Ethics and Information Technology (2021) 23:505–525
https://doi.org/10.1007/s10676-021-09593-z

ORIGINAL PAPER

Artificial Intelligence Regulation: a framework for governance

Patricia Gomes Rêgo de Almeida[1,2] · Carlos Denner dos Santos[1,3] · Josivania Silva Farias[1]

Accepted: 7 April 2021 / Published online: 21 April 2021

Abstract
This article develops a conceptual framework for regulating Artificial Intelligence (AI) that encompasses all stages of modern public policy-making, from the basics to a sustainable governance. Based on a vast systematic review of the literature on Artificial Intelligence Regulation (AIR) published between 2010 and 2020, a dispersed body of knowledge loosely centred around the "framework" concept was organised, described, and pictured for better understanding. The resulting integrative framework encapsulates 21 prior depictions of the policy-making process, aiming to achieve gold-standard societal values, such as fairness, freedom and long-term sustainability. This challenge of integrating the AIR literature was matched by the identification of a structural common ground among different approaches. The AIR framework results from an effort to identify and later analytically deduce synthetic, and generic tool for a country-specific, stakeholder-aware analysis of AIR matters. Theories and principles as diverse as Agile and Ethics were combined in the "AIR framework", which provides a conceptual lens for societies to think collectively and make informed policy decisions related to what, when, and how the uses and applications of AI should be regulated. Moreover, the AIR framework serves as a theoretically sound starting point for endeavours related to AI regulation, from legislation to research and development. As we know, the (potential) impacts of AI on society are immense, and therefore the discourses, social negotiations, and applications of this technology should be guided by common grounds based on contemporary governance techniques, and social values legitimated via dialogue and scientific research.

Keywords Ethics · Artificial Intelligence · Regulation · Governance · Framework

Introduction

The widespread use of AI in our daily actions and in an unnoticeable fashion (Cerka et al., 2015) has introduced unprecedented ethical issues to a broad and complex social system (Cave et al., 2019).

From the same perspective, the complexity of data treatment in the design and development process of a machine learning solution increases the likelihood of ethical surprises, which demands a wider evaluation of the ethical and social impacts (Butterworth, 2018).

Based on this reflection, this work has sought to conduct a vast search for literature that is relevant in terms of Artificial Intelligence Regulation, processing and grouping it into a set of purposes presented as frameworks or guidelines for a framework based on ethical principles. Their main contributions have been customised as a framework based on the Design and Action Theory (Gregor, 2006) that allows for reflections and actions aimed at regulating and governing operations and relationships between natural and legal persons on one side, and AI-embedded systems on the other.

✉ Patricia Gomes Rêgo de Almeida
patricia.almeida@camara.leg.br

Carlos Denner dos Santos
carlosdenner@unb.br

Josivania Silva Farias
josivania@unb.br

[1] University of Brasilia (UnB) – Department of Administration, Brasília, Brazil

[2] Chamber of Deputies of Brazil – Directorate of Innovation and Information Technology, Brasília, Brazil

[3] LATECE, University of Quebec at Montreal (UQAM), Montreal, Canada

Reasons to regulate AI

Since the term was coined in 1956, Artificial Intelligence has been associated with a wide range of concepts (Cerka et al., 2017; Jackson, 2019) based on a thinking human being and on rational behaviour, which could be synthetised as: systems that think and act like humans and systems that think and act rationally (Cerka et al., 2015; Russell & Norvig, 1995). Equally wide is the variety of different names associated with whatever utilises AI technology: robots, smart systems, intelligent systems, intelligent agents, AI agents, AI algorithms, intelligent algorithms, and autonomous systems, to mention a few.

For the purpose of avoiding misunderstandings regarding AI, the High-Level Expert Group established by the European Commission has defined AI systems 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, 2019a). Considering the difficulty of defining AI in a way that could fit all approaches needed for regulation and governance actions with clear communication among all stakeholders, this article adopts the definition of AI established by the High-Level Expert Group on AI systems.

The responsibilities, security, intellectual property, and privacy associated with different systems for medical robots, drones, autonomous cars, among several "intelligent solutions" offered every day have been questioned.

Illustrating the level of risk-related indeterminacy, machine learning has been combined with game theory (Conitzer et al., 2017) in cases where developers were using game theory to help teach strategic defence to algorithms. A game between two algorithms predicted that one would kill the other only in case of an absolute scarcity of resources. However, when a more intelligent algorithm was introduced, it immediately killed the weaker ones (Firth-Butterfield, 2017). This case reinforces the idea that an autonomous system will inevitably find itself in a situation in which it needs not only to obey a certain rule or not, but also to make a complex ethical decision (Dennis et al., 2016).

Facing the risks compels us to explore their causes and effects. Although the effects of AI are not yet known, a large amount of them can currently be classified. Firstly, those coming from the undesired effects, such as biases, discrimination, loss of privacy, false positives and false negatives, loss of autonomy, (psychological, financial, or physical) damage, loss of control, difficulty identifying liabilities, losses or decreases in human rights, unemployment, misjudgements, and concentration of power and wealth in a few companies. Secondly, some risks are the result of intentional misuses, such as fake news, deep fakes, cyberattacks, terrorism, warfare, weapons, people manipulation, espionage, low level of democracy (Beltran, 2020; Benjamins & Garcia 2020; Borgesius, 2018; Jackson, 2020; Jobin et al., 2019; Mika et al., 2019).

Considering all those risks, establishing best practices for delegating and defining new moral responsibility attribution models is crucial to leverage the opportunities created by AI (Taddeo & Floridi, 2018). Risk assessment models can provide support and flexibility for Big Data and AI applications (Mantelero, 2018), and stakeholders who develop and deploy AI-based systems must enhance their knowledge of the values protected by human rights and how those rights apply to their own actions (Smuha, 2020).

Despite being a huge challenge, finding a way to deal with ethical issues must be a constant target of research, for what we need to join all our forces (Bostrom, 2014), and AI regulation is on the right path to get there (Carter, 2020).

The reasons to regulate include: manufacturers' need to comprehend a legal framework within which they can operate reliably; consumers' and society's need to be protected from devices that may harm or adversely affect them; and the need for business opportunities (Holder et al., 2016a).

In industries still lacking regulation, the general approach observed is that innovation is freely allowed, but those in charge should bear the consequences in case certain types of damage are caused (Reed, 2018).

Faced with the challenge of minimising those risks, a combination of strategy and actions must be put to practice during the entire lifecycle of AI systems, in order not only to identify damages and responsibilities, but also, and especially, to avoid them.

Seeking the best way to regulate

Sometimes, when used to denote an attempt to standardise behavioural patterns, the term "regulation" assumes the meaning of a law (Hildebrandt, 2018).

However, on a broader approach, regulation is a sustained attempt to modify behaviours of others according to defined standards or purposes in order to produce the desired outcomes. This can involve standard-setting, information-gathering, and behaviour modification mechanisms (Black, 2002), especially in cases evolving ethical issues, whose understanding is complex when applied to a real world. Therefore, law is just one way of regulating society, while other alternatives to regulate human behaviour may also be widely used (Hildebrandt, 2018).

Disruptive innovation always challenges regulatory strategies due to the reactive nature of traditional regulation (Kaal & Vermeulen, 2017). In the case of innovation by AI, the challenge is amplified, since it is strongly related to ethical issues and its results could be unpredictable in some situations, bringing about unforeseen social impacts. In addition, if AI adoption and implementation are conducted in a reckless manner, social and political instability could ensue, thus threatening freedom, self-determination, human rights, and fundamental values (Caron & Gupta, 2020). As human behaviour encompasses decisions from an ethical perspective, the regulation should also consider it. While norms as instruments of regulation relate to what is good or bad from society's point of view, ethics concerns itself with the nature of the principles upon which those norms are founded (Pedro, 2014).

A few laws have been resorted to in an attempt to settle damages caused by AI-supported products and services judicially. If, on the one hand, the number of cases is multiplying, on the other, the legislative branch seems to be moving at a negligible speed compared to the technological advancements enforcing the perception that traditional regulation does not fit in this challenge (Cerka et al., 2015; Larsson, 2020; Villaronga & Heldeweg, 2018). Part of this increasing gap between laws and technology is caused by the lack of a thorough and accurate definition of AI (Firth-Butterfield, 2017; Larsson, 2020), which is aggravated by the fact that the definition changes as the technology evolves (Fjeld et al., 2020). Considering this issue, the concept of dynamic regulation could fit in the field of AI, as it is based on learning by doing and continuity of regulatory relationships (Kaal & Vermeulen, 2017; Lewis & Yildirim, 2002).

A yet-to-be-solved equation is the breadth of laws dealing with globally produced and commercialised technologies (Holder et al., 2016a) and robot-generated inventions (Holder et al., 2016b). The problem reaches even broader dimensions when one considers the complex networks established in the technology industry, making it possible for products to be subjected to learning from data scattered across the world (Lenardon, 2017).

Large-scale data analyses have revealed that the key challenge related to the AI regulation dilemma is demonstrating it is produced and deployed appropriately (Butterworth, 2018). One of the most advocated strategies is transparency, an opening of the entire production process, especially the decision-making rules, the method, and the data utilised when training the intelligent system (Buiten, 2019; Butterworth, 2018; Tutt, 2017). However, on certain occasions, even in case the AI algorithm is open, full transparency cannot be ensured, as there is a difference between seeing the whole code and understanding all of its potential effects (Firth-Butterfield, 2017). A similar strategy to open data is the Explainable Artificial Intelligence (XAI) standard for the creation of coding models oriented towards a global comprehension (Adadi & Berrada, 2018; Taddeo & Floridi, 2018). In addition to the concerns related to the development process of an AI system, data governance has been recognised as being key to AI governance (Hilb, 2020; UK Government, 2018).

Some of the AI regulation theories that have been proposed are based on contractual and extracontractual liability, or on strict liability, and adopt a liability model in which the moral responsibility is distributed among designers, regulators, and users. The attempt to hold robots accountable for their actions has led a few countries to consider the possibility of granting a legal identity to each unit. One could argue that if parties in a contractual relationship may be legally represented by another entity, then so can systems (Cerka et al., 2017). As a counterargument, the term "robot liability" should be replaced with "indirect liability over the robot", given the impossibility of claiming damages from a robot, i.e., it cannot be held criminally liable. Thus, the impact of such products on society should also be a liability (Jackson, 2019; Nevejans, 2016). Although this latter understanding tends to be more acceptable from a global perspective, a liability model is still an essential and complex variable to be defined through an AI regulation strategy.

Also among the concerns that motivate AI regulation is the approach aimed at minimising the disruption of the work model with the goal of fighting job loss (Wright & Schultz, 2018).

Drawing attention to the domain of what is to be regulated, attempts to legislate digital technologies without proper knowledge for doing so have been criticised (Reed, 2018). With the intention of minimising those risks, a gradual regulation strategy (Villaronga & Heldeweg, 2018) can be used. When mitigating risks, regulatory agencies could bar the introduction of certain algorithms into the market until their safety and efficacy have been proven by means of tests (Tutt, 2017) founded on ethics (Arkin, 2011).

In 2017, the European Parliament Committee on Legal Affairs released a report recommending the creation of a European agency for robotics and AI, suggesting a combination of both hard and soft laws, given the complexity associated with the evolution of the regulatory model. It would put regulators and external experts together to monitor AI trends and study standards for best practices (Cath et al., 2017; Nevejans, 2016). After approving the study of the High-Level Expert Group on AI, the European Commission recommended upgrading the European Framework to one especially designed for AI Governance (European Commission, 2019). In the same direction, the House of Lords (2018) has recommended the creation of an AI regulatory framework.

Another effort observed in the US has resulted in S.3891, which defines conditions for advancing Artificial Intelligence research, including the development of technical standards (US Congress, 2020), and in H.Res. 153, which aims to support the development of guidelines for the ethical dsevelopment of Artificial Intelligence (US Congress, 2019).

In a parallel effort, many self-regulatory private-sector initiatives have been created, and research has been carried out to discuss ethical issues on AI development and use, such as the Partnership on AI to Benefit People and Society (AI4People, 2018; Partnership on AI, 2016; The Future of Life Institute, 2019b), The Montreal Declaration for a Responsible Development of Artificial Intelligence (University of Montreal, 2018), and The Toronto Declaration (Toronto, 2020).

At the government level, ethical principles were considered when the national AI-oriented strategies of a few countries were drawn up, as happened in Japan (Japanese Cabinet Office, 2019), France (French PM, 2018), Germany (German Federal Government, 2018), United Arab Emirates (Dubai, 2019), India (Aayog, 2018), and Singapore (Monetary Authority of Singapore, 2019). Additionally, several countries have shown their intention to create policies and laws to regulate the development and use of AI (Future of Life Institute, 2019a). Similar concerns have served as the basis for recommendations regarding ethical principles by a few transnational organisations, such as the Council of Europe (2018) and the Organisation for Economic Cooperation and Development (2019).

As the major concern regarding both self-regulation and government initiatives kickstarted the debate on AI governance through ethical principles, a set of core topics was comprised in each one of them: privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values. However, different principles can be observed under the same topic, which illustrates the lack of unanimity (Fjeld et al., 2020).

Standardisation documents are also part of the efforts associated with the AI regulation challenge. A good example is the ad hoc technical committee on Artificial Intelligence established within the International Organisation for Standardisation (ISO), whose plan includes two dozen standards on AI and Big Data (Neznamov, 2020).

The huge gap between ethical guidelines and laws, apart from the great number of potential situations in which stakeholders (developers, deployers, etc.) fail to apply such ethical principles, draws attention to the need to shift from principles to processes when it comes to AI governance (Larsson, 2020). Thus, there is a long path to be paved through a connected network of processes including several stakeholders in a way to keep on pace with the society's values.

Method

With the goal of surveying the relevant scientific literature on AI regulation, we have systematically searched for and organised papers to summarise the corpus and perform a qualitative analysis to understand the evolution and current state of the science.

We have compiled papers published between 2010 and 2020 containing the following expressions: ("ARTIFICIAL INTELLIGENCE" and "ETHICAL USE"), ("ARTIFICIAL INTELLIGENCE" and "REGULATION"), or ("ARTIFICIAL INTELLIGENCE" and "GOVERNANCE"). This search then resulted in title and subject matches on the ScienceDirect, JSTOR, SpringerLink, PROQUEST, IEEE, Scopus, DOAJ, and Google Scholar databases. Only peer-reviewed research articles in English have been compiled.

The selection of papers was later refined by reading all abstracts with the goal of removing case-specific discussions, as well as those in which regulation was not the main topic under debate. In addition, new papers on AI-related laws and government strategies that presented arguments on ethical principles have also been included in the sample.

This final corpus of literature has been classified according to specific parameters: year of publication, journal, author, author's institution, author's field of study, country, keywords. Summaries of each paper have also been developed to include: concepts, findings, contributions, agenda, approach, method, and researched subject. The following terms were considered when classifying the articles: "ethics/ethical principles", "how to regulate/existing regulation", "government strategies", and "framework or guidelines similar to a framework". After analysing the abstracts, a sample comprising 109 documents was selected for further reading and inclusion.

Results

In chronological terms, it is worth highlighting that 88.1% of the papers were published after 2015, with a growing production every year following that.

The sample reflects the evolution in the fields of research that take an interest in AI regulation, which is desired (Floridi et al., 2020). Although Artificial Intelligence as a subject of study traditionally pertains to Information Technology (Computer Science and Engineering), there has been a growing interest in its regulation by other areas, such as Law, Business Administration, and Philosophy. Out of the entire sample, researchers from the field of IT (combined or not with other fields) represent 47.7% and researchers exclusively from IT represent 27.5%, whereas researchers from other fields (excluding IT) represent 52.3%. In some cases, the same article is co-authored by researchers from different areas (22%).

Special heed has been paid to the analysis of the main object of the sample, non-exclusively divided into: "ethics and ethical principles" (45.9%), "how to regulate and existing regulation" (47.7%), "government strategies" (9.2%) and "framework or guidelines similar to a framework" (19.3%). It is worth noting that discussions on how to regulate only became significant in 2016. Concerning the discussions on AI regulatory frameworks, AI guidelines based on ethical principles with orientations similar to a framework have also been considered when they go beyond a description of ethical principles and actually provide orientations concerning how to apply those principles. From that perspective, 21 unique models have been found, which will be presented and analysed below.

Model for ethical issues in experimental technologies (Amigoni & Schiaffonati, 2018)

Based on the premise that a robot is an experimental technology, this model intends to minimise the ethical dilemmas associated with decisions made by autonomous systems (Poel, 2016). The proposal supports decision-making processes based on 16 conditions for deploying experimental technologies built to anticipate potential ethical issues as robots interact with people and the environment. Split into three groups, the conditions are aimed at: preventing damages (non-maleficence conditions: means for gaining knowledge of risks and benefits, monitoring of data and risks, possibility and willingness to adapt or terminate the experiment, risk mitigation, consciously scaling up, flexible setup, and avoidance of locking-in and undermining resilience); good-doing (beneficence conditions: expectation of social benefits, clear distribution of responsibilities); and respect for autonomy and justice (experimental subjects informed, approval by democratically legitimised bodies, possibility of experimental subjects influencing the project, possibility of withdrawing subjects from the experiment, special treatment given to vulnerable experimental subjects, fair distribution of potential hazards and benefits, reversibility of compensation of harm).

The model is an approach to regulation through a development process that would be part of a gradual interactive strategy set forth during the design stage. One can find among the outputs the epistemological role of exploratory experiments, while acquiring the knowledge of how robots behave in a real-world scenario. The authors highlight the prediction of "red button" conditions for situations in which the risk of harming people cannot be securely avoided during the experiment.

The 16 conditions proposed by Amigoni and Schiaffonati's "Ethical Framework for Robot Systems" seem to fit perfectly in standardised processes built by regulatory agencies as they test all the technologies submitted by the industry and service providers. The proposal can also be incorporated through risk analyses conducted by scholars for society as a whole.

Interactive regulatory governance model (Villaronga & Heldeweg, 2018)

Considering that regulatory actions cannot keep up with the speed of technology, and that top-down regulation approaches require mature laws, the authors have identified the need for a hybrid approach to start regulating AI technologies. They argue that bottom-up mechanisms can help develop the legislation and produce knowledge of AI development processes.

Focusing on a balance between regulation/legislation-in-progress and technology-in-progress, the proposal is based on an interactive governance model for technological development and law formulation processes in which the attributions of stakeholders are highlighted through process descriptions. The need for continuous learning and a gradual evolution of the legal framework is noteworthy, using such expressions as "Regulatory Innovation" and "Temporary Experimental Legislation", and considering the proper sequence of actions among agents at the maturity stage of an innovation's lifecycle.

The proposed model includes components such as:

- A Regulatory-to-Technology (R2T) macro-process to guide the creation of a new conceptual model for robots in accordance with the existing legislation, considering how it affects the way intelligent systems are built and used. It enables the creation of an AI technology impact assessment encompassing ethical, legal, and societal consequences. It focuses on legal opportunities or constraints that could have an impact on a new or existing robot. The result of the analysis considers a range of alternatives, from "abort development", "adjust plans", "go-ahead and lobby for legal change", or "take risks".
- A Technology-to-Regulatory (T2R) macro-process to adjust the law to the needs that result from the evolution of technology or the relationship between intelligent systems and society. It allows for the implementation of a regulatory impact assessment.
- A Governance Committee to rule on the reports related to the impact of both R2T (*ex-ante* robot) and T2R (*ex-post* robot) processes.

- A data repository shared by R2T and T2R in order to gather data about whether each AI technology (planned or in use) is in compliance with the law.

Among the main benefits of this hybrid AI Governance Model, it is worth highlighting the integration of top-down and bottom-up regulatory actions in an incremental strategy, thus minimising the risk posed when regulating a new, constantly changing object.

The proposed Interactive Regulatory Governance Model helps to raise awareness regarding the lack of a continuous resource to connect both worlds—technology and legislation—while being iteratively developed and improved. Since the legislative branch is in charge of the legislation (in most democratic countries), it can be associated with the R-side of processes. When looking for the most ideal entity to act as the T-side of processes, the tasks of a regulatory agency can be identified.

Connecting both sides, R2T and T2R processes would be a strategy to establish a closer relationship between the legislative branch and the regulatory agency.

Ethics model for AI development and deployment (Schrader & Ghosh, 2018)

Founded upon philosophical principles and the dimensions associated with safeguarding human rights and well-being, the proposed ethical framework for AI development and deployment has been designed to implement core functions to represent ethical activities and the outcomes from both the philosophical and ethical perspectives.

The ethical perspectives are split into six categories: Rights (deontological ethics); Damages and Goods (teleological ethics); Virtue (aretaic ethics); Community (community ethics); Dialogue (communication ethics); and Flourishing (flourishing ethics).

The recommended core functions to be considered when developing AI systems are:

- Identifying ethical issues of AI—fairness, transparency, equity, goodness, beneficence, social utility, happiness, and protection of humans.
- Raising human awareness of AI—a clear understanding of how AI systems work within each product and how the industry develops algorithms.
- Collaborating with AI—dialogical interaction, listening, and understanding between humans and AI.
- Accountability of AI—guaranteeing the ethical compliance of AI systems and their designers.
- Integrity of AI—maintaining the AI system limited to the purpose for which the technology was intended.

A matrix combining the five core functions with the six perspectives has been built as a guideline to be followed during the AI project. As a proactive action in the design, development, and use of products and services that utilise AI, the model seeks to reflect the nature of social changes demanded by a new ethical thought.

Although they do not associate the framework with any specific organisation or institution, the authors' contribution can be applied by a regulatory agency when auditing the industry, as well as in its internal processes, to better understand the impacts technology has on the stakeholders.

Competency-based AI regulation model (Scherer, 2016)

Considering the competencies, strengths, and weaknesses of each state power, the proposal of an AI Regulatory Model (AIDA—Artificial Intelligence Development Act) is based on the distribution of responsibilities without losing sight of the mission goals. The model acknowledges the regulatory role of the executive, legislative, and judicial powers as agents in the regulatory process.

In the proposed model, the legislative branch would provide a statute placing a regulatory agency in charge of certifying AI products and services with regard to user and social safety. In general, legislators have limited knowledge of AI systems, their only support being a few committee meetings with experts. In order to solve this problem, legislators would delegate the responsibility for policy-making to the regulatory agency.

Supported by groups of researchers, the regulatory agency would comprise two main areas: policy-making and certification. Such an agency would be expected to be more agile and competent to monitor the evolution of technology, identify risks in the intelligent learning process and use of AI, issue technical recommendations, and verify that the technology is being applied for its intended purposes. A certificate would be given to designers, manufacturers, and service providers after being approved through the agency's processes. Pre-certification rules would also be made public to the industry and service providers. In case of an accident with certified products, the agency would publish a report to society, explaining the circumstances behind its occurrence and which certification rules/processes would therefore be modified.

Due to their *ex-post* nature, courts would judge cases considering whether or not a certification exists. Courts would judge companies for any losses and damages caused, considering the situation in which those organisations find themselves when it comes to certification. If a company's products or services cause any damages, if certified, the company would be judged based on more lenient rules, whereas uncertified companies would be subjected to more rigid norms.

The proposed model takes into consideration the natural attributions of each entity within the government. Agility is required for the actions performed by the regulatory agencies, which would give them a prominent role in the regulation process. This is key to enable the evolution of technology while the legislation takes its time to mature.

Regulation model sustained by society (Rahwan, 2017)

Inspired by the Social Contract Theory (Rousseau, 2016), the Regulatory Model Sustained by Society adjusts the "human-in-the-loop (HITL)" to the "society-in-the-loop (SITL)" model.

The use of HITL thinking in AI has been largely applied to help an algorithm learn from humans' contributions. The agility and effectiveness of a HITL interactive learning machine stem from user feedback, thus enriching the knowledge that gets generated.

From a regulation perspective, the author argues that it is not sufficient to only adjust HITL to use a human to monitor an AI system and correct it in case of misbehaviour. By doing so, the regulation would rely on the judgment of an individual or group of individuals that subject the whole process to a narrow analysis. If we want to deal with a system that has an impact on the values of an entire society, that society must be included in the analysis, giving it a broader approach. It would not only avoid biased judgments, but also balance the competing interests of different stakeholders.

It is suggested that SITL be used in a process characterised by human-based government and citizen channels. On one side, the government's AI products and services would be run and, on the other side, citizens would evaluate those smart systems based on their own values. This would allow the government to understand how social behaviour and values change. Therefore, society-in-the-loop would become a governance tool for society to control and proactively identify those elements. Conflicts among safety-, privacy-, and justice-related concepts would benefit from this model. This relationship can be summed up as: society-in-the-loop = human-in-the-loop + social contract. The model also recommends auditing mechanisms to tackle the possibility of fake data manipulated by social groups at the learning stage, as well as results that would affect regulations.

For the purpose of using the proposed model as part of a broader AI governance model, both society and academia can be considered in terms of society's role when answering an agency's inquiry regarding the ethical behaviour of AI systems.

Principles of robotics (Boden et al., 2017)

After pinpointing the responsibilities of all agents involved in robotics, five principles were established in a guideline for robot designers, manufacturers, and users. The main goal of the rules is to emphasise that robots are tools, whereas humans are the actual responsible agents. The proposed rules are:

a. Robots should not be designed as weapons, except in the interests of national security.
b. Robots should be designed and operated to comply with existing laws, including those dealing with privacy.
c. Robots should be designed to be safe and secure.
d. Robots should not be used to exploit vulnerable users by pretending to feel emotions.
e. It should be possible to find out who is responsible for any particular robot.

Aiming to encourage responsibility within the robot-related research and the industrial community, seven messages have been created to highlight the responsible innovation spirit needed to abide by the rules.

The opportunity to use this proposal in audits performed by regulatory agencies can be identified, and that need must be reflected in the legislation to be adapted or created.

Agile AI governance (Wallach & Marchant, 2018)

Aware of the concerns regarding AI impacts exceeding the regulatory scope, capabilities, and jurisdiction of an agency or nation, the authors propose a model to address this governance challenge.

The model predicts actions performed by a Governance Coordinating Committee at the national level and a Global Governance Coordinating Committee. The main goal is a soft-law strategy that mitigates risks while the legislation is being drawn up. The soft governance part involves industry standards, social codes, labs, certification practices, procedures, and programmes. The hard governance part concentrates on laws, regulations, and regulatory groups.

A national committee would coordinate the efforts of a governance process encompassing stakeholders to produce recommendations, reports, and roadmaps, while monitoring those actions at the same time. This national forum would also be a perfect structure to enforce soft governance mechanisms as a necessary complement to the hard ones.

On the international level, a global committee would not only coordinate agreements among countries, but also establish a common understanding of which international standards should be used as a soft governance strategy. The international approach is also advocated to bring some balance to the several countries that are not yet participating in the AI regulation dynamics, considering that the current situation makes them more vulnerable.

The proposed model takes a relationship network into account to address AI in a way that bolsters the formulation of actual standards while the legislation matures. The agile meaning of this governance is its incremental approach, which allows for continuous inputs. This would be an alternative to the problem posed by the temporal mismatch between formal regulatory actions and the production and commercialisation of deep machine learning-based products and services around the world. The success of this proposal depends on the amount of effort put into it by the market, academia, government, insurance companies, and organised civil society.

Sustainable AI development (Djeffal, 2018)

Considering the closer connection between sustainable development and governance, the author highlights that governance mechanisms are built to be continuously improved. The proposal concerns the entire lifecycle of an AI-based solution as the main foundation of a Sustainable AI Development (SAID) framework.

Analysed under the lens of a governance structure, SAID is stratified into the following layers: Technological, Social, and Governance.

At the base, the technology layer is in charge of specific applications involving architecture, data, and algorithm design.

Focusing on the impacts systems have on society, the social layer deals with the process of inserting technology into real life. It encompasses an analysis of the potential consequences of using AI in the social sphere.

Highlighting the importance of a broad treatment, the governance layer looks at the way algorithms influence both national and international decisions.

SAID gathers the different approaches examined in the various frameworks and somehow materialises the perception that, in order to be effective, AI regulation demands actions by IT and Social Sciences (Law, Business Administration, Philosophy, and Psychology) professionals alike. It also reminds us that, due to the topic's complexity, an AI governance model must include different process tiers.

Ethical framework for automation using robotics (Wright & Schultz, 2018)

Concerned with the integration between several stakeholders and automation using AI, this framework integrates the Stakeholders Theory with the Social Contract Theory in an attempt to find ethical grounds for developing, providing, and utilising AI.

The proposal considers as stakeholders: workers, the market, governments, the economy, and society in general. The impacts on the job market, from an ethical perspective, and the relationships among those stakeholders are highly emphasised.

The framework is based on a set of steps ranging from the identification of stakeholders, analysis of the social contracts among them, an assessment of how stakeholders are impacted, and lastly, actions aimed at mitigating the risk of terminating or breaching work contracts. An important target to be reached is increasing the benefits for stakeholders.

It is worth noting that this proposal considers as stakeholders those workers whose jobs or occupations will be modified with the introduction of AI into products and services. Due to the complexity of interests among stakeholders and all the labour concerns, the framework fits in the government policy-making process. The impact of such public policies on the country's economy may result in the need for laws, which means the legislative branch must be included as a stakeholder.

Intelligent model to regulate learning algorithms (Buiten, 2019)

Focused on a strategy to fight intelligent services that contain biases, this model postulates that an algorithm should assess the essential elements of a machine learning process (data, testing algorithms, and decision models). The proposal is founded on the thesis that the transparency of a code is insufficient to guarantee an unbiased solution and admits that it is still possible to find biases, even when learning from vast amounts of data.

In the data domain, all data samples are assumed to include some built-in biases that need to be considered. The data must be checked to ensure their validity, reliability, and proper data dependency.

Regarding the testing algorithms, the model recommends using a variety of algorithms and comparing their performance. However, that must be done only after discovering the quality of the available data.

The decision-making process is seen as a delicate phase in which developers must be aware of the correlations between variables, because hidden relationships may obscure a biased orientation. It also acknowledges the difficulty of identifying those problems automatically as algorithms grow in complexity.

Universal declaration of human rights as a framework (Donahoe & Metzger, 2019)

This model is founded on the argument that the several different frameworks related to each specific area of ethics are insufficient to regulate AI on an international scale, both in the private sector and within the government. Due to that gap, the Universal Declaration of Human Rights (Kunz, 1949) has been considered a mature approach that different cultures have been adopting for decades. Modern adjustments were made by the UN Human Rights Council in 2011, published as the UN Guiding Principles on Business and Human Rights (United Nations, 2011), which highlight the roles and responsibilities of private-sector businesses in the protection of human rights.

Under the human rights framework, governments have the duty to protect citizens from violations and infringements of their rights by other governments and non-State actors, including the private sector. Donahoe and Metzger's proposal deals with the centrality of the human person as the focal point of governance and society. It seeks to address the potential impacts of AI, such as:

- The right to equal protection and non-discrimination—avoiding biases in the data and ensuring fairness in machine-based decisions.
- The right to life and personal security—concerning autonomous weapons that move beyond human control.
- The right to an effective remedy for violations and infringements of rights—transparency, fairness, and accountability in cases where AI systems impact people's rights.
- The right to privacy—addressing the loss of privacy in data-driven societies and the need to protect personally identifiable data.
- The rights to work and to enjoy an adequate standard of living—guiding governance decisions around the displacement of human workers by AI.

Software requirement model for the ethical assessment of robots (Millar, 2016)

Considering ethics as a social enterprise, the proposal puts forth a set of general specifications to be considered in a system aimed at assessing robots during their construction. To that effect, five major rules have been built:

- Balancing designer and user requirements, considering the potential damages.
- Utilising a user-centred ethical evaluation tool for AI systems, which must use design methodologies that are able to identify the impacts on human values in use contexts.
- Including the psychology of user-robot relationship variables in the ethical evaluation tool to identify variables such as the user's emotional state.
- Compliance with the Human-Robotics Interaction Code of Ethics (Riek & Howard, 2014).
- Designers' understanding of both acceptable and unacceptable design features, which could be implemented by including ethicists in design teams.

It seems the proposal may be utilised by the industry and regulatory agencies alike. In both cases, it could be the first red flag signalling the need for a red button in robot projects (Arnold & Scheutz, 2018).

Ethical judgement model for codes (Bonnemains et al., 2018)

Considering that (a) an ethical framework allows us to deal with situations involving ethical dilemmas, (b) one framework alone is not efficient enough to compute an ethical decision, and (c) tackling ethical decisions is better than avoiding them, the author proposes a formal logical model that can be implemented by an agent facing an ethical dilemma, with the ability to both make decisions and explain those decisions. It assumes that formal expression analyses are especially useful to identify the subjectivity of a decision.

Different judgements on possible decisions have been studied according to three ethical frameworks: consequentialist ethics, deontological ethics, and the Doctrine of Double Effect. In the path toward a refined and final framework, various ethical dilemmas have been formalised in judgment functions that return three possible results: acceptable ($\top$), unacceptable ($\perp$), or undetermined (?). The concepts of 'decision', 'event', and 'effect' were taken into account when building the model's functionalities.

Those analyses can be appreciated when we judge someone or something based on particular moral theories.

Asilomar AI principles (Future of Life Institute, 2019b)

The governance model proposed by the Asilomar Conference resulted in 23 AI Principles undersigned by thousands of experts (Kozuka, 2019). Grouped under "Research Issues", "Ethics and Values", and "Longer-Term Issues", those principles encompass the lifecycle of an AI-embedded product or service—from motivation and funding to the assessment of benefits and judgement criteria concerning its impacts.

In the Research Issues dimension, the recommendations are to: research goals and funding, establish a connection between researchers and policymakers, research the culture of cooperation, and promote synergy to avoid corner-cutting when devising safety standards.

In the Ethics and Values dimensions, the orientations are to: maintain AI systems secure during their entire lifecycle, make them transparent in case of failure as well as in judgment results, consider designers and builders of advanced AI systems as stakeholders in the responsibility chain, align AI systems' values with their users', design AI systems to be compatible with human rights and cultural diversity, preserve personal privacy, share their benefits as much as possible, and make it possible for a human to take control of AI systems, if so desired.

Finally, in the Longer-Term Issues sphere, the principles are to: be cautious when making decisions without a consensus, build a mitigation plan to deal with the risks, plan a recursive type of self-improvement, and develop AI systems based only on widely shared ethical ideals.

European ethics guidelines for trustworthy AI (AI HLEG 2019b)

With the goal of creating guidelines to orient a new AI governance, a team of experts entitled High-Level Expert Group on Artificial Intelligence has drawn up the Ethics Guidelines for a Trustworthy AI for the European Commission based on a structure supported by values that should be considered throughout the system's lifecycle: lawful, ethical, and robust AI.

Based on the European Union Charter of Fundamental Rights (EU Parliament, 2012), the model establishes trustworthy AI as a key element for a governance framework (Kozuka, 2019) has been built using a three-tier structure.

The highest tier addresses ethical principles based on fundamental human rights: respect for human autonomy, prevention of damages, fairness, and explicability. To ensure fairness in a society with different interests and objectives, it defends an explicable decision-making process. It should consider traceability, auditability, and transparent communications regarding system capabilities. It also recommends that particular attention be paid to vulnerable groups and situations characterised by asymmetries of power or information (employers and workers, or businesses and consumers).

The second tier includes the key requirements necessary for implementing an AI-based system or service throughout its lifecycle: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; social and environmental wellbeing; and accountability. All requirements are connected to one another through a full-mesh relationship where each one of them has the same weight.

Special attention is suggested to the oversight as part of a governance mechanism that could use human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC). A strong connection can also be found between "privacy and data governance" and "diversity, non-discrimination, and fairness", due to the need for mechanisms to avoid inadvertent historical biases, incompleteness, and inadequate data governance models. Regarding the accountability concerns, a recommendation is given to carry out an impact assessment prior to and during the development.

Defending a trustworthy AI implementation throughout the lifecycle of an AI system, the model demands a process-oriented approach that encompasses both technical and non-technical methods when implementing the requirements. Within the non-technical approach, one can find legislation and corporate guidelines encompassing codes of conduct, policies, performance indicators, and agreed-upon standards. Those standards consider AI users, consumers, organisations, research institutions, and governments as stakeholders. They also include a certification granted to organisations that produce transparent, accountable, and fair AI systems in accordance with the established standards. The entity in charge of the certification could play an important role in the communications with "industry and/or public oversight groups, sharing best practices, discussing dilemmas, or reporting emerging issues of ethical concerns."

For the base tier, a list of recommendations directed at the operationalisation of the key requirements in the upper tier for each specific system has been formulated.

Ethically aligned design (IEEE 2019)

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems has proposed five general principles for AI systems and a guideline recommending actions to establish ethical and social implementations for intelligent and autonomous systems that prioritise human wellbeing.

According to that model, the ethical design, development, and implementation of AI systems should consider the following principles: human rights, wellbeing, accountability, transparency, and awareness of misuse.

Focusing on personal data rights, the wellbeing promoted by the effects on the economy, the legal frameworks for accountability and transparency, and the education and awareness policies, recommendations were made to a wide set of stakeholders.

To governments: the governance framework should include standards and regulatory agencies, provide society with ethics education and security awareness regarding the potential risks, improve digital literacy, use multiple metrics as wellbeing indicators, and implement a wellbeing impact assessment.

To industries: programmatic levels of accountability should be provided to address culpability in legal matters, transparency by design, intelligibility of a system's operation and decisions, damage mitigation strategies, assessment starting at the design phase, understanding how each jurisdiction would treat the damage caused by a given AI system.

To the legislative sphere: responsibility, culpability, liability, and accountability issues should be classified.

General recommendations: certification of AI systems, identification and prioritisation of standards for each category of AI systems, continuously updating the standards, metrics utilised to assess AI systems, agreements on moral decisions, evaluation by third parties, applying the classical methodologies of deontological and teleological ethics to machine learning, adherence to the code of conduct by the AI production team, and bridging the language gap between technologists, philosophers, and policymakers.

At the international level: establishing a global multistakeholder dialogue to determine the best practices, facilitating AI research and development in developing nations, and using indicators to assess AI-related technological interventions in those countries.

Government and industries: identifying the types of decisions and operations that should never be delegated to AI systems.

A special guideline for implementing an ethical culture in organisations (IEEE, 2020) has also been built, encompassing a strategy to assess the level of each dimension to be developed (lagging, basic, advanced, and leading).

Avoiding biases and discrimination (Lin et al., 2020)

In order to amplify the effectiveness of bias-reduction intervention procedures in cases of implicit biases, the framework explores an innovative AI-assisted intervention based on a bidimensional approach.

In the first dimension, the different types of information AI provide to users are captured: the current state of affairs (descriptive information), the likelihood of future states (predictive information), and the expected utility of an action (prescriptive information). It considers that all interventions are prescriptive, and the knowledge-based systems (KBS) will decide to intervene depending on how they simulate the results.

In the second dimension, an AI system can intervene in different phases of the decision-making process (input-based interventions, output-based interventions, and cognition-based interventions) as part of an interactive process.

It is a case of regulation by software, which could be used by the industry and service providers as part of their internal process.

Standardisation exchange model (Lewis et al., 2020)

Considering the importance of standardisation in a regulation strategy, the model proposes a process among functional entities in the AI value chain through which information related to standards is exchanged among them.

Classified by their functional roles, the actors—data providers, AI system creator, AI system operator, AI user, oversight authority, and associate stakeholder—change standards focusing on a trustworthy AI.

The benefits of each exchange are presented, as well as the potential topics for new standardisations. Most of them concern issues to be considered in an AI product certification process.

Although the focus is on the industry, the model considers the importance of the government in the whole process and the need for an international community to discuss the standards.

Algorithmic impact assessment (Canadian Government 2020)

Aiming to help public and private-sector companies assess and mitigate the impacts of deploying an automated decision-making system, the Canadian Government has developed the Algorithmic Impact Assessment (AIA) based on the Government Directive on Automated Decision-Making. The AIA questionnaire considers the reasons for using AI on decision-making processes, the capabilities encompassed by the system, algorithm transparency and explainability, system category (health, social assistance, economic, etc.), development and training process, system and data architecture, stakeholders, and risk mitigation measures.

The impact assessment addresses the four levels according to how the decisions impact the rights, health, or well-being, the economic interests of individuals or communities, and the ongoing sustainability of an ecosystem. Thus, levels I, II, III, and IV are each related to a certain impact, namely, reversible brief, reversible in the short term, difficult to reverse, and irreversible.

The Directive on Automated Decision-Making was designed by the Canadian Government to make its administrative decisions compatible with core administrative law principles, such as transparency, accountability, legality, and procedural fairness.

The requirements considered by the Directive on Automated Decision-Making are distributed between two pillars: transparency and quality assurance. Among the transparency requirements, it establishes that:

- Notice on relevant websites must be issued before decisions are made,
- Meaningful explanations must be provided to affected individuals regarding the decisions made,
- The Government of Canada has the right to access all components of the system.

Among the quality assurance requirements, there are rules to ensure testing and monitoring outcomes, data quality, peer review, employee training, contingency, security, compliance with the law, and human intervention.

AI governance by human rights-centred design, deliberation and oversight (Yeung et al., 2019)

Considering international human rights-based standards as the most promising governance framework to deal with ethical standards, Yeung et al. (2019) have proposed the Human Rights-Centred Design, Deliberation, and Oversight model to deal with AI-related ethical issues with legal support. Based on a global approach, the proposed model integrates a suit of technical, organisational, and evaluation tools and techniques involving many stakeholders.

The proposal presents norms based on human rights as the foundation for ethical standards with which AI systems must demonstrably comply:

a. Design and development that take stakeholders' opinions into account. In case an assessment has resulted in "high" or "very high" risks to human rights, a redesign should be pursued.
b. Formal assessment and testing to evaluate their compliance with human rights-based standards. It would occur regularly during the entire life cycle of a system's development—design, specification, prototyping, development, and implementation. A systematic and periodic post-implementation monitoring would be established, through which the AI system would be submitted for review by sending out the related documentation and reports to a public authority.
c. Independent oversight by an external, technically competent entity invested with legal investigation and sanction powers.
d. Auditability supported by traceability and by evidence that the AI system is operating as desired and that it was properly documented during its entire life cycle of development.

The authors highlight the need for laws and norms encompassing all steps covered by the model.

Good AI society (AI4People, 2018)

Focused on the establishment of a good AI society, the proposal joins ethical principles and specific recommendations to enable stakeholders to seize opportunities and avoid or minimise risks.

The model encompasses five ethical principles: Beneficence, Non-maleficence, Autonomy, Justice, and Explicability.

The recommendations are categorised as: assessment, development, incentivisation, and support.

- Assessing institutions on their capacity to reduce the mistakes made by AI systems.
- Considering existing legislation, using participatory mechanisms to align with social values, and assessing tasks/decision-making that should not be delegated to AI systems.
- Assessing current regulations to provide a legislative framework that could keep pace with technological developments.
- Developing a framework to enhance the explicability of AI systems.
- Developing legal procedures to permit the scrutiny of algorithmic decisions in court.
- Developing auditing mechanisms for AI systems to identify unwanted consequences.
- Developing a process to remedy or compensate for damage caused by AI.
- Developing agreed-upon metrics for the trustworthiness of AI products and services.
- Developing a new EU oversight agency responsible for the scientific evaluation and supervision of AI products and services.
- Developing a European observatory for AI.
- Developing legal instruments to prepare and adjust the work environment to the changes brought about by AI.
- Financially incentivising a socially preferable development and use of AI.
- Financially incentivising cross-disciplinary cooperation in the fields of technology, social issues, legal studies, and ethics.
- Incentivising a regular review of the legislation to foster socially positive innovation.
- Financially incentivising the use of lawfully special zones for empirical testing and development.
- Financially incentivising research on the public perception of AI.
- Supporting self-regulatory codes of conduct for data- and AI-related professionals.
- Supporting corporate boards of directors to take responsibility for the ethical implications of AI technologies in their organisations.

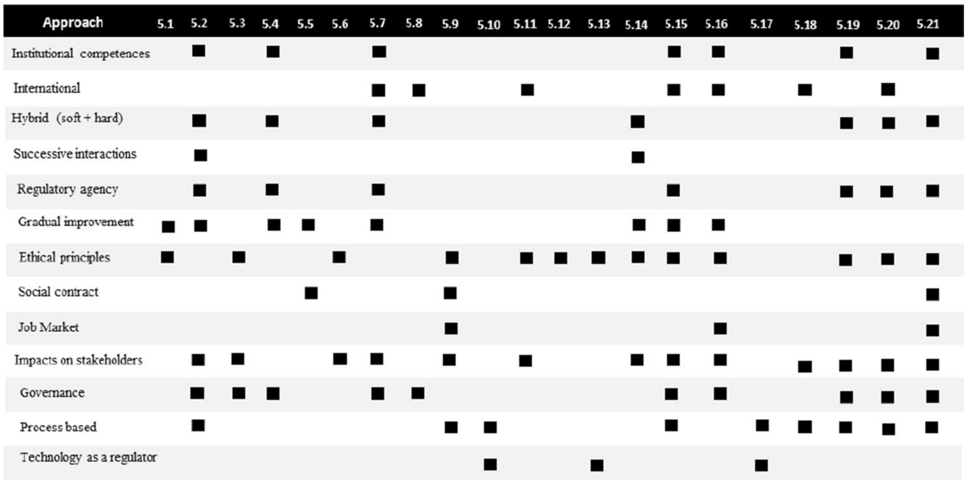
Framework approaches

An analysis of the approaches adopted by each of the 21 frameworks proposed in the sample resulted in Table 1.

The fact that ethical guidelines exist is not enough to have any effect on the software development industry. Thus, models that are strongly grounded on ethical principles require legal mechanisms to fulfill those recommendations (Hagendorff, 2019).

Frameworks that encompass the competencies of government institutions have also foreseen the existence of a regulatory agency, as well as the need for mechanisms to

Table 1 Comparative table of the approaches explored in the frameworks, compiled by author.



Approach	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	5.10	5.11	5.12	5.13	5.14	5.15	5.16	5.17	5.18	5.19	5.20	5.21
Institutional competences		■		■			■								■	■			■		■
International							■	■			■				■	■		■		■	
Hybrid (soft + hard)		■		■			■							■					■	■	■
Successive interactions		■												■							
Regulatory agency		■		■			■								■				■	■	■
Gradual improvement	■	■		■	■		■							■	■	■					
Ethical principles	■		■			■			■		■	■	■	■	■	■			■	■	■
Social contract					■				■												■
Job Market									■							■					■
Impacts on stakeholders		■	■			■	■		■		■			■	■	■		■	■	■	■
Governance		■	■	■			■	■							■	■			■	■	■
Process based		■							■	■					■		■	■	■	■	■
Technology as a regulator										■			■				■				

help the legislative branch speed up its law-making process, aiming for a safer and faster AI regulation.

Frameworks that take the social contract into account rank among the most open to society's participation in a co-production with the government. Those models consider citizens as outstanding stakeholders. Concerns over the impacts on the job market are also a way to assess the impact on stakeholders.

The main argument that proposes a gradual deployment of the regulation is a risk mitigation strategy, but it could also be combined with successive interactions between the legislative branch and the regulatory agency, thus enabling continuous improvement during the legislative procedure.

The interactive regulatory governance model, the agile governance, the ethics guideline for trustworthy AI, the ethically aligned design, the algorithmic impact assessment, the good AI society, and the AI governance by human rights-centred design, deliberation, and oversight proposals encompass a larger number of topics. The AI HLEG proposal highlights that a trustworthy AI must be lawful, ethical, and robust. The others explore the relationship among all parties involved in the regulation process and the attempt to find balance between more or less rigid or flexible mechanisms. It is worth noting that the agile governance proposal does not exclude conventional actions for a formal regulation—the interactive regulatory governance model and the competency-based regulatory model, both of which involve the legislative branch. Therefore, this configures a transitional situation in which consensual standards would be agreed upon and enforced, and the risks would be mitigated until legal mechanisms are made official, which is very similar to the concept of Dynamic Regulation, in which feedback serves as a basis for the maturity of the regulatory instrument (Kaal & Vermeulen, 2017).

When analysing several movements advocating the establishment of criteria for best using AI, studies identified an opportunity to develop a competition around a technological reform (Greene et al., 2019). Pondering over the need to find synergy among global AI regulation-oriented actions, a few proposals rely on a worldwide effort, which sometimes is described as an international committee, while other times just as a joint effort by governments and multinational companies.

Despite the small number of existing software-based regulation models, similar models are likely to arise, since the increasing complexity of AI solutions results in more system rules (Lamo & Calo, 2018; Liu, 2017; Prakken, 2017; Verheij, 2016), which in turn means a higher likelihood of conflicts among those rules in combined systems (Bench-Capon & Modgil, 2017).

AI regulatory and governance framework

The supplementary nature of some of the models confirms the perception that the impacts of AI would demand a combination of design, laws, and education (Calo, 2011). When debating over the complexity of a framework to address such a multidisciplinary topic (Bonnemais et al., 2018) embedded into the political and social context (Leitner & Stiefmueller, 2019), an AI regulatory and governance framework—AIR—was built to include the main contributions from each model in the examined sample (Fig. 1).

Focused on reducing the gap between ethical principles and actions by each stakeholder and making the relationship among them in different dimensions of knowledge clearer, the AIR framework is based on a wide governance process.

Although the government's exclusive competencies are highlighted, aiming for more accuracy in its actions, the power of the State has been distributed among the legislative, the executive, and the judicial branches. This

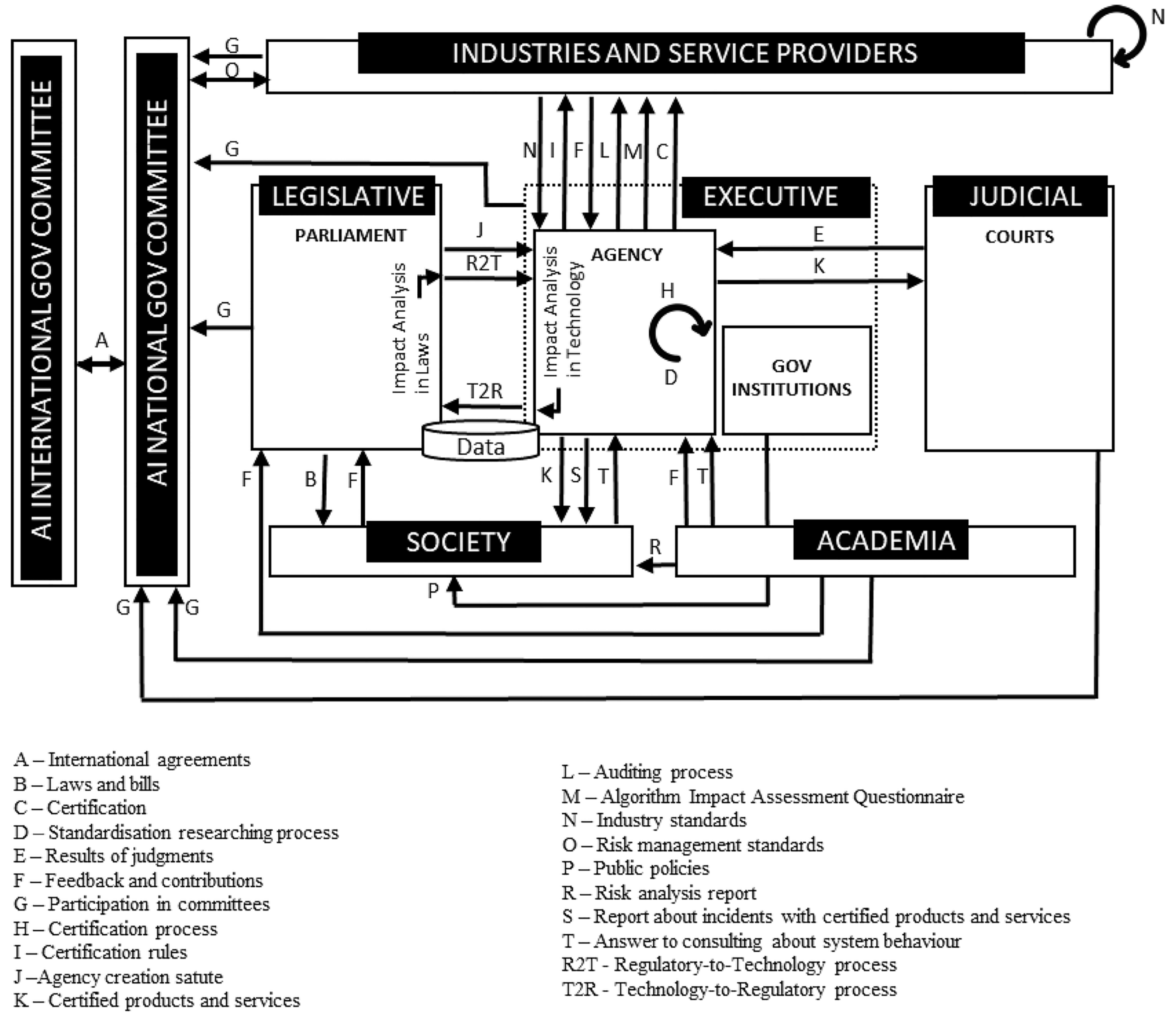

INDUSTRIES AND SERVICE PROVIDERS
N
AI INTERNATIONAL GOV COMMITTEE
AI NATIONAL GOV COMMITTEE
G
O
G
G
A
N I F L M C
LEGISLATIVE
PARLIAMENT
Impact Analysis in Laws
J
R2T
T2R
Data
EXECUTIVE
AGENCY
Impact Analysis in Technology
H
D
GOV INSTITUTIONS
JUDICIAL
COURTS
E
K
F B F
K S T
F T
SOCIETY
R
ACADEMIA
P
G G
A – International agreements
B – Laws and bills
C – Certification
D – Standardisation researching process
E – Results of judgments
F – Feedback and contributions
G – Participation in committees
H – Certification process
I – Certification rules
J –Agency creation satute
K – Certified products and services
L – Auditing process
M – Algorithm Impact Assessment Questionnaire
N – Industry standards
O – Risk management standards
P – Public policies
R – Risk analysis report
S – Report about incidents with certified products and services
T – Answer to consulting about system behaviour
R2T - Regulatory-to-Technology process
T2R - Technology-to-Regulatory process

Fig. 1 AIR framework

segmentation is used by many countries as a functional way to distribute power, according to which the legislative creates the laws, the executive enforces those laws, and the judicial is in charge of solving whatever conflicts arise to guarantee justice and law abidance (Maluf, 1995).

Apart from making laws, it is crucial to maintain the legislative branch open so that its bills (B) can be discussed with society, receiving constant feedback and contributions not only through e-participation systems, but also through a special channel established with scholars, who could also attend the legislative committee meetings (F).

The Parliament or Congress, as an instance of the legislative branch, would approve a statute (J) to create an AI regulatory agency as part of the Federal Government (executive branch). This could be a good moment to define AI, or at least to demand that the agency do it.

Upon its creation, the regulatory agency would establish a strong relationship with the Parliament as part of an ongoing process in which the legislative would survey the impact on the legislation and its evolution based on the knowledge obtained from the regulatory agency (T2R—Technology-To-Regulatory), much like the regulatory agency structures its internal work processes based on the legislation discussed and approved by the legislative (R2T—Regulatory-To-Technology).

The T2R is necessary, at least until each new category of AI systems has been deeply studied by the regulatory agency. Due to the complexity and specificity of AI services and products, laws could potentially be created for each specific field. The natural evolution of the former would also cause the latter to evolve in the long term. A practical way to implement the T2R flow is through the regulatory agency frequently attending the legislative committee meetings to discuss AI regulation.

As a complement to T2R, the R2T flow would be started at least when a new version of a bill is discussed at the legislative committee meetings and when a new law is approved. R2T also feeds other internal processes of the regulatory agency in order to update them with the legislative understanding of what can be regulated by law, which can trigger three reactions: (a) alerts regarding the limitations that the bill/law brings to the ongoing projects of the industries and service providers; (b) opportunities to expand the standards by discussing them with the industries and service providers; and (c) updating certification and auditing processes with new compliance issues.

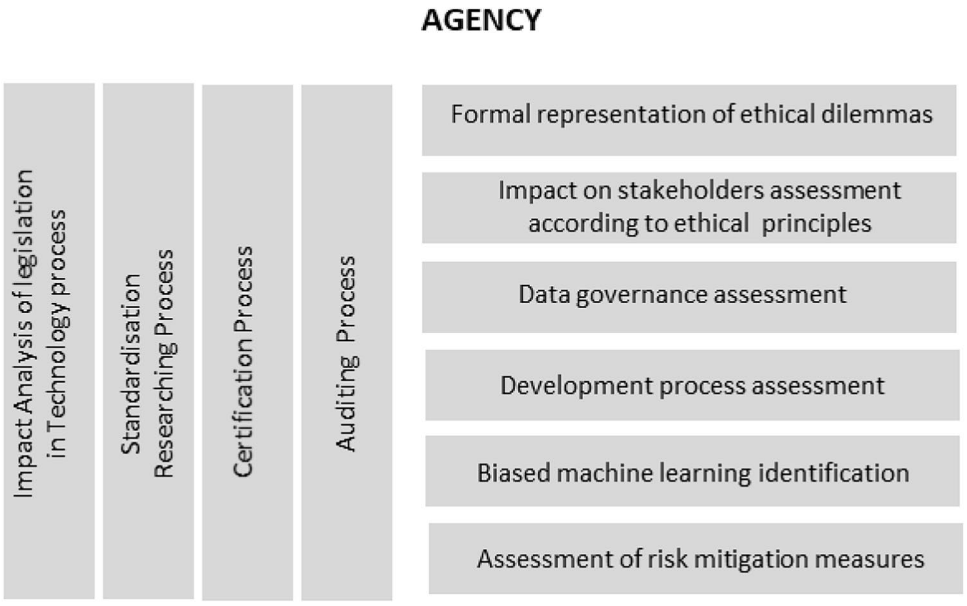
Among the regulatory agency's competencies, a couple of processes require speed and synergy: analysing how the legislation affects the technology process and standardising the research, certification, and auditing processes. In order to be effective, those processes must consider a huge number of variables involved in the entire lifecycle of an AI system: design, prototyping, development, testing, deployment, commercialisation, and use. The efficiency and knowledge of the regulatory agency are expected to possess depend on mechanisms that support those processes (Fig. 2): formal representation models for ethical dilemmas, impact on stakeholders' assessment according to ethical principles, data governance assessment, development process assessment, systems to identify biased machine learning, and assessment of risk mitigation measures.

Being responsible for a closer interaction between the Parliament and the regulatory agency due to the R2T flow, the analyses of how the legislation affects the technology process must be corroborated with information structures that are able to represent the law based on a technical mindset. Those involved must be skilled in both areas of knowledge. The quality and efficiency of this synchronicity of mixed mindsets are strengthened by means of a data repository shared by the Parliament and the regulatory agency. Examples of such data would include: issues related to whether AI projects/products comply with the law, AI project/product impact assessment, legislation/regulation collected overtime across AI projects/products, ethical committee decisions upon approval requests, AI project/product and regulatory assessments under both private and public guidelines.

In order to offer controlled autonomy to the AI industry, technical standards must be established while bills are being discussed. An agile interaction between the "industries and service providers" and the regulatory agency is supported by the standardisation of the research process (D). Despite being a process inside the regulatory agency, standardising research actions entails an in-depth study of ethical and safe mechanisms to make the new projects seen in the AI market feasible. A very technically skilled staff must be allocated to that task, which requires robust laboratories.

As a strategy to motivate the AI market to follow the best practices, standards, and laws (when they exist), the regulatory agency would certify products and services using a certification process (H). Companies that submit their products to the regulatory agency, after a successful appraisal, would receive a certificate (C) within their field of action (transport, healthcare, entertainment, education, military,

Fig. 2 Regulatory agency in the AIR framework


AGENCY
Impact Analysis of legislation in Technology process
Standardisation Researching Process
Certification Process
Auditing Process
Formal representation of ethical dilemmas
Impact on stakeholders assessment according to ethical principles
Data governance assessment
Development process assessment
Biased machine learning identification
Assessment of risk mitigation measures

etc.). The strictness and nature of the assessment process could be different for each of those fields. It is also a way to communicate to society where people can place their trust when buying or using an AI system.

Through a quick process, the industries and service providers would need to receive the regulatory agency's certification rules stated as clearly as possible (I), while providing feedback (F) on the conditions that preclude the development process required by the regulatory agency from moving forward. Accountability requirements would be assured if those lists would show not only new products and services that have been certified, but also those that have lost their certification.

The issuance of certificates could be a strategy to be applied before laws are passed, since they already inform society, in a transparent fashion, about the safety levels and risks of the products and services it consumes. Advertising campaigns by the government and certified companies would also strengthen that strategy.

A robust strategy to avoid fake certifications would be desirable, such as a blockchain mechanism implemented by the agency containing the updated certification list for a given country. Aiming to increase citizen trust in AI certification, certificates could be issued using a digital signature in the system's code, setting an attribute associated with that specific code version. In case someone wants to know if the version of a commercialised AI system is updated, all they need to do is compare the digital signature with the one that is available on the regulatory agency's website.

The regulatory agency could make an "algorithm impact assessment questionnaire" (M) available to the industries and government institutions in order to offer a simulation tool through which they could know, in advance, their level of compliance. It would also fit as a preparation stage for a certification submission.

And finally, an auditing process (L) would be supported by the regulatory agency to check companies demanding certification and certified companies that need to update their certification, as well as to verify issues demanded by courts in case of sentences related to damages supposedly caused by AI systems. This audit would take place in five dimensions: impact on stakeholders based on ethical principles, data governance, development process models, identification of biased machine learning, and risk mitigation measures. The auditing process should be part of regular monitoring through which not only internal changes in companies, but also future problems coming from new arrangements in society could be identified.

Any failures or damages noticed in a certified AI product or service must trigger an internal audit process to identify whether there were problems or limitations in other agency processes that could be a reminder of internal improvement. In a broader, more transparent fashion, the agency should publish the audit results and the next steps (S).

The regulatory agency's processes are interconnected through a knowledge stemming from the mechanisms shown in Fig. 2, which should be handled as much as possible by a skilled multidisciplinary team, since the ethical and technological dimensions are mixed.

Mechanisms for formally representing ethical dilemmas are important to create a transparent communication channel between ethicists and technical profiles. It could also help distinguish between the part of the decision-making algorithm that is related to a dilemma and the rest of the code in relation to which there is a consensus regarding the best decision. This representation model is expected to be continually improving as society changes and new dilemmas are identified. This analysis is interconnected with the impact on the stakeholders' assessment according to ethical principles.

As each company has its own system development process, the regulatory agency must have a process to guarantee a broad system development process assessment, probably by attempting to measure the sample against the best practices and the risks related to each step that does not follow them.

Since a biased machine learning can result from problems with data collection, testing algorithms, or decision models, the regulatory agency must consider all those phases in its development process assessment models. A data governance assessment is an important analysis that is connected to the system development process as well as to the biased machine learning identification process.

The results of the regulatory agency's analysis materialise the total sum of all risks identified in an evaluated AI product or service for which there should be a risk mitigation plan.

As the regulatory agency is a natural actor to create and communicate the best practices to the industries and service providers, the agency must be aware of all projects and trends in the AI market, otherwise companies will not adopt those practices. An alternative to mitigate that risk is to strengthen the dialogue with the industries and service providers on the purpose of contributing to industry standards (N), thus allowing technology to improve its development while the legislation is still under debate, or in case it is not necessary. On the industries' and service providers' side, in order to increase the probability of a successful investment, a gradual strategy supported by a governance model should be behind the implementation of those good practices. Industry standards (N) must incorporate all parameters that are needed for communication among the "industries and service providers" along the entire value chain of an AI system.

As happens in Parliament, an open practice by the regulatory agency is likewise desirable, receiving feedback from academia (F). That feedback and those contributions, among other information, could be how society perceives ethical behaviours. A partnership between academia and the regulatory agency, combining scholars and researchers in the agency's staff, could be a sustainable alternative for maintaining a highly skilled team of professionals dealing with many processes simultaneously.

At an advanced level of an AI governance model, a "society-on-the loop" mechanism could be structured to collect the evaluation of a certain category of AI systems based on their behaviour using an ethical approach. Both civil society and academia could accomplish this. The answers (T) would feed the regulatory agency in the form of a survey to identify potential opportunities for improvement in its internal processes.

Regardless of the existence of a "society-on-the loop" mechanism, academia is always a good, reliable source of risk analysis reports (R) to be published periodically.

Law enforcement by courts would also undergo a continuous learning process with regard to interpretations based on the legislation in effect, as well as on new laws. In countries where the certification is incorporated into laws, decisions on cases involving uncertified companies would be treated differently from those involving certified companies. Thus, society and the courts would need to have up-to-date information about each company's certified products and services (K). Considering a continuous learning process, the regulatory agency would receive the judgment decisions of all cases involving AI systems (E), which would then be stored in the data repository shared with the Parliament. The decisions on cases may indicate types of AI technology use that the regulatory agency has not researched yet, and they may also indicate the need for changes in the legislation. A significant challenge would be to identify when an incident is avoidable or not. In those situations, experts must be involved in the investigation to find out the purpose of supporting the courts.

In order to balance the equation that rules the job market on the path to a digital economy, the government may create public policies (P) to make it feasible to implement in a timely manner the changes required in employer and student skills. Public policies might also be necessary to maintain an advertising campaign to inform people about the importance of certification and standards for AI products and services, helping them to identify when there is a potential case of an AI-embedded system.

As usual, public policies are a long-term strategy that may require actions by different government institutions, but there are many alternatives for implementing them, depending on the country. The regulatory agency may also provide government institutions with information about where and how those changes are needed. In some cases, by means of the T2R flow, the agency may notify the Parliament that a law is lacking that better regulates public policies.

On a national level, discussions to facilitate priority actions and the recognition of industry standards would be enabled through an AI Governance Committee, bringing together the public and private sectors (G). The synergy of efforts for the benefit of all stakeholders must be established, since many variables are considered. Beyond the regulatory agency, other government institutions would probably participate in this national committee, due to the wide impact its decisions could have. For instance, building human capacity and preparing the labour market transformation is a decision that might require a strategy that impacts many ministries and state governments. Adjustments to the current legislation related to many different subjects should probably be made to support the whole transition.

We should not forget the committee's governance approach, which requires working with indicators, i.e. data produced by its stakeholders. Therefore, a national AI governance committee would require at least collection, storage, and analysis processes within other institutions and businesses.

The agreed-upon standards (N) make it possible to move forward in some technological dimensions, while the Parliament discusses adjustments to the legislation when necessary. The risk management criteria (O) related to the use of those standards would be negotiated between the national committee and the industries and service providers, since each standard could impact a long productive chain.

The plethora of components in AI services and products of global reach imposes actions that would be agreed upon in an International Governance Committee comprising representatives from each country's committee (A). On many occasions, transparency in production processes is only feasible through complex international agreements, because corporate trade policies must adapt to different countries. A global strategy could be established to facilitate the production and delivery of standards, as well as the dissemination of best practices in undeveloped countries, since without that help the gap between them and the countries in which an AI governance has been established would increase hugely, putting them in a fragile position. In that regard, one should keep in mind that international standards are not limited to technological issues. Further, those standards also incorporate ethical principles, despite any cultural differences. The Universal Declaration of Human Rights could be a global base to engage governments to face the challenge of dealing with differences among national legislations.

The expert skills and engagement power of self-regulated organisations are a rich contribution to the international AI governance committee.

A possible adjustment entails the segmentation of tasks in charge of the regulatory agency, sharing them with or transferring them to other government institutions. For instance, the audit process could also be implemented by different government institutions in charge of auditing cases of discrimination using personal data, or investigations related to the development of autonomous weapons in that country. Hence, it is important to highlight that laws such as the EU GDPR (2016) only affect personal data. Nonetheless, AI discrimination risks have a wider reach than personal data.

Sharing the standardisation process with specialised private-sector organisations could also be an alternative. In that case, the connection between the standardisation process and the other regulatory agency processes should be maintained.

Despite being represented as a unique institution, the regulatory agency could be materialised as a group of agencies distributed across the country. To that effect, partnerships among countries could also allow for the creation of a set of agencies sharing resources, processes, and knowledge. In both cases, agencies could specialise in different categories of AI products and services. Although the certification issued for a specific category is independent of the certificate issued for another category, a communication process among the agencies is needed to increase the knowledge of how each AI product/service behaves and evolves over time.

Another adjustment to how the AIR framework is interpreted relates to what can be classified as “industries and service providers”. Private-sector companies are considered first. However, since any organisation that develops AI systems or offers services based on AI systems would fall in that category, public organisations may also be included.

Conclusion

The need and urgency to regulate Artificial Intelligence seem indisputable. The complexity of the topic is also evident, whether due to the advanced nature of technology or because its impacts structurally affect social standards. This combination materialises the perception of a problem that is yet to be completely defined.

A study of the literature through a sample comprising 109 documents (articles, laws, and government strategies) revealed significant efforts to identify and scale the risks and ethical dilemmas related to AI, as well as to seek a model for regulating AI based on different methodologies.

The heterogeneous nature of the professional profiles involved in the debate evinces the complexity and maturity with which the topic is being studied. Such an in-depth approach, on the one hand, may have caused certain delays in research, but on the other, it has prevented inappropriate regulatory solutions from being made official.

We had also seen the birth of a reshaped perception of the legislation, as had occurred with disruptive innovations in the past, when legislative efforts focused on adapting laws to the new paradigms brought about by electricity, telephone, and computers. Since this is a more difficult challenge, AI lawmakers will consider that we are still starting to discover the applications of smart algorithms. Therefore, a balance must be kept between a rigid damage prevention and technological development strategy (Gurkaynak et al., 2016).

Despite all efforts being directed to AI regulation and governance, there is still an expressive gap between ethical principles and a functional model that is able to encompass all areas of knowledge that are necessary to deal with the required complexity. The 21 proposed models found in the sample are based on supplementary approaches and are therefore insufficient when analysed separately. Due to the heterogeneous nature of those skills and interests, an ideal model should harmonise interests, offering benefits to all stakeholders during the entire lifecycle of an AI product or service.

The consolidation and process orientation approach proposed by the AIR framework (Fig. 1) seems to be the most adequate strategy for the deployment of an AI governance, given the existence of several agents and the laterality of the topic, which intertwines different areas of knowledge. The expanded view of the presented AIR framework will enable all agents involved to identify their role in the governance process, while establishing a roadmap for a gradual and uninterrupted deployment.

It also contributes to the creation of a new reward and punishment model to balance out this new reality (Bryson, 2018; Waser, 2015), taking into account the world as it will be (Lin et al., 2011).

On the path to improve each component of the AIR framework, more than bringing them closer together, there needs to be a synchronisation of stakeholders towards a sustainable regulation. Along that journey, an alliance between scholars and the government’s three agents (the executive, legislative, and judicial branches) is crucial for the macro-process of regulation.

The countries leading the debate are probably ready to coordinate the partnerships and agreements among institutions that are necessary for a comprehensive and effective governance, as well as to initiate a regulation process. Nonetheless, the launch of AI-embedded products in countries that have advanced regulation models, in and of itself, does not guarantee the same safety levels for countries that are still unripe in this regard.

Much is yet to happen in the formulation of solutions using real-case scenarios to enable an empirical analysis and studies of the evolution of the models presented in the examined sample. To that effect, the AIR framework can make it tangible and feasible to synchronise all the stakeholders’ efforts to achieve an effective result, thus culminating in the creation of a reference model of AI governance in which maturity levels would be established that could be monitored by international bodies in a collaborative action. The way we and future generations will live our lives depends on that cooperation.

References

Aayog, N. (2018). National Strategy for Artificial Intelligence: #AI for All (Discussion Paper) https://www.niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf. Accessed 30 July 2020.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access Review, 6*, 52138–52160

AI HLEG - High-Level Expert Group on Artificial Intelligence. (2019a). A definition of AI: Main capabilities and disciplines. Definition developed for the purpose of the AI HLEG's deliverables.

AI HLEG - High-Level Expert Group on Artificial. (2019b). Ethics guidelines for trustworthy AI. High-level expert group on artificial intelligence for the European Commission.

AI4People. (2018). Ethical framework for a good society: opportunities, risks, principles, and recommendations. *Atomium – European Institute for Science, Media and Democracy.* http://www.eismd.eu/wp-content/uploads/2019/02/Ethical-Framework-for-a-Good-AI-Society.pdf. Accessed 21 June 2019.

Amigoni, F., & Schiaffonati, V. (2018). Ethics for robots as experimental technologies. *IEEE Robotics & Automation Magazine, 25*, 30–36

Arkin, R. C. (2011). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, 121–128.

Arnold, T., & Scheutz, M. (2018). The big red button is too late: an alternative model for the ethical evaluation of AI systems. *Ethics and Information Technology, 20*, 59–69

Beltran, N. (2020). Artificial intelligence in Lethal Autonomous Weapon Systems: What's the problem? Uppsala University – Department of Theology.

Bench-Capon, T., & Modgil, S. (2017). Norms and value based reasoning: Justifying compliance and violation. *Artificial Intelligence & Law Review, 25*, 29–64

Benjamins, V. R. & García I. S. (2020). Towards a framework for understanding societal and ethical implications of Artificial Intelligence. *Vulnerabilidad y cultura digital* by Dykinson. pp 87–98.

Black, J. (2002) Critical reflections on regulation. *Australian Journal of Legal Philosopy, 27*, 1–35. http://www.austlii.edu.au/au/journals/AUJlLegPhil/2002/1.pdf. Accessed 30 July 2020.

Bonnemais, V., Saurel, C., & Tessier, C. (2018). Embedded ethics: Some technical and ethical challenges. *Ethics and Information Technology, 20*, 41–58

Buiten, C. M. (2019). Towards intelligent regulation of artificial intelligence. *European Journal of Risk Regulation, 10*(1), 41–59

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., Newman, P., Parry, V., Pegman, G., Rodden, T., Sorrell, T., Wallis, M., Whitby, B., & Winfield, A. (2017). Principles of robotics: Regulating robots in the real world. *Connection Science, 29*(2), 124–129

Bonnemains, V., Saurel, C., & Tessier, C. (2018). Embedded ethics: Some technical and ethical challenges. *Ethics Information Technology., 20*, 41

Borgesius, F. Z. (2018). Discrimination, artificial intelligence, and algorithmic decision-making. Study for the Council of Europe.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Bryson, J. J. (2018). Patiency is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology, 20*, 15–26

Butterworth, M. (2018). The ICO and artificial intelligence: The role of fairness in the GDPR framework. *Computer Law & Security Review, 34*, 257–268

Calo, M. R. (2011). Peeping hals. *Artificial Intelligence Review, 175*, 940–994

Calo, M. R. (2015). Robotics and the lessons of cyberlaw. *California Law Review, 103*(3), 513–563

Caron, M. S., & Gupta, A. (2020). The social contract for AI. Cornell University. https://arxiv.org/abs/2006.08140v1 Accessed 6 Dec 2020.

Canada Government. (2020). Algorithmic Impact Assessment. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html. Accessed 15 Dec 2020.

Carter, D. (2020). Regulation and ethics in artificial intelligence and machine learning technologies: Where are we now? Who is responsible? Can the information professional play a role? *Business Information Review., 37*(2), 60–68

Cath, C., Watcher, S., Mittelsadt, B., Taddeo, M., & Floridi, L. (2017). Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. https://ssrn.com/abstract=2906249 or https://doi.org/10.2139/ssrn.2906249. Accessed 21 June 2019.

Cave, S., Nyrup, R., Vold, K., & Weller, A. (2019). Motivations and risks of machine ethics. *Proceedings of the IEEE, 107*(3), 562–574

Cerka, P., Grigiene, J., & Sirbikite, G. (2015). Liability for damages caused by artificial intelligence. *Computer Law & Security Review, 31*(3), 376–389

Cerka, P., Grigiene, J., & Sirbikyte, G. (2017). Is it possible to grant legal personality to artificial intelligence software systems? *Computer Law & Security Review, 33*(5), 685–699

Conitzer, V., Sinnott-Armstrong, W., Borg, J. S, Deng, Y., & Kramer, M. (2017). Moral decision making for artificial intelligence. *AAAI Publication, 31º Conference on Artificial Intelligence*

Council of Europe. (2018). European commission for the efficiency of justice, 'European ethical charter on the use of artificial intelligence in judicial systems and their environment. https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c. Accessed 30 July 2020.

Davis, E. (2015). Ethical guidelines for a superintelligence. *Artificial Intelligence Review, 220*, 121–124

Dennis, L., Fisher, M., Slavkovik, M., & Webster, M. (2016). Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems, 77*, 1–14

Djeffal, C. (2018). Sustainable AI Development (SAID): On the road to more access to justice. https://ssrn.com/abstract=3298980 or https://doi.org/10.2139/ssrn.3298980. Accessed 30 July 2020.

Donahoe, E., & Metzger, M. M. (2019). Artificial intelligence and human rights. *Journal of Democracy, 30*(2), 115–126

Dubai (2019). Smart Dubai. Artificial intelligence principles and ethics. https://smartdubai.ae/initiatives/ai-principles-ethics. Accessed 20 July 2020.

EU GDPR. (2016). European Parliament. General Data Protection Regulation. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679. Accessed 30 July 2020.

EU Parliament. (2012). Charter of Fundamental Rights of the European Union (2012/C 326/02), *Official Journal of the European Union*, 2012 C 326, (pp. 391).

European Commission. (2019). Communication from the Commission to the European Parliament, the Council, The European Economic and Social Committee and the Committee of the Regions. Brussels. https://www.eea.europa.eu/policy-documents/communication-from-the-commission-to-1. Accessed 30 July 2020.

Firth-Butterfield, K. (2017). Artificial Intelligence and the Law: More questions than answers. *Scitech Lawyer, 14*, 28–31

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. C., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. Berkman Klein Center for Internet & Society.

Floridi, L., Cowls, J., King, T., & Taddeo, M. (2020). How to design AI for social good: seven essential factors. *Science and Engineering Ethics, 26*, 1771

French, P. M. (2018). For a Meaningful Artificial Intelligence: Toward a French and European Strategy. Mission assigned by the French Prime Minister. https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf. Accessed 30 July 2020.

Future of Life Institute. (2019a). National and International AI Strategies. https://futureoflife.org/national-international-ai-strategies/. Accessed 20 September 2019.

Future of Life Institute. (2019b). Ansilomar AI Principles. https://futureoflife.org/ai-principles/. Accessed 20 September 2019.

German Federal Government. (2018). German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labor and Social Affairs. Artificial Intelligence Strategy. https://www.ki-strategie-deutschland.de/home.html. Accessed 30 July 2020.

Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. *Hawaii International Conference on System Sciences* 52nd, 2019.

Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly, 30*(3), 611–642

Gurkaynak, G., Yilmaz, I., & Haksever, G. (2016). Stifling artificial intelligence: Human perils. *Computer Law & Security Review, 32*(5), 749–758

Hagendorff, T. (2019). The ethics of AI ethics: An evaluation of guidelines. CoRR, abs/1903.03425.

Hilb, M. (2020). Toward artificial governance? The role of artificial intelligence in shaping the future or corporate governance. *Journal of Management and Governance.*

Hildebrandt, M. (2018). Algorithmic regulation and the rule of law. *Philosophy Transactions of the Royal Society,* 376 (2128).

Holder, C., Khurana, V., Harrison, F., & Jacobs, L. (2016a). Robotics and law: Key legal and regulatory implications of the robotics age (Part I of II). *Computer Law & Security Review, 32*(3), 383–402

Holder, C., Khurana, V., Hook, J., Bacon, G., & Day, R. (2016b). Robotics and law: key legal and regulatory implications of the robotics age (Part II of II). *Computer Law Secure Review, 32*, 557–576

House of Lords. (2018). AI in the UK: Ready, willing and able? *Select Committee on Artificial Intelligence*, Report of Session 2017–19. 13 March 2018.

IEEE. (2019). Ethically Aligned Design. Committees of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2nd version. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf. Accessed 20 July 2020

IEEE. (2020). a call to action for business using AI—Ethically aligned design for business. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead/ead-for-business.pdf. Accessed 20 July 2020.

Jackson, B. W. (2019). Artificial Intelligence and the Fog of Innovation: A deep-dive on governance and the liability of autonomous systems. 35 *Santa Clara High Tech*. L.J. 35

Jackson, B. W. (2020). Cybersecurity, privacy, and artificial intelligence: An examination of legal issues surrounding the European Union General Data Protection Regulation and Autonomous Network Defense, 21 *Minnesota Journal of Law, Science & Technology*, *21*

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*, 389–399. https://doi.org/10.1038/s42256-019-0088-2Accessed20July2020

Japanese Cabinet Office. (2019). Social principles of human-centric artificial intelligence. Council for science, technology and innovation, https://www8.cao.go.jp/cstp/english/humancentricai.pdf. Accessed 20 July 2020

Kaal, W. A., & Vermeulen, E. P.M. (2017). How to regulate disruptive innovation: From facts to data. *Jurimetrics*, *57*(2).

Kozuka, S. (2019). A governance framework for the development and use of artificial intelligence: Lessons from the comparison of Japanese and European initiatives. *Uniform Law Review, 24*, 315–329

Kunz, J. (1949). The United Nations declaration of human rights. *American Journal of International Law, 43*(2), 316–323

Larsson, S. (2020). On the governance of artificial intelligence through ethics guidelines. *Asian Journal of Law and Society*, 1–23.

Lenardon, J. P. A. (2017). The Regulation of Artificial Intelligence. *Master Thesis. Tilburg Institute for Law, Technology and Society*. Netherlands.

Lewis, D., Hogan, L., Filip, D., & Wall, P. J. (2020). Global challenges in the standardization of ethics for trustworthy AI. https://doi.org/10.5281/zenodo.3516525. Accessed 30 July 2020.

Lamo, M. & Calo, R. (2018). Regulating Bot Speech. *UCLA Law Review 2019*, July 16, 2018.

Leitner, C., & Stiefmueller, C. M. (2019). Disruptive technologies and the public sector: The changing dynamics of governance. In A. Baimenov & P. Liverakos (Eds.), *Public service excellence in the 21st century.* (pp. 238–239). Palgrave Macmillan.

Lewis, T., & Yildirim, H. (2002). Learning by doing and dynamic regulation. *The RAND Journal of Economics*, *33*(1), 22–36. www.jstor.org/stable/2696373 Accessed 20 July 2020.

Lin, P., Abney, K., & Bekey, G. (2011). Robot ethics: Mapping the issues for a mechanized world. *Artificial Intelligence Review, 175*, 942–949

Lin, Y., Hung, T., & Huang, L. T. (2020). Engineering equity: How AI can help reduce the harm of implicit bias. *Philosophy & Technology*. https://doi.org/10.1007/s13347-020-00406-7

Liu, H. (2017). Irresponsibilities, inequalities and injustice for autonomous vehicles. *Ethics Information Technology Journal, 19*, 193–207

Maluf S. (1995). *Teoria Geral do Estado*. 23ª ed., 205–208. Editora Saraiva. São Paulo.

Mantelero, A. (2018). AI & Big Data: A blueprint for human rights, social and ethical impact assessment. *Computer Law & Security Review, 34*(4), 754–772

Mika, N., Nadezhda, G., Jaana, L., & Raija, K., (2019). Ethical AI for the governance of the Society: Challenges and opportunities. *CEUR Workshop Proceedings*, *2505*, 20–26. http://ceur-ws.org/Vol-2505/paper03.pdf. Accessed 20 July 2020.

Millar, J. (2016). An ethics evaluation tool for automating ethical decision-making in robots and self-driving cars. *Applied Artificial Intelligence, 30*(8), 787–809

Monetary Authority of Singapore. (2019). Monetary Authority of Singapore. Principles to promote fairness, ethics, accountability and transparency (FEAT) in the use of artificial intelligence and data analytics in Singapore's Financial Sector. https://www.mas.gov.sg/~/media/MAS/News%20and%20Publications/

Monographs%20and%20Information%20Papers/FEAT%20Principles%20Final.pdf. Accessed 20 July 2020

Nevejans, N. (2016). European civil law rules in robotics. Study requested by the European Parliament's Committee on Legal Affairs. *Policy Department Citizens' Right and Constitutional Affairs*.

Neznamov, A. V. (2020). Regulatory landscape of artificial intelligence advances in social science, education and humanities research*, volume* 420 pp 201–204. XVII *International Research-to-Practice Conference 2020*. Atlantatis Press.

Organisation for Economic Co-operation and Development (2019). 'Recommendation of the Council on Artificial Intelligence'.

Partnership on AI to Benefit People and Society. (2016) https://www.partnershiponai.org/about/. Accessed 12 July 2020.

Pedro, A. P. (2014). Ética, moral, axiologia e valores: confusões e ambiguidades em torno de um conceito comum. *Kriterion*, vol. 55. Belo Horizonte, nº 130, Dez./2014, 483–498.

Poel, I. V. (2016). An ethical framework for evaluating experimental technology. *Science and Engineering Ethics, 22*(3), 667–686

Prakken, H. (2017). On the problem of making autonomous vehicles conform to traffic law. *Artificial Intelligence & Law, 25*, 341–363

Rahwan, I. (2017). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology., 20*, 5–14

Reed, C. (2018). How should we regulate artificial intelligence? *Philosophy Transactions of the Royal Society, 376*, 2128

Riek, L. D., & Howard, D. (2014). A code of ethics for human-robot interaction profession proceedings of we robot, 2014. SSRN: https://ssrn.com/abstract=2757805. Accessed 20 July 2020.

Rousseau, J. (2016). *The Social Contract*. (202–230). ISBN: 978911495741. London: Sovereign.

Russell, S., & Norvig, P. (1995). *Artificial Intelligence. A Modern Approach*. (pp. 4–5). Prentice Hall.

Scherer, M. U. (2016). Regulating artificial intelligence systems: Risks, challenges, competences and strategies. *Harvard Journal of Law & Technology, 29*(2), 354–398

Schrader, D., & Ghosh, D. (2018). Proactively protecting against the singularity: Ethical decision making AI. *IEEE Computer and Reliability Societies Review, 16*(3), 56–63

Smuha, N. A. (2020). *Beyond a human rights-based approach to AI governance: Promise*. Philosophy & Technology.

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good: An ethical framework will help to harness the potential of AI while keeping humans in control. *Science Review, 361*(6404), 751–752

Toronto. (2020). The Toronto declaration: Protecting the right to equality and non-discrimination in machine learning systems. https://www.torontodeclaration.org/. Accessed 20 July 2020

Tutt, A. (2017). An FDA for algorithms. *Administrative Law Review, 69*(83), 83–123

UK Government. (2018). Government response to House of Lords Artificial Intelligence Select Committee's Report on AI in the UK: Ready, Willing and Able? https://www.gov.uk/government/publications/ai-in-the-uk-ready-willing-and-able-government-response-to-the-select-committee-report. Accessed 31 December 2020

United Nations. (2011). *UN guiding principles on business and human rights*. (p. 2011). UN Human Rights Council.

University of Montreal. (2018). Montreal Declaration for a Responsible Development of Artificial Intelligence. https://www.montrealdeclaration-responsibleai.com/the-declaration Accessed 20 July 2020

US Congress. (2019). H.Res.153 - Supporting the de7velopment of guidelines for ethical development of artificial intelligence. https://www.congress.gov/bill/116th-congress/house-resolution/153?q=%7B%22search%22%3A%5B%22ARTIFICIAL+INTELLIGENCE%22%5D%7D&s=2&r=4

US Congress. (2020). s.3891 – Advancing Artificial Intelligence Research Act of 2020. https://www.congress.gov/bill/116th-congress/senate-bill/3891?q=%7B%22search%22%3A%5B%22ARTIFICIAL+INTELLIGENCE%22%5D%7D&s=3&r=7

Villaronga, E. F., & Heldeweg, M. (2018). Regulation, I presume? Said the robot: Towards an iterative regulatory process for robot governance. *Computer Law & Security Review*, 21 June, 2018.

Verheij, B. (2016). Formalizing value-guided argumentation for ethical systems design. *Artificial Intelligence & Law Review, 24*(4), 387–407

Yeung, K., Howes, A., & Pogrebna, G. (2019). AI governance by human rights-centred design, deliberation and oversight: An end to ethics washing (June 21, 2019). Forthcoming in M Dubber and F Pasquale (eds.) *The Oxford Handbook of AI Ethics*, Oxford University Press (2019), https://doi.org/10.2139/ssrn.3435011. Accessed 15 December 2020.

Wallach, W., & Marchant, G. E. (2018). An agile ethical/legal model for the international and national governance of ai and robotics. *Association for the Advancement of Artificial Intelligence*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6191666/. Accessed 20 July 2020

Waser, M. (2015). Designing, implementing and enforcing a coherent system of laws, ethics and morals for intelligent machines (including humans). *Procedia Computer Science, 71*, 106–111

Wright, S. A., & Schultz, A. (2018). The rising tide of artificial intelligence and business automation: Developing an ethical framework. *Business Horizons, 61*(6), 823–832

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.