Volume **5** (1) 2023 https://doi.org/10.5204/lthj.2682

Risky Artificial Intelligence: The Role of Incidents in the Path to AI Regulation

Giampiero Lupo
ISASI-CNR (Institute of Applied Sciences and Intelligent Systems - National Research Council of Italy), Italy

Abstract

The history of high-tech regulation is a path studded with incidents. Each adverse event allowed the gathering of more information on high technologies and their impacts on people, infrastructure, and other technologies, posing the bases for their regulation. With the increasing diffusion of artificial intelligence (AI) use, it is plausible that this connection between incidents and high-tech regulation will be confirmed for this technology as well. This study focuses on the role of AI incidents and an efficient strategy of incident data collection and analysis to improve our knowledge of the impact of AI technologies and regulate them better. To pursue this objective, the paper first analyses the evolution of high-tech regulation in the aftermath of incidents. Second, the paper focuses on the recent developments in AI regulation through soft and hard laws. Third, this study assesses the quality of the available AI incident databases and their capacity to provide information useful for opening and regulating the AI black box. This study acknowledges the importance of implementing a strategy for gathering and analysing AI incident data and approving flexible AI regulation that evolves with such a new technology and with the information that we will receive from adverse events—an approach that is also endorsed by the European Commission and its proposal to regulate and harmonise rules on AI.

***Keywords**:* Artificial Intelligence; AI incidents; AI regulation; AI ethical documents.

Introduction

The introduction of new, complex technology, such as the automobile or civilian nuclear power, raises many questions regarding its safety, risks, impact on society and the environment. The phenomenon of complex technology's diffusion and interaction with different social contexts is often accompanied by a contrast between actors enthusiastic about the novelty and detractors who fearfully view its risks. Also, the coexistence of new and old forms of technology, such as cars and bicycles,[1] poses new challenges and makes regulating such coexistence necessary.[2]

In recent decades, several application contexts have witnessed an ever-greater diffusion of artificial intelligence (AI) technology. This certainly represents a new experience of introducing complex technology that brings concerns about its safety and implications for the values and functioning of its different areas of application. This change is also more 'revolutionary' than ever because AI refers to autonomous entities by definition and involves technologies that can act as intelligent agents that receive perceptions from the external environment and perform actions autonomously.[3] As has happened for other high technologies, fear about safety and technological risks and the uncertainties from the abrupt change to the expected and loved order has, in the words of Georges Canguilheim,[4] encouraged a rush towards the regulation of the new technology to restore or

[1] Pinch, "Social Construction of Technology," 165–186.
[2] Thomas, "Uneasy Coexistence," 71–98; Pinch, "Social Construction of Technology."
[3] Russell, Artificial Intelligence; see also Santosuosso, "Intelligenza Artificiale."
[4] Canguilhem states that order means a familiar state of relationship between individuals and their environment. The disruption of that order—for instance, with the introduction of a new technology or in case of an adverse event—represents an environmental challenge to individuals' mental and cognitive orientation to the world. Canguilhem, Normal and the Pathological.

safeguard the normal sense and order.[5] As stated by Michel Foucault, norms have the capacity to curb technological change and limit its borders and its implications for the pre-established normal order.[6]

As a reaction to AI diffusion, there is a proliferation of soft laws in the form of normative frameworks, guidelines and collection of ethical principles disciplining the application of AI in different contexts.[7] The drafting of soft laws[8] represents a flexible practice to cope with the unpredictable effects of emerging technologies, unlike law-making, which is more rigid and time consuming.

Also, legislative institutions are gearing up to define a legislative framework that may regulate the use of AI in different contexts in line with human rights and previous fundamental laws. For example, see the proposal for a regulation laying down harmonised rules on AI (Artificial Intelligence Act) drafted by the European Commission (EC) (hereafter, the AI Act).[9]

It is not just the uncertainty about the impact of a technology and its risks that drives high-tech regulation. An interesting pattern describes high-tech regulation as the result of information gathering in the aftermath of undesirable or unfortunate happenings that occurred unintentionally, resulting in harm, injury, damage or loss (i.e., 'incidents'). For instance, see the history of regulation in civil aviation, automobile transportation, nuclear energy or pipeline industries.[10]

This link between incidents and high-tech regulation exists due to the extreme complexity of some technologies that do not facilitate the preliminary identification of weak points in terms of safety and the clarification of their real impact on individuals, society, the environment, existing laws and old technologies. Additionally, in some cases, the 'black box'[11] of high technology's functioning may be opened only in the aftermath of unwanted and harmful happenings when safeguarded by particularly protective strategies to protect intellectual property, such as trade secrets or employee confidentiality obligations.[12]

AI technology was only recently applied in several contexts, and the count of unfortunate happenings may grow in the future. Also for AI, the empirical evidence of incidents may provide fundamental information on the functioning of this technology, bring the debate on the risks of AI to the attention of the public and policymakers and influence the regulatory processes that will affect AI in the future.

This paper focuses on the relationship between AI incidents and regulation and investigates the role of incident analysis in providing information about the impact of AI technologies that can be useful for drafting binding regulation. To pursue this objective, this study analysed AI incident databases freely available online to shed light on what information they can provide and their potentialities and limits for law-making and law amendments. Additionally, to investigate agreement between the emerging AI legislative instruments and the empirical evidence of AI incidents from available data from the databases, this study relied on my analysis of AI ethical guidelines described in the paper 'Ethics of Artificial Intelligence'[13] and on a qualitative analysis of the EC AI Act. The analysis of these legislative instruments also investigated the orientation of drafting bodies towards using incident analysis as a strategy for gathering data on AI impact and improving regulations. The results of this study acknowledge the importance of an effective strategy to gather incident data for investigating and analysing AI impact and for drafting effective AI regulation.

[5] Angelides, "Disorder," 10–20; Lanzara, Capacità Negativa.
[6] Foucault, "Historia de la Medicalización."
[7] Lupo, "Regulating (Artificial) Intelligence," 75–96.
[8] van Dijk, "Ethification" of Privacy.
[9] European Commission, "Proposal for a Regulation of the European Parliament."
[10] Downer, "Trust and Technology," 83–106; Norton, "Four Paradigms," 319–334; Perrow, "Meltdown"; Perrow, Normal Accidents; Dahle, "Major Accidents."
[11] Rai, "Explainable AI," 137–141.
[12] Several scholars addressed the concept of high-tech innovations protected by trade secrets or patent laws and the relationship between these strategies of intellectual property protection, on the one hand, and technological information disclosure, transparency and collaborative research and development, on the other (Choi, "Opening," 192–203; Cammarano, "Importance of Possessing Knowledge," 101–127; Tschider, "Beyond," 683; Adams, "Industrial R&D Laboratories," 99–107; Rai, "Explainable AI," 137–141). Other scholars investigated the potential of incident analysis for opening high-tech black box and improve transparency and innovation disclosure (Jung, "First-Year Analysis," 122–127; Rosenberg, Inside the Black Box; Marabelli, "Light," 351–374; Kowalick, Fatal Exit). This aspect is utterly important for AI: algorithms may be so complex that an AI's developer may not understand how it makes decisions, making them likely candidates for trade secrets instead of public IP protection, such as patent or copyrights (Katyal, "Paradox," 1183; Wachter, "Counterfactual Explanations," 841). Therefore, incident analysis may represent an important strategy for supporting disclosure and transparency. An in-depth analysis of these topics is out of the scope of the paper; however, the paper will briefly deal with the issue of AI black box opening through incident analysis in the next sections. Menell, "Intellectual Property Law," 1473–1570.
[13] Lupo, "Regulating (Artificial) Intelligence"; Lupo, "Ethics of Artificial Intelligence."

The paper is structured as follows. The next section briefly discusses the study's methodology and the working definitions. The following section includes a brief dissertation on the evolution of high technologies' regulation through incidents, revealing relevant patterns that may also affect AI. The next section discusses the recent developments regarding AI soft and hard laws by focusing on the analysis of AI ethical documents and the EC AI Act. The final section describes and analyses the publicly available AI incident databases, the AI incidents' relationship with AI regulation, and the incidents' capacity to provide information on AI impact. The concluding remarks section summarises the results of the study.

Methodology and Definitions

The study presented here is based on an interdisciplinary approach and mixed methodology involving quantitative and qualitative analysis techniques. The investigation of the relationship between high-tech regulation and incidents is based on a literature review of the main publications on the topic.[14]

To shed light on the relationship between AI soft laws and incidents, this study refers to my previous work that investigated AI ethical documents through content analysis to put in evidence worth-mentioning patterns.[15] The investigation of the connection between AI regulation and AI incidents also involved qualitative analysis of the EC proposal for AI regulation (AI Act).[16]
The mentioned incident databases have been investigated through quantitative techniques of analysis.[17] Not many databases exist. Therefore, the choice of databases to be analysed did not require a stringent selection. The selected databases share the following characteristics: they are created by non-profit organisations, they are publicly available for consultation and analysis, and they gather data on each cases' attributes (e.g., AI sector or type of AI technology involved). The incident databases selected for analysis are as follows.

1. The Where in the World is AI? Map[18] is an incident database that is the basis of an interactive web visualisation tool that provides information on existing AI systems with their geolocation. The database is managed by RAII (Responsible AI Institute), a member-driven non-profit organisation building tangible governance tools for trustworthy, safe and fair AI.[19] The database includes 430 cases (gathered from 2006 to 2021) and includes several attributes, such as AI domain and location.
2. The AI Incident Database[20] is a project of the Responsible AI Collaborative,[21] a non-profit organisation that aims to identify, define and catalogue AI incidents. The AI Incident Database includes 2,052 reports (gathered from 2017 to 2022) categorised on the basis of specific taxonomies.
3. The AI, Algorithmic, and Automation Incidents and Controversies (AIAAIC) repository[22] is a repository that details incidents and controversies involving AI, algorithms, and automation. Incidents are also catalogued on the basis of specific attributes. The repository is managed by a set of editors with different backgrounds, from computer science to social science, and it includes 871 cases gathered from 2019 to 2022.

In all three cases, inputs are based on media coverage of AI incidents coming from several sources. Data are inputted voluntarily by the public and checked for appropriateness by the organisation managing the dataset.

The analyses described in the next sections relied on two working definitions. First, the Artificial Intelligence definition. Given that there is no shared definition of AI, this study will rely on one of the most inclusive definitions, comprising technologies from self-driving vehicles to data analysis: AI includes machines mimicking cognitive functions associated with the human mind, including learning, problem solving and natural language processing.[23] Second, 'incident' also needs defining. Also in this case, to allow the study to take advantage of the maximum amount of information, the working definition is considerably inclusive: 'AI incident' is defined as a situation in which AI systems caused, or nearly caused, real-world harm.

[14] Snyder, "Literature Review," 333–339; Paré, "Methods"; Rosenthal, "Meta-Analysis," 59–82.
[15] Lupo, "Ethics of Artificial Intelligence," 614–653. See this publication for more information on the content analysis methods utilized in this study.
[16] Mitchell, "Analyzing the Law," 102–113; Coutin, "Qualitative Research," 50; Taekema, "Theoretical and Normative Frameworks."
[17] Gorard, Quantitative Methods; Pole, Practical Social Investigation; Baškarada, "A Philosophical Discussion."
[18] AI Global, "Where in the World."
[19] Responsible AI, "Responsible Artificial Intelligence."
[20] AIID, "AI Incident Database."
[21] AIID, "Founding Report."
[22] AIAAIC, "Understanding the risks."
[23] Russell, Artificial Intelligence.

The Regulation of High Technology through Incidents

Incidents involving high technologies may not only cause harm to organisations, humans and the environment. In the history of high-tech regulation, it is evident the impact of incidents on high-tech regulatory and organisational frameworks. Internally, new organisational structures, roles and management technologies may be created in the wake of crises and disasters, while externally, pressures arise to create or reform regulatory regimes and their programs for risk prevention, reaction and resilience.[24] Moreover, new standards of practice may be suggested, new stakeholders and communities of interest in risk management may be created, and new mandates for regulatory organisations may be proposed.[25] This interconnection between major incidents and high-tech regulatory regimes has already received researchers' attention for a long time, at least since the 1997 publication of Roger Cooter and Bill Luckin's landmark edited collection *Accidents in History*.[26] The literature[27] acknowledges that the influence of high technology major incidents on the regulation frameworks depends on several factors: the entity of the social amplification of exposed risk due to lost lives and environmental and physical damages, the spread of information, the role of experts in disseminating information and the entity and intensity of media coverage of the event.[28]

Major crises and incidents facilitate gathering information on the functioning of a newly developed technology and on the consequences of its use in particular conditions.[29] As Perrow[30] and the Normal Incident Theory has demonstrated, high technologies are characterised by a certain amount of complexity and by the loose coupling of technological and organisational components. This means that some incidents are unavoidable and unforeseeable. Often, only in the aftermath of a crisis are we able to identify weak points in terms of safety to clarify the real impact of technology on individuals, society, the environment, existing laws and old technologies and to open the 'black box'[31] of high technology's functioning when protected by trade secrets and stringent intellectual property strategies.[32]

The evolution of regulatory regimes through incidents interests most in the high technology field. For instance, consider the case of road safety regulation. From its first outing on public roads, the motor vehicle was a contested technology that provoked a range of responses from enthusiasm to opposition and concern.[33] Most oppositions and concerns originated from the deaths and injuries in which automobiles—and their drivers—were implicated, a toll counting hundreds of millions of people globally from the late nineteenth century to the present day.[34] No other technology (that was not specifically designed to cause harm) had such an impact on human life and death in the same short span.

High risks and numerous casualties also had an impact on regulatory regimes that tried to improve the safety of the use of automotive technology.[35] For instance, the 50,000 lives lost per year in 1966 in the United States contributed to the change of paradigm from 'auto-safety' to 'crashworthiness'.[36] The auto-safety paradigm was based on the assumption that as soon as nobody hits each other, no one will get hurt; therefore, this approach focuses on the 'three Es': (1) engineering roads to limit the possibility of collisions and equipping vehicles with reliable brakes and steering, (2) educating drivers and pedestrians to avoid collisions and (3) drafting and enforcing rules of the road to discipline drivers' behaviour.[37] In contrast, the 'crashworthiness' paradigm diffused since the late 1960s considered that a number of incidents on the road are unavoidable; therefore, car manufacturers had to design and implement technologies like seat belts and airbags that limit the impact of incidents on the human body. This also represented a shift of responsibility for the consequences of incidents from drivers to the technology's developers.[38] As happens with AI, automotive technology is characterised by a complex interaction between technology and human agents, so the question of responsibility in case of failure is complex: who can be blamed in the case of an incident, the driver or the car's manufacturer? In complex technology contexts, such as in aviation and road traffic safety (and potentially in the AI context), the 'blame game' may be one of the major obstacles to effective prevention.[39] It biases

[24] Dahle, "Major Accidents."
[25] Hutter, Organizational Encounters.
[26] Cooter, Accidents in History; Esbester, "Introduction."
[27] Kasperson, "Social Amplification of Risk"; Pidgeon, "Role of Social."
[28] French, "Aggregating Expert Judgement"; Skjong, "Expert Judgment"; Rae, "Forecasts or Fortune-Telling."
[29] Clare, "Learning from Incidents."
[30] Perrow, Normal Accidents.
[31] Rai, "Explainable AI."
[32] Wang, "Protecting the Intellectual Property," 619; Tan, "Embarrassingly Simple Approach."
[33] Tingvall, "History of Traffic Safety," 489–492.
[34] Moraglio, "Knights of Death."
[35] Wetmore, "Delegating to the Automobile."
[36] Norton, "Hell on Wheels," 141–142.
[37] Norton, "Four Paradigms."
[38] Cooter, Accidents in History.
[39] Voas, "IoT Blame Game," 69–73; Phillips, "Case Study."

information, hides prevention initiatives and draws attention away from the fact that incidents originate from a complex interaction of technological, human and organisational factors.[40]

The automotive case also sheds light on the difficulty of regulating new technologies when they must interact or coexist with pre-existing technologies. Again, consider the case of cars and bicycles.[41] In the beginning of automotive history, the coexistence of these transportation technologies happened in a regime of anarchy: only the empirical evidence of incidents provided the necessary push forward and the necessary information to draft effective road safety regulations disciplining this issue.[42]

The aviation industry is another complex technology context in which the regulatory framework owes its accuracy, safety performance and complexity to the numerous incidents that occurred since the first fixed-wing scheduled airline was started on 1 January 1914, from St. Petersburg, Florida, to Tampa, Florida, operated by the St. Petersburg–Tampa Airboat Line.[43] The literature on the relationship between aviation incidents and regulation[44] confirmed that modern incident analysis is at the basis of the theorisation of models explaining the patterns leading to disasters, the drafting of strict rules regulating technology's implementation and use, and the training of human agents involved.

For instance, the aviation incident analysis contributed to the design of the 'Swiss cheese' model that explains the interconnection of conditions contributing to major incidents. The model developed by Reason[45] describes the factors contributing to an incident as gaps or weaknesses in the defensive layers of a system. Each defensive layer represents a barrier against unsafe occurrences. A set of layers and relative gaps can be constituted by the following factors: (a) unsafe acts (e.g., a pilot starts to take off without receiving clearance from the control tower), (b) preconditions for unsafe acts (e.g., a pilot or controller is suffering from mental or physical fatigue), (c) unsafe supervision (e.g., an airline pairs an inappropriate captain and first officer for a flight) and (d) organisational influences (e.g., an air traffic control centre has insufficient staffing). The gaps are continually changing position and size, and when gaps in all layers are aligned, it is possible for an incident trajectory to pass through all the defences, like a skewer passing through the holes in slices of Swiss cheese.

Another example of the amendment of aviation regulation as a consequence of incident analysis is the program implemented in 2008 by the International Civil Aviation Organization (ICAO) to improve the language proficiency of pilots and air traffic controllers around the world.[46] This program was based on the development of language proficiency requirements (LPRs) and a six-level language proficiency rating scale for aviation personnel that member states were required to comply with by 5 March 2011.[47] ICAO initially intended that all pilots and controllers involved in international flights demonstrate proficiency at level 4 or higher. This program and relative regulation were based on the analysis of seven incidents that occurred between 1976 and 2001 and which resulted in the deaths of 1,460 people.[48]

The analysis of aviation's regulatory path sheds light on patterns that may potentially affect the regulation of AI despite the evident differences between the two high technologies. An example is the role of regulatory bodies. Due to the complexity of the aviation context, societies manage their relationship with technology through expert mediators, usually state regulatory bodies like the FAA (Federal Aviation Administration) in the United States. These are commissions and prominent regulators of complex technologies that frame, promulgate and implement an extensive network of specifications and regulations governing the design, use and manufacture of civil aircraft in the world's most significant aviation market.[49] With the expansion of the AI industry and the diffusion of this technology, it is plausible that such types of bodies will be created in several national and supranational contexts also for AI.[50]

Another pattern relates to the standardisation of aviation regulations. The clear interconnection between national contexts brought by this transportation technology leads to the diffusion of standards and the standardisation of regulations at a global

[40] Clare, "Learning from Incidents."
[41] Moraglio, "Knights of Death."
[42] Clarsen, "Mobile Encounters."
[43] Reilly, "St. Petersburg-Tampa Airboat Line," 4.
[44] Lawrenson, "Regulation or Criminalisation," 251–262; Lagos, "Analysis of the Effect"; Wolfe, *Aviation Industry Regulation*; Valdés, "Learning from Accidents," 786–799.
[45] Reason, "Errors and Violations."
[46] Cookson, "Zagreb and Tenerife."
[47] McCreary, "Human Factors."
[48] Weick, "Vulnerable System."
[49] Cookson, "Zagreb and Tenerife."
[50] Veale, "Demystifying the Draft"; De Sanctis, "Artificial Intelligence."

level. With the diffusion of AI and the creation of cross-border bodies regulating its use, it is plausible that a certain amount of standardisation of regulatory regimes will happen.[51]

Another pattern worth mentioning relates to the phenomenon of regulatory capture. This concept describes the practice of powerful industries that come to dominate the agencies that regulate them.[52] This may happen for various reasons, but it often occurs because of an information imbalance that leaves the regulators dependent on their charges. It is common for organisations developing high-risk technologies to have an active role in their own regulation because they alone possess the necessary technical knowledge. As acknowledged by Downer,[53] in the case of the aviation industry, the FAA and other countries' similar bodies rely on aviation industry experts and engineers to perform a variety of functions, including overseeing tasks such as pilot tests, medical examinations and airworthiness assessments. As worrying as it is, this type of practice may be utterly performative in contexts like the aviation industry, where there is an alignment between regulatory bodies and companies in terms of safety standards and performance requirements. Indeed, unlike the shipping industry—where comprehensive insurance and elaborate bureaucratic prophylactics protect shipping companies from disasters at sea—aviation safety is strongly linked to profitability for companies.[54] Even in the AI industry, such alignment between regulatory bodies and the industry's safety interests is possible, as incidents can impact companies' profits. Therefore, it is plausible that we will register a diffusion of regulatory capture phenomenon in the AI context as well.[55]

Also in high-risk industries, incidents support the information gathering on risks, the establishment of standards and the proposition of new mandates for regulatory organisations.[56] Take into consideration the case of the Piper Alpha (United Kingdom [UK], 1988) oil industry incident. The incident resulted from the condensation of a leak on a pump, causing an explosion and a fire. This failure was mainly due to a lack of communication between shifts. Piper Alpha was also a hub for several other production facilities. Feeding from these continued, and this escalated the fire.[57] The incident resulted in 167 deaths with a total insured loss of about 1.7 billion sterling. The offshore incident resulted in a stronger, more independent regulatory regime in the UK inspired by the Norwegian risk regulation regime within the petroleum industry. The responsibility for safety on the UK continental shelf was transferred from the Department of Energy to the Health and Safety Executive to avoid goal conflicts between safety and production objectives. In the new regulatory regime, all offshore facilities needed to conduct a safety case, based on risk analysis.[58]

Another emblematic case of amendment of high industry regulations in the aftermath of an incident is the European Union (EU) Seveso directive.[59] The Seveso directive, which aimed to improve the safety of industries using large quantities of dangerous substances, received a fundamental push forward to its approval in the aftermath of the Seveso disaster. The Seveso disaster was an industrial incident that occurred in 1976 in a small chemical manufacturing plant in northern Italy and resulted in the highest known exposure to 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) in residential populations.[60]

Scholars focusing on the relationship between risks and regulation of high industry put also in evidence that as production and design disintegrate and become more collaborative–involving numerous and dynamic relations between customers and firms and characterised by complex subsystems and services–the production methods become more innovative but also more hazardous.[61] This phenomenon may also affect the production of AI technology in the future. In such a context, regulators must address the problem of uncertainty by inducing firms to systematically check their practices and identify potential hazards. Also, regulators must foster the institutionalisation of incident reporting procedures, including systems to register failures in products or production.[62] These strategies help to trace and correct incident root causes, alert others in similar situations to the potential risks and make certain that the countermeasures to ensure the safety of current operations are taken and the design requirements for future production are updated accordingly.[63]

[51] Chance, "Effect of Aviation Disasters"; Zielke, "Is Artificial Intelligence Ready."
[52] Peltzman, "Toward"; Posner, "Social Costs"; Carpenter, Preventing Regulatory Capture.
[53] Downer, "Trust and Technology."
[54] Cobb, The Plane Truth.
[55] Cihon, "Should Artificial Intelligence."
[56] Amiri, "Pattern Extraction."
[57] Zhen, "Quantitative Risk Modelling."
[58] Dahle, "Major Accidents."
[59] European Council Directive 96/82/EC, 1996.
[60] De Marchi, "Seveso."
[61] Sabel, "Regulation under Uncertainty."
[62] Coglianese, "Meta-Regulation," 12–13.
[63] Gilad, "It Runs in the Family."

AI and Its Regulation: The Ethification Phenomenon and the AI Act

Given the increasing diffusion of AI, the empirical evidence of incidents that can provide information on risks and modality of AI functioning is still limited, but it can plausibly expand in the future.

Despite this, as has happened for other high technologies, the fear of safety and technological risks and the uncertainties related to the application of AI in the field encourage a rush towards AI regulation even in the absence of all necessary information necessary to foresee the real risks of the new technology. This process is even more anxious for AI because this technological evolution refers to technologies that can act as intelligent agents and autonomously on the basis of data and perceptions received from the external environment.[64] Moreover, autonomous technologies may have relevant implications for the contexts in which they are applied. For instance, these regard the use of data, the protection of privacy, the responsibility and accountability of systems, their reliability as well as compliance with fundamental human rights principles and the rule of law.[65]

Two phenomena are related to this regulatory rush as a reaction to AI's uncertainties and risks. On the one hand, there is a proliferation of soft laws[66] in the form of normative frameworks, guidelines and collections of ethical principles disciplining the application of AI in different contexts.[67] On the other hand, national and supranational legislative institutions are defining and drafting the legislative frameworks that may regulate the use of AI in different contexts in line with human rights and previous fundamental laws.[68]

The phenomenon of drafting ethical framework documents has been termed the 'ethification phenomenon',[69] and it is associated with the growing importance of ethical expertise, ethical committees, and ethical advisory groups and boards. The 'algorithmwatch' list,[70] by identifying and collecting a list of 163 ethical documents drafted by different types of actors and in different languages, acknowledges the entity of AI ethification phenomenon. The diffusion of ethical guidelines responds to the need of rapidly regulating a new technology in fast and constant evolution as AI. While drafting hard laws is more rigid, time consuming and may lag behind technological development, drafting ethical documents is a flexible practice to cope with emerging technologies.[71] The downside is that ethical documents are soft-law tools[72] without binding force. In this sense, ethics may be a regulatory tool favourable to those actors who have no interest in having their behaviour regulated given that 'ethics has no teeth'[73] (i.e., it lacks enforcement methods). Drafting ethical guidelines can represent a means for going beyond, ignoring or avoiding the existing legal frameworks or for ensuring that AI will not be regulated by law: a phenomenon denominated 'ethics washing'.[74] Despite this, AI ethical documents should not be underestimated. On the one hand, they provide a first form of regulation in a regulatory context that has not yet fully addressed the issue of the implications of AI use; on the other hand, they may anticipate the 'proto-constitutional discourse'[75] that leads to the crystallisation of comprehensive and binding laws. Ethical guidelines, given that they are often drafted by actors who have practical experience of the application context to be regulated, may represent a form of attention to reality as it evolves, thus, giving a considerable contribution to law-making.[76]

As mentioned, to pursue the objectives of this study relating to data on AI soft laws, I will take into account the results of my research on ethical frameworks described in 'Ethics of Artificial Intelligence'.[77] The study analysed a set of ethical documents using content analysis techniques with the objective of clarifying on which ethical principles and risk factors the documents converge: 108 documents have been manually coded based on their reference to ethical principles or issues related to the application of AI. The main result of the study is that ethical documents converge to a set of principles and issues related to AI

[64] Russell, Artificial Intelligence; Santosuosso, "Intelligenza Artificiale."
[65] Lupo, "Regulating (Artificial) Intelligence."
[66] The definition of soft law is highly debated, with some scholars even denying the notion and considering it as illogical and redundant (Klabbers, "Redundancy of Soft Law," 167; Dawson, "Soft Law"). An in-depth analysis of the debate on soft law is out of the scope of this paper. For the sake of the argument, in this paper, I utilize the term "soft law" to refer to quasi-legal instruments drafted by public and private bodies that do not have any legally binding force, or whose binding force is weaker than the binding force of traditional law. In contrast, I use "hard law" to describe traditional laws as constitutions or international treaties with binding force and that are authoritative and prescriptive (Handl, "Hard Look"; Boyle, "Soft Law"; Christians, "Hard Law," 1049).
[67] Lupo, "Ethics of Artificial Intelligence"; Van Dijk, "Ethification" of Privacy.
[68] Atabekov, "Legal Status"; Weaver, "America's First AI Legislation," 201; Simbeck, "FAccT-Check on AI Regulation."
[69] Contini, "Artificial Intelligence," 4.
[70] Algorithmwatch, "AI Ethics."
[71] Van Dijk, "Ethification" of Privacy.
[72] Floridi, "Soft Ethics."
[73] Rességuier, "AI Ethics."
[74] Wagner, "Ethics as an Escape"; Lohr, "Legal Practitioners' Approach"; Daly, "AI Ethics."
[75] Gill, Towards Digital Constitutionalism.
[76] Daly, "AI Ethics."
[77] Lupo, "Ethics of Artificial Intelligence."

applications. As shown in Figure 1, the principles and risks of 'transparency', 'no discrimination', 'assessment', 'risk of harm', 'safety mechanisms', 'accountability', 'human rights' and 'judicial values' are mentioned in 60% or more of the documents investigated.[78] The mentioned analysis of framework documents has also acknowledged that scarce attention has been paid by public and private bodies that are drafting ethical guidelines to incident analysis operations with the aim of investigating AI impact. None of the 108 ethical guidelines analysed provide for the creation of a structured framework for incident analysis that includes governance, the body responsible for gathering and analysing incident data and incident assessment methods. Only three documents mentioned the incident investigation as a means for data gathering. For instance, the two documents *Top 10 Principles for Ethical Artificial Intelligence*[79] and *Toward a G20 Framework for Artificial Intelligence in the Workplace*[80] state that AI systems need to be transparent and accountable to incident investigators to make clear the internal processes that led to the incident. The document 'Preparing for the Future of Artificial Intelligence',[81] which focuses specifically on self-driving cars, provides for the reporting of data on incidents and near misses to improve AI in automotive testing and system safety.

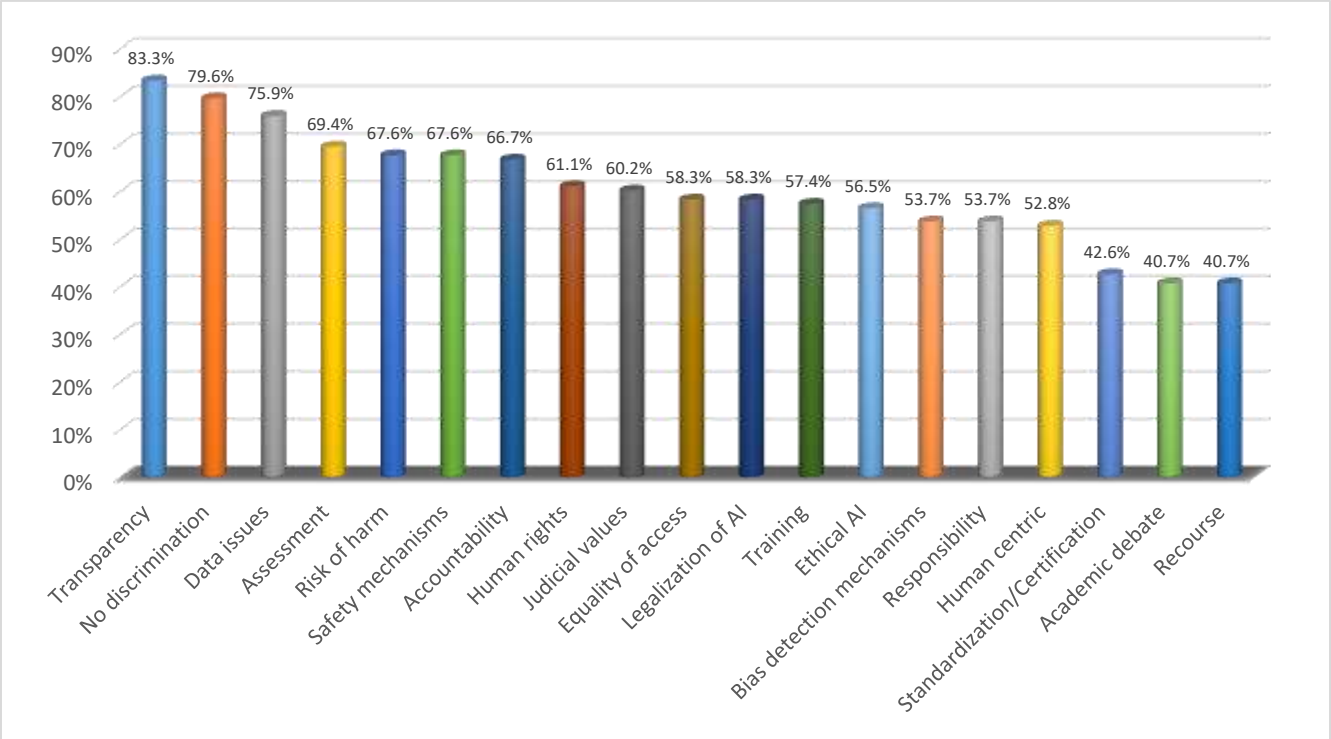

Figure 1: Percentage of documents in which an ethical principle or risk is present (first 20 items). Reproduced from Lupo (2022)[82]

With regards to the drafting of hard laws regulating AI, as mentioned, I focused on the case study of the EC proposal for a regulation on AI.[83] The regulation proposal was drafted in April 2021, and it has been under the scrutiny of the European Parliament (thousands of amendments were submitted) and the Council of the European Union.

The AI Act addresses the risks generated by specific uses of AI through a set of rules affecting developers and users. The legal framework for AI proposes an approach using three different levels of risk according to the type of AI technology: unacceptable risk, high risk and limited risk. The technologies included in the unacceptable risk list are prohibited in the EU or, if developed in a third country, cannot be used in the European Member States. The list of prohibited practices includes all AI systems whose use is considered unacceptable because they contravene EU values and violate fundamental rights.

[78] Lupo, "Ethics of Artificial Intelligence," 627.
[79] UNI Global Union, Top 10 Principles.
[80] Twomey, Toward a G20 framework.
[81] Bundy, "Preparing for the Future."
[82] Lupo, "Ethics of Artificial Intelligence," 627.
[83] European Commission, "Proposal for a Regulation of the European Parliament."

The high-risk AI technologies are included in annex III of the proposal, in a non-exhaustive list that, as stated in the document, the EC may expand within certain predefined areas by applying a set of criteria and risk assessment methodologies. For the high-risk AI category, the proposal foresees a set of provisions that have the aim of safeguarding the health and safety of EU citizens and the respect of EU fundamental rights as well as the EU acquis. Provisions regard data and data governance, documentation and record-keeping, transparency and provision of information to users, human oversight, robustness, accuracy and security. The AI Act also sets horizontal obligations on providers of high-risk systems to implement quality management systems, draft technical documentation of the high-risk AI system and automatically generate logs. Proportionate obligations are also placed on users and other participants across the AI value chain (e.g., importers, distributors and authorised representatives). Title IV of the AI Act addresses the high-risk systems that may pose specific risks of citizen manipulation (i.e., systems that interact with humans or are used to detect emotions). The Act provides for transparency obligations so that citizens are adequately informed when they are interacting with these types of systems.

The EU proposal also sets up a governance system responsible for the AI Act application, introducing a European Artificial Intelligence Board (the 'Board') at the EU level, composed of representatives from the Member States and the EC, and the composition of national competent authorities designated by the Member States at the national level. The AI Act provides monitoring obligations for the EC and national authorities through the establishment of an EU-wide database for high-risk AI systems with fundamental rights implications. A set of rules regarding monitoring and reporting also affect providers.

Even though the Act does not set compulsory rules for non-high-risk systems, it creates a framework for the creation of codes of conduct with the aim of encouraging providers to apply mandatory requirements for high-risk AI systems voluntarily.

Considering the absence of provisions regarding incident analysis in recent AI ethics guidelines, the EC position expressed in the AI Act towards this activity is groundbreaking. The EC AI Act states that technology providers are obliged to report any serious incident or any malfunctioning of a system that constitutes a breach of fundamental rights obligations as soon as they become aware of them. National competent authorities will investigate the incidents or malfunctioning, collect the necessary information and transmit it to the EC with adequate metadata. This information is collected by the EC, which will also conduct a comprehensive analysis of the overall market for AI systems that are like the one affected by a malfunction or an incident.

Opening the AI Black Box: Incidents and Regulations

The analysis of the evolution of high-tech regulation through incidents acknowledges that incident investigation helps to open the 'black box'[84] of complex technologies' functioning in the field, clarifies their impact on individuals, society and the environment and demonstrates their weak points and safety issues.

Also for AI, the practice of incident analysis to learn from mistakes is consolidating. Several actors (e.g., research institutes and experts' associations) are listing and categorising the unwanted and harmful happenings that involve AI technologies and their application in different contexts.

The databases selected for the analysis—Where in the World is AI? Map,[85] AI Incident Database[86] and AIAAIC repository[87]—are among the most searched and utilised online repositories of AI incidents that are publicly available. This study focused on these three repositories to assess these tools' capacity to gather and provide information on AI functioning and related issues and investigate the link between AI incidents and AI laws.

The methodology section already described in depth the three databases. The three databases are open resources based on the voluntary contributions of the public that report and input data on AI incidents. Data are displayed in different modalities as maps or datasheets. In all three cases, experts from the non-profit organisations that are responsible for the databases, check and review public inputs to ensure the correctness of information and coherence in terms of incidents' description and classification. The three databases are utilised mainly by ICT (Information and Communication Technology) system designers, industrial product developers, public relations managers, researchers and public policy researchers. They all have the declared objective of providing information on AI risks and the nature and opacity of AI through incident data gathering.

The first result of the analysis of AI incident repositories worth mentioning concerns the database Where in the World is AI? Map (2020). Unlike the other databases investigated, this database reports not only incidents but also general information

[84] Rai, "Explainable AI," 137–141.
[85] AI Global, "Where in the World."
[86] Responsible AI, "Responsible Artificial Intelligence."
[87] AIID, "AI Incident Database." McGregor, "Preventing."

regarding AI to define the diffusion of these systems in the world. The repository categorised 323 out of 430 (75%) news articles investigated as 'harmful' AI, while only 22% were categorised as beneficial AI. This data may confirm that the media's attitude to focus mainly on 'negative' news[88] may affect AI applications too. The media's approach may contribute to the diffusion of a diffident attitude towards the use of AI that may also influence the evaluation of experts or policymakers involved in law-making. My analysis focusing on the ethical documents disciplining AI[89] confirms this diffusion of diffidence and concern acknowledging that only 37% of the 108 documents investigated include sentences indicating potential positive outcomes caused by AI use as economic and wellbeing improvement. Additionally, the Where in the World is AI? Map database analysis also suggests the importance of gathering data not only on harmful and opaque AI but also on the investigation of relevant examples of AI beneficial for their context of application and on best practices. The strategy of best practice analysis is also present in the AI Act proposed by the EC,[90] and it is also diffused in the ethical guidelines analysed in my study on the ethics of AI:[91] the study demonstrated that 32.4% of 108 documents investigated provide for a best practice analysis strategy.

The analysis of the AI Incident Database allows to focus on the data regarding the types of AI technologies that are more involved in unwanted occurrences. Table 1 shows the distribution of AI incidents registered in the AI Incident Database in terms of the type of technology. The data show that the systems more often involved in incidents are image elaboration systems (9.40%), autonomous vehicles (8.72%) and natural language processing (8.72%).[92]

[88] Patterson, "Bad News," 17–20.
[89] Lupo, "Ethics of Artificial Intelligence."
[90] The EC proposal provides for the creation of a European Artificial Intelligence Board (the "Board"), composed of representatives from the Member States and the Commission. Aside from facilitating a smooth, effective, and harmonized implementation of the regulation by contributing to the effective cooperation of the national supervisory authorities and the Commission, the Board will also collect, analyze, and share best practices among the Member States.
[91] Lupo, "Ethics of Artificial Intelligence."
[92] Excluding the generic category "Other" (20.8%).

Table 1: Distribution of incidents by type of AI technology

Type	Percentage
Image elaboration	9.40%
Autonomous vehicles	8.72%
Natural language processing	8.72%
Decision-making	8.05%
Facial recognition	8.05%
Recommendation engine	6.04%
Content manipulation	5.37%
Environmental sensing	4.03%
Data analytics	3.36%
Biometrics	2.01%
Chatbot	2.01%
Statistical projection	2.01%
Voice recognition	2.01%
Forecasting	1.34%
Interpreting traffic patterns	1.34%
Procedural content generation	1.34%
Risk assessment	1.34%
Robotics	1.34%
Speech recognition	1.34%
Virtual assistant	1.34%
Other	20.81%

Source: AI Incident Database.

This distinction in terms of risks related to the different types of AI technologies is also present in the EU's first attempt at AI regulation. The EC AI Act indicates a list of high-risk technologies that must be subjected to stricter norms. The EC list (see Table 2) is very inclusive and it reflects an extremely cautious attitude towards AI technologies and their risks even before the setting of an incident data gathering and analysis strategy that may help to quantify and assess AI risks. It is interesting to note that some technologies that the AI Incident Database acknowledged as risky and subject to incidents (e.g., autonomous vehicles) are not listed in the high-risk category of the EC.

Table 2: High-risk technologies identified in the European Commission proposal for AI regulation

1	Biometric identification	10	Assessment of the emotional state of a natural person
2	Critical infrastructure management	11	Detect deep fakes
3	Education and vocational training (access)	12	Evaluation of the reliability of evidence
4	Education and vocational training (assessment)	13	Profiling of natural persons in law enforcement
5	Employment and worker management	14	Crime analytics
6	Eligibility assessment of persons for public (and private) services	15	Law and case law examination
7	Creditworthiness assessment	16	Risk assessment migration
8	Emergency first response (services eligibility)	17	Authenticity of travel documents (assess)
9	Criminal risk assessment	18	Application for asylum examination

Source: European Commission proposal for AI regulation.[93]

Focusing on the AIAAIC repository, it is possible to investigate the different types of sectors most affected by AI incidents. Figure 2 shows the distribution of incidents collected in the AIAAIC repository by sector of AI application. The data demonstrate that aside from the generic technological sector, which coherently registers the higher rate of incidents, the second sector most affected by incidents is the government sector (21.54%). For instance, an example of an incident regarding AI applied in governmental services (immigration policies) is the recent case of facial recognition technology utilised on members of the public without consent by Government of Canada immigration officials at Toronto Pearson International Airport in 2016.[94] The high risk registered for AI applied in a governmental context does not reflect the diffused concern of private and public bodies trying to regulate AI through soft laws. The AI ethical documents study previously quoted acknowledged that only 1.85% of the documents investigated address AI applied in public administration.[95] This is not the case with the AI Act proposed by the EC, which categorised several types of AI applied in governmental services and operations in the high-risk category, including crime analytics AI, systems for law and case law examination, and AI utilised for the assessment of a person's eligibility for public services. It is not possible in this paper to assess how much the empirical reality of AI incidents in the government sector has influenced the EC strategy. However, it is plausible that highly publicised events may have somehow affected the inclusion of some systems in the high-risk category. For example, the COMPAS (a risk assessment system used in different United States jurisdictions to evaluate alternative measures of detention and the relative risk of recidivism of convicts)[96] case has been largely reported by the media and in academic publications, which have focused on the various problems in terms of discrimination bias that characterised the system. In particular, COMPAS has been accused of associating high risks of recidivism with belonging to an ethnic minority.[97] This case and the relative academic and public debate may be part of the motivations that encouraged the EC to include criminal risk assessment systems in the high-risk category.

[93] European Commission, "Proposal for a Regulation of the European Parliament."
[94] Cardoso, "Ottawa Tested."
[95] Lupo, "Ethics of Artificial Intelligence."
[96] Blomberg, *Validation of the COMPAS.*
[97] Hong, "Racism," 79–84.

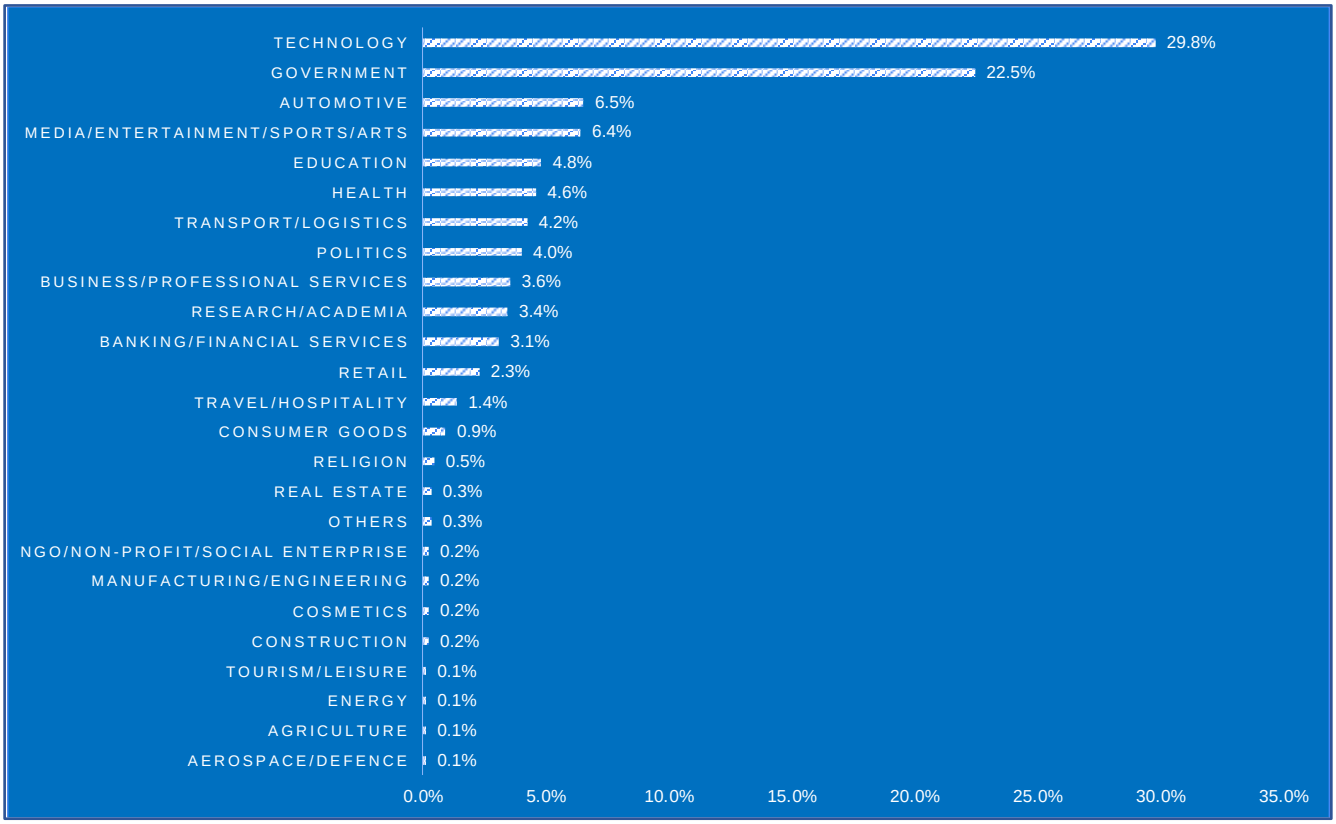

Figure 2. Distribution of incidents by sector of AI application[98]

The results relative to the analysis of AIAAIC data on AI incidents in the different government sectors are utterly interesting (see Table A.1 in the Appendix). The data show that most incidents take place in the police sector (39.8%), while there are few incidents in other sectors, such as the justice sector (11 incidents, 5.6%), probably due to the scarce diffusion of AI in this area. Despite this, the EC AI Act strictly regulates different types of AI systems applied in the judiciary by including them in the high-risk category. This shows an evident concern, partially corroborated by the empirical reality of incidents, towards such systems.

The elaboration of AIAAIC data also allows a discussion on the types of incidents that more diffusely affect AI (see Figure 3). The data confirm that incidents principally involved the reliability of systems (18.6%), respect of privacy and protection of personal data (17.7%), different types of discrimination (12.2%), safety from harms (9.5%) and issues related to surveillance (7.7%). Interestingly, the issues mentioned are also largely quoted and disciplined in the ethical frameworks disciplining AI: reliability is quoted in 68%, discrimination in 80% and transparency in 83% of the 108 documents investigated in my previous study on ethical documents.[99]

[98] AIID, "AI Incident Database."
[99] Lupo, "Ethics of Artificial Intelligence."

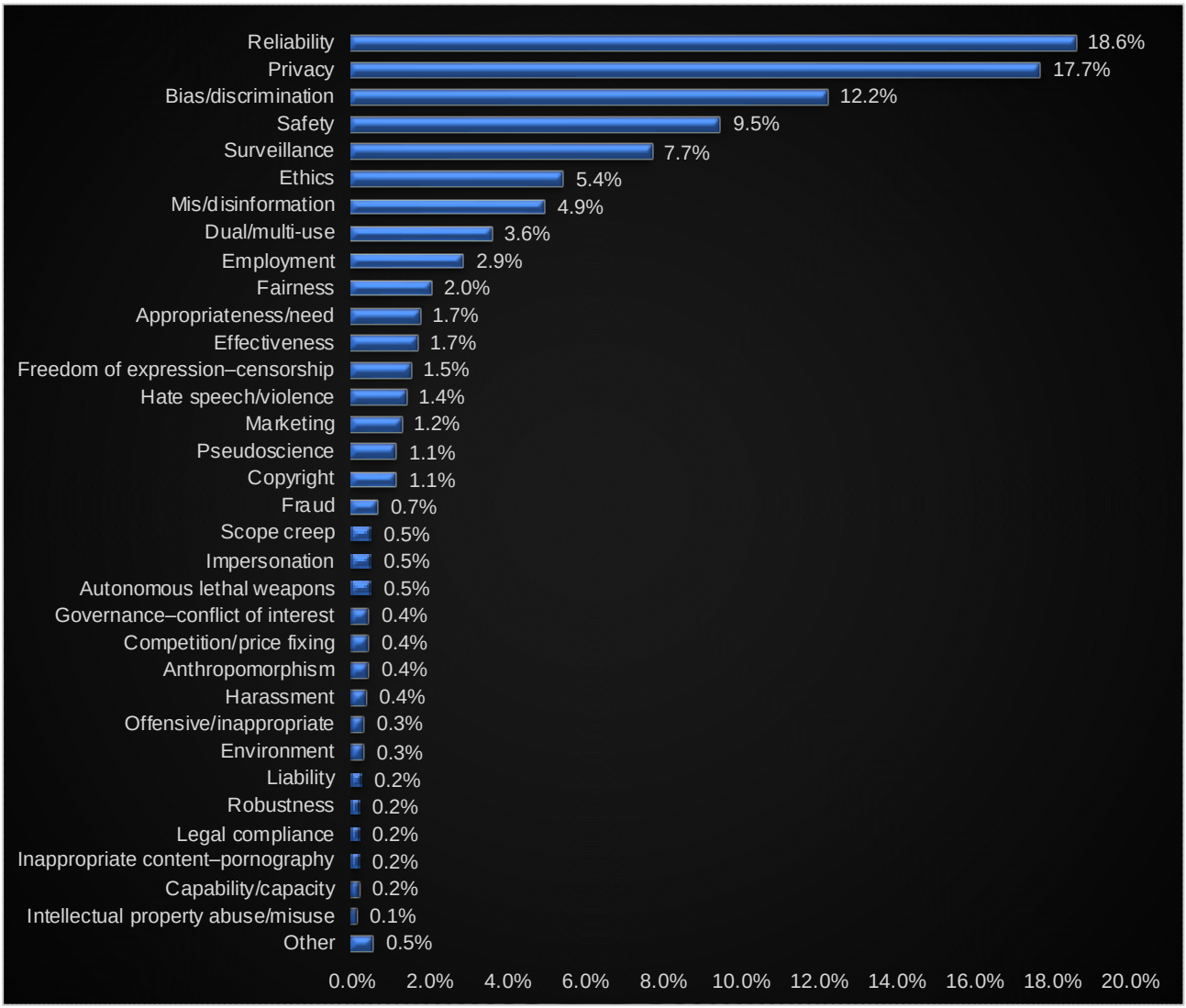

Figure 3: Distribution of types of AI incidents[100]

Despite the high number of incidents regarding illegitimate surveillance (129 of the 871 incidents registered by AIAAIC), the study on ethical documents[101] has shown drafting bodies paid relatively little attention to this issue: only 13% of the AI documents investigated focused on the topic of avoiding or strictly regulating AI based surveillance systems. The types of incidents registered in the AIAAIC database are also mentioned in the EC proposal for regulating AI. The AI Act seeks to limit risky AI implications, including issues related to the reliability of systems, discrimination bias, transparency, and protection of privacy. Also, the EC proposal does not mention issues related to the use of AI for surveillance, probably confirming an ambivalent attitude of public (and private) institutions that consider AI surveillance technologies potentially useful for ensuring security and supporting home affairs operations.[102]

The analysis of AIAAIC data also allows a discussion on the cross-sectorial quality of AI risks. The data illustrate that the same incidents and issues interest different application contexts. Consider the three more-diffused types of AI incidents based on the AIAAIC data: reliability incidents, privacy incidents and discrimination incidents. Reliability and privacy incidents affected 65% of the 25 sectors that used AI applications, and discrimination incidents affected 56% of sectors. Considering discrimination bias as an example, these types of incidents affected very different contexts of AI application: government (64 incidents), automotive (32 incidents) and health care (15 incidents). The cross-sectorial quality of AI risks may facilitate data

[100] AIID, "AI Incident Database."
[101] Lupo, "Ethics of Artificial Intelligence," 636.
[102] Veale, "Demystifying the Draft"; Almeida, "Ethics of Facial Recognition," 377–387.

gathering on AI incidents and implications and the regulation of AI for those areas where AI is not considerably diffused, such as the justice sector.

The analysis of incident databases presented in this paper demonstrated that these types of repositories can be very useful in gathering and providing data on AI risks. The datasets investigated provide valuable content on attributes of AI systems involved in incidents, such as the sector of deployment, type of technology and geolocation of systems. These data help to understand the real impact of systems on citizens, their fundamental rights, the environment and other technologies. The analysis of these repositories is potentially useful for developers and policymakers and can be the basis of AI regulation.

It is worth emphasising that the quality and reliability of publicly available databases are moderate. The databases suffer from the arbitrariness of data input, input assessment, AI incident classifications and a lack of external accountability.

As mentioned, the AI Act provides for a strategy of incident data gathering and analysis that involves AI providers, national competent authorities and the EC. It is plausible that this method of incident analysis will provide more accountable and high-quality data useful for understanding the impact of AI and opening the 'black box'.

The analysis confirms the link between high-tech regulations and diffused information on incidents in the AI sector and provides evidence of some gaps in terms of AI risk limitation affecting actual attempts to regulate AI through soft laws or 'hard' regulations (such as the AI Act). AI technology is undergoing ever-greater development and diffusion and is evolving rapidly; the related risks are also in an evolutionary phase. Therefore, legislation must be flexible and adaptable and evolve rapidly and with technology. The strategy of the AI Act seems to follow this direction, for instance, by establishing the governance and method for the rapid modification of the high-risk list based on AI risk assessment.

Concluding Remarks

This study confirms the importance of creating an effective strategy of incident investigation for analysing AI impact in several contexts and supporting the approval of effective AI laws that are in line with technological evolution. This also affects those areas of AI application characterised by scarce diffusion of systems and, therefore, by a scarcity of empirical data on harms and damages. The analysis confirmed that the risks related to AI use are often cross-sectorial; that is, they interest different contexts of application. Thus, the analysis of incidents in different sectors may facilitate the regulation of AI in areas where AI is not considerably diffused, such as the justice sector.

Currently, AI incident data are collected only by private initiatives, with several flaws, such as the lack of accountability for who is responsible for inputting, categorising and monitoring the data. With the future approval of the EC AI Act, there will be a working procedure for collecting information on incidents and investigating them that involves AI providers and public bodies at the national and EU level. The future will tell us how much this strategy will help us understand AI, open the black box and support the development and use of a more trustworthy and responsible AI.

Bibliography

Adams, James D. "Industrial R&D Laboratories: Windows on Black Boxes?" In *Essays in Honor of Edwin Mansfield: The Economics of R&D, Innovation, and Technological Change*, edited by Albert N. Link and F. M. Scherer, 99–107. Boston: Springer, 2005.

AIAAIC. "Understanding the Risks and Harms of AI, Algorithms, and Automation." https://www.aiaaic.org/.

AI Global. "Where in the World is AI?" https://map.ai-global.org/.

AIID. "AI Incident Database." https://incidentdatabase.ai.

AIID. "Founding Report." https://incidentdatabase.ai.

Algorithmwatch. "AI Ethics Guidelines Global Inventory." https://inventory.algorithmwatch.org/.

Almeida, Denise, Konstantin Shmarko, and Elizabeth Lomas. "The Ethics of Facial Recognition Technologies, Surveillance, and Accountability in an Age of Artificial Intelligence: A Comparative Analysis of US, EU, and UK Regulatory Frameworks." *AI and Ethics* 2, no 3 (2022): 377–387. https://doi.org/10.1007/s43681-021-00077-w.

Amiri, Mehran, Abdollah Ardeshir, Mohammad Hossein Fazel Zarandi, and Elahe Soltanaghaei. "Pattern Extraction for High-Risk Accidents in the Construction Industry: A Data-Mining Approach." *International Journal of Injury Control and Safety Promotion* 23, no 3 (2016): 264–276. https://doi.org/10.1080/17457300.2015.1032979.

Angelides, Steven. "Disorder as "Pseudo-Idea". *Atlantis: Critical Studies in Gender, Culture & Social Justice* 35.2 (2012): 10-20.

Atabekov, Atabek and Oleg Yastrebov. "Legal Status of Artificial Intelligence across Countries: Legislation on the Move." *European Research Studies* 21, no 4 (2018): 773–782.

Baškarada, Saša and Andy Koronios. "A Philosophical Discussion of Qualitative, Quantitative, and Mixed Methods Research in Social Science." *Qualitative Research Journal* 18, no 1(2018): 2–21. https://doi.org/10.1108/QRJ-D-17-00042.

Blomberg, Thomas, William Bales, Karen Mann, Ryan Meldrum, and Joe Nedelec. *Validation of the COMPAS Risk Assessment Classification Instrument*. Tallahassee: College of Criminology and Criminal Justice, Florida State University, 2010.

Boyle, Alan. "Soft Law in International Law-Making." In *International Law*, edited by Malcolm D. Evans, 119–137. Oxford: Oxford University Press, 2014.

Bundy, Alan. "Preparing for the Future of Artificial Intelligence." *AI & Society* 32, no 2 (2017): 285–287. https://doi.org/10.1007/s00146-016-0685-0.

Cammarano, Antonello, Vincenzo Varriale, Francesca Michelino, and Mauro Caputo. "The Importance of Possessing Knowledge on Black-Box Components: The Case of Smartphone OEMs." *Journal of Engineering and Technology Management* 67 (2023): 101727. https://doi.org/10.1016/j.jengtecman.2022.101727.

Canguilhem, Georges. *The Normal and the Pathological*. New York: Zone Books, 1991.

Cardoso, Tom and Colin Freeze, "Ottawa Tested Facial Recognition on Millions of Travellers at Toronto's Pearson Airport in 2016." *The Globe and Mail*, July 19, 2021. https://www.theglobeandmail.com/canada/article-ottawa-tested-facial-recognition-on-millions-of-travellers-at-torontos/.

Carpenter, Daniel and David A. Moss. *Preventing Regulatory Capture: Special Interest Influence and How to Limit It*. New York: Cambridge University Press, 2013.

Chance, Don M. and Stephen P. Ferris. "The Effect of Aviation Disasters on the Air Transport Industry: A Financial Market Perspective." *Journal of Transport Economics and Policy* 21, no 2 (1987): 151–165.

Choi, Kanghwa, Ram Narasimhan, and Soo Wook Kim. "Opening the Technological Innovation Black Box: The Case of the Electronics Industry in Korea." *European Journal of Operational Research* 250, no 1 (2016): 192–203.

Christians, Allison. "Hard Law & Soft Law in International Taxation." *Wisconsin International Law Journal* 25, no 2 (2007): 1049.

Cihon, Peter, Matthijs M. Maas, and Luke Kemp. "Should Artificial Intelligence Governance Be Centralised? Design Lessons from History." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 228–234. New York: ACM, 2020.

Clare, James and Kyriakos I. Kourousis. "Learning from Incidents in Aircraft Maintenance and Continuing Airworthiness: Regulation, Practice and Gaps." *Aircraft Engineering and Aerospace Technology* 93, no 2 (2021): 338–346.

Clarsen, Georgine. "Mobile Encounters: Bicycles, Cars and Australian Settler Colonialism." *History Australia* 12, no 1 (2015): 165–186. https://doi.org/10.1080/14490854.2015.11668558.

Cobb, Roger W. and David M. Primo. *The Plane Truth: Airline Crashes, the Media, and Transportation Policy*. Washington, DC: Brookings Institution Press, 2004.

Coglianese, Cary and Evan Mendelson. "Meta-Regulation and Self-Regulation." In *The Oxford Handbook of Regulation*, edited by Martin Cave, Robert Baldwin, and Martin Lodge, 145–168. Oxford: Oxford University Press, 2010.

Contini, Francesco. "Artificial Intelligence and the Transformation of Humans, Law and Technology Interactions in Judicial Proceedings." *Law, Technology and Humans* 2, no 1 (2020): 4–18. https://doi.org/10.5204/lthj.v2i1.1478.

Cookson, Simon. "Zagreb and Tenerife: Airline Accidents Involving Linguistic Factors." *Australian Review of Applied Linguistics* 32, no 3 (2009): 22.1–22.14.

Cooter, Roger and Bill Luckin, eds. *Accidents in History: Injuries, Fatalities and Social Relations*. 41: Amsterdam: Rodopi, 1997.
Coutin, Susan Bibler. "Qualitative Research in Law and Social Sciences." *Arts, Social Sciences* 1 no 1 (2012): 50.
Dahle, I. B., G. Dybvig, G. Ersdal, T. Guldbrandsen, B. A. Hanson, J. E. Tharaldsen, and A. S. Wiig. "Major Accidents and Their Consequences for Risk Regulation." In *Advances in Safety, Reliability and Risk Management*, edited by Grall Bérenguer and Guedes Soares, 33–41. London: CRC Press, 2012.
Daly, Angela, S. Kate Devitt, and Monique Mann. "AI Ethics Needs Good Data." Submitted February 15, 2021. https://arxiv.org/abs/2102.07333v1.
Dawson, Mark. *Soft Law and the Rule of Law in the European Union: Revision or Redundancy?* EUI Working Papers RSCAS 2009/24. Italy: European University Institute, 2009.
De Marchi, Bruna. "Seveso: From Pollution to Regulation." *International Journal of Environment and Pollution* 7, no 4 (1997): 526–37. https://dx.doi.org/10.1504/IJEP.1997.028318.
De Sanctis, Fausto Martin. "Artificial Intelligence and Innovation in Brazilian Justice." *International Annals of Criminology* 59, no 1 (2021): 1–10. https://doi.org/10.1017/cri.2021.4.
Downer, John. "Trust and Technology: The Social Foundations of Aviation Regulation." *The British Journal of Sociology* 61, no 1 (2010): 83–106. https://doi.org/10.1111/j.1468-4446.2009.01303.x.
Esbester, Mike and Jameson M. Wetmore. "Introduction: Global Perspectives on Road Safety History." *Technology and Culture* 56, no 2 (2015): 307–318. https://doi.org/10.1353/tech.2015.0059.
European Commission. "Proposal for a Regulation of the European Parliament and of Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts," COM/2021/206 final.
Floridi, Luciano. "Soft Ethics and the Governance of the Digital." *Philosophy & Technology* 31 (2018): 1–8. https://doi.org/10.1007/s13347-018-0303-9.
Foucault, Michel. "Historia de la Medicalización." *Educación Médica y Salud* 11, no 1 (1977): 3–25.
French, Simon. "Aggregating Expert Judgement." *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales. Serie A. Matemáticas* 105, no 1 (2011): 181–206.
Gilad, Sharon. "It Runs in the Family: Meta-Regulation and Its Siblings." *Regulation & Governance* 4, no 4 (2010): 485–506. https://doi.org/10.1111/j.1748-5991.2010.01090.x.
Gill, Lex, Dennis Redeker, and Urs Gasser. *Towards Digital Constitutionalism? Mapping Attempts to Craft an Internet Bill of Rights*. Berkman Klein Center for Internet & Society Research Publication 2015-15. Cambridge, MA: Harvard University, 2015.
Gorard, Stephen. *Quantitative Methods in Social Science Research.* New York: Continuum, 2003.
Handl, Gunther F, W. Michael Reisman, Bruno Simma, Pierre Marie Dupuy, and Christine Chinkin. "A Hard Look at Soft Law." In *Proceedings of the Annual Meeting (American Society of International Law)*, 371–395. Cambridge: Cambridge University Press, 1988.
Hong, Joo-Wha and Dmitri Williams. "Racism, Responsibility and Autonomy in HCI: Testing Perceptions of an AI Agent." *Computers in Human Behavior* 100 (2019): 79–84. https://doi.org/10.1016/j.chb.2019.06.012.
Hutter, Bridget and Michael Power. *Organizational Encounters with Risk.* Cambridge: Cambridge University Press, 2005.
Jung, James J., Peter Jüni, Gerald Lebovic, and Teodor Grantcharov. "First-Year Analysis of the Operating Room Black Box Study." *Annals of Surgery* 271, no 1 (2020): 122–27.
Kasperson, Jeanne X., Roger E. Kasperson, Nick Pidgeon, and Paul Slovic. "The Social Amplification of Risk: Assessing Fifteen Years of Research and Theory." In *The Feeling of Risk: New Perspectives on Risk Perception*, edited by Paul Slovic, 345–372. Hoboken: Taylor and Francis, 2013.
Katyal, Sonia K. "The Paradox of Source Code Secrecy." *Cornell Law Review* 104 (2018): 1183.
Klabbers, Jan. "The Redundancy of Soft Law." *Nordic Journal International Law* 65, no 2 (1996): 167–182.
Kowalick, Thomas M. *Fatal Exit: The Automotive Black Box Debate.* Hoboken: John Wiley & Sons, 2005.
Lagos, Alfredo, Majid Motevalli, Vahid Motevalli, and Nobuyo Sakata. "Analysis of the Effect of Milestone Aviation Accidents on Safety Policy, Regulation, and Technology." Paper presented at the 46th Annual Transportation Research Forum, Washington, DC, March 6–8, 2005.
Lanzara, Giovan Francesco. *Capacità Negativa: Competenza Progettuale e Modelli di Intervento Nelle Organizzazioni.* Bolonga: Il Mulino, 1993.
Lawrenson, Anthony J. and Graham R. Braithwaite. "Regulation or Criminalisation: What Determines Legal Standards of Safety Culture in Commercial Aviation?" *Safety Science* 102 (2018): 251–262. https://doi.org/10.1016/j.ssci.2017.09.024.
Lohr, Jason D., Winston J. Maxwell, and Peter Watts. "Legal Practitioners' Approach to Regulating AI Risks." In *Algorithmic Regulation*, edited by Karen Yeung and Martin Lodge, 224–247. Oxford: Oxford University Press, 2019.
Lupo, Giampiero. "The Ethics of Artificial Intelligence: An Analysis of Ethical Frameworks Disciplining AI in Justice and Other Contexts of Application." *Oñati Socio-Legal Series* 12, no 3 (2022): 614–653. https://doi.org/10.35295/osls.iisl/0000-0000-0000-1273.

Lupo, Giampiero. “Regulating (Artificial) Intelligence in Justice: How Normative Frameworks Protect Citizens from the Risks Related to AI Use in the Judiciary.” *European Quarterly of Political Attitudes and Mentalities* 8, no 2 (2019): 75–96.
Marabelli, Marco, Sean Hansen, Sue Newell, and Chiara Frigerio. “The Light and Dark Side of the Black Box: Sensor-Based Technology in the Automotive Industry.” *Communications of the Association for Information Systems* 40, no 1 (2017): 351–374. https://doi.org/10.17705/1CAIS.04016.
McCreary, John, Michael Pollard, Kenneth Stevenson, and Marc B. Wilson. “Human Factors: Tenerife Revisited.” *Journal of Air Transportation World Wide* 3, no 1 (1998): 23–32.
McGregor, Sean. “Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database.” In *Proceedings of the AAAI Conference on Artificial Intelligence*, 15458–15463. Palo Alto: AAAI Press, 2021.
Menell, Peter S. and Suzanne Scotchmer. “Intellectual Property Law.” In *Handbook of Law and Economics* Vol. 2, edited by A. Mitchell Polinsky and Steven Shavell, 1473–1570. Amsterdam: Elsevier, 2007.
Mitchell, Matthew. “Analyzing the Law Qualitatively.” *Qualitative Research Journal* 23, no 1 (2023): 102–113.
Moraglio, Massimo. “Knights of Death: Introducing Bicycles and Motor Vehicles to Turin, 1890–1907.” *Technology and Culture* 56, no 2 (2015): 370–393.
Norton, Peter. “Four Paradigms: Traffic Safety in the Twentieth-Century United States.” *Technology and Culture* 56, no 2 (2015): 319–334. https://doi.org/10.1353/tech.2015.0065.
Norton, Peter D. “Hell on Wheels: The Promise and Peril of America’s Car Culture, 1900–1940 by David Blanke.” *Michigan Historical Review* 34, no 2 (2008): 141–42.
Paré, Guy and Spyros Kitsiou. “Methods for Literature Reviews.” In *Handbook of eHealth Evaluation: An Evidence-Based Approach* [Internet], edited by Francis Lau and Craig Kuziemsky. Victoria, BC: University of Victoria, 2017.
Patterson, Thomas E. “Bad News, Period.” *PS: Political Science & Politics* 29, no 1 (1996): 17–20.
Peltzman, Sam. “Toward a More General Theory of Regulation.” *The Journal of Law and Economics* 19, no 2 (1976): 211–40. https://doi.org/10.1086/466865.
Perrow, Charles. “The Meltdown Was Not an Accident.” In *Markets on Trial: The Economic Sociology of the US Financial Crisis: Part A (Research in the Sociology of Organizations, Volume 30)*, edited by Michael Lounsbury and Paul M. Hirsch, 309–330. Bingley, UK: Emerald Group Publishing Limited, 2010.
Perrow, Charles. *Normal Accidents: Living with High-Risk Technologies.* Princeton: Princeton University Press, 1999.
Phillips, Mark. “Case Study: The Corporate Blame Game: Firestone Tires and the Ford Explorer.” *Christian Business Academy Review* 2 (2007): 67–74.
Pidgeon, Nick, and Baruch Fischhoff. "The Role of Social and Decision Sciences in Communicating Uncertain Climate Risks." *Nature Climate Change* 1, no 1 (2011): 35-41. https://doi.org/10.1038/nclimate1080.
Pinch, Trevor. “The Social Construction of Technology (SCOT): The Old, The New.” In *Material Culture and Technology in Everyday Life: Ethnographic Approaches*, Volume 25, edited by Phillip Vannini, 25-45. New York: Peter Lang, 2009.
Pole, Christopher John and Richard Lampard. *Practical Social Investigation: Qualitative and Quantitative Methods in Social Research.* Harlow, UK: Pearson Education, 2002.
Posner, Richard A. “The Social Costs of Monopoly and Regulation.” *Journal of Political Economy* 83, no 4 (1975): 807–28.
Rae, Andrew and Rob Alexander. “Forecasts or Fortune-Telling: When Are Expert Judgements of Safety Risk Valid?” *Safety Science* 99, part B (2017): 156–165. https://doi.org/10.1016/j.ssci.2017.02.018.
Rai, Arun. “Explainable AI: From Black Box to Glass Box.” *Journal of the Academy of Marketing Science* 48 (2020): 137–141. https://doi.org/10.1007/s11747-019-00710-5.
Reason, James, Antony Manstead, Stephen Stradling, James Baxter, and Karen Campbell. “Errors and Violations on the Roads: A Real Distinction?” *Ergonomics* 33, no 10–11 (1990): 1315–1332. https://doi.org/10.1080/00140139008925335.
Reilly, Thomas. “The St. Petersburg-Tampa Airboat Line: 90 Days That Changed the World of Aviation.” *Tampa Bay History* 18, no 2 (1996): 4.
Responsible AI. “Responsible Artificial Intelligence.” https://www.responsible.ai/.
Rességuier, Anaïs and Rowena Rodrigues. “AI Ethics Should Not Remain Toothless! A Call to Bring Back the Teeth of Ethics.” *Big Data & Society* 7, no 2 (2020): 1–5. https://doi.org/10.1177/2053951720942541.
Rosenberg, Nathan. *Inside the Black Box: Technology and Economics.* Cambridge: Cambridge University Press, 1982.
Rosenthal, Robert and M. Robin DiMatteo. “Meta-Analysis: Recent Developments in Quantitative Methods for Literature Reviews.” *Annual Review of Psychology* 52, no 1 (2001): 59–82. https://doi.org/10.1146/annurev.psych.52.1.59.
Russell, Stuart J. and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Harlow, UK: Pearson Education Limited, 2016.
Sabel, Charles, Gary Herrigel, and Peer Hull Kristensen. “Regulation Under Uncertainty: The Coevolution of Industry and Regulation.” *Regulation & Governance* 12, no 3 (2018): 371–394. https://doi.org/10.1111/rego.12146.
Santosuosso, Amedeo and Dianora Poletti. *Intelligenza Artificiale e Diritto. Perché le Tecnologie di IA Sono una Grande Opportunità per il Diritto*. Milan: Mondadori Università, 2020.
Simbeck, Katharina. “FAccT-Check on AI Regulation: Systematic Evaluation of AI Regulation on the Example of the Legislation on the Use of AI in the Public Sector in the German Federal State of Schleswig-Holstein.” In *FAccT ‘22: 2022*

ACM Conference on Fairness, Accountability, and Transparency, 89–96. New York, Association for Computing Machinery, 2022.
Skjong, Rolf and Benedikte H. Wentworth. "Expert Judgment and Risk Perception." In *The Proceedings of the Eleventh (2001) International Offshore and Polar Engineering Conference*, edited by Jin S. Chung, Mohamed Sayed, Hiroshi Saeki, and Toshiaki Setoguchi, 537–544. California: International Society of Offshore and Polar Engineers, 2001.
Snyder, Hannah. "Literature Review as a Research Methodology: An Overview and Guidelines." *Journal of Business Research* 104 (2019): 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039.
Taekema, Sanne. "Theoretical and Normative Frameworks for Legal Research: Putting Theory into Practice." *Law and Method* (2018).
Tan, Zhi Qin, Hao Shan Wong, and Chee Seng Chan. "An Embarrassingly Simple Approach for Intellectual Property Rights Protection on Recurrent Neural Networks." Submitted October 3, 2022. https://arxiv.org/abs/2210.00743v2.
Thomas, Jayan Jose. "An Uneasy Coexistence: The New and the Old in Indian Industry and Services." In *A New India: Critical Reflections in the Long Twentieth Century*, edited by Anthony D'Costa, 71–98. Cambridge: Cambridge University Press, 2010.
Tingvall, Claes. "The History of Traffic Safety: Describing 100 Years." *Technology and Culture* 56, no 2 (2015): 489–492.
Tschider, Charlotte A. "Beyond the Black Box." *Denver Law Review* 98 (2020): 683. https://doi.org/10.1353/tech.2015.0069.
Twomey, Paul, Olena Ivus, Joanna Wajda, Hector Torres, James M. Boughton, R. Andreas Kraemer, Malcolm D. Knight, Céline Bak, and Jorge Braga de Macedo. *Toward a G20 Framework for Artificial Intelligence in the Workplace.* Ottawa: Centre for International Governance Innovation, 2018.
UNI Global Union. *Top 10 Principles for Ethical Artificial Intelligence*. Nyon, Switzerland: UNI Global Union, 2017.
Valdés, Rosa María Arnaldo and Fernando Gómez Comendador. "Learning from Accidents: Updates of the European Regulation on the Investigation and Prevention of Accidents and Incidents in Civil Aviation." *Transport Policy* 18, no 6 (2011): 786–799. https://doi.org/10.1016/j.tranpol.2011.03.009.
Van Dijk, Niels and Simone Casiraghi. *The "Ethification" of Privacy and Data Protection in the European Union. The Case of Artificial Intelligence.* Brussels Privacy Hub Working Paper Vol. 6 No. 22. Brussels: Brussels Privacy Hub, 2020.
Veale, Michael and Frederik Zuiderveen Borgesius. "Demystifying the Draft EU Artificial Intelligence Act—Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach." *Computer Law Review International* 22, no 4 (2021): 97–112.
Voas, Jeffrey and Phillip A. Laplante. "The IoT Blame Game." *Computer* 50, no 6 (2017): 69–73. https://doi.ieeecomputersociety.org/10.1109/MC.2017.169.
Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR." *Harvard Journal of Law & Technology* 31, no 2 (2017): 841–887.
Wagner, Ben. "Ethics as an Escape from Regulation. From 'Ethics-Washing' to Ethics-Shopping?" In *Being Profiled: Cogitas Ergo Sum*, edited by Emre Bayamlioglu, Irina Baraliuc, Liisa Albertha Wilhelmina Janssens, and Mireille Hildebrandt, 84–89. Amsterdam: Amsterdam University Press, 2018.
Wang, Yumin and Hanzhou Wu. "Protecting the Intellectual Property of Speaker Recognition Model by Black-Box Watermarking in the Frequency Domain." *Symmetry* 14, no 3 (2022): 619. https://doi.org/10.3390/sym14030619.
Weaver, John Frank. "Everything is not Terminator: America's First AI Legislation." *Robotics, Artificial Intelligence & Law* 1 no 3 (2018): 201-207.
Weick, Karl E. "The Vulnerable System: An Analysis of the Tenerife Air Disaster." *Journal of Management* 16, no 3 (1990): 571–593. https://doi.org/10.1177/014920639001600304.
Wetmore, Jameson M. "Delegating to the Automobile: Experimenting with Automotive Restraints in the 1970s." *Technology and Culture* (2015): 440–463. https://doi.org/10.1353/tech.2015.0057.
Wolfe, Harry P. and David A. NewMyer. *Aviation Industry Regulation.* Carbondale, IL: Southern Illinois University Press, 1985.
Zhen, Xingwei, Jan Erik Vinnem, Xue Yang, and Yi Huang. "Quantitative Risk Modelling in the Offshore Petroleum Industry: Integration of Human and Organizational Factors." *Ships and Offshore Structures* 15, no 1 (2020): 1–18. https://doi.org/10.1080/17445302.2019.1589772.
Zielke, Thomas. "Is Artificial Intelligence Ready for Standardization?" In *Systems, Software and Service Processes Improvement. EuroSPI 2020*. Communications in Computer and Information Science, Vol. 1251, edited by M. Yilmaz, J. Niemann, P. Clarke, and R. Messnarz, 259–74. Cham: Springer, 2020.

Primary Legal Material

European Union
Council Directive 96/82/EC of 9 December 1996 on the Control of Major-Accident Hazards Involving Dangerous Substances (Repealed). European Council.

Appendix

Table A.1: Distribution of AI incidents in government sectors (AIAAIC data)

Government sector	Frequency	%
Police	78	39.8%
Immigration	18	9.2%
Welfare	15	7.7%
Municipal	14	7.1%
Health	15	7.7%
Justice	11	5.6%
Defence	12	6.1%
Transport	8	4.1%
Education	6	3.1%
Culture	2	1.0%
Employment	2	1.0%
Energy	2	1.0%
Housing	2	1.0%
Tax	2	1.0%
General	1	0.5%
Agriculture	1	0.5%
Environment	1	0.5%
Finance	1	0.5%
Foreign	1	0.5%
Postal	1	0.5%
Retail	1	0.5%
Security	1	0.5%
Telecommunications	1	0.5%
Total	196	100.0%