Centre on transnational corporations for the current treaty initiative. In S. Deva & D. Bilchitz (Eds.), *Building a Treaty on Business and Human Rights: Context and Contours* (pp. 27–47). Cambridge University Press. https://doi.org/10.1017/9781108183031.003

Hamon, R., Junklewitz, H., Malgieri, G., Hert, P. De, Beslay, L. & Sanchez, I. (2021). Impossible explanations?: Beyond explainable AI in the GDPR from a COVID-19 use case scenario. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 549–559. https://doi.org/10.1145/3442188.3445917

Hamon, R., Junklewitz, H., Sanchez, I., Malgieri, G., & De Hert, P. (2022). Bridging the Gap Between AI and Explainability in the GDPR: Towards Trustworthiness-by-Design in Automated Decision-Making. *IEEE Computational Intelligence Magazine, 17*(1), 72–85. https://doi.org/10.1109/mci.2021.3129960

Hao, K. (2019). In 2020, Let's Stop AI Ethics-washing and Actually do Something. *MIT Technology Review*. https://www.technologyreview.com/s/614992/ai-ethics-washing-time-to-act/

Hazenberg, J. L. J. (2016). Transnational Corporations and Human Rights Duties: Perfect and Imperfect. *Human Rights Review, 17*(4), 479–500. https://doi.org/10.1007/s12142-016-0417-3

Hess, D. (2015). Power, ideology, and technological determinism. *Engaging Science, Technology, and Society, 1*, 121–125. https://doi.org/10.17351/ests2015.010

High-Level Independent Group on Artificial Intelligence (AI HLEG). (2019a). A Definition of AI: Main Capabilities and Disciplines. In *European Commission*. https://ec.europa.eu/digital-single-%0A10.1145/3301275.3302317-

High-Level Independent Group on Artificial Intelligence (AI HLEG). (2019b). *Ethics Guidelines for Trustworthy AI* (Issue December). https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Hildebrandt, M. (2008). Legal and Technological Normativity: More (and less) than twin sisters. *Techne: Research in Philosophy and Technology*, *12*(3), 169–183. https://doi.org/10.5840/techne20081232

Hildebrandt, M. (2011a). Law at a Crossroads: Losing the Thread or Regaining Control? The Collapse of Distance in Real Time Computing. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1331963

Hildebrandt, M. (2011b). Legal Protection by Design: Objections and Refutations. *Legisprudence, 5*(2), 223–248. https://doi.org/10.5235/175214611797885693

Hildebrandt, M. (2015). *Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology*. Edward Elgar Publishing.

Hildebrandt, M. (2016). Law as Information in the Era of Data-Driven Agency. *Modern Law Review, 79*(1), 1–30. https://doi.org/10.1111/1468-2230.12165

Hildebrandt, M. (2017). Saved by Design? The Case of Legal Protection by Design. *NanoEthics, 11*(3), 307–311. https://doi.org/10.1007/s11569-017-0299-0

Hildebrandt, M. (2018). Algorithmic regulation and the rule of law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2128). https://doi.org/10.1098/rsta.2017.0355

Hildebrandt, M. (2019). *Law for Computer Scientists • Law for Computer Scientists 10. "Legal by Design" or "Legal Protection by Design"?*

Hogenhout, L. (2021). A Framework for Ethical AI at the United Nations. *Unite Paper*, 1–23. https://edition.cnn.com/2021/02/16/tech/emotion-recognition-ai-education-spc-intl-

Hopgood, S. (2018). *The Endtimes of Human Rights*. Cornell University Press. https://doi.org/10.7591/9780801469305

Horner, J. (2003). Morality, ethics, and law: Introductory concepts. *Seminars in Speech and Language, 24*(4), 263–274. https://doi.org/10.1055/s-2004-815580

IBM. (2020). *AI Ethics | IBM*. IBM Cloud Learn Hub. https://www.ibm.com/cloud/learn/ai-ethics

Jang, E. (2017). What Companies Are Winning The Race For Artificial Intelligence? *Forbes*. https://www.forbes.com/sites/quora/2017/02/24/what-companies-are-winning-the-race-for-artificial-intelligence/#7f04527ef5cd

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2.

Johnson, D. R., & Post, D. (1996). Law and Borders - The Rise of Law in Cyberspace. *Stanford Law Review*. https://doi.org/10.2307/1229390

Jørgensen, R. F., & Zuleta, L. (2020). Private Governance of Freedom of Expression on Social Media Platforms. *Nordicom Review, 41*(1), 51–67.

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of Parallel and Distributed Computing, 74*(7), 2561–2573. https://doi.org/10.1016/j.jpdc.2014.01.003

Kampourakis, I. (2019). CSR and Social Rights: Juxtaposing Societal Constitutionalism and Rights-Based Approaches Imposing Human Rights Obligations on Corporations. *Goettingen Journal of International Law, 9*(3), 537–569.

Kanalan, I. (2014). Horizontal Effect of Human Rights in the Era of Transnational Constellations: On the Accountability of Private Actors for Human Rights Violations. In *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2539110

Kemp, L., Cihon, P., Maas, M. M., Belfield, H., Seán, D., Leung, J., & Cremer, Z. (2019). UN High-level Panel on Digital Cooperation: A Proposal for International AI Governance. In *UN High-Level Panel on Digital Cooperation*, pp. 1–4.

Kennedy, R. (2021). The Ethical Implications of Lawtech. In D. Dennehy, A. Griva, N. Pouloudi, Y. K. Dwivedi, I. Pappas & M. Mäntymäki (Eds.), *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society* (Vol. 12896). Springer International Publishing. https://doi.org/10.1007/978-3-030-85447-8

Klonick, K. (2018). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review, 131*(6), 1599–1670.

Kooiman, J. (2003). Governing as governance. In *Governing as Governance*. SAGE Publications Ltd. https://doi.org/10.4135/9781446215012

Koops, B. J. & Leenes, R. (2014). Privacy regulation cannot be hardcoded. A critical comment on the "privacy by design" provision in data-protection law. *International Review of Law, Computers and Technology*, *28*(2), 159–171. https://doi.org/10.1080/13600869.2013.801589

Krishnamurthy, V. (2019). Are Internet Protocols the New Human Rights Protocols? Understanding "RFC 8280 - Research into Human Rights Protocol Considerations." In *Business and Human Rights Journal*. https://doi.org/10.1017/bhj.2018.30

Kunz, M. & Ó hÉigeartaigh, S. (2020). Artificial Intelligence and Robotization. In R. Geiß & N. Melzer (Eds.), *Oxford Handbook on the International Law of Global Security*. Oxford University Press.

Kuziemski, M., & Misuraca, G. (2020). AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications Policy, 44*, 6 101976. https://doi.org/10.1016/J.TELPOL.2020.101976

LaGrandeur, K. (2021). How safe is our reliance on AI, and should we regulate it? *AI and Ethics, 1*(2), 93–99. https://doi.org/10.1007/s43681-020-00010-7

Laitinen, A. & Sahlgren, O. (2021). AI Systems and Respect for Human Autonomy. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.705164

Lambrecht, A. & Tucker, C. E. (2018). Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2852260

Lane, L. (2018a). *The Horizontal Effect of International Human Rights Law: Towards a multi-level governance approach*. https://www.rug.nl/research/portal/publications/the-horizontal-effect-of-international-human-rights-law(d6becf0f-de98-45cd-a6ed-39cb4687cd23).html)

Lane, L. (2018b). The horizontal effect of international human rights law in practice: A comparative analysis of the general comments and jurisprudence of selected united nations human rights treaty monitoring bodies. In *European Journal of Comparative Law and Governance* (Vol. 5, Issue 1). https://doi.org/10.1163/22134514-00501001

Lane, L. (2018c). The Horizontal Effect of International Human Rights Law in Practice. In *European Journal of Comparative Law and Governance* (Vol. 5, Issue 1). https://doi.org/10.1163/22134514-00501001

Larsson, S. (2013a). Conceptions, categories and embodiment: Why metaphors are of fundamental importance for understanding norms. In M. Baier (Ed.), *Social and Legal Norms: Towards a Socio-Legal Understanding of Normativity* (pp. 121–139). Ashgate.

Larsson, S. (2013b). Copy Me Happy: The Metaphoric Expansion of Copyright in a Digital Society. *International Journal for the Semiotics of Law*. https://doi.org/10.1007/s11196-012-9297-2

Larsson, S. (2013c). Metaphors, law and digital phenomena: The Swedish pirate bay court case. *International Journal of Law and Information Technology, 21*(4), 354–379. https://doi.org/10.1093/ijlit/eat009

Larsson, S. (2019). The Socio-Legal Relevance of Artificial Intelligence. *Droit et Société*, *N°103*(3). https://doi.org/10.3917/DRS1.103.0573

Larsson, S. (2020). On the Governance of Artificial Intelligence through Ethics Guidelines. *Asian Journal of Law and Society, 7*(3), 437–451. https://doi.org/10.1017/als.2020.19

Latonero, M. (2018). Governing artificial intelligence: upholding human rights & dignity. Data & Society. In *Data & Society* (pp. 1–37). https://datasociety.net/library/governing-artificial-intelligence/

Leenes, R., Palmerini, E., Koops, B. J., Bertolini, A., Salvini, P., & Lucivero, F. (2017). Regulatory challenges of robotics: Some guidelines for addressing legal and ethical issues. *Law, Innovation and Technology, 9*(1), 1–44. https://doi.org/10.1080/17579961.2017.1304921

Leslie, D., Burr, C., Aitken, M., Katell, M., Briggs, M. & Rincon, C. (2022). *Human rights, democracy, and the rule of law assurance framework for AI systems: A proposal*. 10.5281/zenodo.5981676

Lessig, L. (1999). *Code: And Other Laws of Cyberspace*. Basic Books.

Levy, K., Chasalow, K. E., & Riley, S. (2021). Algorithms and Decision-Making in the Public Sector. *Annual Review of Law and Social Science, 17*, 309–334. https://doi.org/10.1146/annurev-lawsocsci-041221-023808

Liu, H. Y., & Maas, M. M. (2021). 'Solving for X?' Towards a problem-finding framework to ground long-term governance strategies for artificial intelligence. *Futures, 126*,. https://doi.org/10.1016/j.futures.2020.102672

Lloyd, P. (2009). Ethical imagination and design. *Design Studies*. https://doi.org/10.1016/j.destud.2008.12.004

Lyytinen, K., & Rose, G. M. (2003). Disruptive information system innovation: The case of internet computing. *Information Systems Journal, 13*(4), 301–330. https://doi.org/10.1046/J.1365-2575.2003.00155.X

Maas, M. M. (2022). Aligning AI Regulation to Sociotechnical Change. In J. Bullock, B. Zhang, Y.-C. Chen, J. Himmelreich, M. Young, A. Korinek & V. Hudson (Eds.), *Oxford Handbook on AI Governance*. Oxford University Press.

MacChi, C. (2018). A treaty on business and human rights: Problems and prospects. In The Future of Business and Human Rights (Issue November 2015, pp. 63–86). Intersentia. https://doi.org/10.1017/9781780686455.005

Maguire James. (2021). *Top Performing Artificial Intelligence (AI) Companies of 2021*. Datamation. https://www.datamation.com/artificial-intelligence/ai-companies/

Martin, D. A., Conlon, E., & Bowe, B. (2021). A Multi-level Review of Engineering Ethics Education: Towards a Socio-technical Orientation of Engineering Education for Ethics. *Science and Engineering Ethics, 27*(5), 1–38. https://doi.org/10.1007/s11948-021-00333-6

McCarthy, J. (2018). *What is Artificial Intelligence?* http://www-formal.stanford.edu/jmc/whatisai.pdf

McGregor, L., Murray, D., & Ng, V. (2019). International human rights law as a framework for algorithmic accountability. *International and Comparative Law Quarterly, 68*(2), 309–343. https://doi.org/10.1017/S0020589319000046

McNamara, A., Smith, J. & Murphy-Hill, E. (2018). Does ACM's code of ethics change ethical decision making in software development? *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 729–733. https://doi.org/10.1145/3236024.3264833

Metzinger, T. (2019). EU guidelines: Ethics washing made in Europe. *Der Tagesspiegel*. https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html

Microsoft. (2017). *Responsible AI principles from Microsoft*. Our Approach. https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3Aprimaryr6

Microsoft. (2020). *Microsoft Global Human Rights Statement*. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4JIiU

Mijatović, D. (2018). Safeguarding human rights in the era of artificial intelligence. In *Human Rights Comments*. https://www.coe.int/en/web/commissioner/-/safeguarding-human-rights-in-the-era-of-artificial-intelligence

Miller, H. T. (2012). *Governing Narratives : Symbolic Politics and Policy Change*. University of Alabama Press.

Miller, K. (2020a). A Matter of Perspective: Discrimination, Bias, and Inequality in AI. In *Legal regulations, implications, and issues surrounding digital data* (pp. 182–202). https://doi.org/10.4018/978-1-7998-3130-3.ch010

Miller, K. (2020b). *A Matter of Perspective*. 182–202. https://doi.org/10.4018/978-1-7998-3130-3.CH010

Minkkinen, M., Zimmer, M. P., & Mäntymäki, M. (2021). Towards Ecosystems for Responsible AI: Expectations on Sociotechnical Systems, Agendas, and Networks in EU Documents. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12896 LNCS*. https://doi.org/10.1007/978-3-030-85447-8_20

Mishra, K. S., Polkowski, Z., Borah, S. & Dash, R. (2021). *AI in Manufacturing and Green Technology: Methods and Applications*. Routledge.

Mitrou, L. (2019). Data Protection, Artificial Intelligence and Cognitive Services: Is the General Data Protection Regulation (GDPR) 'Artificial Intelligence-Proof'? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3386914

MSI-NET. (2017). *Algorithms and Human Rights : Study on the Human Rights Dimensions of Automated Data Processing Techniques (in particular Algorthims) and Possible Regulatory Implications*. Council of Europe Study DGI. https://edoc.coe.int/en/internet/7589-algorithms-and-human-rights-study-on-the-human-rights-dimensions-of-automated-data-processing-techniques-and-possible-regulatory-implications.html

Mueller, M. L. (2013). *Networks and States : the Global Politics of IG*. MIT Press. https://doi.org/10.7551/mitpress/9780262014595.001.0001

Mueller, M. L., & Badiei, F. (2019). Requiem for a Dream: On Advancing Human Rights via Internet Architecture. *Policy and Internet*. https://doi.org/10.1002/poi3.190

Muller, C. (2020a). *The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law*. www.coe.int/cahai

Muller, V. (2020b). *Ethics of Artificial Intelligence and Robotics (Stanford Encyclopedia of Philosophy)*. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/ethics-ai/

Murray, A. (2007). *The Regulation of Cyberspace: Control in the Online Environment*. Routledge.

Murray, S., Wachter, R. & Blog, R. C. (2020). *Discrimination By Artificial Intelligence in a Commercial Electronic Health Record—a Case Study*. Healthaffairs. https://www.healthaffairs.org/do/https://doi.org/10.1377/hblog20200128.626576/full

Mylly, T. (2009). *Intellectual property and European economic constitutional law: The trouble with private informational Power*. Edward Elgar Publishing. IPR University Center.

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2133). https://doi.org/10.1098/rsta.2018.0089

Niederman, F., & Baker, E. W. (2021). Ethics and AI Issues: Old Container with New Wine? In D. Dennehy, A. Griva, N. Pouloudi, Y. K. Dwivedi, I. Pappas, & M. Mäntymäki (Eds.), *Responsible AI and analytics for an ethical and inclusive digitized society* (Vol. 12896, pp. 161–172). Springer International Publishing. https://doi.org/10.1007/978-3-030-85447-8

O'Neil, C. (2016). *Weapons of Math Destruction:How Big Data Increases Inequality and Threatens Democracy*. Crown Publishers.

Organisation for Economic Co-operation and Development (OECD). (2019). Artificial intelligence in society. In *Artificial intelligence in society*. OECD. https://doi.org/10.1787/eedfee77-en

Papagiannidis, E., Enholm, I. M., Dremel, C., Mikalef, P., & Krogstie, J. (2021). Deploying AI Governance Practices: A Revelatory Case Study. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12896 LNCS*. https://doi.org/10.1007/978-3-030-85447-8_19

Partnership on AI. (2020). Partnership on AI. https://partnershiponai.org/

Pedemonte, V. (2020). *AI for Sustainability: An overview of AI and the SDGs to contribute to the European policy-making*. https://ec.europa.eu/futurium/en/system/files/ged/vincent-pedemonte_ai-for-sustainability_0.pdf

Perritt, H. H. (1997). Cyberspace Self-Government: Townhall Democracy or Rediscovered Royalism? *Berkeley Technology Law Journal, 12*(2), 413–482.

Perry, B., & Uuk, R. (2019). Ai governance and the policymaking process: Key considerations for reducing ai risk. *Big Data and Cognitive Computing, 3*(2), 1–17. https://doi.org/10.3390/bdcc3020026

Pery, A., Rafiei, M., Simon, M. & van der Aalst, W. M. P. (2021). Trustworthy Artificial Intelligence and Process Mining: Challenges and Opportunities. *ArXiv Preprint* ArXiv:2110.02707. http://arxiv.org/abs/2110.02707

Pesenti, J. (2021). Facebook's five pillars of Responsible AI. In *Facebook Meta*. https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai/

Petit, N. (2017). Law and Regulation of Artificial Intelligence and Robots - Conceptual Framework and Normative Implications. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2931339

Pichay, S. (2018). *Our Principles – Google AI*. https://ai.google/principles/

Pillay, R. G., & Curiae, A. (2014). The limits to self-regulation and voluntarism : From corporate social responsibility to corporate accountability. *Amicus Curiae, 99*, 10–13.

Pinjušić, T. (2022). *The Atos blueprint for responsible AI - Atos*. ATOS. https://atos.net/en/lp/cybersecurity-magazine-ai-and-cybersecurity/the-atos-blueprint-for-responsible-ai

Pizzi, M., Romanoff, M., & Engelhardt, T. (2020). AI for humanitarian action: Human rights and ethics. *International Review of the Red Cross, 102*(913), 145–180. https://doi.org/10.1017/S1816383121000011

Popkin, H., Pratap, A. & Wolpow, N. (2020). AI 50: America's Most Promising Artificial Intelligence Companies. In *Forbes.com*. http://proxy-tu.researchport.umd.edu/login?ins=tu&url=https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=138673512&site=eds-live&scope=site

Rachovitsa, A., & Johann, N. (2022). The human rights implications of the use of AI in the digital welfare state: Lessons learned from the Dutch SyRI case. *Human Rights Law Review, 22*(2), 1–15. https://doi.org/10.1093/hrlr/ngac010

Radu, R. (2019). Negotiating Internet Governance. In *Negotiating internet governance*. Oxford University Press. https://doi.org/10.1093/oso/9780198833079.001.0001

Radu, R. (2021). Steering the governance of artificial intelligence: national strategies in perspective. *Policy and Society, 40*(2), 178–193. https://doi.org/10.1080/14494035.2021.1929728

Rai, A., Constantinides, P., & Sarker, S. (2019). Next-generation digital platforms: Towards human-AI hybrids. *MS Quarterly, 43*, 3–8.

Ramasastry, A. (2015). Corporate Social Responsibility Versus Business and Human Rights: Bridging the Gap Between Responsibility and Accountability. In *Journal of Human Rights* (Vol. 14, Issue 2, pp. 237–259). https://doi.org/10.1080/14754835.2015.1037953

Raso, F., Hilligoss, H., Krishnamurthy, V., Bavitz, C. & Kim, L. (2018a). Artificial Intelligence & Human Rights: Opportunities & Risks. *The Berkman Klein Center for Internet & Society Research Publication Series No. 2018a-6*, *7641*, 63. https://cyber.harvard.edu/publication/2018a/artificial-intelligence-human-rights

Raso, F., Hilligoss, H., Krishnamurthy, V., Bavitz, C., & Kim, L. Y. (2018b). Artificial Intelligence & Human Rights: Opportunities & Risks. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3259344

Ray, K. (2021). Quest for I (Intelligence) in AI (Artificial Intelligence): A non-elusive attempt. In E. Osaba (Ed.), *Artificial intelligence: Latest advances, new paradigms and novel applications*. https://doi.org/10.5772/intechopen.96324

Redeker, D., Gill, L., & Gasser, U. (2018). Towards digital constitutionalism? Mapping attempts to craft an Internet Bill of Rights. *International Communication Gazette, 80*(4), 302–319. https://doi.org/10.1177/1748048518757121

Reed, C. (2004). *Internet Law*. Cambridge University Press.

Reed, C. (2018). How should we regulate artificial intelligence? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2128). https://doi.org/10.1098/rsta.2017.0360

Reidenberg, J. R. (1997). Lex informatica: The formulation of information policy rules through technology. *Texas Law Review, 553*, 553–593.

Rességuier, A., & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data and Society*, *7*(2). https://doi.org/10.1177/2053951720942541

Rights, H., & Look, T. (2021). Introduction : Putting Flesh on the Bone. In S. Deva & D. Bilchitz (Eds.), *Building a Treaty on Business and Human Rights: Context and Contours* (pp. 1–24). Cambridge University Press.

Risse, M. (2018). Human Rights and Artificial Intelligence: An Urgently Needed Agenda. *Revista Publicum*, *4*(1). https://doi.org/10.12957/publicum.2018.35098

Robbins, S. (2019). A misdirected principle with a catch: Explicability for AI. *Minds and Machines, 29*(4), 495–514. https://doi.org/10.1007/s11023-019-09509-3

Russel, S. & Norvig, P. (2020). Artificial intelligence: a modern approach. In *Choice Reviews Online* (4th ed., Vol. 33, Issue 03). Prentice Hall.

Sætra, H. S. (2021). Challenging the Neo-Anthropocentric Relational Approach to Robot Rights. *Frontiers in Robotics and AI, 8*, 301. https://doi.org/10.3389/frobt.2021.744426

Sander, B. (2019). Freedom of Expression in the Age of Online Platforms: Operationalising a Human Rights-Based Approach to Content Moderation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3434972

Santoni de Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy and Technology, 34*(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Sartor, G. & Lagioia, F. (2020). The impact of the General Data Protection Regulation (GDPR) on artificial intelligence. *European Union*, 100. https://doi.org/10.2861/293

Saslow, K., & Lorenz, P. (2019). *Artificial Intelligence Needs Human Rights*. Think Tank at the Intersection of Technology and Society.

Scantamburlo, T., Cortés, A. & Schacht, M. (2020). Progressing Towards Responsible AI. ArXiv:2008.07326v1. http://arxiv.org/abs/2008.07326.

Scherer, M. U. (2016). Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *Harvard Journal of Law & Technology*, *29*(353). https://doi.org/10.2139/ssrn.2609777

Slee, T. (2020). The incompatible incentives of private-sector AI. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI* (pp. 106–123). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.6

Smith, M. & Miller, S. (2021). Facial Recognition and Privacy. In *Biometric Identification, Law and Ethics* (pp. 21–38). http://www.cdt.org/blogs/harley-geiger/612facial-recognition-and-privacy

Smuha, N. A. (2021a). Beyond a Human Rights-Based Approach to AI Governance: Promise, Pitfalls, Plea. *Philosophy and Technology, 34*, 91–104. https://doi.org/10.1007/s13347-020-00403-w

Smuha, N. A. (2021b). From a 'race to AI' to a 'race to AI regulation': Regulatory competition for artificial intelligence. *Law, Innovation and Technology, 13*(1), 57–84. https://doi.org/10.1080/17579961.2021.1898300

Smuha, N. A., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., & Yeung, K. (2021). How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3899991

Stahl, B. C., Andreou, A., Brey, P., Hatzakis, T., Kirichenko, A., Macnish, K., LaulhéShaelou, S., Patel, A., Ryan, M., & Wright, D. (2021a). Artificial intelligence for human flourishing – Beyond principles for machine learning. *Journal of Business Research, 124*(2), 374–388. https://doi.org/10.1016/j.jbusres.2020.11.030

Stahl, Bernd C., Antoniou, J., Ryan, M., Macnish, K. & Jiya, T. (2021b). Organisational responses to the ethical issues of artificial intelligence. *AI and Society*, *0123456789*. https://doi.org/10.1007/s00146-021-01148-6

Stahl, B. C. (2021). *Artificial intelligence for a better future: An ecosystem perspective on the ethics of AI and emerging digital technologies*. Springer International Publishing. https://doi.org/10.1007/978-3-030-69978-9

Stahl, B. C., Timmermans, J., & Flick, C. (2017). Ethics of emerging information and communication technologies: On the implementation of responsible research and innovation. *Science and Public Policy, 44*(3), 369–381. https://doi.org/10.1093/scipol/scw069

Stark, L., & Hoffmann, A. L. (2019). Data Is the New What? Popular Metaphors & Professional Ethics in Emerging Data Culture. *Journal of Cultural Analytics., 10*(22148/16), 036.

Strobel, J., & Tillberg-Webb, H. (2009). Applying a critical and humanizing framework of instructional technologies to educational practice. In *Learning and Instructional Technologies for the 21st Century* (pp. 1–19). Springer US. https://doi.org/10.1007/978-0-387-09667-4_5

Surden, H. (2020). Ethics of AI in law. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI* (pp. 718–736). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.46

Suzor, N. (2019). *Lawless: The Secret Rules That Govern Our Digital Lives.* Cambridge University Press.

Szczepański, M. (2019). Economic impacts of artificial intelligence (AI). In *EPRS | European Parliamentary Research Service* (Issue July). https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/637967/EPRS_BRI(2019)637967_EN.pdf

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science, 361*(6404), 751–752.

Taeihagh, A. (2021). Governance of artificial intelligence. *Policy and Society, 40*(2), 137–157. https://doi.org/10.1080/14494035.2021.1928377

Taeihagh, A., Ramesh, M., & Howlett, M. (2021). Assessing the regulatory challenges of emerging disruptive technologies. *Regulation and Governance, 15*(4), 1009–1019. https://doi.org/10.1111/rego.12392

Tasioulas, J. (2013). Human Dignity and the Foundations of Human Rights. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.2557649

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. Knopf.

Tharoor, S. (2000). Are Human Rights Universal? *World Policy Journal*, *16*(4), 1–6. https://www.jstor.org/stable/40209657?seq=1

The IEEE Global Initiative. (2017). Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. In *IEEE*. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

Tirosh, N. (2017). Reconsidering the 'Right to be Forgotten' – memory rights and the right to memory in the new media era. *Media, Culture and Society*. https://doi.org/10.1177/0163443716674361

Trajtenberg, M. (2018). AI as the Next Gpt: A Political-Economy Perspective. *Ssrn*. https://doi.org/10.3386/w24245

Troxell, G., & Troxell, W. (2017). A reflective analysis on professional codes of ethics. 2017 ASEE annual conference & exposition proceedings, 2017-June. https://doi.org/10.18260/1-2--27506

Tushnet, M. (2015). Internet Exceptionalism: An Overview from General Constitutional Law. *William and Mary Law Review, 56*(4), 1637.

Ufert, F. (2020). AI Regulation Through the Lens of Fundamental Rights: How Well Does the GDPR Address the Challenges Posed by AI? *European Papers, 5*(2), 1087–1097.

UK Government. (2021). *National AI Strategy*. https://www.gov.uk/government/publications/national-ai-strategy

Umbrello, S. (2022). The role of engineers in harmonising human values for AI systems design. *Journal of Responsible Technology, 10*, 100031. https://doi.org/10.1016/j.jrt.2022.100031

Ruggie, J. (2011). Guiding principles on business and human rights: Implementing the United Nations "Protect, Respect and Remedy" framework. In United Nations. https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf

van Veen, C. & Cath, C. (2018). Artificial Intelligence: What's Human Rights Got To Do With It? *Data & Society: Points*. https://points.datasociety.net/artificial-intelligence-whats-human-rights-got-to-do-with-it-4622ec1566d5

Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the draft EU artificial intelligence act: Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International, 22*(4), 97–112.

Vesnic-Alujevic, L., Nascimento, S. & Pólvora, A. (2020). Societal and ethical impacts of artificial intelligence: Critical notes on European policy frameworks. *Telecommunications Policy*, *44*(6). https://doi.org/10.1016/j.telpol.2020.101961

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, 11(1), 233. https://doi.org/10.1038/s41467-019-14108-y

Vochozka, M., Kliestik, T., Kliestikova, J. & Sion, G. (2018). Participating in a highly automated society: How artificial intelligence disrupts the job market. *Economics, Management, and Financial Markets*, *13*(4), 57–62. https://doi.org/10.22381/EMFM13420185

Wagner, B. (2018). Ethics as an escape from regulation. In *Being rofiled: Cogitas Ergo Sum* (pp. 84–89). Amsterdam University Press. https://doi.org/10.2307/j.ctvhrd092.18

Wyatt, S. (2004). Danger! Metaphors at Work in Economics, Geophysiology, and the Internet. *Science Technology and Human Values, 29*(2), 242–261. https://doi.org/10.1177/0162243903261947

Wettstein, F. (2012). CSR and the Debate on Business and Human Rights: Bridging the Great Divide. *Business Ethics Quarterly, 22*(4), 739–770. https://doi.org/10.5840/beq201222446

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 195–200. https://doi.org/10.1145/3306618.3314289

Winfield, A. F., Michael, K., Pitt, J., & Evers, V. (2019). Machine ethics: The design and governance of ethical ai and autonomous systems. *Proceedings of the IEEE, 107*(3), 509–517. https://doi.org/10.1109/JPROC.2019.2900622

Winfield, A. & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133). https://doi.org/10.1098/rsta.2018.0085

Winner, L. (1977). *Autonomous Technology: Technics-Out-Of-Control As A Theme In Political Thought*. MIT Press.

Wu, T. (2010). Is Internet Exceptionalism Dead? In B. Szoka & A. Marcus (Eds.), *The Next Digital Decade: Essays on the Future of the Internet*. TechFreedom.

Yavar Bathaee. (2018). The artificial intelligence black box and the failure of intent and causation. *Harvard Journal of Law & Technology*, *2*(4), 31–40. https://www.theverge.com/

Yeung, K., Howes, A. & Pogrebna, G. (2019). AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing. In M. D. Dubber & F. Pasquale (Eds.), *The Oxford Handbook of AI Ethics*. Elsevier BV. https://doi.org/10.2139/ssrn.3435011

Zalnieriute, M. (2019). From Human Rights Aspirations to Enforceable Obligations by Non-State Actors in the Digital Age: The Example of IG and ICANN. *Yale Journal of Law & Technology, 21*, 278–336. https://doi.org/10.2139/ssrn.3333532

Zalnieriute, M., & Milan, S. (2019). Internet Architecture and Human Rights: Beyond the Human Rights Gap. *Policy and Internet, 11*(1), 6–15. https://doi.org/10.1002/poi3.200

Zamani, E. D. (2022). The Bitcoin protocol as a system of power. *Ethics and Information Technology*, *24*(1). https://doi.org/10.1007/s10676-022-09626-1

Zamfir, I. (2018). Towards a binding international treaty on business and human rights. In *EPRS | European Parliament Research Service* (Issue November). https://www.europarl.europa.eu/RegData/etudes/BRIE/2018/620229/EPRS_BRI(2018)620229_EN.pdf

Završnik, A. (2020). Criminal justice, artificial intelligence systems, and human rights. *ERA Forum, 20*(4), 567–583. https://doi.org/10.1007/s12027-020-00602-0

Ziemele, I. (2009). Human Rights Violations by Private Persons and Entities: The Case-Law of International Human Rights Courts and Monitoring Bodies. In *EUI Working Papers* (Issue 8). http://hdl.handle.net/1814/11409

Zittrain, J. (2006). A History of Online Gatekeeping. *Harvard Journal of Law and Technology*.

Vasiliki Koniakou graduated in June 2012 from the Faculty of Law of the National Kapodestrian University of Athens and continued her studies at the University of Turku Faculty of Law, in the two-year Law and Information Society (LIS) Program. In 2014 she obtained her Master's Degree in International, European and Comparative Law, majoring Law and Information Society. In 2016 she started her doctoral studies at the University of Turku, with a scholarship from the Academy of Finland. She completed her studies in September 2022, defending her doctoral dissertation ("Re-thinking Internet Governance"). Her dissertation focuses on the excessive privatization of Internet Governance on the content and logical layers of the Internet. Reviewing the current Internet Governance model under the light of human rights, rule of law and good governance, she explores the role of conceptual metaphors and technological determinism in informing and shaping the way the Internet is governed. She is currently a Senior Researcher at the ELTRUN and the IST Lab of the Athens University of Economics and Business, and a legal counsellor at the Athens Centre of Entrepreneurship and Innovation (ACEin). In this new chapter in her research adventures, her focus remains to be law and technology, as well as human rights and rule of law across different technologically mediated contexts. Her research interests include Internet Governance, Artificial Intelligence, human-centric and human rights-aware design of technologies, and the legally interesting applications of blockchain. She is also interested in issues related to the democratization of Technology Governance, and the injection of human rights in design and standardization. She identifies herself within the family of Science and Technology Studies (STS) as well as Law and Technology, Internet researchers, and Digital Constitutionalism supporters.