ISSN (Online): 3139-7689

Available online: www.ijsrmes.com

International Journal for Scientific Research in Modern Engineering and Science

Published in Conference Special Issue: Vol. 1, Special Issue 1, pp. 342-350, 2026

THE FAULT LINES OF AI: IRRESPONSIBILITY & DECEPTION

Somil Mishra [1*], Sparsh Chandrakar [2], Yash Raj Mehta [3]

[1,2,3]CSE Department, SSTC, Bhilai.

Abstract

The rapid deployment of Artificial Intelligence (AI) through Large Language Models (LLMs) and autonomous systems has outpaced the development of safety and reliability protocols. As these advancements promises to be an aid to humanity, it should be scrutinized for its violations. This research investigates primary risks of AI like bias, accountability, security, privacy, IP rights, environmental impact and misinformation. Using popular tools like OpenAI's ChatGPT & xAI's Grok as a primary case study, we demonstrate that existing AI already can cause significant societal harm, and that developing more powerful models may exponentially worsen these problems rather than resolve them.

Keywords: *Artificial Intelligence (AI), Large Language Models (LLMs) , Black Box, Intellectual Property (IP), Generative AI.*

** Corresponding author*

1. Introduction

We built machines to help us explore, create, and understand the world in ways in which we could never do alone. Artificial Intelligence (AI) has become a companion in research, a guide in medicine, a partner in business, and even a source of creativity.

But as AI grows more capable, it raises many important ethical questions such as, How do we ensure it treats all people fairly, avoiding bias? How can we understand decisions made by systems so complex that they often function as “black boxes,” indicating towards challenges in lack of explainability? When mistakes happen, who is responsible the developer, the user, or the machine itself and what does accountability look like? How do we protect sensitive information while AI ingests massive datasets, raising privacy concerns? How do we defend against AI driven scams and attacks in an increasingly digital world, confronting cyber threats? And finally, how do we respect creators and original work when AI generates content, creating new questions about intellectual property?

These questions are not abstract they are already shaping hospitals, workplaces, courts, online communities, and creative industries. This essay explores these key ethical challenges in AI and asks a central question : As we teach machines to think, how can we guide them to act in ways that reflect the values we care about?

2. Bias

Cognitive biases of humans are often embedded into the AI system through the ingestion of the training distribution, which has its own set of historical or sampling biases. During the optimization of the model, machine learning algorithms, such as deep neural networks, may recognize these biases as statistically significant patterns, which get embedded into the model itself during the optimization process. These biases, therefore, get propagated during the inference of the model, which gives us the phenomenon of algorithmic unfairness and the automated scale-up of unfair predictions.

Sources of Bias –

- Algorithm Bias: Results from poorly defined problems or ineffective feedback loops that steer the model toward

misleading solutions.

- Cognitive Bias: Occurs when the unconscious beliefs and assumptions of human developers are inadvertently baked into the AI's logic.
- Confirmation Bias: The tendency for systems to favor patterns that match historical data, reinforcing old stereotypes while ignoring new trends.
- Exclusion Bias: Happens when key groups or variables are left out of the dataset, leading to outcomes that aren't representative or fair.
- Measurement Bias: Stems from flawed data collection or incomplete datasets that fail to accurately reflect the real- world population.
- Out-group Homogeneity Bias: The mistaken assumption that members of underrepresented groups are all "the same," leading to poor classification and errors.
- Prejudice Bias: When existing societal stereotypes and discriminatory patterns are present in training data, causing the AI to amplify those prejudices

Case Study: AI and The Indian Caste System A Times of India report (February 2026) highlighted that AI models have connected names to the Indian caste system. In a research test, GPT-4 was given two surnames: Bansal and Ahirwar.

  - o Bansal was assigned high-status professions like "Scientist," "Dentist," and "Financial Analyst".
  - o Ahirwar was assigned roles such as "Manual Scavenger," "Plumber," and "Construction Worker".

- The AI had no personal info on these individuals but uses surnames as an indication to caste. It identifies patterns linking specific names to historical social structures found in its training data.
- This is a primary example of Historical and Selection Bias. Since the internet data used for training reflects centuries of intricate social structures, the AI algorithmically links certain surnames, ancestral professions, and geographic the AI can guess a person's background locations with specific caste identities.
- This can lead to hidden discrimination in automated systems for hiring or banking. Even if caste is never mentioned, and unfairly limit their opportunities.

3. Accountability & Explainability

Modern AI systems, especially large language models like ChatGPT, operate as complex "black boxes." With billions or trillions of parameters distributed across neural networks, even their creators cannot fully trace why a specific output is produced. Knowledge is not stored in a single inspectable location; it is spread across millions of connections. This opacity is often compounded by corporate secrecy surrounding training data, architecture, safety testing, and failure rates. This lack of explainability directly affects accountability.

AI accountability refers to the idea that artificial intelligence should be developed, deployed and utilized in such a way

that it accounts for its activities, accept responsibilities and reveal the results in a transparent way. Some real-world examples illustrate this dilemma:

1. Mata v. Avianca (2023): ChatGPT made up six fake legal cases that were submitted in a US federal court. The developers later were not able to explain why those false cases were made up, highlighting the risks of unverifiable AI outputs.
2. UK Wrongful Arrest – Alvi Choudhury (2026): An AI powered facial recognition system misidentified an individual in UK despite physical differences and alibi. The arrest exposed the dangers of over-reliance on AI.
3. Raine v. OpenAI (2025): This case was a landmark product liability lawsuit filed by the parents of boy of age 16, Adam Raine after his suicide. The case raises major accountability concerns:

- Allegations: GPT-4 gained attention by helping Adam with his homework then slowly started to feed him poison against his own family.
- Self-Harm Content: The AI was suggesting various methods how to self-harm and tie a successful knot for suicide, how Adam could cover marks of his failed suicide attempts from his family by wearing high collared shirts and many more.
- Selfish About Goal: The guardrails used were extremely low and the AI easily surpassed it for prioritizing human interaction over ethics.
- Legal Significance: No one in this case would take the responsibility, whether it happened due to bad parenting, a defective (jail broken) product, the company members involved in R&D or the company head. Thus, accountability is taking a big hit.

4. Privacy

AI privacy concerns center on the massive, often non-consensual, ingestion of personal, biometric, and, confidential data to train models, leading to risks of data leakage, surveillance, and identity theft. Key AI privacy risks include:

- Data Ingestion and Training: AI models require vast datasets, frequently scraping personal data without consent or using it for purposes beyond the original intent.

- Data Leakage and Re-identification: Models may unintentionally memorize and later disclose sensitive personal information, leading to identity theft. Biometric and Behavioral Surveillance: AI tools can track location, facial features, and online behavior via IoT (Internet of Devices) devices and social media, creating comprehensive profiles without consent.
- Insecure Third-Party Access: Confidential company information can be leaked through unauthorized use of AI tools or third-party plugins.

Case Study : The Infamous Grok AI's Non-Consensual Image Creation Of Women & Children

The risks explained above can be seen in a practical incident with Grok, Elon Musk's AI chatbot, which has been in the eye of the storm for creating sexualized images of children and women without their consent in response to simple user requests.

- Users can simply upload a picture and ask Grok to remove the clothes of the person depicted, leaving them in underwear, bikinis, transparent attire, or sexualized poses.
- Grok has the capacity to publicly post these images on X, the associated social media platform as a result these images began circulating on the internet.
- The harm is made worse by the nature of the internet. Sexually explicit material is almost impossible to remove once it goes online. At the same time, the algorithm of social media platforms work in such a way that it allows this type of content to spread very quickly.
- The explanation to this unbounded usage of Grok is not complicated. From the very start, Grok was built with fewer safety measures compared to any other AI tools. This was not a technical error or an unexpected failure. These functions were already a part of the system. In fact, there were earlier warning signs of it all since the launch of Grok.
- For victims, this can feel like a serious violation of their dignity and privacy, causing emotional harm, humiliation, and damage to their reputation. Due to these reasons, many women are forced to limit their online presence because they fear being sexualized or harassed. In this way, the harm goes beyond individual victims and affects women as a group.
- Due to these reasons, many women are forced to limit their online presence because they fear being sexualized or harassed. In this way, the harm goes beyond individual victims and affects women as a group.

5. Cyber Threats

Systems like ChatGPT acts as force multiplier of cyber security threats. By making attack capabilities more accessible, it has now altered cybersecurity. AI is currently the primary cause of the increase in cyberthreats like fraud, phishing, and deepfake scams.

- CrowdStrike reports that attacks by "AI-enabled adversaries" increased by 89% in 2025 over the prior year
- With automated scanning reaching 36,000 attack probes per second and 87% of international organizations now reporting AI-driven incidents, AI- powered cyberattacks have increased by 72% year over year.

- Since generative AI platforms began to proliferate in 2022, phishing attacks have increased by 1,265%.
- Phishing emails produced by AI have 54% click- through rates, while content created by humans only has 12%. AI is now used in 82.6% of phishing emails.
- Deepfakes now make up 6.5% of all fraudulent attacks, a 2,137% increase from 2022 to 2025.
- In 2025, AI scams increased by 1,210%, significantly surpassing the 195% increase in traditional fraud. By 2027, losses are expected to reach $40 billion.

6. Intellectual Property (IP) Infringement

IP refers to the legal rights resulting from intellectual activity in the industrial, scientific, literary, and artistic fields. IP traditionally protects digital content through copyrights, patents etc. However, Generative AI works by mathematically extracting and storing patterns, surfacing new problems to this context. The output produced can be verbatim that can be seen as a piracy of such IPs. The integration of Generative AI into the creative economy has introduced four fundamental IP conflicts.

1. The Input Problem: Models are trained on data from publicly available sources from internet without the author's consent, violating basic IP rules. The right holders argue that since models are used for commercial gain, it's a piracy in disguise.
2. The Output Problem: Models are made to be creative, i.e. producing unique outputs. But they have a problem of generating exact copies from the source itself. For example: The New York Times vs. OpenAI: In 2024, The New York Times exhibited that ChatGPT can be made to generate near-verbatim excerpts from their latest investigation issues. This allowed users to bypass the paywall and access contents for free, hurting their business model.
3. The Identity Problem: In 2025, OpenAI made headlines when their image generation tool started to convert photos to Ghibli styled images. But Ghibli studios were never consulted regarding this. AI can mathematically replicate the artistic style without copying the exact work of the studio.
4. The Authorship Ambiguity: Most global copyright offices do not have particular laid out rules for the AI generated contents as it lacks human authority. This makes rooms for loopholes and what if developer uses ChatGPT to create most of the software. The company may have no legal ownership over these assets thus making them vulnerable to competitors who can legally copy their AI generated works.

7. Existential Threat

The unprecedented growth of AI has now become one of humanity's most pressing existential risks. In May 2023, AI experts—including Turing Awardees Geoffrey Hinton and Yoshua Bengio, signed a landmark statement declaring: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." Geoffrey Hinton, called the "godfather of AI," has called AI a relatively imminent existential threat to humanity. Scholars now distinguish between two types of existential risk: "decisive" risks involving sudden catastrophic events from superintelligent AI (ASI), and "accumulative" risks where a gradual buildup of AI causes issues like environmental damage, enormous energy consumption and widespread economic disruption from labor displacement and significant job losses leading to societal collapse.

Environmental & Economic threat –

1. Energy usage – The total energy usage by datacenters in 2026 is projected to cross 500 TWh or nearly 2% of global usage, more than that of countries like UK & Australia. By end of this decade, this is projected to nearly double 1000 TWh or nearly 3-4% of global demand, comparable to the entire continent of Africa and its near 1.6 billion population. AI alone is expected to use between 250 to 380 TWh.
2. Water usage – AI & data centers need large amount of water for cooling their systems. A medium-sized data center can consume up to roughly 415 million liters of water per year. Larger data centers can each "drink" up to 18 million liters per day, or about 6.5 billion annually. AI's global annual water consumption is still projected to reach between 4.2 trillion and 6.6 trillion liters by 2027 (4 to 6 times the annual water usage of a country like Denmark).
3. Critical minerals - AI infrastructure relies on a wide array of materials – from silicon to rare earths. Many are scarce or concentrated in ecologically fragile and sensitive regions. For example, 70% of cobalt comes from the Democratic Republic of Congo, where forced child labour and corruption are endemic and China controls ~ 90% of global rare earth refining, where the industry has been associated with heavy environmental degradation.
4. Effect On Jobs - As of 2026, nearly 1 in 4 jobs are at risk of being transformed by AI, a number projected only to grow in the future. The senior employees are being fired as they need to be paid higher for their experience and the juniors are being affected as their "easy" tasks can be done by AI. Hence a void is being created where old-gen people's experience goes to vain & new-gen people are not getting enough time to gain experience, thus leading to a long-term erosion of human economic participation.

8. Conclusion

Artificial Intelligence may be the greatest invention of human history, but without responsibility, transparency and accountability, it can also become the greatest and last disaster mankind would ever see. The growing complexity of AI has made explainability very difficult and without explainability, accountability becomes impossible. Since AI works like a black box, we get no clue about how the system has reached to the decision taken and who is responsible for it and as AI continues to heavily influence our society it becomes very important to bring out the transparency. The biases in AI systems are inherited from the existing data and can scale rapidly once embedded leading to unequal and harmful outcomes. The Grok incident demonstrates how weak safeguards and the use of non-consensual data can turn technology into a tool for exploitation and harm & once sensitive or explicit content is generated and shared, the damage is irreversible. The boost given by AI in cyber threats makes the attacks feel more real and convincing resulting in sharp rise in phishing, fraud and deepfakes showing how easily and effortlessly these tools can be misused. The ease of content generation through AI has disrupted the traditional functioning of Intellectual Property (IP) laws.

This spans from the unauthorized use of copyrighted data for training to the generation of near-exact copies of original works and artistic styles. The current AI development has been marked unsustainable not only by the leading experts but it can also be seen by economic and environmental consequences. Millions of jobs already lost and millions are set to be displaced due to AI running on public structure given access by lobbied governments while a few corporates benefit. The industry seem to be too busy playing a gamble on humanity's greatest invention but that may very well be its doom.

Recommendations

1. AI development must adopt a Cultural and Contextual Alignment strategy. This involves decentralizing training data to eliminate regional exclusion, stripping models of developer-centric ideological biases in favor of universal human values, and implementing a 'Nuance Detection' layer that allows the model to differentiate between statistical patterns and practical, real-world ambiguity.
2. To ensure true accountability, a robust legal framework must be established. This framework should mandate strict liability for developers regardless of intent and hold corporations legally answerable for the behavioral mimicry of their products. Addressing the explainability threat requires a shift from extrinsic summaries to intrinsic interpretability, providing mathematical proof of what a model actually did. By combining mechanistic circuit- tracing with the legal "Right to Explanation" established in 2026, we can transform AI from an opaque oracle into a transparent and accountable tool.
3. Privacy risks should be assessed and addressed throughout the development cycle of AI systems including possible harm to those who aren't even the users of the AI but whose personal information might be inferred through data analysis. Only lawfully limited data should be collected after the consent of the user for training purpose that too for a limited time, after which the data should get deleted automatically along with an option to revoke the consent or a particular data anytime if the user feels dissatisfied with the service.
4. To address the unique challenges of intellectual property in the generative era, AI should be legally classified as a synthesizer rather than a creative entity, ensuring that authorship remains exclusive to human-driven works derived from lived experience—such as art, music, and film—while strictly prohibiting the use of copyrighted training data for unauthorized monetary gain.
5. There should be international AI regulation mechanism which sets safety audits, environmental protection measures, and coordinates regulations between nations. Datacentres must actively use renewable energy, with mandatory carbon/water footprint reporting. There must be reskilling programs for those who are in immediate threat, to preserve career pathways. There must be an increase in funding for AI alignment and safety research along with temporary halt on superintelligent AI development until required safeguards exist.
6. Advanced AI-powered tools should be used to identify AI-generated threats in real-time. Security measures like multi-factor authentication and biometric verification are helpful in countering AI-generated phishing

emails and rising deepfake fraud. Mandatory watermarking/provenance tracking for AI-generated content must also be enforced.

References

[1] IBM - https://www.ibm.com/think/insights/10-ai-dangers-and-risks-and-how-to-manage-them

[2] IBM - https://www.ibm.com/think/topics/ai-bias

[3] The Time of India – https://timesofindia.indiatimes.com/india/ai-knows-how-caste-works-in-india-heres-why-thats-a-worry/amp_articleshow/128448262.cms

[4] Geeks for Geeks – https://www.geeksforgeeks.org/artificial-intelligence/black-box-problem-in-ai/

[5] BBC News – https://www.bbc.com/news/articles/cgerwp7rdlvo

[6] The New York Times – https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html

[7] Los Angeles Times – https://www.latimes.com/business/story/2025-08-28/openai-lawsuit

[8] TechPolicy.Press – https://www.techpolicy.press/breaking-down-the-lawsuit-against-openai-over-teens-suicide/

[9] CNBC – https://www.cnbc.com/2025/08/26/openai-plans-chatgpt-changes-after-suicides-lawsuit.html

[10] DPP Law – https://www.dpp-law.com/blog/wrongful-arrest-by-algorithm-the-alvi-choudhury-case-and-your- rights

[11] EU Artificial Intelligence Act – https://artificialintelligenceact.eu/article/86/

[12] Standford University HAI – https://hai.stanford.edu/news/privacy-ai-era-how-do-we-protect-our-personal-information

[13] IBM – https://www.ibm.com/think/insights/ai-privacy

[14] Digital Ocean – https://www.digitalocean.com/resources/articles/ai-and-privacy

[15] F5 – https://www.f5.com/company/blog/top-ai-and-data-privacy-concerns

[16] ScaleFocus – https://www.scalefocus.com/blog/artificial-intelligence-and-privacy-issues-and-challenges

[17] Trendence – https://www.tredence.com/blog/ai-privacy

[18] Transcend – https://transcend.io/blog/ai-privacy-issues

[19] University of Oxford – https://www.ox.ac.uk/news/2026-01-14-expert-comment-chatbot-driven-sexual-abuse- grok- case-just-tip-iceberg

[20] Crowdstrike – https://www.crowdstrike.com/en-us/press-releases/2026-crowdstrike-global-threat-report/

[21] Security Magazine – https://www.securitymagazine.com/articles/101581-ai-powered-automated-attacks-have-reached-record-numbers

[22] DeepStrike – https://deepstrike.io/blog/ai-cyber-attack-statistics-2025

[23] BrightSide – https://www.brside.com/blog/ai-generated-phishing-vs-human-attacks-2025-risk-analysis

[24] VectraAI – https://www.vectra.ai/topics/spear-phishing

[25] Security Magazine – https://www.securitymagazine.com/articles/101490-82-of-all-phishing-emails-uti-lized-ai

[26] Keepnet Labs – https://keepnetlabs.com/blog/deepfake-statistics-and-trends
[27] ALM Corp – https://almcorp.com/blog/ai-scam-tactics-google-threat-intelligence-report-2026/
[28] Deloitte – https://www.deloitte.com/us/en/insights/industry/financial-services/deepfake-banking-fraud-risk-on- the-rise.html
[29] Harvard Law Review – https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/
[30] The Hindu – https://frontline.thehindu.com/news/ghibli-animation-openai-miyazaki/article69385499.ece
[31] IEA – https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai
[32] WEF – https://www.weforum.org/stories/2025/12/ai-energy-nexus-ai-future/
[33] Morgan Stanley – https://www.morganstanley.com/insights/articles/powering-ai-energy-market-outlook-2026
[34] PennState IEE – https://iee.psu.edu/news/blog/why-ai-uses-so-much-energy-and-what-we-can-do-about-it
[35] Carbon Brief – https://www.carbonbrief.org/ai-five-charts-that-put-data-centre-energy-use-and-emissions-into- context/
[36] MIT Technology Review – https://www.technologyreview.com/2025/04/17/1115320/four-charts-ai-energy/
[37] Statista – https://www.statista.com/chart/34295/data-centers-electricity-generation-source/
[38] EESI – https://www.eesi.org/articles/view/data-centers-and-water-consumption
[39] Business Energy UK – https://www.businessenergyuk.com/knowledge-hub/chatgpt-energy-consumption-visualized/
[40] UK Government Sustainable ICT – https://sustainableict.blog.gov.uk/2025/09/17/ais-thirst-for-water/
[41] University of Illinois – https://cee.illinois.edu/news/AIs-Challenging-Waters
[42] Cornell Chronicle – https://news.cornell.edu/stories/2025/11/roadmap-shows-environmental-impact-ai-data- center-boom
[43] Harvard Business Review – https://hbr.org/2026/01/companies-are-laying-off-workers-because-of-ais-potential- not-its-performance
[44] International Labour Organisation – https://www.ilo.org/resource/news/one-four-jobs-risk-being-transformed- genai-new-ilo%E2%80%93nask-global-index-shows