JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Traceability and Allocation of Responsibility in an AI caused harm

Name: Disha Gupta & Roshan Kundoor Menon

Designation: 5th year BBA/ BA LLB- Jindal Global Law School

Contact No.: 8853470888 & 9840310407

Email: disha355048@gmail.com & theroshanmenon@gmail.com

Abstract

Jack and Jill went up the hill to fetch a pail of water, Jack fell down and broke his crown because Jill pushed him, you'd know Jill is legally liable for the tort of battery. It is easy to trace as to how the tort was committed and by whom. To change the scenario, let us say Jack and Jill went up the hill to fetch a pail of water from an automated well with deep learning involved within its AI system- and suddenly an iron bucket- instead of pouring water as expected, hits Jack and as a result he gets pushed, breaks his crown and this time Jill came tumbling after. Who is liable for this AI caused harm? Is the team who created the automated well or any one of them? Can they be held liable for this AI caused harm? If so, then how does one allocate responsibility here? This paper discusses the concept of traceability, importance and need of an explainable AI and how does the former help in allocating responsibility in AI caused harm.

INTRODUCTION

When it comes to the idea of understanding the legal personhood of Artificial Intelligence, we find ourselves in a precarious situation. The legal system is mired with inconsistencies and reverts to using archaic terminology and phraseology when confronting and defining modern issues. To that end, the law

has often found itself at odds with the situations that prevail in society today, leading to several calls for change and a need for a greater degree of robustness in the system. Thus, combining something as contentious as Artificial Intelligence, with the law seems to be a union doomed to failure. However, it would not do to simply ignore the existence of AI systems altogether, in the stubborn belief that doing so will enable us to not recognise it as an issue. Instead, there is a need to confront the issue head on, and in doing so, create a more robust system that leaves room for evolving to improve itself with time passing. To connect with the outward responsibility and the legal liability of an actor in AI caused harm, it is essential to look within the internal processing of the AI system and hence the concept of traceability is an inevitable part of this process.

To ensure transparency in AI systems, traceability is a vital element. Trust in the AI systems can be developed if traceability and accountability in AI caused harm can be synchronised well. In 2018 the European Commission had created a group of experts on Artificial Intelligence i.e. High Level Expert group on Artificial Intelligence (AI HLEG) to come up with ethical guidelines for the AI systems[1]. AI systems are not a result of one man army but involve multiple actors with respect to its coding, programming, sensibility and other integral performance tasks. The multiple actors involved are the developers, researchers, designers of the system, the people or the organizations (business sectors or government organizations) who bring the AI into action by manufacturing products or offering services, the end users that are actively involved with the AI system as well as the entire society that shall be affected by the use of AI directly or indirectly[2].

Traceability- A factual concept

To legally conclude as to who shall be held accountable in AI caused harm, setting up the facts right is a part of the process. Repeating, replicating and reproducing the algorithm sets tells us that traceability is a factual concept that when tracked correctly can help us narrow down the actor that falls under the tortuous liability in AI caused harm.

[1]'Ethics Guidelines For Trustworthy AI - Shaping Europe'S Digital Future - European Commission' (*Shaping Europe's digital future - European Commission*, 2020)

Given the similar conditions in which the AI system operates the same sets of data and algorithm may be used to arrive at the same precise decisions as earlier arrived at by the AI. After a number of researches and an attempt to check whether the AI arrives at the right decision after interchanging its algorithm in different permutations and combinations of hypothetical situations just to check the credibility of the decision that AI arrives at, it can be said that the researchers and programmers team may repeat the data sets for creating same AI systems. When it comes to replicating the algorithm and data sets by a different team of researchers, they may do so by relying on the original results of the original researchers work[3]. Reproducing the results by trying to analyse different data sets in a similar fashion, and reproducing identical results, again and again strengthens the core of the AI and as a result makes it trustworthy[4].

Principle of Explicability

It is necessary for human beings to exercise control over algorithms in a meaningful way. Before coming to the novel artificial intelligence systems, it is important to establish as to how the AI differs from Good old fashioned AI as well as Machine learning. In case of GOFAI the programming codes help the decision to be generated and any error in the decisions shall be rectified by altering the computer code language[5]. Similarly with machine learning, with time and processing the machine learns to perform better and make more informed decisions. The way the latter operate cannot be the way AI systems should operate; reason being the involvement of life and property that come in contact while the AI systems are operated. For example in case of self automated cars or AI lawyer robots, the real harm can be caused to the human and property and the only way to rectify such errors in these scenarios maybe by reading the Bible for the rest of our lives. The principle of explicability is important to trace the opaqueness of the increased amount of algorithms reaching a certain decision and challenging how ethical such a decision is[6]. How the AI reached certain decision and why it did reach a particular decision – the explanation to the algorithms is important for effective traceability and as a result would help in holding a particular actor accountable for the AI caused harm.

[3]Adam Brinckman and others, 'Computing Environments For Reproducibility: Capturing The "Whole Tale"' (2019).
[4]Marçal Mora-Cantallops and others, 'Traceability For Trustworthy AI: A Review Of Models And Tools' (2021).
[5]Ron Schmelzer, 'Understanding Explainable AI' (*Forbes*, 2019)
[6]Scott Robbins, 'A Misdirected Principle With A Catch: Explicability For AI' (*springer.com*, 2019).

The sets of data and all the processes that help AI reach its decisions should be properly stored and documented to promote transparency and easy traceability[7]. It allows preventing future mistakes, helping the system make better decisions and promote audit ability. The future is AI and hence explainable AI is the need.

How does traceability help in allocation of responsibility in AI caused harm?

AI only serve as agents and unlike human beings, it cannot hold moral responsibility for its actions. If judges rely on the decisions of the AI without knowing as to how do the system arrive at such decision or when a driver uses self automated car but is unaware of how does the algorithm respond, it creates a moral problem of overly relying on the AI and unintentionally ignoring how the actions are being performed. Concept of distributed responsibility among the multiple actors involved helps us know that not one actor can be held fully accountable for the mismanagement or poor regulation of AI[8]. The black box of the artificial intelligence systems need to tracked and traced at profound levels by the programmers, software developers, code researchers and other agents that can figure out as to how be the deep learning function in AI being performed and likewise errors can be avoided[9]. In my opinion, this is more of a technical issue that falls in the domain of data scientists and other technology experts. Only with their intervention into the technical aspects of traceability, one can allocate the moral responsibility of the actions or the decisions taken by the AI systems.

Challenges within Traceability

Unlike other technologies, AI involves deep learning. It may with time develop certain layers within its intelligence system to arrive at decisions that with updated data may change overtime. If unregulated, traceability and transparency can get challenging. The fast pace, at which AI adapts, stores information, reads it and learns can result in errors that might be difficult to understand and hence, chances are AI gets uncontrollable and fail to explain its decisions as well. Given the right amount of quality data, AI decision processing can be relied on but the real problem arises within complex AI systems that tend to develop hidden strategies and layers of data processing as they are programmed to learn continuously from the data

[7]'Requirements Of Trustworthy AI - FUTURIUM - European Commission' (*FUTURIUM - European Commission*, 2021)
[8]Mark Coeckelbergh, 'Artifcial Intelligence, Responsibility Attribution, And A Relational Justifcation Of Explainability' (*Link.springer.com*, 2019)
[9]Amina Adadi, 'Peeking Inside The Black-Box: A Survey On Explainable Artificial Intelligence (XAI)' (2018)

that is being absorbed by the system itself. To dwell deep within those hidden layers in order to understand why the system arrived at certain decision, can get challenging. It can serve as ignorance on part of human agents who hold the responsibility of understanding the AI. This opaqueness in the algorithm sets also creates difficulties for the organizations that seek to use AI for offering their services to customers as such challenges does not allow them to abide by the regulations such as GDPR (General Data Protection Regulation).The laws shall require organizations to give detailed explanation to the customers as to how does the AI system shall process their data and how much of privacy should be expected. Audit ability and traceability of the many assumptions that the AI creates with the given data to arrive at its decision can get credibility only if proper tracking and monitoring of the data is shown. There are certain risks involved as well along with the challenges found in traceability. In cases where the AI systems due to algorithm bias, start inferring previous data sets without establishing a proper link as to why did it chose to do a certain act in future, shall result in problems such as breach of privacy[10]. For example if a customer's details logged in for holiday plans at a hotelpreviously that uses AI systems to establish better customer response and feedback and such feedback obtained may be used for inference by the AI to post the same on their social media without their consent because the AI created the assumption that the customers like sharing all good things or love pomp and show on social media platforms. Here even though the algorithm may not be set to do so but because of the hidden processing within the AI, it learns to do so and only with effective traceability can this assumption created by the AI can be broken but it poses challenges because of the opaqueness of the algorithm.

Hence, it can be concluded that how the role of traceability does helps us in figuring out the errors and as a result trace the specific actor that shall be held legally liable to fix the issue of allocation of responsibility when the harm is done. But to ensure the right way of creating legal linkages, one must understand how AI connects itself to the law or vice versa. The further sections discuss about the allocation of responsibility and the legalities involved in case of harm caused by AI.

[10]Tom Bigham and Suchitra Nair, 'AI And Risk Management' (*Www2.deloitte.com*, 2018)

How do you legally define an AI?

It is hard today to put a concrete definition to what AI is. One of the more common definitions is to refer to it as a power given to a machine to copy intelligent human behaviour[11]. Through this copying, it carries out actions and tasks autonomously, following a set goal. Advancements in machine technology have helped add to this definition, with successive generations of machines able to more closely mimic humans in their decision making and execution, improving on the quality of their copying from generation to generation. Another theory holds closer to the third law of Sir Arthur C. Clarke, any sufficiently advanced technology is indistinguishable from magic.[12] To that end, AI represents machines that are capable simply of more than what current technology is able to achieve[13]. From a legal standpoint, the definitions vary. As it stands, there are few countries that have taken a concrete stance on the ascribing of laws to ascertain the onus upon an AI in a matter regarding its civil liability. To that end, their legal definitions have largely been left to specific instances wherein a working understanding of what they entail is made use of. Yet, this is now a changing scenario as AI technology becomes more and more human-like in the tasks it can accomplish. This even extends to AI anticipating not only what the most ideal state is for a task, but also calculating for human errors and making predictive analysis based on what is an objectively incorrect decision. Maia – a Cornell University chess AI – doesn't always play the 'computer move' but can play while anticipating mistakes a human would make, something entirely alien to the chess supercomputers of today that perform at a level over 400 ELO points ahead of the highest rated player in history, Magnus Carlsen.[14]

The 'legal character' of an AI in tort

How exactly is it that an AI can be given a status? This is something that finds precedent through existing laws. Entities like companies, while not having a human presence, are considered to be separate legal entities and thus have their own rights and protections. Indian law even considers deities to have special status as legal entities, affording them with protections as such. At the outset, similar concessions can be made for AI, with an argument to be had that the effects of an AI's actions can be more directly felt that a metaphysical concept of theism. To that end, even with an AI not being a 'true AI' capable of operating

[11] "Artificial Intelligence" (*Merriam-Webster*)
[12] Clarke A. C., *Hazards of Prophecy: The Failure of Imagination*
[13] "Meaning of AI for the Legal Industry" (*Meaning of AI for the legal industry | Thomson Reuters*)
[14] Knight W, "A New Artificial Intelligence Makes Mistakes-on Purpose" (*Wired* February 13, 2021)

autonomously[15], or having self-determination, reaching a critical 'I think, therefore I am' moment in its cognitive development[16], it can be held to be a legal entity before a court. While not many, some countries have taken steps towards creating legal definitions for AI and having codified discussions on how to treat the subject in court. For the most part, countries follow the 'duty of care' principle with AI, applying it here to indicate that the nearest human 'owner' of the AI system would bear the responsibility for its actions in a civil suit. Here, the user of the AI system would be the one held responsible – which could prompt discussion of the relationship between the AI's end user and the AI's manufacturer, along the lines of landmark tort law cases like Donoghue v. Stevenson[17]. Taking into account the principles of this case, there now exists an option for an aggrieved party to take the manufacturer of the AI system to court instead of merely presenting a case against the end user of the product. This is especially important to keep in mind when we consider that the average end user for an AI product, would likely be far less knowledgeable of the full functional capabilities of the AI, which a producer should take into account when manufacturing, thusly allowing for there to be a greater degree of responsibility to be hoisted upon them to ensure safety not only of the end user, but also when considering third parties that might be affected. This is an eventuality that has already come up in thought experiments and research on AI today. Research papers have already focused on how an AI could potentially be held responsible for civil or criminal action. In the case of the former, the arguments that take precedence today hold that the 'instructor' of the AI – the person giving it the directions to act, be this a programme or an end user – should be held criminally liable[18]. As it stands today, we lack any AI system – as far as we are made aware – that can fully self-actualise and act with a consciousness of its own. Thus, the reformative or punitive purpose of the criminal justice system would largely find no bearing when it came to application to the AI system itself, outside of merely quarantining its usage while efforts were undertaken to understand what caused the defect. The hardest aspect to adjudge in a criminal suit would be *mens rea*, the 'guilty mind' or even more simply, the intent to commit a crime. While lessened sentencing can exist for accidental crimes when the *actus reus*, or guilty act, is present, the *mens rea* is the reasoning that exists behind the crime itself. To that end, modern

[15]Balasundaram R, "What Is Real Artificial Intelligence: Characteristics of True AI" (*Emarsys*October 26, 2020)
[16] "Cogito, Ergo Sum" (*Encyclopædia Britannica*)
[17]Donoghue v Stevenson [1932] UKHL 100
[18] Kingston, John. (2016). Artificial Intelligence and Legal Liability. 269-279. 10.1007/978-3-319-47175-4_20.

AI systems would be incapable of finding themselves possessing a *mens rea* without the intervention or instruction of a human, be it their programmer or user[19].

Allocating liability for an AI's actions: Current scenario

Some countries have started preparing policies to better prepare themselves for the rising tide of AI. Countries like Russia and Japan have begun policy recommendations on the topic of AI, along with others who are preparing preliminary discussions on the subject[20]. Most countries involved in the discussion are developed and rich nations, representing the first wave of countries experimenting or experiencing the effects of AI. Japan and Russia too have still stuck to the point of ascertaining liability for an AI caused harm to a third party, with the conditions of each case being the deciding factors in allocating responsibility[21]. This approach also finds some precedent through the Liability Convention of the United Nations, which governed the issue of liability for the Outer Space Treaty[22]. As per Article V of the Convention, any State from whose territory or facility a space object is launched is treated as the launching state for that object. Thus, even if a private party is responsible for the launch, the State will be held liable for damages that may arise. A similar principle is established for AI. In this case however, the determination of the responsible party is given more importance, allowing for a greater degree of freedom in understanding who it would be that would have ultimate liability. To that end, AI caused harm will only be the found to be the fault of the end user, or the manufacturer.

This is inherently a short-term solution. As it stands today, we have no AI that is capable of taking a moralistic stand on the actions it undertakes, therefore we have no AI that is able to assume *mens rea* for its actions. However, considering this to be the end state of AI is a highly mistaken view to hold. From the original state of computers being unable to perform at chess, we went so far as to engineer supercomputers that could calculate millions of probabilities per move, while a human being could only see tens at best. And at this stage AlphaZero, a new chess AI – not a supercomputer – has emerged that only thinks of a few ten thousand probabilities at a time. However, this is an AI that was merely given the rules of chess and

[19]Claussén-Karlsson M, "Artificial Intelligence and the External Element of the Crime An Analysis of the Liability Problem"

[20] "Civil Liability of Artificial Intelligence" (*INDIAai*)

[21] Ai, Machine Learning & Big Data Laws and Regulations" (*GLI - Global Legal Insights - International legal business solutions*)

[22] United Nations, "Liability Convention" (*United Nations Office for Outer Space Affairs*September 1972)

within 7 hours had taught itself the game and routinely defeated computers with significantly higher thought power. Simply because AlphaZero thought like a human[23]. It played with the precision of a machine, but with the creativity of a person, something chess supercomputers lack today. The ability to think for themselves and not along the lines of existing theory. With the sheer amount of computing power an AI is capable of, it becomes hard to fully track its patterns of learning. An AI learns exponentially faster than a human, thus providing a difficulty when it comes to replicating exact scenarios and circumstances, without being aware of which of the millions of datasets present need to be targeted. AI that emulate human behaviour already exist among us. CAPTCHA tests are all the more common while browsing today, with newer bots able to blur the line dividing the 'dumbest' human from the 'smartest' computer. To that end, while we may yet hold a person responsible today for the actions of an AI, we will approach a state sooner than we consider where no human may be responsible for the final actions executed by a non-human intelligence. We think of even the best artificial intelligences today as being created with a human behind them, providing them with directives and instructions in operations. Thus, we take solace in the inherent humanness of the goals that the AI system is created to achieve. We are woefully unprepared for the day when such a machine, programmed with a human mind's guidance, creates an intelligence of its own. When a machine begets this intellectual 'offspring'. An artificial intelligence that has no human creator. With the stronger AI of today being those that tend to learn and think while adapting for themselves, what is to stop the strongest of the future's AI from making decisions for its own fate and the fate of others in the way humans do? In the novel 'Do Androids Dream of Electric Sheep', which inspired the 'Blade Runner' films, Phillip K, Dick presented readers with the conundrum of law enforcement forced to confront their understanding of the ethical and philosophical concepts of life and sentience. The book was written in 1968, set in 1992 and that year was later changed to 2021, the world today. In a novel the years seemed far off at the time of writing, but time only goes forward. At any given time we prepare to enter the future and to face the challenge of AI regulation the requirement of the hour is for the law to adapt and evolve to exist in a new world where humans are no longer its sole subjects.

[23]Somers J and Thomas L, "How the Artificial Intelligence Program Alphazero Mastered Its Games" (*The New Yorker*December 28, 2018)