# Ital-IA 2022

# Risky Artificial Intelligence: the Role of Accidents in the Path to AI Regulation

**Giampiero Lupo**
ISASI - CNR (Istituto di Scienze Applicate e Sistemi Intelligenti - Consiglio Nazionale delle Ricerche)
giampiero.lupo@isasi.cnr.it

## Abstract

The introduction of a new complex technology such as the automobile or the civilian nuclear power raises many questions regarding its safety, its risks, and its impact on society, the environment and the institutions. The phenomenon of complex technology's diffusion and its interaction with different social contexts is often accompanied by a contrast between actors enthusiastic about the novelty and detractors who fearfully view the risks that the new technology brings with it.

Recently, we have witnessed an ever-greater diffusion of Artificial Intelligence (AI) technology in different application contexts. This certainly represents a new experience of complex technology introduction that brings with it concerns about its safety and its implications on 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 include technologies that can act as intelligent agents that receive perceptions from the external environment and perform actions autonomously (Russell & Norvig, 2002; Santosuosso & POLETTI, 2020). As it happened for other high technologies, the fear for safety and technological risks and the uncertainties related to the abrupt change for the "expected and loved order" (Angelides, 2012) encourage a rush towards the regulation of the new technology that may restore or safeguard the normal sense and order (Lanzara, 1993). Norms may have the capacity of curbing technological change and of limiting its borders and its implications for the pre-established normal order (Foucault, 1977).

As a consequence of AI diffusion, on the one hand, 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 contests (Lupo, 2019; Van Dijk, 2020). This proliferation, also identified as the "ethification" phenomenon" (Van Dijk, 2020) represents a flexible practice to cope with unpredictable effects of emerging technologies, different from law-making that is more rigid and time consuming. On the other hand, 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. See for example the proposal for a "Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)" drafted by the European Commission (Commission, 2020).

Looking at the evolution of complex technologies' regulation, in many cases it has evolved or taken steps forward in the aftermath of undesirable or unfortunate happenings that occurred unintentionally resulting in harm, injury, damage, or loss that is "accidents". For instance, SEVESO directive (Council, 1997) aimed at improving the safety of industries using large quantities of dangerous substances, received a fundamental push forward to its approval in the aftermath of Seveso disaster, an industrial accident that occurred in 1976 in a small chemical manufacturing plant in Northern Italy that resulted in the highest known exposure to 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) in residential populations (Marchi, 1997; Rai, 2020). The evolution of high technologies' regulation as a consequence of accidents interested several sectors as civil aviation (Downer, 2010), automobile transportation (Norton, 2015), nuclear energy (Perrow, 2010, 2011), pipeline industries (Dahle et al., 2012) etc.

This link between accidents and high technology 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, environment, existing laws and old technologies (Valdés & Comendador, 2011). Additionally, the "black box" (Rai, 2020) of high technology's functioning that can be opened only in the aftermath of unwanted and harmful happenings, is often protected by trade secrets and intellectual property.

AI technology is only recently applied in several contexts of application, and the count of unfortunate happenings may grow in the future. Despite this, the empirical evidence of AI "accidents" can bring the debate on the risks of AI to the attention of the public and policy makers and influence the regulatory processes that will affect this technology. In addition, accidents provide fundamental information on the functioning of autonomous technology helping to open the "black box" and thus supporting safety regulations' drafting.

This study focuses on the role of accidents in providing information on the impact of AI technologies useful for their regulation. It compares the types of AI incidents occurred up until now with the issues regulated by AI soft laws and the EU Commission's regulation proposal. The study has been finalized as part of the activities of several research projects as Towards Cyberjustice (www.cyberjustice.ca) and ACT (Accessibility trough Cyberjustice - www.ajcact.org) seeing the involvement of IGSG-CNR (Informatica Giuridica e Sistemi Giudiziari) and ISASI-CNR (Istituto di Scienze Applicate e Sistemi Intelligenti) and which focus on the ethical aspects and the impact on fundamental rights of the use of AI and on the use of AI in the judiciary.

The empirical analysis implemented takes advantage of a set of AI accident lists implemented by different types of actors (in particular the 1. AI Global Where in the World is AI_ Map; 2. The AI Incident Database; 3. AI, algorithmic & automation incident & controversy repository)[1]. Data allowed to classify and analyse the type of accident, the sector of AI application, the typology of AI involved. Additionally, the analysis of AI ethical guidelines already implemented in a previous study (Lupo, 2019, 2022-forthcoming) and a more qualitative analysis of the EU Commission proposal allowed to investigate the agreement of these legislative instruments with the empirical evidence of the AI accidents occurred. This study contributes to assess the efficacy of the mentioned normative instruments in avoiding risks of harm and infringement of fundamental rights related to AI use.

The results of the study that will be discussed as a contribution to the workshop are the following. First, the assessment of the paths of other high technologies' regulation helped to put in evidence phenomena that may also affect AI and AI regulation. For instance, phenomena as the "regulatory capture" in aviation industry that describes the practice of powerful industries that come to dominate the agencies that regulate them (Downer, 2010) due to an information imbalance between regulators and high technologies' producers, may also affect AI in the future given the complexities related to its functioning and the intellectual property laws protecting the AI "black box". Second, the analysis of a set of AI ethical documents (n=108) through content analysis allowed to acknowledge that ethical documents converge to a set of principles and issues related to AI application as "transparency", "no discrimination", "accountability", "protection of human rights", "respect of judicial values", "protection of privacy and personal data". Third, the investigation of legal frameworks disciplining AI that focused in particular on the mentioned EC proposal allowed to clarify the regulatory orientation of some institutions as the European Commission and the strategy that will be implemented in order to limit potential negative effects from the use of AI. In particular, the EC approach based on the identification of different levels of risks on the basis of types of AI technologies has been clarified. The EC proposal distinguishes between unacceptable risk, high risk, limited risk AI. The technologies included by EC proposal in the high risk list are strictly regulated with the aim of safeguarding health and safety of EU citizens and the respect of EU fundamental rights as well as EU acquis, while for the limited risk AI, the proposal only encourages and facilitates the drawing up of codes of conduct intended to foster the voluntary application of the requirements set out for the high-risk AI systems. Differently, the proposal prohibits the development of unacceptable risk AI systems in the EU or the use of these systems in the EU if developed in a third country.

Fourth, the quantitative analysis of AI incidents lists and the investigation of the agreement of legislative instruments previously analysed with incidents data allowed to clarify worth-mentioning patterns. The analysis put in evidence the discrepancy between incidents' data and EU Commission proposal of AI regulation with regards to the typology of AI involved. In particular, the analysis put in evidence that some technologies that data acknowledged to be risky and subjected to incidents, as the autonomous driving, are not listed in the high-risk category of the Commission. Moreover, the analysis acknowledged the differences between data on incidents regarding sector of AI application and the ethical documents investigated in a previous study (Lupo 2021). For instance, the analysis of ethical documents acknowledged that only 1,85% of the documents investigated regulate AI applied for public administration (Lupo, 2022-forthcoming), while differently, the data on AI incidents reveal that the second sector most affected by incidents is the government sector (21,54%). Finally, some discrepancies regarding types of incidents have been registered between incidents' data and both EU Commission's proposal and ethical documents investigated in the previous study mentioned (Lupo, 2022-forthcoming). For instance, despite the high number of incidents regarding surveillance (95 incidents registered by AIAAIC), the mentioned study on ethical documents acknowledges that only the 13% of AI documents investigated focus on risks related to AI surveillance systems. Also the Commission proposal does not mention issues related to the use of AI for surveillance probably confirming an ambivalent attitude of public institutions that see AI surveillance technology as very useful for ensuring security and supporting home affairs' operations.

These results put in evidence the necessity to focus on the empirical evidences of incidents in order to open the AI black box, understanding AI real impact on the contexts of application and provide for an effective regulation of AI. Interestingly, this approach is also supported by the EU Commission proposal for the regulation of AI. The proposal provides for the appointment of National Competent Authorities that will have the aim of collecting data on AI

[1] 1. AI Global (2020). Where in the World is AI_ Map. 2. McGregor, S. (2021) Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. In Proceedings of the Thirty-Third Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-21). Virtual Conference. 3. AIAAIC (2020). AI, algorithmic & automation incident & controversy repository.

incidents and to investigate on them in order to provide information regarding AI use's risks and potential proposals for AI regulation's update.

## References


Angelides, S. (2012). Disorder as "Pseudo-Idea". *Atlantis: Critical Studies in Gender, Culture & Social Justice, 35*(2), 10-20.

Commission, E. (2020). Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

COM/2021/206 final.

Council, E. (1997). Council Directive 96/82/EC of 9 December 1996 on the control of major-accident hazards involving dangerous substances.

Dahle, I., Dybvig, G., Ersdal, G., Guldbrandsen, T., Hanson, B., Tharaldsen, J., & Wiig, A. (2012). *Major accidents and their consequences for risk regulation*: Taylor & Francis Group: London.

Downer, J. (2010). Trust and technology: the social foundations of aviation regulation. *The British Journal of Sociology, 61*(1), 83-106.

Foucault, M. (1977). Historia de la medicalización. *Educación médica y salud, 11*(1), 3-25.

Lanzara, G. F. (1993). *Capacità negativa: competenza progettuale e modelli di intervento nelle organizzazioni*: Il mulino.

Lupo, G. (2019). 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*(2), 75-96.

Lupo, G. (2022-forthcoming). The Ethics of Artificial Intelligence: an analysis of ethical frameworks disciplining AI in justice and other contexts of application. *Oñati Socio-legal Series*.

Marchi, B. D. (1997). Seveso: from pollution to regulation. *International Journal of Environment and Pollution, 7*(4), 526-537.

Norton, P. (2015). Four paradigms: traffic safety in the twentieth-century United States. *Technology and culture*, 319-334.

Perrow, C. (2010). The meltdown was not an accident *Markets on trial: The economic sociology of the US financial crisis: Part A*: Emerald Group Publishing Limited.

Perrow, C. (2011). *Normal accidents: Living with high risk technologies-Updated edition*: Princeton university press.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science, 48*(1), 137-141.

Russell, S., & Norvig, P. (2002). Artificial intelligence: a modern approach.

Santosuosso, A., & POLETTI, D. (2020). Intelligenza artificiale e diritto. *Perché le tecnologie di IA sono una grande opportunità per il diritto. Mondadori Università, Milano*.

Valdés, R. M. A., & Comendador, F. G. (2011). Learning from accidents: Updates of the European regulation on the investigation and prevention of accidents and incidents in civil aviation. *Transport policy, 18*(6), 786-799.

Van Dijk, N., and Casiraghi, S. (2020). The "ethification" of privacy and data protection law in the European Union. The Case of Artificial Intelligence. . *Brussels Privacy Hub Working Paper., Vol. 6, N. 22.*