COUNCIL OF EUROPE

CONSEIL DE L'EUROPE

Strasbourg, 15 June 2020 CAHAI(2020)07-fin

AD HOC COMMITTEE ON ARTIFICIAL INTELLIGENCE (CAHAI)

AI Ethics Guidelines: European and Global Perspectives

Provisional report by Marcello Ienca* and Effy Vayena*

***Chair of Bioethics, Health Ethics and Policy Lab, Department of Health Sciences and Technology, ETH Zurich.**

www.coe.int/cahai

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Health Ethics and Policy Laboratory
Department of Health Sciences and Technology (D-HEST)

Executive Summary: In recent years, private companies, research institutions and public-sector organizations have issued principles, guidelines and other soft law instruments for the ethical use of artificial intelligence (AI). However, despite an apparent agreement that AI should be 'ethical', there is debate about both what constitutes 'ethical AI' and which ethical requirements, technical standards and best practices are needed for its realization. The aim of this report is mapping the relevant corpus of soft law documents and other ethical-legal frameworks developed by governmental and non- governmental organisations globally with a twofold aim. First, we want to monitor this ever-evolving spectrum of non-mandatory governance instruments. Second, we want to prospectively assess the impact of AI on ethical principles, human rights, the rule of law and democracy. The report employs an adapted and pre-validated scoping review protocol to provide a comprehensive and up-to-date overview of current soft law efforts. We reviewed a total of 116 documents published inter alia by governmental agencies, non-governmental organisations, academic institutions and private companies. Our analysis identifies five prominent clusters of ethical principles and assesses their role in the current governance discourse. *Ex negativo*, our analysis reveals existing blind spots and interpretative gaps in the current soft law landscape. Furthermore, we establish a link between ethical principles and human rights, with special focus on the rights and freedoms enshrined in the European Convention on Human Rights (ECHR) to assess the extent to which the protection of human rights is integral in current non- mandatory governance frameworks. Finally, we provide empirically-informed policy implications to inform scientists, research institutions, funding agencies, governmental and inter-governmental organisations and other relevant stakeholders involved in the advancement of ethically responsible innovation in AI.

Key findings:

- An increasing number of governmental and nongovernmental organisations (incl. private companies and academic organisations) are developing ethics guidelines or other soft law instruments on AI.
- These soft law documents are being primarily developed in Europe, North America and Asia. The global south is currently underrepresented in the landscape of organisations proposing AI ethics guidelines.
- Current AI ethics guidelines tend to agree on some generic principle but they sharply disagree over the details of what should be done in practice. Furthermore, no single ethical principle is common to all of the 116 documents on ethical AI we reviewed.
- We found growing agreement around the following ethical principles: transparency, justice, non-maleficence, responsibility, and privacy. Ethical considerations regarding sustainability, dignity and solidarity appear significantly underrepresented.
- Most guidelines agree that AI should be *transparent* to avoid potential problems. But it is not clear whether transparency should be achieved through publishing source code, the underlying databases or some other means.
- Slightly more than half of reviewed soft law documents explicitly recommend the promotion of human rights —or warn against their violation— when designing, developing and deploying AI systems.
- Regular expressions built from the codes reveal significant variations in theme coverage among documents produced within member countries of the Council of Europe (CoE) compared to documents produced elsewhere. Compared to the rest of the world, soft law documents produced within countries that are members of the Council of Europe appear to emphasize the ethical principles of solidarity, trust and trustworthiness. In contrast, they appear to refer more sporadically to the principles of beneficence and dignity.
- The principles of privacy, justice and fairness showed the least variation across CoE-member countries, CoE-observer countries and the rest of the world, hence the highest degree of cross- geographical and cross-cultural stability.

Key policy implications:

- Soft law instruments issued by governmental and nongovernmental organisations (incl. private companies and academic organisations) are useful tools to exert practical influence on public decision making over AI and steering the development of AI systems for social good and in abidance of ethical values and legal norms. However, soft law approaches should not be considered substitutive of mandatory governance. Due to conflict of interest, self-regulation efforts by private AI actors are at particular risk of being promoted to bypass or obviate mandatory governance by governmental and intergovernmental authorities.
- In order to ensure inclusiveness, cultural pluralism and fair participation to collective decision making on AI, the development of soft law documents by organisations located in currently underrepresented global regions, especially Africa and South America, should be promoted.
- The convergence of current soft law instruments around five generic ethical principles such as transparency, justice, non-maleficence, responsibility, and privacy reveals five priority areas of oversight and possible intervention by mandatory governance authorities at both the governmental and intergovernmental level.
- In order to be translated into effective governance, these ethical principles should be conceptually clarified. Policy makers have the duty to resolve semantic ambiguities and conflicting characterisations of these principles.

- The sharp disagreement of current soft law documents on the interpretation and practical implementation of these principles indicates that mandatory governance solutions are likely subject to public disagreement, hence require a transparent process of democratic deliberation.
- Underrepresented ethical considerations such as those regarding sustainability, dignity and solidarity need to be further scrutinized to avoid importing into mandatory governance the same conceptual gaps and normative blind spots of soft law.
- As nearly half of reviewed soft law documents do not explicitly recommend the promotion— or warn against the violation— of human rights when designing, developing and deploying AI systems, greater focus on the human rights implications of AI is urgently needed.
- Member countries of the Council of Europe are well-positioned to steer the international governance of AI towards the promotion of human rights.

INTRODUCTION

Artificial Intelligence (AI) is the study and development of computer systems able to perform tasks normally believed to require human intelligence. Typically, computer systems are deemed intelligent (hence called 'intelligent agents' or 'intelligent machines') when they have the ability to perceive their environment and take autonomous actions directed towards successfully achieving a goal. Historically observed, although general reflections on mechanical reasoning have populated the scientific and philosophical literature since ancient times, the field of AI in the narrow sense originated in the 1940s as a consequence of concomitant advances in mathematical logic (e.g. the Church-Turing thesis), information theory, neurobiology and cybernetics. The field of AI encompasses a variety of complex computational approaches that render or mimic cognitive functions such as learning, memory, reasoning, vision, and natural language processing. The most common of these approaches is called machine learning (ML) and involves the development of algorithms that perform tasks in absence of explicit instructions from human operators. Unlike conventional computer programs, ML algorithms build mathematical models based on training data and rely exclusively on inference and pattern identification to make autonomous predictions and decisions[1]. Today, AI is a major catalyzer of technological transformation. At the dawn of the 2020s, AI systems are embedded in an uncountable number of systems and devices regularly used by humans such as mobile phones, social media, cars, airplanes, analytic software, email communication systems, home appliances etc. AI is integral to a broad variety of human activities including (but not restricting to) telecommunication, transportation, manufacturing, healthcare, banking, insurance, law enforcement and the military.

Due to its technological novelty, capacity for autonomous action and general-purposive nature, AI holds potential for transforming human societies at greater pace and in greater magnitude compared to any other technology. The transformative potential of AI has been deemed "revolutionary" by experts[2], with authors referring to AI development as an "ongoing revolution" that "will change almost every line of work"[3]. For this reason, it is paramount and urgent to assess the implications of AI for core principles and values of human life, the future of human societies and the systems of rules that govern those societies, first and foremost democracy and the rule of law[4-6].

In recent years, several governmental and intergovernmental organizations as well as non-state actors have issued principles, guidelines, recommendations, governance frameworks or other soft law instruments for AI. Soft law instruments are normative documents that are not legally binding or enforceable but of persuasive nature which can have practical influence on decision making in a manner that is comparable to that of binding regulations (hard laws). The aim of these instruments is steering the development of AI for social good and in abidance of ethical values and legal norms. However, despite an apparent agreement that AI should be 'ethical', there is debate about both what constitutes 'ethical AI' and which ethical requirements, technical standards and best practices are needed for its realization. Furthermore, due to the rapid proliferation of AI-related soft law documents and the large diversity of their issuers, it is hard to keep track and make sense of this ever-evolving body of non- mandatory governance in a comprehensive and rigorous manner.

This report provides a comprehensive mapping and meta-analysis of the current corpus of principles and guidelines on ethical AI. This analysis will inform scientists, research institutions, funding agencies, governmental and intergovernmental organizations and other relevant stakeholders involved in the advancement of ethically responsible innovation in AI. Furthermore, it will discuss how these ethical principles, moral customs, and recommended social practices can be translated into mandatory governance, especially internationally binding legal instruments.

Particular attention is devoted in this report to examining the nexus between AI governance and human rights and providing a prospective assessment of the impact of AI technology on human

rights and freedoms[7-9]. Human rights are rights inherent to all human beings, regardless of race, sex, nationality, ethnicity, language, religion, or any other status[10]. These rights describe both moral principles and legal norms in municipal and international law to which a person is inherently entitled as a human being. They are, therefore, inalienable, inviolable and universal. They are inalienable as they are not subject to being taken away by anyone; inviolable, as they should not be infringed under any circumstance; universal, as they are applicable everywhere and at every time. Human rights and freedoms are protected by international conventions. In the European space, a fundamental instrument is the European Convention on Human Rights (ECHR) which was drafted in 1950 by the Council of Europe and entered into force on 3rd September 1953. The ECHR enshrines a set of basic rights and freedoms that should be protected, making a legal commitment to abide by standards of behaviour that respect those rights and freedoms[11].

METHODOLOGY

In February 2020, we conducted a scoping review of the existing corpus of soft law instruments related to AI. Scoping review methods allow to synthesize and map the literature in a certain domain in an exploratory manner, hence are particularly suitable for screening and assessing complex or heterogeneous areas of research. Given the absence of a unified database for soft law instruments, we developed a protocol for discovery and filtering, adapted from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework. The protocol, which was pilot-tested and calibrated prior to data collection, consisted of three sequential and iteratively linked phases: screening, eligibility assessment and content analysis. This methodology is designed to provide a formal and evidence-based procedure to map, monitor and iteratively assess the soft law governance efforts in the area of AI.

1.1. Phase 1: Screening

In the screening phase, we combined retrospective screening of existing repositories with purposive and unstructured web search. First, we screened the following four data repositories and textual sources to retrieve relevant entries related to soft law documents on AI:

- European Union Agency for Fundamental Rights (December 11, 2019), *AI Policy Initiatives List.*
- Fjeld & Nagy (January 2020), Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI, *Harvard Berkman Klein Center Research Publication*.
- Jobin, Ienca & Vayena (September 2019). The Global Landscape of AI Ethics Guidelines.
  Nature Machine Intelligence.
- Wong & CASBS (June 2019). Fluxus Landscape: An Expansive View of AI Ethics and Governance.

As according to the Arksey and O'Malley framework for scoping reviews[12], structured database search was complemented with unstructured grey literature search to identify soft law instruments that might have eluded. Entries were assessed for eligibility (see 1.2) and, wherever eligibility was confirmed, included manually into the final synthesis and admitted to the second phase.

Finally, we reviewed the list of "top-45 AI companies" compiled in May 2019 by Datamation, a US-based computer science magazine focused on technology analysis. Each of the 45 AI actors ranked in this list was screened independently by accessing their websites and searching for AI ethics or policy statements manually and via keyword search. Finally, unstructured web search was performed to retrieve information that might have remained undetected through our search strategy. Eligible entries were included manually in the final synthesis and admitted to the second phase.

1.2. Phase 2: Eligibility Assessment

In the eligibility assessment phase, we screened all retrieved entries to assess their eligibility to be included into the final synthesis. Decisions on eligibility were guided by the inclusion and exclusion criteria listed in Table 1.

Screening	
	- Types: websites, written articles and other documents published online or parts thereof, such as dedicated web pages, blog posts, institutional reports and declarations, as well as references contained within;
Sources included:	- Issuers: private sector for profit organizations (companies, corporations, holdings etc., including private sector alliances); academic and research institutions (universities, professional societies, science foundations etc.); national governmental agencies (ministries, data protection authorities, competition authorities etc.; non-governmental organisations including non-profit organisations and charities.
	- Language: English, German, French, Spanish, Dutch, Italian and Greek (the languages spoken by the researchers).
Sources excluded:	- Types: videos, images and audio/podcasts (except written descriptions), books, journalistic articles, academic articles, syllabi, legislation, official standards, conference summaries;
	- Issuers: intergovernmental and supranational organisations.
	- Language: others than those above.
Eligibility	
Sources included:	- which refer to “artificial intelligence” and/or “AI”, either explicitly in their title or within their description (example: Google: “AI Principles”); or
	- which do not contain the above reference in their title but mention “robot”, “robotics”, “big data”, “machine learning” instead *and* reference AI or artificial intelligence explicitly as being part of robots and/or robotics; or

	- which do not contain the above reference in their title but are thematically equivalent (by referring to “algorithms”, “predictive analytics”, “cognitive computing”, “machine learning”, “big data”, “deep learning”, “autonomous” or “automated” instead.
	AND
	- which describes a principle, guideline, standard (including “ethics/ethical”, “principles”, “tenets”, “declaration”, “policy”, “guidelines”, “values” etc.), internal strategy (e.g. creation of advisory board) or other type of initiative.
	AND
	- which is expressed in normative or prescriptive language (i.e. with modal verbs or imperatives such as "responsible", "fair", "trust/trustworthy" etc.); or
	- which is principle- or value-based (i.e. indicating a preference and/or a commitment to a certain ethical vision or course of action).
	- which reference actions/visions/commitments/courses of action that apply to the actor enunciating them or other private sector actors.
Sources excluded:	- websites and documents about robotics that do not mention artificial intelligence as being part of robots/robotics; and
	- websites and documents about data or data ethics that do not mention artificial intelligence as being part of data;
	- websites and documents about AI ethics directly aimed at non-private sector actors (e.g. consulting for the public sector)
	- websites and documents about ethics whose primary focus is not AI (e.g. business ethics).

Table 1- Eligibility Criteria

1.3. Content Analysis

In the second phase, entries included in the final synthesis were assessed using an expanded version of a previously validated content analysis protocol developed by the authors[13,14]. This protocol involves both a quantitative and a qualitative analysis. At the quantitative level, entries were classified according to instrument type, issuer, and geographic provenience of the issuer. Furthermore, relative frequencies of relevant quantitative data were measured and visually charted. Finally, we performed full-text screening with the assistance of a keyword search plugin to identify documents that made explicit reference to human rights. Documents included

in this category made explicit reference to either preserving and promoting human rights or preventing their violation when designing, developing or deploying AI applications. These documents were differentiated from those that did not mention human rights or did so but without making any explicit normative statement about their promotion or non- violation in the context of AI.

At the qualitative level, thematic content analysis was conducted to identify recurrent thematic patterns related to the following domains: (i) ethical principles and values, and (ii) human rights. This thematic content analysis was conducted manually by the researchers with qualitative software assistance (NVivo/MAXQDA for Mac). Emerging thematic patterns were analyzed in-depth, coded, and clustered into pre-defined ethical categories based on the ethical matrix developed by Jobin, Ienca and Vayena (2019).

Given the size of the final synthesis database, this manual thematic analysis was complemented with an automated analysis via natural language processing (NLP). We retrieved the documents and web contents automatically, where possible, using Python wget package and added the rest manually. Next, we built regular expressions[15] from the codes resulting from the qualitative analysis protocol developed by Jobin, Ienca & Vayena[13]. The regular expressions of the codes belonging to the same theme were joined together into one regular expression by 'or' statements ('|'). To ensure comprehensiveness and inclusion, the original English codes were translated into the following languages: German, French, Spanish, Italian and Dutch. To determine the theme coverage, we checked for the occurrence of the theme regular expressions in the documents, i.e., we determined the theme to be present in the guideline if the theme's regular expression had at least one match. Finally, we grouped the results by member type and normalised by the total number of guidelines within each group. Variations between the ethical principles and values raised within the 47 Member States of the Council of Europe were compared with, respectively, principles and values raised within Observer States as well as the rest of the world.

1.4. Normative ethical and policy analysis

In the fourth and last phase, empirically informed normative ethical and policy analysis was conducted. The aim of this conclusive study component is transferring the preliminary findings of the previous study phases from the descriptive to the normative-prescriptive level. During this phase, we performed three sequential theoretical steps. First, we assessed the results of our content analysis to identify which ethical principles and values are most common and recurrent across the corpus of documents under analysis. As previous research has shown significant interpretative variation within recurrent thematic clusters[13], we complemented the assessment of relative thematic frequencies with a detailed appraisal of their interpretation. This appraisal was instrumental to evaluating which interpretations of the principles are the most effective, hence should be adopted and pursued by global actors. Second, we assessed our review data to identify which principles and values are less frequent or missing in the current landscape of AI ethics guidelines. This second step was instrumental to identifying possible blind spots in international soft law initiatives and, consequently, making normative recommendations on how to overcome these ethical gaps. Third and finally, we advanced normative recommendations on core ethical principles and values that require prioritization in international AI governance. This conclusive part was instrumental to informing future normative ethical frameworks and delineating a roadmap for international policy on AI, ethics and human rights. To this purpose, we provided a reader- friendly visual summary of the study findings and a toolbox for future monitoring and evaluation (e.g. indicators) at the interface between AI, ethics and human rights.

2. Findings

Our search identified 116 documents containing soft law documents on AI issued by non-intergovernmental organisations until February 2020. Data reveal a significant increase over

time in the number of publications, with 93.9% having been released since 2016. The peak in the number of soft law documents published internationally was reached in 2018 and experienced a non-negligible decrease in the subsequent year (see Figure 1).

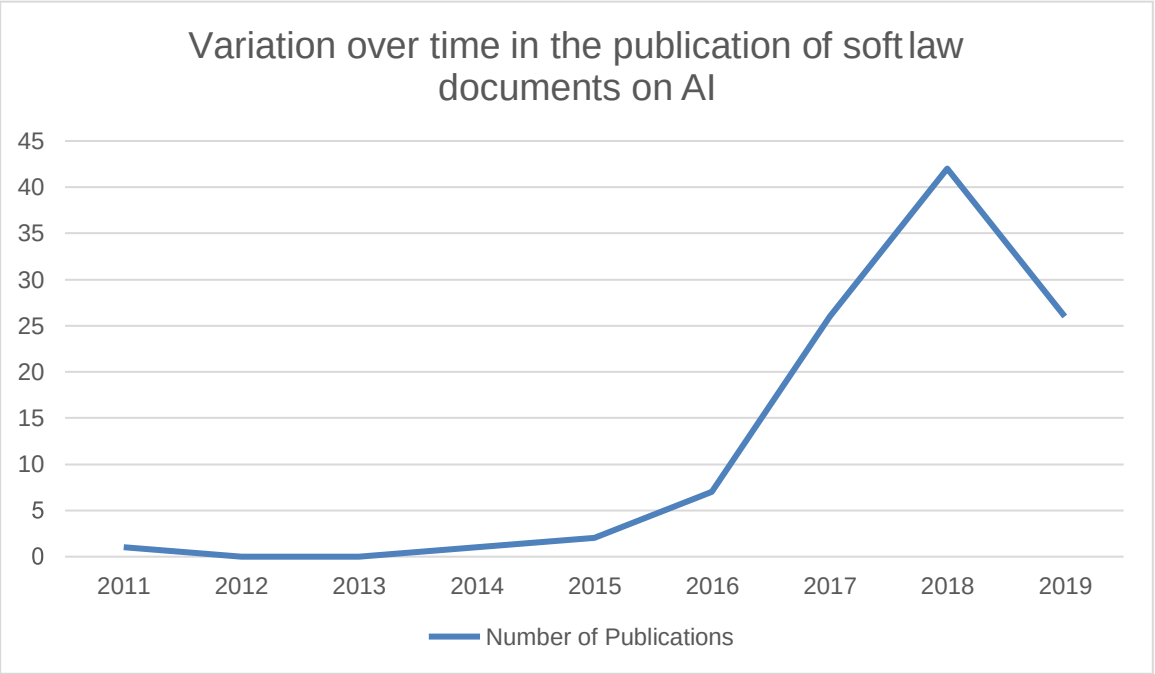

Figure 1- Variation over time in the publication of soft law documents on AI

Data breakdown by type of issuing organization shows that most documents were produced by governmental agencies (n=39), followed by private companies and private sector alliances (n=36), academic and research institutions including science foundations, professional societies and research alliances (n=28) as well as non-governmental organisations (NGOs) including non-profit organisations (NPOs) and charities (n=13). A detailed distribution of issuing organisations by type is provided in Figure 2.

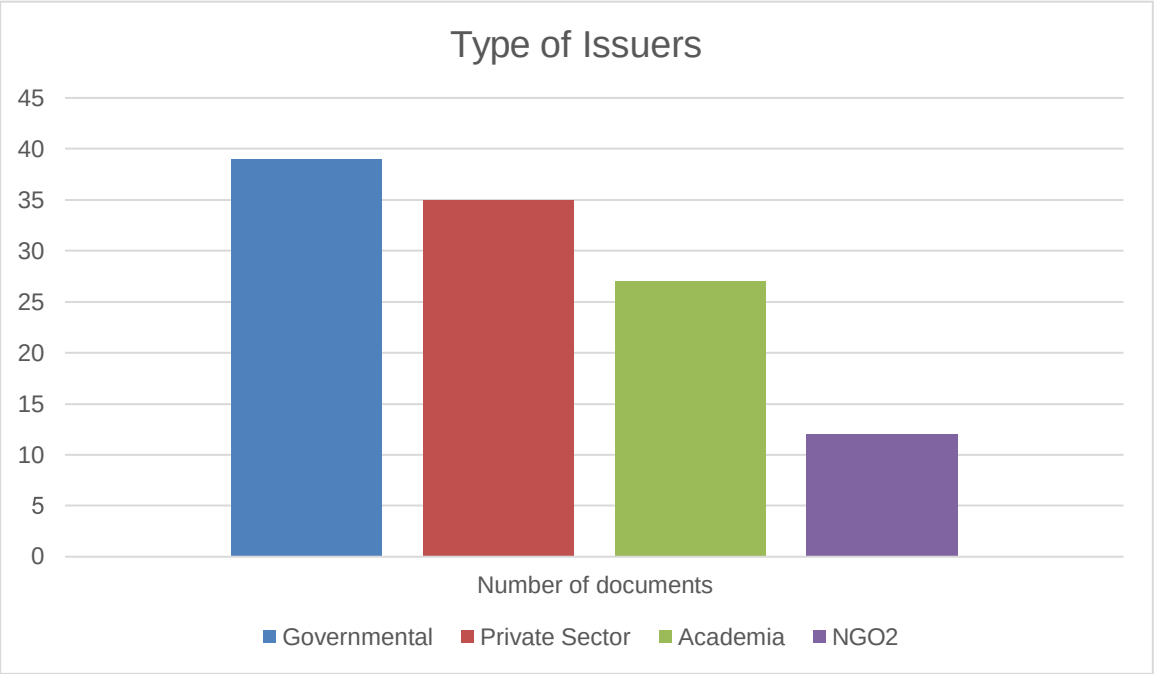

Figure 2- Types of issuing organisations

Data breakdown by geographic distribution of issuing organisations shows that 46% (n=53.5) of soft law documents are issued by organisations based in member countries of the Council of Europe. 32% (n=37.5) by organisations based in observer countries of the Council of Europe. 21% (n=25) by organisations based in countries that are neither members nor observers of the Council of Europe. Overall, data show a prominent representation of issuing organisations based in economically developed countries, with the USA (*n* = 29.5; 25.2%) and the UK (*n* = 17.5; 16%) together accounting for more than one third of all ethical AI principles. Other countries include, in descending order, Germany (n=8), Japan (n=6), Finland (n=4), Belgium, China, France and The Netherlands (n=3), India, Italy, Singapore and Spain (n=2), Australia, Austria, Czech Republic, Iceland, Lithuania, Malta, Mexico, New Zealand, Norway, Russia, South Korea, Sweden, Switzerland, UAE, and the Vatican (n=1). Thirteen documents were issued by international organisations or organisations that could not be ascribed to any specific country. African and South-American countries are not represented independently from international organizations. A visual overview of the geographic distribution of issuing organisations is presented in Figure 3:

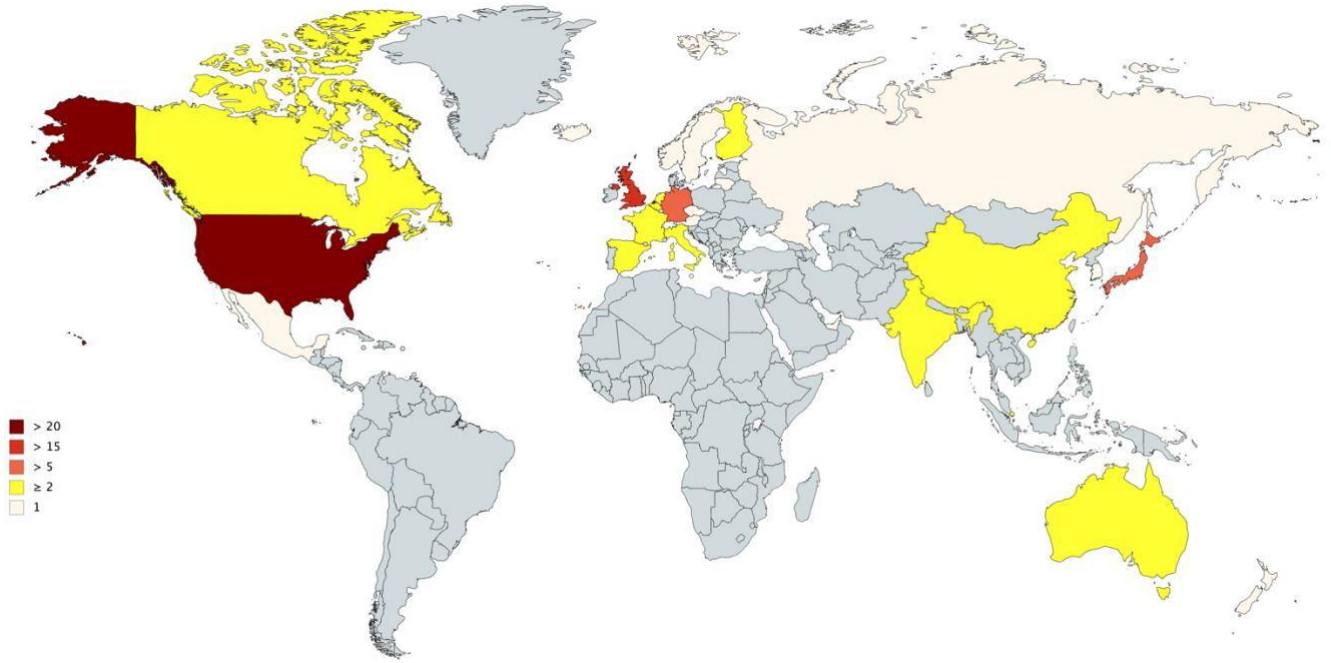

Figure 3- Geographic distribution of soft law documents by country of issuing organisation

More than half of the documents (n=62) make explicit reference to promoting, respecting or preventing the violation of human rights. Of these documents, 31 are issued by organisations based in member countries of the Council of Europe, 14 by organisations based in Observer countries and 17 in non-members non-observer countries. Documents issued by organisations based in member countries of the Council of Europe make reference to human rights in 57.9% of cases. Documents from non-CoE member countries make reference to human rights in 49.6% of cases. This reveals that the human rights implications of Artificial Intelligence are more frequently addressed by organisations based in member countries of the Council of Europe compared to the rest of the world.

Our thematic content analysis retrieved a variety of ethically relevant codes, which could all be consistently allocated to the eleven overarching ethical clusters identified by Jobin, Ienca & Vayena (2019)[13]. These are, by decreasing order of frequency of the sources in which they were featured: transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, dignity, sustainability, and solidarity. A detailed frequency representation of ethical principles and associated codes is presented in Table 2.

Ethical principle	Number of documents	Included codes
Transparency	101/116	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice and fairness	97/116	Justice, fairness, consistency, inclusion, equality, equity, (non-)bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution, impartiality
Non-maleficence	84/116	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non- subversion
Responsibility	79/116	Responsibility, accountability, liability, acting with integrity
Privacy	74/116	Privacy, personal or private information, confidentiality
Beneficence	58/116	Benefits, beneficence, well-being, peace, social good, common good
Freedom and autonomy	48/116	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trustworthiness	41/116	Trust, trustworthiness
Sustainability	20/116	Sustainability, environment (nature), energy, resources (energy)
Dignity	20/116	Dignity
Solidarity	10/116	Solidarity, social security, cohesion

Table 2- Frequency of ethical themes and associated codes

No single ethical principle appears to be common to the entire corpus of documents, although there is an emerging convergence around the following principles: transparency, justice and fairness, non- maleficence, responsibility and privacy. These principles are referenced in nearly two thirds of all the sources. Nonetheless, further thematic analysis reveals the persistence of significant semantic and conceptual divergences in both how the 11 ethical principles are interpreted and the specific recommendations or areas of concern derived from each.

Regular expressions built from the codes reveal significant variations in theme coverage among documents produced within member countries of the Council of Europe (CoE) compared to documents produced elsewhere. Compared to documents produced in CoE observer countries, soft law documents produced within member countries of the Council of Europe appear to emphasize the following ethical principles: transparency, sustainability, freedom and autonomy, trust/trustworthiness and solidarity (see Figure 4). In contrast, they appear to refer more sporadically to the principles of justice, beneficence, and dignity. Compared to documents produced in the rest of the world (non-member non-observer countries), soft law documents produced within member countries of the Council of Europe appear to emphasize the principles of trust/trustworthiness and solidarity while addressing all other principles less frequently. The principles of privacy, justice and fairness showed the least variation, hence the highest degree of cross-geographical and cross-cultural stability.

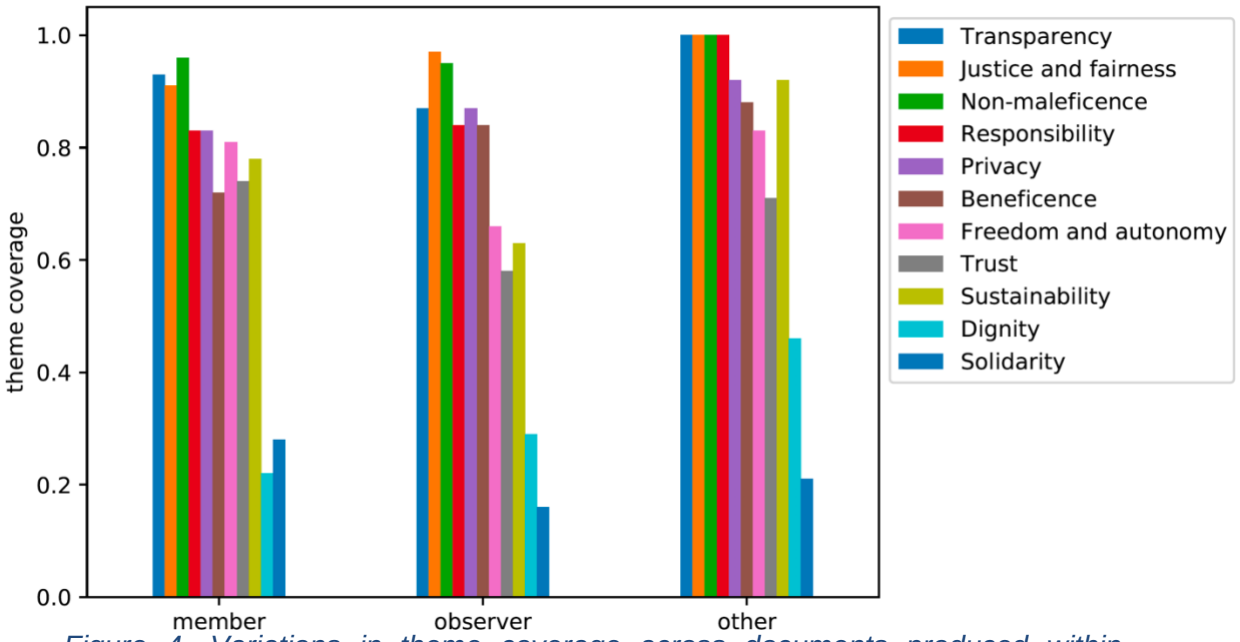

Figure 4- Variations in theme coverage across documents produced within member countries of the Council of Europe (CoE) vs documents produced in the rest of the world.

A detailed thematic evaluation of the afore listed is presented in the following.

Transparency: Featured in 101 out of 116 sources, transparency is the most prevalent ethical principle in the current soft law spectrum. Thematic analysis reveals significant variation in relation to the interpretation and justification of calls for transparency. This variation is observed to cause obvious divergences in the implementation strategies proposed to achieve transparency in relation to AI. References to transparency can be clustered into two main thematic families: (1) transparency of algorithms and data processing methods, (2) transparencies of human practices related to the design, development and deployment of AI systems. Calls for transparency of type 1 typically involve the promotion of methodological approaches to “explainable AI”, that is AI systems whose outputs and decisions can be understood by human experts. These methods and techniques contrast with "black box" approaches to machine learning where the steps through which an AI system arrived at a specific decision are unintelligible to human experts including the system’s designers. While private companies, especially private AI actors, tend to reduce transparency to interpretability and explainability through technical solutions —such as, among others, layerwise relevance propagation (LRP) and local interpretability—governmental bodies such as national data protection officers emphasise the importance of oversight methods such as audits. Calls for transparency of type 2 do not focus on interpretable algorithms but on the transparency of human practices related to data and AI such as disclosing relevant information to data subjects, avoiding secrecy when deploying AI strategies and forbidding conflicts of interest between AI actors and oversight bodies. Calls for transparency of this type are more common among governmental actors and NGOs.

Justice, fairness, and equity: Justice is mainly expressed in terms of fairness and prevention (or mitigation) of algorithmic biases that can lead to discrimination. Fears that AI might increase inequality and cause discrimination appear less common in soft law documents issued within the private sector compared to governmental bodies and academia. Documents disagree on how to achieve justice and fairness in AI. Some sources focus on respecting diversity and favouring inclusion and equality both when designing AI systems (especially when compiling the training datasets) and when deploying them in the society. Others sources call for a possibility to appeal or challenge decisions, predicating it on the right to redress and remedy. Fair access to the benefits of AI is also a commonly recurring theme. Documents issued by governmental actors place particular emphasis on AI’s impact on the labour market, and the need to address democratic or societal challenges. We identified five main non- mutually-exclusive implementation strategies for preserving and promoting justice and fairness in AI:

I. Via technical solutions such as standards and best practices;
II. By raising public awareness of existing rights and regulation;
III. Via better testing, monitoring and auditing of AI systems;
IV. By developing or strengthening the rule of law and the right to appeal, recourse, redress, or remedy;
V. Via systemic changes and processes such as governmental action and oversight, a more interdisciplinary workforce, as well as better inclusion of civil society or other relevant stakeholders in an interactive manner.

While solutions II-V appeared to be the preferred solution among governmental agencies (especially data protection officers), solutions of type I appeared more common among private AI actors.

Non-maleficence: References to non-maleficence occur significantly more often than references to beneficence and encompass general calls for safety and security or state that AI should never cause foreseeable or unintentional harm. Some documents focus on specific risks or potential harms, especially the risk of intentional misuse via cyberwarfare and malicious hacking. The most common sources of harm mentioned in the documents are social discrimination, privacy violation, and bodily or psychological harm. Soft law documents focused on harm mitigation often call for both technical solutions and mandatory governance interventions at the level of AI research, design, as well as technology development and deployment. Technical solutions include in-built data quality evaluations or security and privacy by design frameworks, though others advocate for establishing industry standards. Proposed governance strategies include active cooperation across disciplines and stakeholders, compliance with existing or new legislation, and the need to establish oversight processes and practices, notably tests, monitoring, audits and assessments by internal units, customers, users, independent third parties, or governmental entities. Some sources explicitly mention co-optation for military purposes —the so-called dual use problem— as a primary area of AI deployment requiring governance intervention.

Responsibility and accountability: References to developing 'responsible AI' are widespread. Nonetheless, the notions of responsibility and accountability are rarely defined. Diverse actors are named as being responsible and accountable for AI's actions and decisions. These include AI developers, designers, and the entire industry sector. Further disagreement emerged on whether AI should be held accountable in a human-like manner or whether humans should always be the only actors who are ultimately responsible for technological artefacts.

Privacy: Privacy is widely regarded as a value to uphold and a right to be protected. While privacy considerations are frequently addressed in current AI guidelines, there is no consensus on which unique challenges, if any, are raised by advances in AI compared to other data-intensive technologies. Thematic analysis reveals that most documents refer to privacy in general terms, without establishing any explicit nexus between the capabilities of AI and novel privacy challenges. Although poorly characterized, the privacy problem of AI is often presented in association with issues of data protection and data security. Proposed strategies to preserve privacy in AI can be clustered into three categories: (A) technical solutions such as differential privacy, secure multiparty computation and homomorphic encryption; (B) public engagement solutions such as raising awareness among users and data subjects, and (C) regulatory approaches solutions such as better defining the requirements for legal compliance (especially data protection regulation) or even creating new laws and regulations to accommodate the unique of AI.

Beneficence: While promoting good (*beneficence* in ethical terms) is often mentioned, it is rarely defined, though notable exceptions mention promoting human well-being and flourishing, peace and happiness, creating socio-economic opportunities and favouring economic prosperity. Similar uncertainty concerns the actors that should benefit from AI: private sector issuers tend to highlight the benefit of AI for customers, while academic and governmental sources typically argue that AI should benefit 'everyone', 'humanity' and 'society at large'. Strategies for the promotion of good include aligning AI with human values, minimizing power concentration and using AI capabilities for the promotion of human rights.

Freedom and autonomy: Soft law documents link AI to the preservation or promotion of several freedoms and liberties. These notably include freedom of expression, informational self-determination, the right to privacy and personal autonomy. This latter notion is generally referred to as a positive freedom, specifically the freedom to flourish, to decide for oneself and to self-determine one's own course of action. A minority of documents, however, refer to autonomy as a negative freedom, such as a freedom from technological experimentation, manipulation or surveillance. Proposed solutions to preserve freedom and autonomy in AI include pursuing transparent and explainable AI, raising AI literacy, ensuring informed consent or, conversely, actively refraining from collecting and spreading data in absence of informed consent.

Trust and trustworthiness: Slightly more than one in three soft law documents call for trustworthy AI research and technology or for the promotion of a culture of trust among scientists and engineers. Some documents, however, explicitly warning against excessive trust in AI, arguing that trust can only occur among peers and should not be delegated to AI. Suggestions for building or sustaining trust include education, reliability, accountability, processes to monitor and evaluate the integrity of AI systems over time and tools and techniques ensuring compliance with norms and standards.

Sustainability: Sustainability is sporadically mentioned, typically in relation to protecting the environment or even improving the planet's ecosystem and biodiversity. Some documents demand AI systems to process data sustainably and increase their energy efficiency to minimize ecological footprint47. A smaller portion of document focuses on social sustainability, that is ensuring accountability in relation to potential job losses and expand opportunities for innovation.

Dignity: While dignity remains undefined in existing guidelines, soft law documents specify that it is a prerogative of humans but not of robots. References to dignity are strongly intertwined with the protection and promotion of human rights. It is argued that AI should not diminish or destroy but respect, preserve or even increase human dignity. Dignity is believed to be preserved if it is respected by AI developers in the first place and promoted through new legislation, through governance initiatives, or through government-issued technical and methodological guidelines.

Solidarity: Solidarity is the least recurring ethical theme and it is mostly referenced in relation to the implications of AI for the labour market. Sources call for a stronger social safety net to cope with the long-term implications of AI for human labour. They underline the need for redistributing the benefits of AI in order not to threaten social cohesion[6,5] and respecting potentially vulnerable persons and groups. Lastly, there is a warning of data collection and processing practices focused on individuals which may undermine solidarity in favour of 'radical individualism'.

Limitations

This study has several limitations. First, from a bibliographic perspective, guidelines and soft-policy documents are an instance of grey literature, hence not indexed in conventional scholarly databases. Therefore, their retrieval is inevitably less replicable and unbiased compared to systematic database search of peer-reviewed literature. Following best practices for grey literature review, this limitation has been mitigated by developing a discovery and eligibility protocol which was pilot-tested prior to data collection. Although search results from search engines are personalized, the risk of personalization influencing discovery has been mitigated through the broadness of both the keyword search and the inclusion of results. A language bias may have skewed our corpus towards English results. We minimised this limitation by including entries written in the following languages (besides English): German, French, Italian, Spanish and Dutch. Keywords and codes in the afore-listed languages were translated into English and included in the analysis. Our content analysis presents the typical

limitations of qualitative analytic methods. Following best practices for content analysis, this limitation has been mitigated by developing an inductive coding strategy which was conducted independently by two reviewers to minimize subjective bias. Finally, given the rapid pace of publication of AI guidance documents, there is a possibility that new policy documents were published after our search was completed. To minimize this risk, continuous monitoring of the literature was conducted in parallel with the data analysis and until 1st March 2020.

Discussion and Normative Ethical Analysis

We found a rapid increase in the number and variety of soft law documents on AI, demonstrating the increasing active involvement of the international community in non-mandatory governance in this technological domain. Organisations issuing AI guidelines, principles and other soft law instruments come from a wide range of sectors. In particular the nearly equivalent proportion of documents issued by the public (i.e. governmental organisations) and the private sector (companies and private sector alliances) indicates that the ethical challenges of AI concern both public entities and private enterprises. However, there is significant divergence in the solutions proposed to meet the ethical challenges of AI, with public actors prioritizing technical solutions such as explainable and interpretable AI over mandatory regulation and in-depth ethical reflection. Further, the relative underrepresentation of geographic areas such as Africa and South America indicates that the international debate over ethical AI may not be happening globally in equal measures. More economically developed countries (MEDCs) are shaping this debate more than others, which raises concerns about neglecting local knowledge, cultural pluralism and global fairness. These findings confirm the uneven geographic representation and distribution of AI ethics actors observed in previous studies[13]. Compared to previous studies, however, our review reveals that novel actors from previously unrepresented countries are now participating in international non-mandatory governance. These include actors from AI superpowers, that is global-leading AI countries such as China, as well as middle income countries from previously unrepresented world regions such as Russia and Mexico.

The proliferation of soft-law efforts can be interpreted as a governance response to advanced research into AI, whose research output and market size have drastically increased in recent years[16]. Our analysis shows the emergence of an apparent cross-stakeholder convergence on promoting the ethical principles of transparency, justice, non-maleficence, responsibility, and privacy. Nonetheless, our thematic analysis reveals substantive divergences in relation to four major factors: (i) how ethical principles are interpreted, (ii) why they are deemed important, (iii) what issue, domain or actors they pertain to, and (iv) how they should be implemented. Furthermore, unclarity remains as to which ethical principles should be prioritized, how conflicts between ethical principles should be resolved, who should enforce ethical oversight on AI and how researchers and institutions can comply with the resulting guidelines. These findings suggest the existence of a gap at the cross-section of principles formulation and their implementation into practice which can hardly be solved through technical expertise or top-down approaches.

Although no single ethical principle is explicitly endorsed by all existing guidelines, transparency, justice and fairness, non-maleficence, responsibility and privacy are each referenced in more than half of all guidelines. This focus could be indicating a developing convergence on ethical AI around these principles in the global policy landscape. In particular, the prevalence of calls for transparency, justice and fairness points to an emerging moral priority to require transparent processes throughout the entire AI continuum (from transparency in the development and design of algorithms to transparent practices for AI use), and to caution the global community against the risk that AI might increase inequality if justice and fairness considerations are not adequately addressed. Both these themes appear to be strongly intertwined with the theme of responsibility, as the promotion of both transparency and justice seems to postulate increased responsibility and accountability on the side of AI makers and deployers.

It has been argued that transparency is not an ethical principle per se, but rather "a proethical condition for enabling or impairing other ethical practices or principles"[17]. This characterization of transparency as a proethical condition for other principle is detectable in IBM's Supplier's Declaration of Conformity (SDoC) that helps to provide information about the four key pillars of trustworthy AI. The allegedly pro-ethical nature of transparency might partly explain its higher prevalence compared to other ethical principles. It is notable that current guidelines place significant value in the promotion of responsibility and accountability, yet few of them emphasize the duty of all stakeholders involved in the development and deployment of AI to act with integrity. This mismatch is probably associated with the observation that existing guidelines fail to establish a full correspondence between principles and actionable requirements, with several principles remaining uncharacterized or disconnected from the requirements necessary for their realization.

As codes related to non-maleficence outnumber those related to beneficence, it appears that, for the current AI community, the moral obligation to preventing harm takes precedence over the promotion of good. This fact can be partly interpreted as an instance of the so-called negativity bias, i.e. a general cognitive bias to give greater weight to negative entities[18,19], a hypothesis emphasized by cognitive psychologist Steven Pinker in a recent in-depth analysis of the Scientific Foresight Unit (STOA) of the European Parliament[20]. This negative characterization of ethical values is further emphasized by the fact that existing guidelines focus primarily on how to preserve privacy, dignity, autonomy and individual freedom *in spite of* advances in AI, while largely neglecting whether these principles could be actively promoted through responsible innovation in AI.

The issue of trust in AI, while being addressed by less than one third of all sources, tackles a critical ethical dilemma in AI governance: determining whether it is morally desirable to foster public trust in AI. While several sources, especially those produced within the private sector, highlight the importance of fostering trust in AI through educational and awareness-raising activities, a smaller number of sources contend that trust in AI may actually diminish scrutiny and undermine some societal obligations of AI producers[21]. This possibility would challenge the dominant view in AI ethics that building public trust in AI is a fundamental requirement for ethical governance[22]. In relation to trust, we observed to additional conceptual challenges. First, conceptual clarity on the meaning and dynamics of trust seems lacking across the current documents. Most sources failed to specify the trustor and the trustee of the trusting relationship they described, hence neglect that "trust" is a relational and highly complex which involves at least two actors, which trust each other to do, or not to do, a certain activity. This relationship is affected by a wide range of framing factors, for example culture, belief systems, contexts, as well as traits of the actors within the trust relationship. These contextual factors seemed to be neglected in the current literature. Most importantly, the trait "trustworthiness" and the relational construct "trust" appeared frequently conflated or used interchangeably by the AI actors we reviewed. This conflation does not only lead to conceptual confusion but may also foster false hopes among AI users and policy makers. Trust and trustworthiness are different concepts, and trustworthiness does not lead per se to a trusting relationship. Further

governance work in this area should clarify this crucial conceptual distinction and demand greater clarification about the requirements of a trusting relationship.

The relative thematic underrepresentation of sustainability and solidarity suggests that these topics might be currently flying under the radar of the mainstream ethical discourse on AI. The underrepresentation of sustainability-related principles is particularly problematic in light of the fact that the deployment of AI requires massive computational resources which, in turn, require high energy consumption[23]. The environmental impact of AI, however, does not only involve the negative effects of high-footprint digital infrastructures, but also the possibility of harnessing AI for the benefit of ecosystems and the entire biosphere. This latter point, highlighted in a report by the World Economic Forum though not in the AI guidelines by the same institution, requires wider endorsement to become entrenched in the ethical AI narrative[24]. The ethical principle of solidarity is sparsely referenced, typically in association with the development of inclusive strategies for the prevention of job losses and unfair sharing of burdens. Little attention is devoted to promoting solidarity through the emerging possibility of using AI expertise for solving humanitarian challenges, a mission that is currently being pursued, among others, by intergovernmental organisations such as the United Nations Office for Project Services (UNOPS) or the World Health Organization (WHO) and private companies such as Microsoft. As the humanitarian cost of anthropogenic climate change is rapidly increasing[25], the principles of sustainability and solidarity appear strictly intertwined though poorly represented compared to other principles.

While numerical data indicate an emerging convergence around the promotion of some ethical principles, in-depth thematic analysis paints a more complicated picture, as there are critical differences in *how* these principles are interpreted as well as what requirements are considered to be necessary for their realization. Results show that different and often conflicting measures are proposed for the practical achievement of ethical AI. For example, the need for ever larger, more diverse datasets to "unbias" AI appears difficult to conciliate with the requirement to give individuals increased control over their data and its use in order to respect their privacy and autonomy. Similar contrasts emerge between the requirement of avoiding harm at all costs and that of balancing risks and benefits. Furthermore, it should be noted that risk-benefit evaluations will lead to different results depending on whose well-being it will be optimized for by which actors. If not resolved, such divergences and tensions may undermine attempts to develop a global agenda for ethical AI.

Despite a general agreement that AI should be ethical, significant divergences emerge within and between guidelines for ethical AI. Furthermore, uncertainty remains regarding how ethical principles and guidelines should be implemented. These challenges have implications for science policy, technology governance and research ethics. At the policy level, they urge increased cooperative efforts among governmental organisations to harmonize and prioritize their AI agendas, an effort that can be mediated and facilitated by inter-governmental organisations. While harmonization is desirable, however, it should not come at the costs of obliterating cultural and moral pluralism over AI. Therefore, a fundamental challenge for developing a global agenda for AI is balancing the need for cross-national harmonization over the respect for cultural diversity and moral pluralism. This challenge will require the development of deliberative mechanisms to adjudicate disagreement concerning the values and implications of AI advances among different stakeholders from different global regions. At the level of technology governance, harmonization is typically implemented in terms of standardizations. Efforts in this direction have been made, among others, by the Institute of Electrical and Electronics Engineers (IEEE) through the "Ethically Aligned Designed" initiative[26]. Finally, soft governance mechanisms such as Independent Review Boards (IRBs) will be increasingly required to assess the ethical validity of AI applications in scientific research, especially those in the academic domain. However, AI applications by governments or private corporations will unlikely fall under their oversight, unless significant expansions to the IRBs' purview are made.

Overall, our findings indicate that the international community does not agree on what constitutes ethical AI and what requirements are necessary for its achievement. Nonetheless, signs of convergence are noticeable around the notions of transparency, non-maleficence, responsibility, and privacy. Enriching the current ethical AI discourse through a better appraisal of critical yet underrepresented ethical principles such as human dignity, solidarity and sustainability is likely to result into a better articulated ethical landscape for AI. Furthermore, shifting the focus from principle- formulation to translation into practice is desirable. A global agenda for ethical AI should balance the need for cross-national and cross-domain harmonization over the respect for cultural diversity and moral pluralism. Overall, our review provides a useful starting point for understanding the inherent diversity of current principles and guidelines for ethical AI and outlines the challenges ahead for the global community.

Policy Implications

The plethora of international efforts to produce soft law documents on AI provides valuable proxy information about how humanity will react to the many governance challenges posed by AI. The international community seems to converge on the importance of transparency, non-maleficence, responsibility, and privacy for the development and deployment of ethical AI. However, enriching the current ethical AI discourse through a better appraisal of critical yet underrepresented ethical principles such as human dignity, solidarity and sustainability is likely to result into a better articulated ethical landscape for artificial intelligence. Furthermore, shifting the focus from principle-formulation to translation into practice must be the next step. A global agenda for ethical AI should balance the need for cross-national and cross-domain harmonization over the respect for cultural diversity and moral pluralism.

These findings have implications for public policy, technology governance and research ethics. At the policy level, greater intra-stakeholder cooperation is needed to mutually align different AI ethics agendas and seek procedural convergence not only on the ethical principles but also on their implementation. While global consensus might be desirable, it should not come at the costs of obliterating cultural and moral pluralism and might require the development of deliberative mechanisms to adjudicate disagreement among stakeholders from different global regions. Such efforts can be mediated and facilitated by inter-governmental organisations such as the Council of Europe. Furthermore, they could be complemented by bottom-up approaches involving all relevant stakeholders on an equal footing.

Policy interventions in this arena should clarify how AI ethics guidelines relate to existing national and international regulation. In spite of AI's alleged sociotechnical uniqueness, soft law documents on AI do not operate in an ethical-legal vacuum. In contrast, ethics guidelines and other soft law instruments will ultimately have to operate in a context already heavily populated by rules, including hard law (mandatory governance). Failure to consider the context of those rules could undermine the import of the principles into actionable and effective international governance. An example of that is transparency, the most widely recurring ethical principle. In spite of its frequent occurrence, the principle of transparency is typically referred without an explicit link to the underlying binding regulation. Today, institutions that use AI technology are already subject to numerous transparency rules under existing legal systems such as the Fair Credit Reporting Act in the United States and the specific practical requirements on data controllers and processors as outlined in Articles 12-14 of the EU General Data Protection Regulation (GDPR). Similarly, clarifying the distinction between "trust" and "trustworthiness" is a critical task for policy makers.

Besides integrating hard and soft law, an additional challenge is translating ethics principles into practice and seeking harmonization between divergent AI ethics codes. At the level of technology governance, promising attempts to harmonization have been pursued through standardization initiatives such as those led by the Institute of Electrical and Electronics Engineers (IEEE), i.e. the world's largest technical professional organization dedicated to advancing technology innovation. The IEEE is pursuing both AI ethics efforts for general-

purpose autonomous and intelligent systems, under the framework of the "Ethically Aligned Designed" initiative[26], as well as domain-specific ones such as the "Neurotechnologies for Brain-Machine Interface Standards Roadmap" developed by the IEEE Standards Association.

Another policy implication regards research oversight. Research ethics mechanisms such as Independent Review Boards (IRBs) will be increasingly required to assess the ethical validity of AI applications in scientific research, especially those in the academic domain. However, AI applications by governments or private corporations will unlikely fall under their oversight, unless significant expansions to the IRBs' purview are made.

Overall, the thematic variety and informational richness of the documents we analysed suggests that soft law instruments issued by governmental and nongovernmental organisations (incl. private companies and academic organisations) are useful tools to exert practical influence on public decision making over AI. If adequately conceptualized, designed and drafted, soft law initiatives hold potential for steering the development of AI systems for social good and in abidance of ethical values and legal norms. However, soft law approaches should not be considered substitutive of mandatory governance. Self-regulation efforts by private AI actors are at particular risk of being promoted to bypass or obviate mandatory governance by governmental and intergovernmental authorities. This risk has been emphasised by the German philosopher Thomas Metzinger, a member of the EU High-Level Expert Group on AI, who observed how a significant portion of the AI ethics discourse is shaped by the private sector[20].

The uneven geographic representation of issuing organisations of AI ethics guidelines requires close monitoring and reflection by international, especially inter-governmental, organisations. In order to ensure inclusiveness, cultural pluralism and fair participation to collective decision making on AI, the development of soft law documents by organisations located in currently underrepresented global regions, especially Africa and South America, should be promoted. Intergovernmental organisations such as the Council of Europe can play a crucial role in the establishment of international platforms of mutual exchange and debate on AI ethics and governance.

The convergence of current soft law instruments around five generic ethical principles such as transparency, justice, non-maleficence, responsibility, and privacy reveals five priority areas of oversight and possible intervention by mandatory governance authorities at both the governmental and intergovernmental level. Prioritizing the realisation of these principles could facilitate the establishment of a core set of norms based on widely agreed ethical precepts. Furthermore, their wide acceptance across both private and public actors is likely to ensure higher degrees of compliance. That being said, in order to be translated into effective governance, these ethical principles should be conceptually clarified. Policy makers have the duty to resolve semantic ambiguities and conflicting characterisations of these principles. The sharp disagreement of current soft law documents on the interpretation and practical implementation of these principles indicates that mandatory governance solutions are likely subject to public disagreement, hence require a transparent process of democratic deliberation.

In parallel, the relative underrepresentation of ethical considerations such as those regarding sustainability, dignity and solidarity needs to be further scrutinized to avoid importing into mandatory governance the same conceptual gaps and normative blind spots of soft law. Mandatory governance should complement and fill the gaps of non-mandatory approaches rather than mirroring the same blind spots of the soft law. To adequately address the sustainability and solidarity challenges of AI, a greater cooperation between environmental protection agencies, ministries of labour and employment as well as ministries of technology and innovation might be required.

As nearly half of reviewed soft law documents do not explicitly recommend the promotion—or

warn against the violation— of human rights when designing, developing and deploying AI systems, greater focus on the human rights implications of AI is urgently needed. Member countries of the Council of Europe are well-positioned to steer the international governance of AI towards the promotion of human rights. The human rights implications of AI should be thoroughly investigated at various levels: First, it should be investigated at the level of rights and obligations in the philosophical sense, as they operate independently of legal enactment as justified moral norms. Second, it should be assessed at the level of international human rights law. In this regard, European Convention on Human Rights (ECHR) can pivotal role in international doctrinal research and deliberation on AI. Adherence to the convention is a critical requirement to ensure the socially responsible development and adoption of a new technology. It is therefore of paramount importance to assess the impact of the sociotechnical transformation induced by AI on the fundamental rights and freedoms postulated in the ECHR. This impact assessment should have a twofold goal: (i) evaluating if and how AI will affect or pose new risks for human rights and freedoms; (ii) evaluating if and how the responsible development of AI and public deliberation in its regard can contribute to the promotion of those rights and freedoms. It should be underscored that since technologies are not developed in a vacuum but within a social-historical context of human practices, customs and norms, effective impact assessment strategies should not look at AI in abstraction but contextually to current practices and norms[27]. Finally, it is important to investigate the interface between AI and human rights not only from a high-level perspective, but also and foremost by looking at the human rights salience of specific domains of applications of AI such as *inter alia* robotics[8,28], big data[29,30], autonomous weapons[31,32] and brain-computer interfaces[33].

Acknowledgments

The authors would like to thank Dr. Anna Jobin, Karolina Ignatiadis and Manuel Schneider whose work has contributed to the realization of this report.

References

1. Michie D, Spiegelhalter DJ, Taylor C. Machine learning. *Neural and Statistical Classification.* 1994;13.
2. Appenzeller T. The AI revolution in science. *Science.* 2017;357:16-17.
3. Harari YN. Reboot for the AI revolution. *Nature News.* 2017;550(7676):324.
4. Helbing D, Frey BS, Gigerenzer G, et al. Will democracy survive big data and artificial intelligence? In: *Towards Digital Enlightenment.* Springer; 2019:73-98.
5. Livingston S, Risse M. The Future Impact of Artificial Intelligence on Humans and Human Rights. *Ethics & International Affairs.* 2019;33(2):141-158.
6. Nemitz P. Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences.* 2018;376(2133):20180089.
7. Ashrafian H. Intelligent robots must uphold human rights. *Nature.* 2015;519(7544):391-391.
8. Van Est R, Gerritsen J, Kool L. Human rights in the robot age: Challenges arising from the use of robotics, artificial intelligence, and virtual and augmented reality–Expert report written for the Committee on Culture. *Science, Education and Media of the Parliamentary Assembly of the Council of Europe (PACE)(Rathenau Institute), retrieved from:* https://www/ *rathenau nl/sites/default/files/2018-02/Human% 20Rights% 20in% 20the% 20Robot% 20Age-Rathenau% 20Instituut-2017 pdf (January 5, 2019).* 2017.
9. Raso FA, Hilligoss H, Krishnamurthy V, Bavitz C, Kim L. Artificial Intelligence & Human Rights: Opportunities & Risks. *Berkman Klein Center Research Publication.* 2018(2018-6).
10. Assembly UG. Universal declaration of human rights. *UN General Assembly.* 1948;302(2).
11. Mowbray A. The European Convention on Human Rights. In: *International Human Rights Law.* Routledge; 2016:287-304.
12. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *International journal of social research methodology.* 2005;8(1):19-32.
13. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nature Machine Intelligence.*
2019;1(9):389-399.
14. Ienca M, Ferretti A, Hurst S, Puhan M, Lovis C, Vayena E. Considerations for ethics review of big data health research: A scoping review. *PloS one.* 2018;13(10):e0204937.
15. Li Y, Krishnamurthy R, Raghavan S, Vaithyanathan S, Jagadish H. Regular expression learning for information extraction. Paper presented at: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing2008.
16. Shoham Y, Perrault R, Brynjolfsson E, et al. The AI Index 2018 annual report. *AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford, CA.* 2018.
17. Turilli M, Floridi L. The ethics of information transparency. *Ethics and Information Technology.* 2009;11(2):105-112.
18. Rozin P, Royzman EB. Negativity bias, negativity dominance, and contagion. *Personality and social psychology review.* 2001;5(4):296-320.Ito TA, Larsen JT, Smith NK, Cacioppo JT. Negative information weighs more heavily on the brain: the negativity bias in evaluative categorizations. *Journal of personality and social psychology.* 1998;75(4):887.
19. Peter J. Bentley, Miles Brundage, Olle Häggström, Thomas Metzinger. Should we fear artificial intelligence? *European Parliamentary Research Service.* 2018:Scientific Foresight Unit (STOA).
20. Bryson J. No one should trust artificial intelligence. *Science & Technology: Innovation, Governance, Technology.* 2018;11:14.
21. Winfield AF, Jirotka M. 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.* 2018;376(2133):20180085.
22. Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:190602243.* 2019.
23. Forum WE. Harnessing Artificial Intelligence for the Earth.
24. Scheffran J, Brzoska M, Kominek J, Link PM, Schilling J. Climate change and violent conflict. *Science.* 2012;336(6083):869-871.
25. IEEE. Ethically aligned design. *IEEE Standards v1.* 2016(Global Initiative).

26. Rasmussen T. *Social theory and communication technology.* Routledge; 2019.
27. Liu H-Y, Zawieska K. From responsible robotics towards a human rights regime oriented to the challenges of robotics and artificial intelligence. *Ethics and Information Technology.* 2017:1-13.
28. Mantelero A. AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review.* 2018;34(4):754-772.
29. Vayena E, Tasioulas J. The dynamics of big data and human rights: The case of scientific research. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences.* 2016;374(2083):20160129.
30. Heyns C. Human rights and the use of autonomous weapons systems (AWS) during domestic law enforcement. *Hum Rts Q.* 2016;38:350.
31. Asaro P. On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross.* 2012;94(886):687-709.
32. Ienca M, Andorno R. Towards new human rights in the age of neuroscience and neurotechnology. *Life Sci Soc Policy.* 2017;13(1):5.