Can artificial intelligence accountable and liable?

Alfredo M. Ronchi
Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, 20133, Italy
Tel: +39 393 0629373, Email: alfredo.ronchi@polimi.it

Abstract: One of the most impacting technologies recently disseminated within our society is Artificial Intelligence instantiated as powerful LLMs. This long-lived technology rooted in the 1950s was considered too cryptic and too far from citizens' practical use. But the exponential dissemination of AI platforms like ChatGPT and the immediate familiarity acquired by any typology of citizens boosted the implementation of AI in a wide range of products like autonomous driving cars, drones and documents and visual content created by Gen AI. This wide implementation of AI and Gen AI issued the need to face the juridical position of AI in case of generated harm and relative assignation of liability. These poses posed the focus on some of the potential concerns in assigning accountability and liability in case of AI and Gen AI managed systems.

Keywords: AI, Gen AI, Accountability, Liability, Society, Ethics, Human Rights, Biases, Digital Transformation

Preamble

Time ago Ranjit Makkuni, one of the chief researchers at Xerox PARC in the 1970s, wrote a paper entitled "The Betrayed Revolution" observing that "When the original vision of the Dynabook (later to become the laptop) was conceived by Alan Kay's team at Xerox PARC in the 1980s, computational tools were envisioned as a tool for creativity and leisure. Indeed, the first slides of the Dynabook showed people freeing themselves from the shackles of their offices, lying in sprawling natural landscapes, pursuing their artistic creativity. Nearly 50 years later, that vision had become a nightmare ….". This scenario is potentially replicated nowadays as a potential side effect of digital transformation.

Is or will the sudden evolution or revolution of the cyber world, now driven by artificial intelligence and related technologies, be fully beneficial to humanity, or are there unforeseen and unwanted side effects?
The added value and results achieved thanks to cyber technology are not in question, just think of the use of artificial intelligence to identify hidden patterns in DNA or collect and summarize a huge number of documents; this analysis is not due to a jurisprudent but to a computer scientist looking at artificial intelligence from a humanistic perspective with particular attention to responsibility and accountability in the field of jurisprudence, considering the medium and long-term impact on society and the potential disadvantages

or unwanted side effects[1] with reference to human rights, security, ethics, social, cultural, economic and environmental aspects, possibly suggesting mitigation strategies.

The origin – it was a game changer

Following the studies on Turing[2] machine in 1950 a first release of the Touring Test was published by A. M. Turing on Mind[3], the incipit of the article is "*I propose to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. …".* For the first time the idea to break the rule of rigid procedures and algorithms unleashing the hidden power to identify "invisible" patterns and create associations of items.
Almost thirty years later artificial intelligence, expert systems and fuzzy logic become some of the keywords identifying cutting edge technologies. A.I. advances captured the interest of journalists being considered the seed of the "Big Brother" or the ignition of the progressive slavery of men ruled by machines. General artificial intelligence design and Generative AI will improve itself to overcame human intelligence[4]? Are we going to face what several times was depicted in the science fiction movies?

Now we face the widespread diffusion of this technology that after more than ten years of developments in the 1980s come to the fore, one first times, in the 1996 with the chess challenge between Garry Kasparov and Deep Blue[5], the first game was in favour of Deep Blue, but Kasparov recovered and won the match 4–2.
We don't know if the concerns often related to AI, considered a potential cyber-Leviathan dominating humanity thanks to its high "intelligence far exceeding any human one", is due to the term that was associated long time ago to the early studies and applications in this field. The term "artificial intelligence" probably ignited the understanding that there were two "intelligences" one human and anther artificial potentially competing to "rule" the world, as some Sci Fi movies use to depict like "*2001: A space Odyssey*" due to HAL[6], War Games, Eagle Eye or even Lucy[7].
AI and ML can offer relevant benefits to humanity but, after a first phase of free evolution and identification of positive evolutionary trends, they must be carefully framed in their potentiality and regulated. This regulatory action will address both development and deployment of their applications to avoid potential drawbacks. As an example, citizens consider Generative AI able to express "creativity" how can we ensure that its "creation" comply with ethical principles and will not generate harm?

Ethics and rules like Asimov's laws applied to AI

[1] Ronchi Alfredo M., (2021) – "From Ingsoc to Skynet it is not only science fiction: From novels and science fiction to quasi-reality" – Proceedings UNESCO IFAP conference in Khanthi Mansjisk

[2] Alan Mathison Turing (1912-1950), British mathematician Alan Turing, c. 1930s. Turing did the earliest work on AI, and he introduced many of the central concepts of AI in a report entitled "Intelligent Machinery" (1948)

[3] Alan M. Turing, Computing machinery and intelligence, Mind, 59 October, pp. 433-460, 1950. https://academic.oup.com/mind/article/LIX/236/433/986238

[4] Science fiction movies already proposed similar scenarios e.g. Wargames (1983 American techno-thriller film directed by John Badham) or Eagle Eye (2008 American action-thriller film directed by D. J. Caruso)

[5] https://www.ibm.com/history/deep-blue

[6] A Space Odyssey is a 1968 epic science fiction film produced and directed by Stanley Kubrick based on an Arthur C. Clark novel – HAL acronym is formed by the chars preceding IBM - The spaceship officer Dr. Dave Bowman used to interact with HAL by voice calling "Hello, HAL. Do you read me, HAL?"—HAL: "Affirmative, Dave. I read you.", Dr. Dave Bowman: "Open the pod bay doors, HAL", HAL: "I'm sorry, Dave. I'm afraid I can't do that" . . . We all know what happened later.

[7] Science fiction movies already proposed similar scenarios e.g. Wargames (1983 American techno-thriller film directed by John Badham) or Eagle Eye (2008 American action-thriller film directed by D. J. Caruso), Lucy (2014 French Sci Fi movie written and directed by Luc Besson)

This theme grounds its roots in the well-known problem related to autonomous drive cars but due to the proliferation of AI managed systems and devices has gained a key role in systemic liability allocation models that ensure accountability. without stifling innovation. The role of followers assigned to regulations is usually due to the will to do not influence or shape the evolution of science and technology. The initial trend from humans to AI systems, usually associated with “black boxes,” considered human supervision to be not mandatory and even unnecessary, as it created a significant time lag compared to decisions and actions that AI could execute on the fly. This results in critical decisions being entrusted to automated decision-making software, with no room for human intervention. As we all know, this approach has already led to some inconveniences and even disasters. Tokyo stock exchange, Boeing 373 Max, some automatic driving systems, etc.

Following the actual trend to embed AI at decision level in a broad range of fields the legal system will soon face the problem to assign responsibility when AI systems will cause harm. Taking as an example to well-known car collision case, apart from potential ethics and cultural differences (east and west), the analysis will take into account the design and performance of the hardware side (sensors, actuators, schema, redundance, etc.) and the logic side trying to understand the logic schema followed by the system usually without human supervision and last but not least a mix of negative circumstances due to a negative combination of exceptions on both sides. All these aspects not mentioning other vehicles involved in the case, environmental conditions, physical condition of the read, etc.
If we consider some cases in a field, like aeronautics, that is more advanced, it is evident that several additional parameters will make the assignation more difficult: regular updates, compliance with requested mandatory updates or notification forms, etc.

How will jurisprudence define a proper methodology to evaluate AI decision-making case? If a usual algorithmic decision-making system can be analysed and evaluated or amended, the dynamic outcome of the AI system will make the case more complicated.
This also applies to tort-based approaches that address the unintended behaviour of a decision-making system that, due to an unpredictable set of inputs, has deviated from the correct course of action. Tort-based approaches involve using the tort law framework to assign responsibility for harm and provide remedies to victims, often focusing on corrective justice. In such a case described above we potentially face the case of a collection of negative circumstances that all together unpredictably caused a disaster. Again, we are basically back to the chain of the fact checking list and probably to try to identify the origin of the problem and assign the responsibility.
The use of algorithmic procedures in already digitised working environments usually splits the accountability and liability problems into two sections. Some operational fields use to submit to authorities for checking and archive the source code of the application and in any case assign responsibility to the responsible person within the organisation. In this situation, whether the source code is provided or not, the liability falls on the person guaranteeing the results. In the event of damage, ordinary case law applies, without any specific guidance regarding the software, except for potential legal recourse against the software developer and the loss of confidence in the product. This is what usually happens, for instance, in case of structural analysis or mechanical designing tools. This is true for algorithmic software: if we add a little AI or fully rely on AI systems, the approach could potentially differ. If the AI module provides output to a human supervisor, we likely revert to the previous "algorithm-based" approach, where the responsibility rests with the human. If the AI module connects directly to the operational side, responsibility is likely attributed to the human in charge of the operation, but in both cases, there is no feedback to the technology. In both cases, there is no responsibility on the AI, so ideally, the same or similar damage can happen again. Of

course, if the damage is made public and the related information shared, trust in the AI system will suffer, but is this enough?
As happens in cases of uncertain attribution of liability, should we envisage mandatory insurance for any potentially harmful AI system as a partial remedy for victims? In such a situation, how can insurance companies assess the risks and allocate the corresponding policy costs?

How can risk managers consider the role and behaviour of AI and Gen AI?
Recently, thanks to the appeal of AI-based applications, a range of management and decision-making tasks are being entrusted to AI systems, not always supervised by humans. This new approach makes risk assessment much more difficult, as this specific aspect is usually not included in the scenario analysed by the intelligent application, which is often responsible for optimizing one or more resources. Some artificial intelligence applications are tasked with optimizing project management plans, without providing explanations of the logic behind them or the specific aspect deemed to need optimization. Risk managers receive process output from a sort of black-box and must analyse it to identify potential risks associated with that specific configuration. This often leads to a “snapshot” of the process rather than analysing the actual full “video”. Research labs are investing time and resources to “open” the black-boxes and provide explainability concerning the outcomes.
There is a clear need to develop systemic liability allocation models that ensure accountability. without stifling innovation. As it already happened in similar situations one of the first steps is to identify if there is any potential relation to traditional situations to reconnect the specific case to traditional jurisprudence. One of the relevant differences is the lack of a pre-casted algorithmic procedure that can be analysed and amended. At a first it looks much more like a concourse in liability between human and AI.

Furthermore, is there any possibility that in the future such cases will be directly managed by “AI judges”, there are already several AI systems supporting attorney in law starting from the provision of the analytics of the huge set of documents, court rulings and laws pertaining similar cases. Are there some plans to solve low impact cases by AI as a concrete opportunity to solve the overcharge of duties in some courts?

Key players like Google and Microsoft are upping their game. Anthropic[8], Mistral[9], DeepSeek[10] and OpenAI are continuing their winning streaks disseminating their AI platforms in every-day smart companions, Palantir[11] already developed and tested AI solutions to fight crime and foresee potential criminal actions stating that human rights and ethics are fully respected. Which will be at the "end of the day” the juridical status of AI and potentially assigned responsibilities? what about Gen AI and the black box that generate the outcomes?

Wrapping up

Digital technology was usually conceived to benefit society in different fields, health, labour, safety, security, achievements and more. In the last twenty years some potential drawbacks become visible, sometimes digital technology is framing us, we faced the potential “dual use” of such friendly technology, Nowadays the rapid exploitation of AI and its dissemination in almost any sector of our daily life poses different questions and concerns, among them:

[8] https://www.anthropic.com
[9] https://mistral.ai
[10] https://www.deepseek.com/en
[11] Palantir a US company founded in 2003 in Denver, with the support of different agencies https://www.palantir.com

Is AI always providing benefits or is there any drawback? Because of AI role of decision maker how can we assign accountability and liability in case of harms?
This document posed the focus on some of the potential concerns in assigning accountability and liability in case of AI and Gen AI managed systems.
Nowadays, in the light of actual trends Moral Philosophy / ethics must be the always the key reference point in developing, customising and using technologies. There is a strong need to keep humans in the centre of the loop and respect human rights and ethics.