HUMAN RIGHTS VIOLATIONS: ASCERTAINMENT OF LIABILITY ON AI

- Krishna Agarwal

Artificial Intelligence ("**AI**") has revolutionized the age with rapid advancement with the help of two machine learning algorithms- Support Vector Machines ("**SVMs**") and Deep Neural Networks ("**DNNs**"), which have paved the way for innovation through their accuracy and efficiency.[1] However, AI can be perceived as black box wherein the output generated cannot always be completely explained to humans due to the embedded complexity in it.[2] SVMs help in the classification of data by ascertaining geometric patterns which are not perceived by humans.[3] Whereas, DNNs utilize the hierarchical level of artificial neural networks to imitate the functions of a human brain.[4] A DNN consists of several nodes, densely interconnected to each other, that are often organized in several layers.[5]

Repeatedly, questions are raised against the appropriateness of the data, the varied features of algorithms and the verification of the output.[6] A black box is said to be created when humans are not able to comprehend the predictions made by complex algorithms, such as SVMs and DNNs.[7]

---

[1] *See* Brett Lantz, MACHINE LEARNING WITH R: EXPERT TECHNIQUES FOR PREDICTIVE MODELING, 205 (1st ed., 2014) (describing neural networks and support vector machines as powerful machine learning algorithms with wide application that have inner workings that are obfuscated for a variety of reasons and are therefore referred to as back boxes).

[2] *Explainability*, C3.ai, https://c3.ai/glossary/machine-learning/explainability (last visited Nov. 14, 2021).

[3] Michael E. Mavroforakis et. al., *A Geometric Approach to Support Vector Machine (SVM) Classification*, 17 IEEE TRANSACTIONS ON NEURAL NETWORKS, no. 3, 2006, at 671, 671.

[4] Ed Burns et. al., *What is Deep Learning and How Does it Work?*, TECHTARGET, https://searchenterpriseai.techtarget.com/definition/deep-learning-deep-neural-network (last visited Oct. 1, 2021).

[5] Seyed Morteza Nabavinejad et. al., *An Overview of Efficient Interconnection Networks for Deep Neural Network Accelerators*, 10 IEEE J. ON EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYS., no. 3, 2020, at 268, 268.

[6] Neil Raden, *The Problem of AI Explainability - Can We Overcome it?*, DIGINOMICA (Feb. 6, 2020), https://diginomica.com/problem-ai-explainability-can-we-overcome-it.

[7] Oscar GR, *"Black Box". There's No Way to Determine How the Algorithm Came to Your Decision.*, TOWARDS DATA SCIENCE (May 27, 2020), https://towardsdatascience.com/black-box-theres-no-way-to-determine-how-the-algorithm-came-to-your-decision-19c9ee185a8.

A balance has to be maintained between the explainability and the complexity of the model as the explainability diminishes with the increasing complexity of models.[8]

In this piece, *firstly,* I have analyzed the effects of black boxes, which reflect the tussle between human rights and innovation through AI. *Secondly,* I have proposed a solution of ascertaining liability by ensuring that both the human rights and the use of AI aren't compromised.

VIOLATION OF HUMAN RIGHTS

In the *Republic,* Plato provided an allegory to people living in a cave, who see shadows cast on the walls of the cave due to their campfire.[9] In the allegory, shadows represent the world, and the one who can recognize the difference between shadows and true reality is the one capable of ruling.[10] Similarly, in the context of AI, shadows are cast by the output of AI in that there is a high probability that AI output is biased or not fully representative of reality. This increasingly impacts the rights of individuals, as the results of the algorithms may be detrimental. Further, due to the creation of the black box, it is unclear how the input of AI becomes the output. Therefore, it is imperative to develop a mechanism so that liability can be imposed on the developer/user to ensure certain accountability in the system, which will prevent violation of human rights and promote efficient working of the system. One cannot blindly trust AI, as Plato's cave people trusted the shadows.

The revolution and development through AI may satisfy utilitarianism by leading to the overall development of society. But, at the same time, human rights are being jeopardized. A prominent example of this is the application of COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), a risk assessment algorithm used by judges to estimate the probability of recidivism.[11] It is alleged that COMPAS is biased against African Americans with a limited

[8] Pantelis Linardatos et. al., *Explainable AI: A Review of Machine Learning Interpretability Methods*, 23 ENTROPY, no. 1: 18, 1 (2020), https://www.mdpi.com/1099-4300/23/1/18/pdf.
[9] David Macintosh, *Plato: A Theory of Forms*, PHILOSOPHY NOW, https://philosophynow.org/issues/90/Plato_A_Theory_of_Forms (last visited Nov. 20, 2021).
[10] *Id.*
[11] Ed Yong, *A Popular Algorithm is No Better at Predicting Crimes than Random People,* THE ATLANTIC (Jan. 17, 2018), https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/.

accuracy of 65%.[12] This fact is worrisome. Before placing confidence in such algorithms to make decisions that will severely affect the quality of life of an individual, it is necessary to determine the transparency and accuracy of the software. Datasets being unrepresentative of minority groups may revamp the social, cultural, and institutional bias, leading to discrimination which is contrary to the spirit of conventions such as UN SDGs and the CEDAW.[13]

The Artificial Intelligence Global Surveillance Index noted that 75 out of 176 countries used enabling technologies such as Internet of Things (IoT), cloud computing for surveillance irrespective of the nature of the government.[14] A controversy was raised in 2019 when Amazon disclosed that several employees heard conversations of the customers recorded by Echo Smart Speaker.[15] The use of tools like facial recognition and smart policing violates the Harm Principle where human rights including liberty are disregarded.[16]

Therefore, the challenge is to incentivize the use of explainable AI without compromising innovation. This can be done if one shifts the approach from Caveat Emptor ('Buyer Beware') to Caveat Venditor ('Seller Beware').[17]

SLIDING SCALE APPROACH: ASCERTAINING LIABILITY

I have used the Cost-Benefit Analysis method[18] to determine the different thresholds of liabilities against different characteristics of AI. This is because ascertainment of liability in case of AI is

---

[12] *Id.*

[13] *Artificial Intelligence and Gender Equality*, UNESCO, 4 (2020), https://en.unesco.org/system/files/artificial_intelligence_and_gender_equality.pdf.

[14] Steven Feldstein, *The Global Expansion of AI Surveillance*, CARNEGIE ENDOWMENT FOR INT'L PEACE (Sep. 17, 2019), https://carnegieendowment.org/2019/09/17/global-expansion-of-ai-surveillance-pub-79847.

[15] Anthony Cuthbertson, *Amazon Admits Employees Listen to Audio from Echo Devices*, THE INDEPENDENT (Apr. 11, 2019),https://www.independent.co.uk/life-style/gadgets-and-tech/news/amazon-alexa-echo-listening-spy-security-a8865056.html.

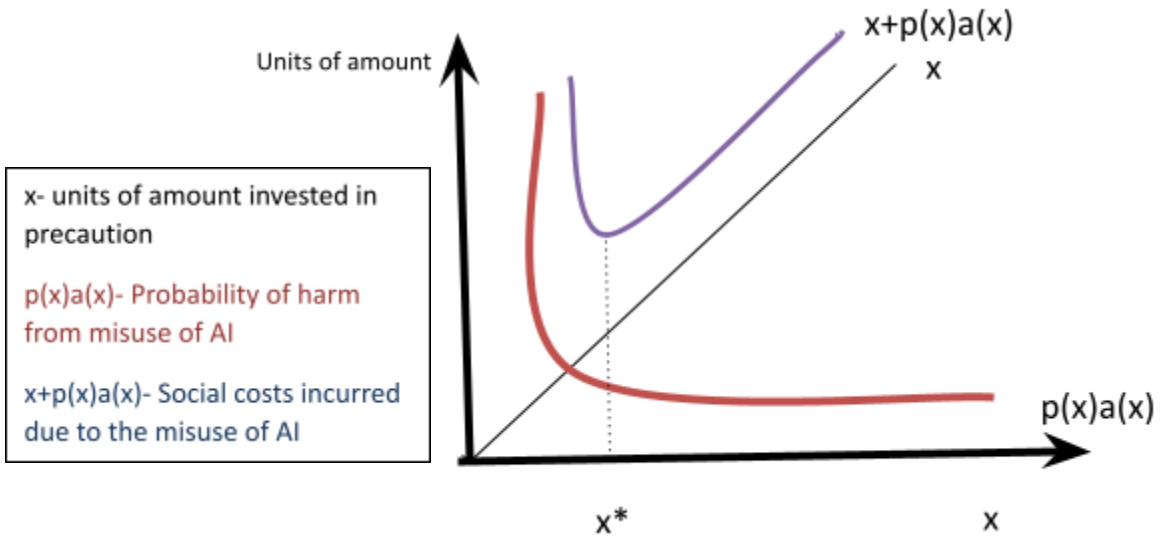
[16] David Brink, *Mill's Moral and Political Philosophy*, STAN. ENCYCLOPEDIA OF PHIL. (Aug. 21, 2018), https://plato.stanford.edu/entries/mill-moral-political/ (providing a description for the harm principle).

[17] *Caveat Emptor Law and Legal Definition*, USLEGAL, INC., https://definitions.uslegal.com/c/caveat-emptor/ (last visited Nov. 20, 2021); *Caveat Venditor Law and Legal Definition,* USLEGAL, INC., https://definitions.uslegal.com/c/caveat-venditor/ (last visited Nov. 20, 2021).

[18] Thomas J. Miceli, THE ECONOMIC APPROACH TO LAW, 42–46 (2004) (describing details of the Cost Benefit Analysis Method applied to AI below).

difficult due to its non-explainable characteristics.[19] As AIs have different degrees of explainability and supervision, the liability on the developer should be determined using the sliding-scale approach.[20] This means that different principles of negligence, strict liability, traditional tests etc. should be used depending on the specific characteristics of the AI such as transparency, monitoring, use etc.

Let's assume 'x' units of amount are invested in precaution and care by the injurer. Let P(x) be the probability of an accident which is a decreasing function as it decreases when one takes more precaution, and let a(x) represent harm from misuse of AI, such as release of confidential information, infringements on liberty, and promotion of data bias.[21] As can be seen in the below figure, P(x)a(x) is a decreasing function of x as the amount of harm caused reduces with the amount of investment. The expected social cost curve is x+ p(x)a(x) which is a U shape curve having the minimum value at the lowest point of the curve where x=x*. At x*, the social costs incurred due to AI will be minimum. The value of x* reflects the "value of precaution at minimum costs and avoidance of accidents."



[19] Danny Tobey, *Explainability: Where AI and Liability Meet*, DLA Piper (Feb. 25, 2019), https://www.dlapiper.com/en/germany/insights/publications/2019/02/explainability-where-ai-and-liability-meet/.
[20] Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 Harv. J. of L. & Tech. 890, no. 2, 2018, at 932–33.
[21] Mike Thomas, *7 Dangerous Risks of Artificial Intelligence*, Built In (Jul. 28, 2021), https://builtin.com/artificial-intelligence/risks-of-artificial-intelligence (describing risks associated with AI).

The rationale behind using the optimal value of x* as the standard for due care in this context is comprehensible when one understands that AI cannot be fully regulated because of the tradeoff between innovation and developer liability.[22] The above graph explains two consequences of imposition of liability. *First*, in a simple model, if no liability is imposed, AI may increase social costs by not valuing human dignity when it violates liberty and equality. *Second*, in the case of negligence, the accused tortfeasor has to have met the requisite standards for due care to avoid liability.[23] So, when x is greater than or equal to x*, no liability is imposed on the accused tortfeasor. However, if $x < x^*$, then the liability will be imposed and the total costs will be $x+ p(x)a(x)$.

Complications arise in the application of the standard because the x* standard for due care can only be calculated only if AI can be monitored by the developers, and the efficient value of the x* standard will vary with different characteristics of AI. For example, if the probability of background risk is 'b' and 'a' is the damage,then, it is necessary that the accused tortfeasor minimizes $[x + \{(p(x))/(p(x) + b)\} a(x)]$ at the optimal value of x* where the probability that an accident may occur when a background risk is involved is calculated by $(p(x))/(p(x) + b)$.

Imposition of strict liability of the x* standard may incorporate the cost incurred by the victim wherever supervision is possible.[24] However, certain errors in AI may be unforeseeable because of bug, bad data or a runaway feedback loop.[25] Hence, in AI it is not always possible to justify strict liability on account of 'opacity' in the model, which is due to the model's high complexity. Applying strict liability of the x* standard may therefore disincentivize innovation of high complexity models with 'opacity' and thus hamper development. [26]

---

[22] George Maliha et. al, *To Spur Growth in AI, We Need a New Approach to Legal Liability*, Business Law (Jul. 13, 2021), https://hbr.org/2021/07/to-spur-growth-in-ai-we-need-a-new-approach-to-legal-liability.

[23] Thomas J. Miceli, The Economic Approach to Law, 44 (2004).

[24] *See* Bathaee, *supra* note 20, at 931–32 (considering the potential effects if a strict liability framework were imposed to alleviate issues arising from black box AI).

[25] Andre D. Selbst, *Negligence and AI's Human Users*, 100 B.U. L. Rev. 1315, 1336–37 (2020) (discussing errors that can occur in the context of the US of AI in finance).

[26] Bathaee, *supra* note 20, at 931.

When there is absolute opacity, we cannot determine if the accused tortfeasor is negligent or not or at what point or at what degree it became negligent. Accordingly, broad scope liability can be imposed to indemnify the users of the software. If the users are risk-averse, they will be willing to have a premium equivalent to $\{(p(x))/(p(x) + b)\}$ a(x) so that the value of x is equivalent to x* to achieve a risk-neutral state. This is because, risk averse individuals avoid uncertainty and risks, and paying a premium amount ensures that security irregardless of the loss caused.[27] The advantages of first-party insurance have often been advanced by economists where victim-protection is the main concern.[28] However, with regards to the functioning of DNNs, neither the user nor the developer can explicitly determine the functioning of the algorithm. [29]

I propose that to maintain a balance, it is necessary that the users are obligated to purchase the insurance of the software developers and corporates irrespective of first-party or third-party insurance. Third-party insurance can be developed by setting up a regulatory authority which may either be a government or private organization in the age of globalization and heightened dependency on transnational companies. [30]

BURDEN OF PROOF- ON USER OR THE DEVELOPER?

The burden of proof can lie either on the plaintiff (user) or on the defendant (developer). However, if the burden of proof is placed on the plaintiff, then she may not have the requisite resources and expertise to prove the alleged claim, and, if the burden of proof is shifted on the developer, then it may discourage innovation or use of AI.[31]

---

[27] J. Singh, *Risk Aversion and Insurance (Explained with Diagram)*, Economics Discussion, https://www.economicsdiscussion.net/articles/risk-aversion-and-insurance-explained-with-diagram/1419.

[28] Michael G. Faure, *Economic Criteria for Compulsory Insurance*, 31 The Geneva Papers 149, 149 (2006).

[29] Will Knight, *The Dark Secret at the Heart of AI*, MIT Tech. Rev. (Apr. 11, 2017), https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/

[30] Ram Shankar Siva Kumar et. al, *The Case for AI Insurance*, Harv. Bus. Rev. (Apr. 29, 2020), https://hbr.org/2020/04/the-case-for-ai-insurance.

[31] *See* Bruce L. Hay et. al., *Burdens of Proof in Civil Litigation: An Economic Perspective*, 26 The J. of Legal Stud. 413, 413 (1997) (stating that giving one party the burden of proof saves the opponent costs that they would have otherwise incurred).

Hence, I suggest that the burden of proof should vary on a case to case basis depending on the trade-off between explainability and technological efficiency of AI. For instance, the victim might claim that her human rights, such as liberty, privacy or equality, have been jeopardized due to the use of AI. Now the Court, through the information 'B' presented in the petition, perceived some facts about the commission of the alleged act 'A'.

P(A) denotes the general likelihood of the alleged incident 'A', implying that if the frequency of the commission of an alleged act is high, then the probability of it being present in a specific case also rises. Consequently, P(B|A) and P(B|~A) reflects the probability that the court will take note of information B if A were to occur and the probability that the court will take note of B if A were not to occur respectively.

The following condition must be satisfied if burden of proof has to be established on the plaintiff: - **P(B|A) P(A) * plaintiff's cost < P(B|~A) P(~A) * developer/corporate's costs**.[32] This implies that the burden of proof will be on the plaintiff only if the costs of the plaintiff on the basis of probability that A occurred given B is present is less than the costs of the developer on the basis of the probability that A did not occur when B is present. In the case of complex DNNs, the plaintiff's cost in litigation will be higher due to non-availability of any proper evidence and the serious infringement of rights if human rights are affected. Therefore, the onus should shift from the plaintiff to the corporations. The shifting of the onus should depend on the facts and circumstances by evaluating if the plaintiff's costs are substantially higher or not.

CONCLUSION

Artificial Intelligence, with its rapid impact on societal development, has seemingly fueled the Fourth Industrial Revolution. But, at the same time, it is clouded with problems of incomplete information and asymmetric information due to lack of explainability and reliability of the

[32] *See id.* at 423 (providing the basic probabilistic model for burden of proof assignment).

output.[33] It is necessary to evaluate the problems to bridge the gap between human rights and the outputs realized by DNNs and SVMs. The rights and liabilities of using and developing AI have to be ascertained in a way that augments allocative efficiency, productive efficiency, and distributive efficiency of both the user and the developer.

[33] *See* Hal Varian, *Artificial Intelligence, Economics, and Industrial Organization*, *in* THE ECONOMICS OF ARTIFICIAL INTELLIGENCE: AN AGENDA, 401 (2019) (considering problems of AI operation in the gaming context in applying deep network techniques in situations with incomplete or asymmetric information).