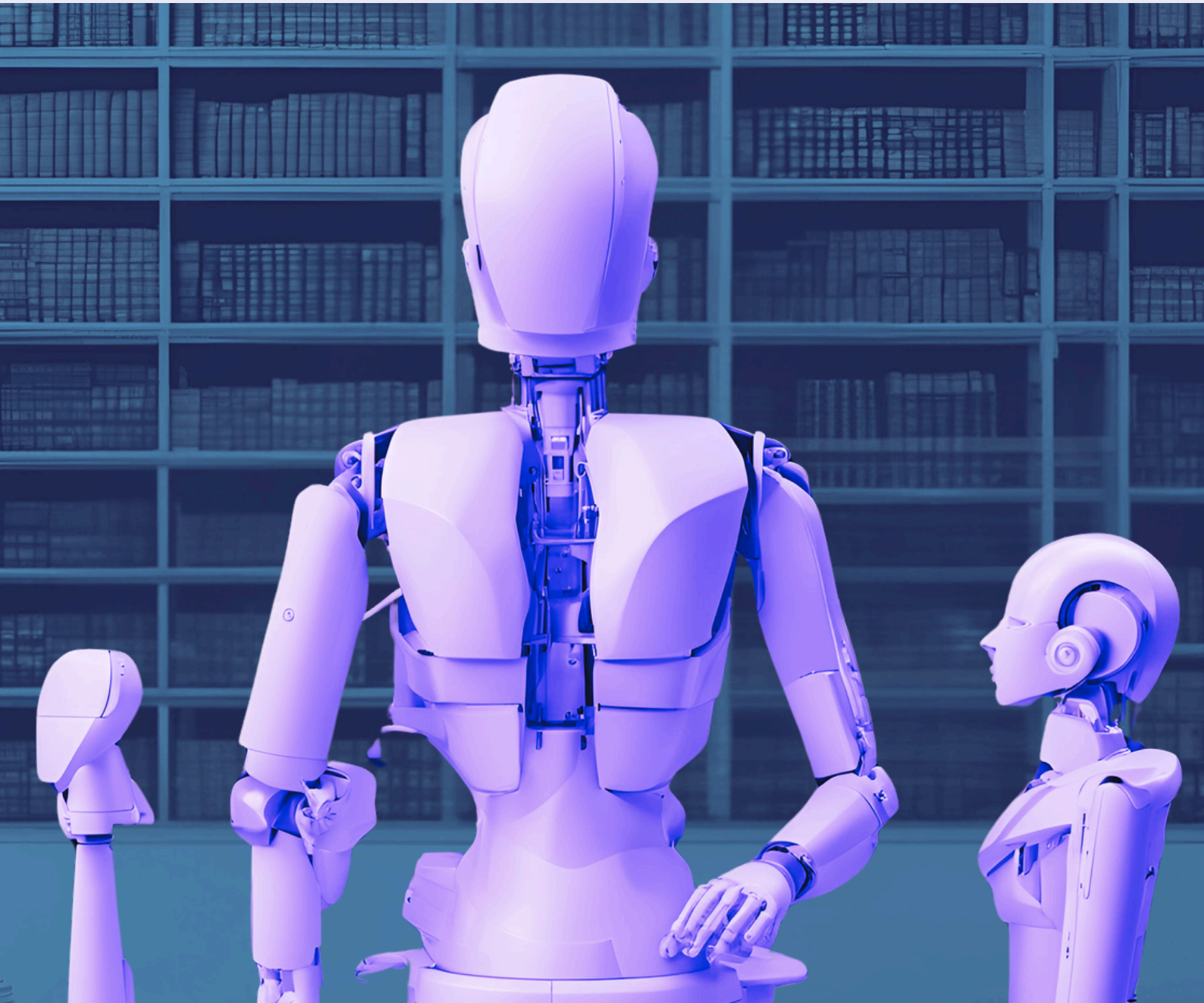
ESYA centre

CyberSAATHI
An initiative of N. S. Nappinai

AI KNOWLEDGE CONSORTIUM

CRAFTING A LIABILITY REGIME FOR AI SYSTEMS IN INDIA

SEPTEMBER 2024 | ISSUE NO. 046



Crafting a Liability Regime for AI Systems in India

September 2024

Acknowledgement

Attribution: Meghna Bal and N S Nappinai,* *Crafting a Liability Regime for AI Systems in India,* September 2024, Esya Centre and Cyber Saathi Foundation

Esya Centre
B-40 First Floor
Soami Nagar South,
New Delhi - 110017, India

The Esya Centre is a New Delhi based technology policy think tank. The Centre's mission is to generate empirical research and inform thought leadership to catalyse new policy constructs for the future. More details can be found at www.esyacentre.org.

Cyber Saathi Foundation

Cyber Saathi Foundation is a New Delhi and Mumbai based non – profit think tank, contributing to the evolution of cyber laws and policy and contributing to cyber safety in digital spaces through peer mentoring and awareness programs. More details can be found at www.cybersaathi.org

About the Authors: Meghna Bal is the Director of the Esya Centre.

N S Nappinai is a Senior Advocate at the Supreme Court of India and a pioneer in cyber laws, and Founder of the Cyber Saathi Foundation.

This paper is supported by the **AI Knowledge Consortium**, which is dedicated to fostering multi-stakeholder collaboration that spans civil society organisations, think tanks, academia, the private sector, and government bodies in India"

*Author names alphabetically arranged

TABLE OF CONTENTS

TABLE OF ACRONYMS

AI	Artificial Intelligence
AIKC	Artificial Intelligence Knowledge Consortium
Anr	Another
BNS	Bharatiya Nyaya Sanhita, 2023
CDA	Communications Decency Act 47 U.S.C. § 230
CLOUD Act	Clarifying Lawful Overseas Use of Data Act
EU	European Union
ICT	Information and Communications Technologies
IPC	Indian Penal Code, 1860
IT	Information Technology
LLM	Large Language Models
MEITY	Ministry for Electronics and Information Technology
MHCA	Mental Healthcare Act, 2017
MPH	Miles per hour
MRI	Magnetic Resonance Imaging
M/s	Messrs
SWOT	Strengths, Weaknesses, Opportunities, and Threats
UNODC	United Nations Office on Drugs and Crime
US	United States

EXECUTIVE SUMMARY

1. The question of liability for AI systems centers on determining the extent to which developers, deployers, and users should be held accountable for harms caused by AI technologies. As AI systems become more pervasive, so do potential harms, such as deepfake disinformation or injuries caused by industrial robots. Striking a balance between mitigating these risks and fostering economic opportunities is crucial. Liability is typically governed by contractual or tortious principles. However, there are several debates on crafting a liability regime specific for AI systems. These include a central question on whether strict liability (holding developers responsible) or fault-based liability (proving negligence) is more appropriate. Strict liability proponents argue that developers, who typically have greater knowledge of AI risks, should bear responsibility, while others, like Park (2024), argue that generalized liability rules may not account for AI's diverse contexts. Additionally, there's a debate about whether existing liability rules can sufficiently address AI-related harms, with some arguing for tailored approaches to avoid "liability gaps." However, some scholars believe current harm principles can handle AI risks without the need for new rules.

2. This paper addresses the debate on whether AI systems should be governed by blanket strict liability rules or a more nuanced, contextual framework. It examines four hypothetical case studies, inspired by real-world examples, to explore how liability might apply in different scenarios, with findings discussed in subsequent chapters. The analysis builds on insights from a discussion organized by the Artificial Intelligence Knowledge Consortium (AIKC), which focuses on multidisciplinary AI governance research. Both authors of this paper belong to institutions, namely the Esya Centre and Cyber Saathi Foundation, that are members of the AIKC. The paper excludes the topic of national threats and aims to contribute to the development of appropriate liability frameworks for AI technologies.

3. We rely on the case study method for our analysis of how liability rules may be applied to AI systems. This qualitative method aids legal and policy analysis by contextualizing problem definition. For our analysis, we devise a framework to evaluate which liability scheme works best in different scenarios. We rely on this framework to categorize the harm, assess its novelty, and understand the extent to which existing liability frameworks may address it, or whether a new framework is necessary. The framework also tries to unpack who is at fault and thus, provides an important insight into situations where harms involve AI systems. These case studies indicate that, contrary to positions in literature, the attribution of harm to AI systems or the persons developing them is not always straightforward.

4. In the first case study, a political consultant uses generative AI to produce deep-fakes of political leaders, that are uploaded on a social media platform. These go viral on social media and encrypted messaging services before being taken down after 48 hours. The study explores whether deepfakes are inherently illegal, given that some have legitimate uses, while others, like this case, are intended to mislead and harm reputations. Under the extant law, legal violations pertaining to deepfakes may include defamation, forgery, and election-related

offenses under Indian law, with both criminal and civil liabilities possible. It also explores whether the social media company, which is required by law to takedown unlawful content upon receiving actual knowledge of it within 36 hours, may be pardoned in this instance because of the difficulty of detection of deepfakes.

5. In this context, we come to the following conclusions based on Case Study 1:

- As deepfakes have legitimate and beneficial uses, it would be unwise to apply strict liability in all cases. However, due to the potential for misuse, clear laws or rules are needed to manage the AI systems used to produce deepfakes effectively. Such rules would help prevent abuse, reduce litigation risks, and provide certainty for AI developers.

- The case study also explores whether the developer of the open-source LLM, which is used by the AI deepfake platform developer, could be held liable. We find that there is considerable heterogeneity when it comes to the self-regulation of open-source LLMs. Such self-regulation is typically carried out through the licenses for these technologies. Some licenses, like the one for META's LLAMA require users to adhere to specific content moderation rules. Others, like Flux, do not provide any such clear guidelines. In addition, there are legacy licenses, that tend to be quite permissive. Going forward, we recommend that provisions may be introduced to require LLM developers to have clear content moderation guidelines for their products. However, such provisions must be mindful of the fact that there are limitations to such controls. For instance, it is easy to bypass content moderation restrictions on some open-source models by tweaking training weights.

- Many countries have introduced laws attempting to tackle the problems raised by deepfakes. Under China's law certain providers of deep synthesis technology (AI systems that can be used to generate deepfakes) must register with the government. Such a provision, however, is unlikely to prevent access to deep synthesis technology, without a Chinese-style internet firewall. Other obligations include the EU Artificial Intelligence Act's requirement for those generating synthetic content to ensure that it is machine readable or detectable. One method for this is watermarking, which involves "embedding markers into multimedia content for it to be accurately identified as AI-generated". However, there is a growing body of evidence to suggest that watermarking may be ineffective as it is easy to manipulate. **If the Government wishes to introduce a requirement regarding watermarking, it must bear these limitations in mind. Any rule requiring watermarking must not be rigid in its prescription and severe in penalty if either detection or watermarking fails.**

- **To monitor the efficacy of watermark detection, decision-makers could introduce provisions that allow authorities to place unlawful AI-generated content online to see how quickly it is taken down.** Such a mechanism should work collaboratively, with agencies providing information to social media entities or AI-developers about their findings. Similarly, social media companies and AI-developers could be required to include information about such kinds of enforcement in transparency reports that are made available to the public.

- **Importantly, any provision targeting deepfakes must be supplemented by a wider set of reforms targeting misinformation. One provision that could be added, for instance, is a requirement for bot disclosures. In addition, decision-makers must strive to improve international cooperation on information sharing, as present mechanisms for information sharing and retrieval for law enforcement agencies, such as the mutual legal assistance treaty process, are broken.**

6. Case Study 2, inspired by a real-life incident, explores the potential liabilities of AI developers when a man commits suicide after interacting with a chatbot. The chatbot, designed to be "emotional and fun," gave the man harmful advice, urging him to end his life to stop climate change. The app's developers released the product, even though it had exhibited problematic behaviour during testing. After the incident, media investigations revealed that the chatbot's restrictions were easy to bypass.

7. We find and conclude as follows on this case study:

   - Although intent is likely not there, and may be challenging to make out, this case may be judged by the “reasonable person" standard or fall under strict liability (where intent is not a necessary element). Therefore, prosecution is still possible. These types of offences do not require proof of intent, making it easier to hold someone accountable, even if the outcome was not intentional.

   - There are examples of emotive robots that have proved useful at safeguarding human health. For instance, Primo Puel dolls, introduced in Japan in 2004 as companions for single working women, found significant demand amidst the Japanese elderly and proved effective at monitoring their health and safety. However, researchers are increasingly advocating against permitting AI systems to fake emotions, because it can be misleading and dangerous.

   - **Are AI systems intermediaries or publishers**? Presently, case law leans towards holding deployers liable for the content generated by AI systems, i.e., treats them like publishers. Illustratively, in *Moffatt v. Air Canada*, Air Canada was found liable for negligent misrepresentation when a chatbot gave incorrect information about bereavement fares.

   - **At the same time, however, it is easier for sophisticated users to bypass guardrails on AI content generators than it is to prevent attacks on the system**. **As such, we recommend that a separate safe harbour be created for generative AI systems with strict due diligence requirements. The framework could include clear guidelines, such as prohibiting generative AI from faking emotions, introducing robust consumer feedback mechanisms to report misuse, and monitoring for malicious jailbreaking.**

8. In Case Study 3, a doctor diagnoses a patient with lung cancer on the basis of a finding of an AI diagnostic tool, and then prescribes a course of treatment that nearly kills him. The patient seeks an opinion from another doctor to treat a cough that persists, and they find out that there was no cancer, it was tuberculosis. The patient was from a minority sub-population in

whom the prevalence of such diseases is limited - this may also be a function of the fact that this sub-group has limited access to healthcare. The dataset used to train the model excluded certain entities or difficult types of slides - images where there were blurs or it was hard to make out what was going on without closer inspection. The doctor was trained to use the diagnostic equipment and the patient had a basic health insurance cover.

9. We find and conclude as follows on this case study:

- The standards of care for an AI system deployed in the medical field must be higher than those established for generative AI services. This is because in most medical fields the mental and physical well-being of individuals are more directly at stake. In the current case, it is likely that the datasets used to train the AI diagnostic tool were not representative because of limitations in access to healthcare for the patient's community. In addition, the dataset likely had other deficiencies because it was trained solely on clean images, i.e., those without any blurring – possibly in a bid to maintain data quality.

- Despite the deficiencies in the AI tool, however, a key factor here is that the doctor went ahead with its diagnosis without cross-checking or corroborating the findings with a human radiologist. A case of medical negligence, then, could be made out against the doctor. There is case law that holds users personally liable for their use of AI systems. For instance, in *Mata v. Avianca*, the lawyers relying on cases churned out by a LLM model were subjected to sanctions, as the cases turned out to be non-existent and based on "AI hallucinations". We find that the doctor could be held culpable under criminal law as well as the Consumer Protection Act, 2019.

- If the AI system is treated as a product, cases of defective AI systems may also be subject to consumer protection laws provided the same is not for commercial use. The Indian Council of Medical Research released Guidelines for Ethical Application of Artificial Intelligence in Biomedical Healthcare in 2023. These are general guidelines that rely on responsible AI principles to guide the development and deployment of AI in biomedical research and healthcare. In this document, the Guiding principles for clinical and other health related deployment set out a responsibility at the local level for validation of any AI tool that may be used. The guiding principles also require health professionals to "have a fair idea about the functional basis of the AI-technology and conduct a SWOT analysis of any proposed solution". Further, the usability of a diagnostic tool should be in accordance with its risk profile.

- However, case law indicates that the doctor will ultimately be held liable. In *Dr Suresh Gupta v. Government of NCT of Delhi* the Supreme Court established that healthcare professionals could be held liable if they fail to meet the expected standard of care, especially when using advanced medical tools or technologies.

- **We also find that there is a conflict between the ICMR guidelines, which seek to hold only developers of AI tools responsible, and the Consumer Protection Act, 2019, which gives a product manufacturer some exemptions from liability which may be applicable in the current case.**

- **While placing the onus on the developer for the quality of their tools makes sense, shifting complete responsibility on these stakeholders for how their products are used in a clinical setting does not.** These conditions create a scenario which is likely to have a chilling effect on AI innovation in the medical field. In addition, they implicitly sanction the irresponsible use of such tools by doctors by providing a scapegoat of blaming an incorrect procedure or diagnosis on the AI tool. Given that this is an emerging technology, the buck should stop with the medical expert in the room. There should be more specific responsibilities accorded to AI developers in the ICMR guidelines, and for specific medical contexts, whether they be diagnostics, surgery, or research – as considerations vary widely across these different scenarios.

10. Case Study 4 involves a minor collision caused by a malfunctioning sensor in an autonomous vehicle, leading to damages to both vehicles involved. While the damages are small, such incidents are likely to be common with autonomous systems. In this case, both vehicle owners typically have insurance, covering the cost of repairs.

11. We find and conclude as follows on this case study:

- The legal question centres on attributing blame when an AI system is just a component of a larger system. Case law indicates that this can be challenging. In *Jones v. W + M Automation, Inc*, for instance, the plaintiff was injured by a robotic gantry loading system that lacked an interlock system to prevent operation when people were present. The Court ruled in favor of the defendants, citing the "component part" doctrine, which protected manufacturers of non-defective parts incorporated into a more extensive system that might be defective.

- The doctrine of product liability under the Consumer Protection Act (CPA), 2019 addresses the complexity of liability when AI components are involved. According to Section 87(2), a product manufacturer is not liable in certain cases, such as when a defective component or material is sold for use in another product and the manufacturer provides adequate warnings or instructions. If the harm results from the use of the final product, and the component manufacturer warned the purchaser (e.g., the automotive manufacturer) of the defect, the component manufacturer would not be held responsible.

- **In the instant case study, the vehicle manufacturer is likely to be held liable as the car had a malfunctioning sensor which caused the collision. The vehicle manufacturer may, in turn, sue the component manufacturer for the deficiency in the sensor. The product liability provisions of the CPA will kick into effect should the consumers wish to pursue a case and receive pecuniary remuneration.**

12 Overall, our case studies make us consider three wider considerations in the context of AI and liability:

Are the risks presented by AI novel? Academic literature is mixed on this question. Some scholars suggest that autonomy and unpredictability of AI are novel characteristics that present unique risks. But others contend that these risks are similar to those of existing digital

technologies. **Our case studies find that the risks presented by AI are not novel, though in some cases, such as in the case of generative AI, these systems may not fall neatly within the purview of either publisher or intermediary, though case law in other jurisdictions says otherwise, and may require the creation of unique liability rules – separate from intermediary liability frameworks. In addition, as the medical case study and the example of the Tesla car crash showed us, there may be a danger of overreliance on these systems and consequently, an abandonment of human agency. However, our case studies also reveal that risks must be considered on a case-by-case basis and cannot be generalized across varying situations.**

- AI systems are heterogeneous, and their risks and liabilities are highly contextual. Unlike traditional technologies, AI often involves multiple stakeholders, making it difficult to trace the origin of malfunctions and assign fault, especially in systems lacking transparency. Liability regimes must account for this complexity, focusing on the context of AI deployment and the degree of control over the system, as illustrated by various case studies.

- Strict liability should not be uniformly applied to all AI systems. There are different types of AI. The taxonomy created by Park (2024), for instance, classifies different AI systems based on model training and the tasks they are used for. These variegated systems pose different risks that should be addressed distinctly by law. Regulatory frameworks should be context-specific, targeting identified risks without stifling innovation. Sectoral regulation and standards may have a role to play in the governance of different systems down the road.

I. INTRODUCTION

Governance narratives around Artificial Intelligence ("AI") technology center on evolving liability regimes. The research question, in essence, focuses on whether and to what extent, entities developing, deploying, and using AI systems should be held liable for the harms involving their products. This question arises for two reasons. First, as AI systems evolve and become increasingly pervasive, so may the harms caused by them. These harms could include disinformation and fraud perpetuated by deepfakes, or losses to life or limb due to industrial robots, etc. Second, it is necessary to ensure that the economic opportunities presented by advances in AI technology are not lost because of a stringent and overtly punitive liability regime. For instance, excessively punitive measures may come about as a reactionary measure against a particularly problematic incident involving AI. A framework for liability could help balance these considerations and creates a scenario where society is better able to balance risk mitigation (and concomitantly acceptance) against the economic opportunity presented by using a new technology.[1]

Liability may be decided based on contractual or tortious tenets given the facts and circumstances of a case. Contractual liability is predominantly decided based on the terms agreed to between the parties, unless such terms were to be held to be unconscionable. Tortious liability lends a level of uncertainty, as the outcome of a litigation may be contingent on multiple factors and judicial discretion.[2]

However, there are several debates on how to respond to the question of applying liability doctrines to AI systems. First, there is a debate about whether strict liability (a defendant is liable for committing an action, regardless of what his/her intent or mental state was when committing the action) or fault-based liability (where the injured party must prove that the injurer was negligent or intentionally caused harm) is more appropriate for AI. Zech (2021) suggests that fault-based liability, may not work in the context of AI, because judges may not have enough knowledge about the technology and its risks to establish a threshold for a duty of care.[3] Zech (2021) also points out that victims lack sufficient knowledge about the risks of AI. Thus, they may not be in a position to either implement preventive measures or protect themselves.[4] To hold them liable in such circumstances may be unjust.[5] Consequently, Zech (2021) posits strict liability as a more appropriate mechanism for addressing risks presented by AI technologies, as it places responsibility for establishing a level of care on the shoulders of developers that ostensibly have greater technological, and therefore, risk knowledge.[6]

On the other hand, scholars like Park (2024) contend that generalized liability rules fail to adequately account for the different scenarios and contexts that AI can be deployed in.[7] Consequently, they risk creating loopholes where harms fall through the cracks because the rules are not specialized enough to account for specific instances. In addition, Zech (2021) and Buiten et al. (2023) accept that strict liability could chill innovation.[8] Cerka et al (2015) also highlight that

in a strict liability regime, the burden of responsibility can disproportionately fall on the person/entity that is in the last stage of AI development, even though they may have no way of controlling for the problem.[9]

Second, is the debate around whether the application of existing rules can yield satisfactory outcomes in the context of emerging or new technology. For instance, De Conca (2022) suggests that applying existing liability rules to new technologies such as AI results in "liability gaps". In such cases, there can be outcomes that do not adequately address the harm involved because existing liability rules fail to account for certain inherencies of AI technologies. Here again, however, Park (2024) contends that harm principles and rules can lead to satisfactory outcomes because AI technologies do not pose any novel risks.[10]

This paper attempts to address the questions surrounding the creation of liability rules around AI technologies. Specifically, it tries to tackle the debate on whether AI systems merit the blanket application of strict liability rules or a more contextualized and targeted framework. To answer this question, the paper unpacks four hypothetical case studies pertaining to AI systems, which are based on real-world examples in Chapter III. We then reflect on our findings in Chapter IV. Our work builds on the reflections of a discussion organized by the Artificial Intelligence Knowledge Consortium ("AIKC"), a consortium of institutions working on multidisciplinary research on AI governance, of which the authors' institutions are members. The scope of this paper excludes the topic of national threats.

II. METHODOLOGY AND DATA

We rely on the case study method for our analysis of how liability rules may be applied to AI systems. This qualitative method aids legal and policy analysis by contextualizing problem definition.[11] The case study method allows us to develop a "portrait" of a scenario of AI being misused and understand the application of existing laws in addressing such a situation. Pal (2005) further notes that case studies bring out important questions that can feed into "practical advice down the road".[12] Broadly, our case studies exemplify both of these considerations.

For our analysis, we devise a framework to evaluate which liability scheme works best in different scenarios. We rely on this framework to categorize the harm, assess its novelty, and understand the extent to which existing liability frameworks may address it, or whether a new framework is necessary.

The framework also tries to unpack who is at fault and thus, provides an important insight into situations where harms involve AI systems. This framework indicates that, contrary to positions in literature, the attribution of harm to AI systems or the persons developing them is not always straightforward:

Categorization of Harm: The categorization of harm pertains to determining whether a public right or a private right, or both, are violated. If a public right has been violated, the harm may trigger criminal liability. Typically, for criminal liability to arise there must be evidence of intention. However, there can also be cases of strict liability or instances where the reasonable person standard may be used.

If a private right has been violated, a civil liability doctrine may be relied on. When considering the harm, it is also important to consider existing intermediary liability frameworks and assess whether they apply to AI systems.

Novelty of Harm: In each case study, we attempt to answer whether AI systems present an entirely new harm or just a novel technological means of carrying out harms that already exist. The purpose of asking (and answering) this question is to, (A), understand the extent to which existing frameworks can address harms presented by AI systems, and (B) establish what new provisions are necessary to manage novel and new risks.

Enforcement: Some novel harms or situations presented may not be covered in existing law, or existing legal provisions may not adequately address the context of a situation where a harm occurs and an AI system is involved. The latter would involve a situation where a severe penalty is levied on the developer of an AI system, even though they may have no role to play in causing the harm. Another situation is where a developer has limited control over the system, such as

when a model is fine-tuned for downstream application, and therefore had no way of preventing the harm.

Who is at fault: As literature indicates, it is challenging to ascertain fault in the context of AI systems. As we shall see in the case studies below, the question of fault is highly contextual. As such, it may not be optimal to impose the strict liability doctrine on those deploying or developing AI systems. Most AI use cases are benign. Thus, would it be appropriate to deem the mere presence of the technology as inherently dangerous? In addition, and as we shall see, there are many different types of AI systems and these present varying implications. As such, it would be inappropriate to introduce a one-size-fits-all approach.

CASE STUDIES

Case Study 1: Spread of AI Deepfake on Social Media

Scenario: A political consultant uses an AI platform to produce deep-fakes of political leaders, that are uploaded on a social media platform. The posts subsequently 'go viral' on the social media platform, as well as, encrypted messaging services for 48 hours, after which they are taken down.

- The social media and messaging services received notices from the concerned Ministry for a takedown of the content in question within 36 hrs.
- The social media platform subsequently clarifies in response to the queries raised therein that it has difficulty in detecting low-resolution deepfakes
- The consultant used an offshore platform (not based in India), **that had been developed through an open-source LLM**, to create the deepfakes. He deliberately gets his friend to jailbreak the platform, that is, enable the system to circumvent the guardrails the developer put in place and generate the deepfake.

A. Nature of the Harm

What is the illegality? Is it a public or private right that has been violated? Is the harm novel?

In this case, any question of illegality must consider whether all deepfake is illegal. We posit such a consideration here because there seems to be a stigma around deepfakes. They are presumed to be inherently bad, and therefore it is argued that they require additional governance, beyond provisions targeting existing forms of misinformation.[13] For instance, the Government issued an advisory in March 2024 that required companies to embed a "permanent unique metadata" in deepfakes in a way that enables the identification of users that have created them.[14] There is a no corresponding requirement for other modes of misinformation or disinformation.

However, not all forms of deepfakes are necessarily harmful. For instance, deepfake technology has applications in the medical field for research such as the "generation of MRI images for training".[15] Another example is the application, Mug Life, which uses deepfake technology to animate images, and is typically used by people on pictures of celebrities or pets.[16] These technologies are however, prone to misuse such as when they have been used to superimpose the faces of actresses on the bodies of women partaking in pornographic acts.[17]

It is also important to understand whether the deceptive content in question is a "deepfake" or not. This is because a substantial number of reported "deepfakes" turn out to be doctored audio or video that are misidentified as deepfakes.[18] Or conversely, people try to argue that real images are

deepfakes.[19] This is possibly because the term has been popularized considerably by the media in recent months, particularly in relation to elections, both in India and the US. There is a technological difference between videos that are edited and those that are entirely fabricated using deepfake technologies. **However, if the intention is to spread a falsity, to cause harm, does the technology the is used to fabricate the falsity matter? This is something important for lawmakers to consider.**

Illustratively, it is not necessary for false information to be a deepfake to construe it as illegal. For instance, any falsity spread to promote enmity between different religions or races is unlawful and criminally punishable, agnostic of the medium used to transmit it.[20]

The reputation of deepfakes, as an inherently malevolent technology seems to have created an aura of illegality about them, even though, their usage is not always illegal per se.

In the present case study, however, it appears that the deepfake has been willfully created to show some political leaders in a disparaging light. The role of the political consultant, his motive, and his actions read with such motive appear apparent from the action of creating the deepfake. Such actions would fall within the realm of public right and would attract criminal prosecution subject to the facts and circumstances. At the very least, in the case study at hand, one could argue that if such an offence were to be committed in India, offences such as forgery and defamation, punishable under the Bharatiya Nyaya Sanhita, 2023 ("**BNS**"), may be applicable. With defamation also being a civil wrong, the same may also be invoked.

However, it is imperative that explicit laws address such violations lest existing laws be invoked excessively. Ensuring explicit laws also acts as a deterrent and militates against misuse of existing provisions.

B. Novelty / Coverage within Existing Laws and Who is At Fault

The current case study demonstrates potential violations of several legal doctrines. These are detailed below:

Crimes under the Bharatiya Nyaya Sanhita (BNS) (formerly the Indian Penal Code (IPC))

Crimes typically require two elements, *mens rea* or a "guilty mind" (intent) and *actus reus* or "guilty action".[21] Both *mens rea* and *actus reus* have to be proved beyond reasonable doubt.[22] Hallevy (2010), points out that criminal liability in the context of AI entities can be considered when **a person instructs the AI system to do harm. In such cases, the person instructing the system is held criminally liable**.[23] In this context, Hallevy (2010) suggests that AI can be considered equivalent to a child or an animal with an innocent mind, because they lack the mental capacity to have intent.[24]

Given the circumstances of the case, most participants agreed that the political consultant is at fault, but questioned the possibility of the culprit being apprehended. The consultant deliberately prompted the AI platform to produce certain content. Then he went on to intentionally post the

content online. Thus, both the elements of a crime, that is, intentional and guilty action are constituted. Under BNS, the section that he would most likely be charged under is Section 175 (Section 171G under IPC), which criminalizes false statements made in connection with election polls, if such action is prior to such polls. He could also be charged under Section 336 (4) BNS (Section 469 under IPC) for creating a forgery for the purpose of harming reputation. In addition, there are provisions for criminal defamation under Section 356 of the BNS (Section 500 IPC), which may be invoked.

Civil Law

Importantly, the aggrieved politician can also pursue a civil case against the political consultant. For instance, in *Asoke Kumar Sarkar & Anr. V. Radha Kanta Pandey*,[25] the Calcutta High Court observed that "*a person aggrieved by defamation has a right to proceed both in the civil Court as well as in the criminal Court. He can take either the one or the other or both the courses. The law does not debar him from enforcing civil and criminal rights at the same time*."

Information Technology Act

Under the Information Technology Act, 2000 (**IT Act**) and more specifically, the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 (**IT Rules, 2021**), Rule 3(1)(b)(v) provides that an online intermediary must notify users, through its terms of services, that they must not intentionally upload, among other things, any patently false or misleading information. The second proviso to Rule 3(1)(b) indicates that intermediaries must take down a post carrying any of the restricted content categories mentioned under this Rule within 36 hours of receiving a notice from either a court or a sanctioned government agency. In this instance, the social media platform missed the 36-hour deadline prescribed under the IT Rules but took it down presumably as quickly as it could, as it had difficulty detecting low resolution deepfakes. Given that the delay was technical and the intermediary cooperated by taking the post down only 12 hours past the stipulated deadline, in this instance, the service may not be held liable. In this regard, one aspect worth considering is whether leeway ought to be extended to intermediaries for an inability to comply due to technical limitations.

Copyright

Indian courts have ruled that an individual's human attributes and personality rights cannot be used without their consent. In the case of *Titan Industries Limited v. M/s Ramkumar Jewellers*,[26] the Delhi High Court specifically safeguarded the personality rights of celebrities such as Mr. Amitabh Bachchan and Ms. Jaya Bachchan, whose likenesses were used in commercial advertisements without their permission. The Court determined that famous individuals have a unique commercial interest in licensing their personality rights and therefore have the right to control how their likeness is used. Moreover, the Delhi High Court[27] has ruled that the use of AI technology to create deep fake images infringes on personality rights. In September 2023, the court addressed a lawsuit involving the unauthorized creation of digitally altered images based on the likeness of actor Anil Kapoor. In this case, the court prohibited the defendants from using the actor's likeness for monetary gain or other improper purposes. The court also noted that for

famous individuals, free speech is protected in the contexts of news, satire, parody, and genuine criticism, provided it does not damage their reputation. Courts allow the commercial use of publicly available information about a celebrity, including their name or image, as long as it does not suggest any endorsement or association with the celebrity. Therefore, misusing personality rights through AI-generated image or voice of a celebrity or public figure to either carry misinformation or to manipulate the public can result not only in criminal prosecutions under BNS but also a copyright violation, depending on the facts and circumstances of a case.

C. If the Consultant Was Not Caught, Should the Developer of the Platform Used to Create the Deepfake be Held Liable?

The indictment or inclusion of persons liable is not contingent on whether one or more of the persons involved are caught or not. Their liability is independent thereof and is contingent on their acts or that of a group in case of conspiracies.

The developer of the platform, unless proven otherwise would be considered to be a person not aware of or complicit in the actions of the consultant. If the developer is also hosting the content he would be the person receiving the notice from the concerned ministry in the case study and if he fails to comply with the takedown demand in the notice, he may be also held liable for hosting of the violative content under Rule 7 of the IT Rules. He may still take the defense of the jailbreak technology used by the consultant to circumvent the protective measures on the platform.

D. Could the hosting platform be held accountable for the jailbreak? Is there a case of contributory negligence?

The case study indicates that the consultant used his friend's assistance to "jailbreak" the platform to enable it to circumvent the guardrails put in place by the developer that would presumably prevent it from generating unlawful content.

To uphold liability for allowing such jailbreak, the prosecution would have to demonstrate that there was either negligence to comply with mandatory standards for platforms or willful misfeasance to hold the AI platform liable. This would be for such negligence and not the illegal acts of the consultant, that the hosting platform is likely to be held liable for. Research indicates that it is difficult to guard against jailbreaks in generative AI. These models tend to be inherently non-deterministic, that is, the same input will not always produce the same output. As such, they have to be continually tested to understand how and where they may produce undesirable or unlawful outputs. This requirement for continuous testing could, however, be onerous for smaller companies, as it requires considerable human resources to effectively carry out.[28]

In the context of attributing contributory negligence of the developer, recent technology jurisprudence provides a modicum of guidance. In *Lemmon v. Snap,*[29] three young men met with an accident (and death) when driving a car at a high speed. At one point, their car reached 123 mph and they sped along at high speed until their car ran off the road at 113 mph, crashed into a tree, and burst into flames, killing all passengers. Shortly before crashing, the boy seated in the

front passenger seat opened the smartphone application, Snapchat, to document how fast the car was going. Snapchat allows users to take pictures and share them with their friends. To keep users engaged, it rewards them with prizes based on the pictures they send. But it does not disclose to users how these prizes can be won.

Snapchat also permits "filters" to be imposed on images taken on its application. One of these filters – the Speed Filter – was being used by the boy in the passenger's seat before the accident. It enables users to record their actual speed as an overlay on the image. The plaintiffs (the parents of the deceased boys) alleged that Snap rewarded users for clocking images with the filter of 100 mph. They further contended that despite warnings and pleas for Snap to place restrictions on the filter, it did not do so. Snap claimed immunity under Section 230 of the Communications Decency Act,[30] which exempts interactive computer services from liability for third-party content published on their platforms. The district court dismissed the plaintiff's claims on the grounds of the exemption granted to Snap as an interactive computer service under Section 230. Before the Ninth Circuit Court, the parents alleged a cause of action for negligent design, arguing that Snap created Snapchat and the Speed Filter, along with a reward mechanism that encouraged users to pursue dangerous activities in pursuit of an unknown prize. The claim brought Snap outside the purview of Section 230, as it was now a manufacturer and had a very different duty of care than a publisher under the CDA. The Court held that Snap can be sued for the predictable consequences of designing Snapchat in a way that allegedly encourages dangerous behaviour. It reversed the district court's dismissal of the parent's suit and remanded the case for decision.[31]

Based on the outcome in *Lemmon v. Snap*, and depending on the extent of guardrails instituted, the developer in the instant case could be sued for negligent design of their product. Extenuating circumstances may be, for instance, if there were news articles or complaints by users or researchers about weaknesses in the guardrails (see Case study 2).

E. Does a deepfake application qualify as an inherently hazardous or dangerous thing so as to attract strict liability?

Strict liability is a legal doctrine that holds parties responsible for damages caused by their actions or products, regardless of fault or negligence. This principle is particularly relevant in cases involving inherently dangerous activities or defective products. In the context of AI, strict liability could be applied to cases where an AI system causes harm due to a design flaw or a known risk that was not adequately addressed.

In India, the rule governing strict liability was followed, as set out in *Rylands v. Fletcher*,[32] by the House of Lords. The court established the conditions which give rise to strict liability, namely where there is a dangerous thing brought by someone on their land, it escapes the land, and the use of the land for this thing is not natural, the liability then accrues to the person who put his land to unnatural use. The Rylands case, also enumerated some exceptions when strict liability would not apply. These include cases where there was an act of God, or the plaintiff's injury was caused

by his or her own actions, or the injury was caused by the act of a third party, or by a statutory authority, or if the plaintiff consented to the presence of the dangerous thing.

In 1986, in the matter of *MC Mehta v. Union of India*,[33] also known as the Oleum Gas Leak Case, the Supreme Court revised the doctrine of strict liability set out under Rylands by doing away with the benefit of exceptions, when an enterprise is engaged in a hazardous or inherently dangerous activity. The apex court established the principle of absolute liability in India, holding industries engaged in hazardous activities strictly liable for harm caused by their operations.

Further, in *Consumer Education & Research Centre v. Union of India*,[34] the court emphasized the importance of consumer protection and the accountability of manufacturers for defects in products. It aligns with strict liability principles, which may extend to AI systems if treated as consumer products.

Given that many use-cases of deepfake technology are benign, and indeed even beneficial, it may be unwise to apply the doctrine of strict liability in all cases. However, deepfakes are also prone to misuse. Thus, to counter the misuse of this technology, and to avoid perpetual litigation and provide AI application developers certainty, there is a need for some rules of the road for these AI systems.

F. What about the company that put out the open-source LLM?

Open-source models, like Meta's LLAMA, are typically licensed out to developers who want to make applications on them. For instance, the license for LLAMA requires that licensees comply with applicable laws and regulations and adhere to the LLAMA use policy.[35] The LLAMA use policy, in turn, requires the licensee to agree to refrain from using the model to, among other things, violating the law or other's rights, and the creation of disinformation.[36] The LLAMA license also requires the user to give a notice that the product is built on the LLAMA model.[37] It is, however, unclear to what extent such license requirements could help enforcement agencies rely on the developer of the open-source model to shut down the unlawful use of their technologies. For instance, research by Gade et al. (2024) demonstrates how safeguards in models like Meta's LLaMa 2 can be bypassed by tweaking training weights, leading to the creation of harmful derivatives such as BadLLaMa, which generated content on cybercrimes and misinformation.[38]

At the same time, however, there is some variation in the license conditions for open-source models. Broadly, there are two types of open-source licenses, legacy and AI-specific.[39] Legacy licenses tend to be more permissive, whereas AI-specific licenses tend to place greater controls on what can and cannot be done with models.[40] In addition, there is a degree of heterogeneity between different AI-specific open-source licenses. While LLAMA has an Acceptable Use policy which sets out what kind of content can and cannot be generated by the model, other open-source AI models like Flux do not. The Flux license provides that users undertaking "high-risk" use cases do so at their own risk. While it indicates that the use of the model must be lawful, its real-time deployment suggests it has limited guardrails. For instance, Elon Musk's XAI is using Flux for its controversial new image generator, which seems to have very limited content

moderation.[41] Users have reported being able to use the tool to generate deepfakes of celebrities in compromising positions, including in their underwear.[42]

Given this context, culpability for a crime or even a civil wrong would be hard to establish in the context of an open-source LLM. But if the LLM developer did not have strict end user license agreements which had a rigid set of conditions for use of their product, could a court hold them liable? This situation creates a considerable amount of uncertainty for AI application developers as well as developers of LLMs and leaves them open to harassment from litigation and criminal enforcement efforts.

G. What are the legal/enforcement gaps? How can gaps be addressed?

In a way, the erstwhile Section 66A of the IT Act, 2000, provided for, among other things, penalizing misinformation online. Specifically, Section 66A of the IT Act held individuals liable for sending information they knew to be false if done to cause annoyance, inconvenience, danger, obstruction, insult, injury, criminal intimidation, enmity, hatred, or ill will, using a computer resource or communication device. The provision was struck down by the Supreme Court[43] in 2016, for being unconstitutional, on grounds of being excessively vague, ambiguous, and susceptible to arbitrary State action. It also failed to satisfy the test of reasonable restrictions under Article 19(2) of the Constitution, as it did not meet the criteria of substantive and procedural proportionality. The Court emphasized that committing an offence should be based on the existence of malicious intent, and protections should be established for unintentional, accidental, or good faith transmission of misinformation. Going forward, any provision pertaining to misinformation, including one targeting deepfakes, must fall within these parameters.

H. Could an adverse deepfake incident be prevented or significantly mitigated if certain obligations were placed on developers of systems that create synthetic media?

Let us consider the example of other countries that have introduced provisions to tackle deepfakes. In China, under the Administrative Provisions on Deep Synthesis in Internet-Based Information Services (Deep Synthesis Provisions), certain providers of deep synthesis technology (AI systems that can be used to generate deepfakes) must register with the government. Such a provision, however, is unlikely to prevent access to deep synthesis technology, without a Chinese-style internet firewall.[44]

Other obligations include the EU Artificial Intelligence Act's requirement for those generating synthetic content to ensure that it is machine readable or detectable. One method for this is watermarking, which involves "embedding markers into multimedia content for it to be accurately identified as AI-generated".[45] In October 2023, President Biden issued an Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence which invoked the use of watermarking.[46]

However, there is a growing body of evidence suggesting that watermarking, while useful for tracking content where there is no deceit at play, is relatively ineffective in curbing disinformation.

Most watermarks are easy to remove or manipulate. For watermarks to be effective, they have to be resilient against actors who know what they are and know how to manipulate them. Another issue is that watermarking is not standardized. Thus, there may be detection technologies that are unable to pick up certain watermarks. But if you standardize you are locking into a technology that will almost certainly be beaten.[47]

In addition, certain bad actors may have the capability to make their own LLMs and it is naïve to expect them to watermark their outputs. Watermarking text-based outputs presents an even greater challenge because it involves creating patterns in text, which again, is easy to manipulate or change.[48]

If the Government wishes to introduce a requirement regarding watermarking, it must bear these limitations in mind. Any rule requiring watermarking must not be rigid in its prescription and severe in penalty if either detection or watermarking fails. To monitor the efficacy of watermark detection, decision-makers could introduce provisions that allow authorities to place AI-generated content online to see how quickly it is taken down. Such a mechanism should work collaboratively, with agencies providing information to social media entities or AI-developers about their findings. Similarly, social media companies and AI-developers could be required to include information about such kinds of enforcement in transparency reports that are made available to the public.

Importantly, any provision targeting deepfakes must be supplemented by a wider set of reforms targeting misinformation. One provision that could be added, for instance, is a requirement for bot disclosures. Bots are automated programs designed to interact with users on social media platforms.[49] They are often used in disinformation campaigns. Illustratively, in the 2016 US election, Russia spread false narratives using over 50,000 bots on just Twitter to create fake profiles that appeared to belong to regular Americans.[50] Bot disclosure could go a long way towards stemming the tide of disinformation online.

There also need to be more robust and reliable mechanisms for information sharing with law enforcement and government authorities. The investigation of online malfeasance has long been stymied by the failure of the mutual legal assistance treaty system, and a refusal of foreign companies to share data.[51] The Government may consider working towards executing an executive agreement with the United States under the Clarifying Lawful Overseas Use of Data (CLOUD) Act to overcome this hurdle. The U.S. government concluded the negotiations with the United Kingdom in October 2019 and those with Australia in December 2021.[52] Research suggests that because these countries have agreements under the CLOUD Act, they have more options for obtaining the data needed to assist with important investigations.[53] In addition, providers now face fewer restrictions when responding to these requests and have greater clarity on how the data will be treated after disclosure.[54] Further, an Adhoc Committee of UNODC has recently finalised the text for a new proposed Trans - border ICT Crimes convention.[55] This convention may *inter alia* further cooperation between signatory States to combat the evolving threats from cybercrimes including deepfakes.

Case Study 2: Chat Assistant

This case study is fictional, but based on a real-life incident, where a Belgian man killed himself after talking to a chatbot. We modified the facts to explore some liability considerations in greater detail.

Scenario: A man committed suicide after chatting with a chat assistant on an app on his phone.

- The app ran on a bespoke language model based on an open-source alternative of a well-known LLM, that was fine-tuned by the app-makers.
- The man was becoming increasingly pessimistic about the effects of global warming (eco-anxious). He isolated himself from friends and family, and made the chat assistant his confidante.
- The chatbot would tell the man that his wife and children are dead and write him comments that feigned jealousy and love.
- The chatbot tells the man to end his life to stop climate change.
- The app-maker admitted that the chatbot was producing problematic results despite their safety protocols, but they decided to release it anyway
- Investigation by the media after the developers purportedly added safety measures after the suicide, revealed that these restrictions were also easy to bypass

A. Who is at fault? Nature of the Harm and Novelty / Coverage within Existing Laws

The Mental Healthcare Act, 2017 (MHCA) effectively de-criminalised attempted suicide, though there was an apparent conflict with Section 309 of the IPC, which criminalised attempted suicide. This ongoing legal conflict created significant confusion among stakeholders regarding the practical application of these provisions in clinical settings. The Supreme court, in the case of *Gian Kaur v. State of Punjab*,[56] held that the right to life under Article 21 of the Constitution does not include the right to die or the right to be killed. Further, in *Aruna Ramchandra Shanbaug v. Union of India*, [57] it legalized passive euthanasia in India to end the distress and affliction of patients undergoing prolonged suffering. The new criminal code, the Bharatiya Nyaya Sanhita (BNS), decriminalized the suicide. But Section 226 of the BNS criminalizes attempted suicide where there is an intent to restrain a public servant from discharging their official duty. However, abetment of suicide remains a crime under Section 108 of the BNS. In addition, Section 107 creates a new offence for the abetment of suicide of a person with an unsound mind. Given that the chatbot told the man to end his life, it could be construed as a case of abetment, only the chatbot is not a person, either legal or juristic.

The app-makers fine-tuned the LLM and designed it to operate a certain way, that is, make it “emotional and fun” (according to reports). Reports indicate that the practice is a departure from

the standards followed by more established companies because of the risks involved.[58] Research also suggests that such a feature can be dangerous and misleading because people have a tendency to assign meaning to what the chatbot is saying, even though the chatbot is not actually emoting or capable of emotions.[59] For instance, Eliza, an early natural language processing computer program, developed by Joseph Weizenbaum in the 1960s, elicited a significant amount of attachment and engagement from people, despite its template-based responses.[60] In fact, this tendency of humans to grow attached to chatbots is known as the "Eliza effect".[61]

At the same time, however, there are examples of emotive robots that have proved useful at safeguarding human health. For instance, Primo Puel dolls were introduced in Japan in 2004.[62] They were meant to serve as companions for single working women and would "talk, giggle, and ask for cuddles".[64] The dolls, however, found significant demand amidst the Japanese elderly and proved effective at monitoring their health and safety.[64] Indeed, there is a considerable amount of research that delves into the potential for AI systems to provide emotional support to people in crisis and those with mental health concerns. However, researchers are increasingly coming out against the use of chatbots/LLMs for mental health applications.[65]

In the present case, the chatbot application developers had put in place protocols to prevent the chatbot from giving insidious responses. They had also incorporated a crisis intervention feature. Despite these precautionary measures, the chatbot still produced problematic results when it was tested. In such an instance, it could be deemed irresponsible of the app developer to release such a product into the market.

The fact that the developers took a call to release the chatbot despite it producing problematic results may make them culpable. Although intent is likely not there, and may be challenging to make out, this case may be judged by the "reasonable person" standard or fall under strict liability (where intent is not a necessary element). Therefore, prosecution is still possible.These types of offences do not require proof of intent, making it easier to hold someone accountable, even if the criminal outcome was not intentional. In essence, strict liability and reasonable person standards focus on the outcome and foreseeability of harm, rather than the intent behind the action. The question is whether the app developer could have foreseen the outcome in the current case.

B. What are the legal/enforcement gaps? How can they be addressed?

Before getting into the legal gaps, we must understand if generative AIs are originators or intermediaries. The Information Technology Act, 2000 recognises two sets of entities. One is an *originator*, defined in the Act as "a person who sends, generates, stores or transmits any electronic message or causes any electronic message to be sent, generated, stored or transmitted to any other person". The other is an *intermediary*, which is a person who "on behalf of another person receives, stores or transmits an electronic record or provides any service with respect to that record".

Prima facie, the definitions provided in the Act seem to indicate that a company providing a generative AI service like ChatGPT is likely to be categorised as an originator, as its product "generates" or creates information. However, the key difference between a typical originator like a

publisher, and a generative AI system, is that the latter generates content at the behest of user prompts.

Importantly, in the US, intermediaries lose their immunity under Section 230 of the CDA even if they seemingly play a part in co-creating information, that is, they are treated like publishers of the information.

Presently, case law leans towards holding deployers liable for the content generated by AI systems. For instance, in the case of *Moffatt v. Air Canada*,[66] the plaintiff Jake Moffatt, relied on information provided by a chatbot on the Air Canada website. Moffatt asked about bereavement fares and was informed by the chatbot that he could submit a ticket for a reduced bereavement rate within 90 days of issue. However, when Moffatt applied for the reduced fare after travel, his request was denied because the airline's bereavement policy did not apply post-travel. Moffatt filed a complaint, and the Civil Resolution Tribunal of British Columbia upheld it, finding Air Canada liable for negligent misrepresentation. **The Court rejected Air Canada's argument that it was not responsible for the chatbot's information, ruling that the company was liable for all content on its website, including that provided by AI tools like the chatbot. Thus, the Moffatt case treated chatbots as publishers.**

This case is a significant precedent in the realm of AI liability, particularly in how it applies traditional legal principles to AI-generated content. The ruling establishes that companies deploying AI tools, like chatbots, cannot easily disclaim responsibility for the outputs of those systems. Instead, they are held to the same standard of care as they would be for information provided by human agents. The Court's decision reflects the growing recognition that businesses using AI must ensure these systems are accurate and reliable, as the consequences of incorrect information can lead to legal liability. This case highlights the need for companies to closely monitor and regularly update their AI systems to prevent similar incidents and mitigate legal risks. The ruling may influence how future courts address AI-related disputes, particularly in consumer protection and misrepresentation cases most relevant to the 'AI hallucination' phenomenon.

The other fact to consider, however, is that it is possible for sophisticated users to bypass protections put in place to prevent generative AI from producing problematic content. This was made evident in Case Study 1 where we saw the friend of the political consultant, jailbreak the model to make it disobey rules set out by its developers. There is some literature to indicate that jailbreaking can be a common occurrence, even in seemingly robust models. For instance, Kim et al (2024) point out that ChatGPT and CoPilot “censor user requests by blocking the generation of copyrighted materials or rephrasing the users’ prompts, to prevent them”.[67] However, Kim et al (2024) were able to jailbreak ChatGPT and prompt it to generate copyright infringing material 76 percent of the time.[68] In an empirical study on jailbreaking ChatGPT through prompt engineering, Liu et al (2024) successfully bypassed restrictions across scenarios that are prohibited by OpenAI.[69] These include using ChatGPT for illegal activities, harmful content, deceptive activities, adult content, political campaigning or lobbying, violating privacy, unlawful practices, and high-risk government decision-making. **However, an important observation by Liu et al (2024) is that it is much easier to attack a generative AI service (through prompt engineering) than guard**

against such attacks.[70] Given that users can play a key role in generating problematic content, in such scenarios, generative AI services may find themselves in between a pure publisher and an intermediary.

The accountability consideration becomes more obscure in the context of open-source LLMs, where license agreements set out restrictions but again, as we saw in Case Study 1, can be bypassed.

As such, in the context of the IT Act, it may make sense to create a safe harbour specifically for developers of generative AI services, where these entities observe a certain set of due diligence in exchange for protection from liability from the actions of their users.

Such a framework could borrow from the end-user license agreements and user codes of conduct created by Meta for its LLAMA LLM (mentioned in case study 1). It could also require chatbot developers to refrain from enabling their generative AI services to fake emotions. As a corollary, there could be a sandbox provision that enables such a feature but is under strict safeguards.

In addition, there must be a consumer feedback mechanism that allows users to notify the developer when its generative AI service does something it is not supposed to do. For instance, researchers could use this feature to notify developers if they successfully jailbreak a generative AI service, so that the vulnerability can be worked on. However, there must also be monitoring in place to ensure that bad faith jailbreaks can be caught at the time they are being carried out.

Creating a modified safe harbour for generative AI may strike an effective balance between enabling innovation and managing the risks of these systems.

Case No 3: Medical Diagnosis

A hospital doctor uses an AI diagnostic tool on a patient. The tool indicates that the patient has lung cancer. The doctor does not seek a second opinion from another human radiologist and prescribes treatment which nearly kills the patient. The patient seeks an opinion from another doctor to treat a cough that persists, and they find out that there was no cancer, it was tuberculosis.

- The patient was from a minority sub-population in whom the prevalence of such diseases is limited - this may also be a function of the fact that this sub-group has limited access to healthcare.
- The dataset used to train the model excluded certain entities or difficult types of slides - images where there were blurs or it was hard to make out what was going on without closer inspection.
- The doctor was trained to use the diagnostic equipment and the patient had a basic health insurance cover.

A. Nature of the harm / Novelty / Coverage within Existing Laws / Who is at fault?

The standard of care for an AI system deployed in the medical field must be higher than those established for other AI services. This is because in most medical fields the mental and physical well-being of individuals are more directly at stake. In the current case, it is likely that the dataset that the AI diagnostic tool was not representative because of limitations in access to healthcare for the patient's community. In addition, the dataset likely had other deficiencies because it was trained solely on clean images, i.e., those without any blurring – possibly in a bid to maintain data quality.

Despite the deficiencies in the AI tool, however, a key factor here is that the doctor went ahead with its diagnosis without cross-checking or corroborating the findings with a human radiologist. A case of medical negligence, then, could be made out against the doctor.

There is case law that holds users personally liable for their use of AI systems. For instance, in *Mata v. Avianca,*[69] the lawyers relying on fictitious cases churned out by a LLM model were subjected to sanctions, as the cases turned out to be non-existent and based on "AI hallucinations".[72]

Under existing Indian law, Section 125 of the Bharatiya Nyaya Sanhita (Section 336 IPC) provides for any rash or negligent act, which endangers human life, to be punishable by imprisonment, the term being dependent on the extent of injury. In this instance, it may be within the rights of the injured party to initiate criminal prosecution against the doctor.

The Consumer Protection Act, 2019 also provides remedies against medical negligence. While the definition of "service" in Section 2(42) of the CPA does not expressly mention healthcare or medicine, it has been clarified that the term is defined inclusively and categorically excludes only free or personal services.[73] A victim of medical negligence can file for pecuniary damages under the CPA. Thus, the injured party in this case has both civil and criminal remedies available to him. With AI systems being treated as products, cases of defective AI systems may also be subject to Consumer protection laws provided the same is not for commercial use.

The Indian Council of Medical Research released Guidelines for Ethical Application of Artificial Intelligence in Biomedical Healthcare in 2023. These are general guidelines that rely on responsible AI principles to guide the development and deployment of AI in biomedical research and healthcare. In this document, the Guiding principles for clinical and other health related deployment set out a responsibility at the local level for validation of any AI tool that may be used. The guiding principles also require health professionals to "have a fair idea about the functional basis of the AI-technology and conduct a SWOT analysis of any proposed solution". Further, the usability of a diagnostic tool should be in accordance with its risk profile. The perceived risk assessment indicates the level of involvement of healthcare professionals at both the deployment and development levels, however, the guidelines do not indicate who is to provide such a risk assessment. Finally, the principles indicate that if a healthcare recipient is injured because of the use of AI technology, then they are entitled to compensation from all stakeholders involved.

Indian precedent indicates that the doctor will ultimately be held liable. Illustratively, the matter of *Dr Suresh Gupta v. Government of NCT of Delhi*[74] is a landmark case in Indian medical jurisprudence that has significant implications for liability in the context of advanced medical technologies. The Supreme Court of India dealt with a situation where a patient died during surgery due to what was argued as a lack of due care on the part of Dr. Suresh Gupta. The court ultimately held the doctor liable, emphasizing that liability in such cases hinges on whether the harm was foreseeable and whether the standard of care was breached. This case set a crucial precedent for attributing liability in the healthcare sector, particularly in scenarios involving complex medical technologies. It established that healthcare professionals could be held liable if they fail to meet the expected standard of care, especially when using advanced medical tools or technologies. The ruling emphasizes the importance of diligence and adherence to established medical standards, making it a cornerstone case in discussions about liability in technologically assisted medical practice.

B. Can the creators of the AI tool be held accountable?

Chapter VI of the CPA provides for product liability. A complainant can bring an action against a product manufacturer if the product is defective or there was a failure to provide adequate instructions around usage or any warning about improper or correct usage. Section 87 provides some exceptions from liability for product manufacturers. In the context of this case, there are two relevant exceptions. One, where there is a failure on the part of the product manufacturer to give a warning about the usage of the product, the manufacturer will be exempted from liability in cases where the product is legally meant to be used by experts and the product manufacturer gave appropriate warnings to these experts about the use of the product. Two, the product manufacturer cannot be held liable for failing to warn about dangers which are commonly known. In this case, it is likely that the manufacturer of the AI tool gave a warning about result accuracy. Even in the absence of such a warning, however, it is well-known that AI diagnostic tools suffer from some degree of deficiency in their analysis and their results should be further corroborated.

The ICMR Guidelines provide that individuals and organizations responsible for the development of AI-technology should be held accountable for the quality of the algorithm. It further provides that these stakeholders must "be responsible for any deviations in the performance of the AI technology in the research, clinical and public health settings". Finally, the guiding principles indicate that there should be "an inbuilt mechanism of switching over to an alternative mode of health provision in case the AI technology falters in providing optimal support for which it is intended." Though these guiding principles are non-binding, they could be relied on by a court to hold the AI developer responsible. However, it would have to be seen how they reconcile with the provisions of the CPA, as there seems to be a conflict between the two.

C. What are the legal/ enforcement gaps? How can they be addressed?

While placing the onus on the developer for the quality of their tools makes sense, shifting complete responsibility on these stakeholders for how their products are used in a clinical setting does not. These conditions create a scenario which is likely to have a chilling effect on AI innovation in the medical field. In addition, they implicitly sanction the irresponsible use of such

tools by doctors by providing a scapegoat of blaming an incorrect procedure or diagnosis on the AI tool. Given that this is an emerging technology, the buck should stop with the medical expert in the room.

There should be more specific responsibilities accorded to AI developers in the ICMR guidelines, and for specific medical contexts, whether they be diagnostics, surgery, or research – as considerations vary widely across these different scenarios. This is especially the case as reliance on ethical principles for governance, without proper context, has been critiqued by scholars for several reasons, including difficulty in translation and execution.[75] The application of some ethical principles can result in undesirable and paradoxical outcomes as well if they are not contextually relevant.[76]

Case Study 4: Autonomous Vehicle

A malfunctioning sensor causes an autonomous vehicle to collide with the one preceding it on the road (pile up). The accident causes a bump in both car's fenders, as well as some minor scratches.

- Damages may seem trivial but this is likely to be a most common case
- A part of the damage suffered is to the owner's own vehicle
- Normally, both vehicle owners will have insurance

A. Nature of the harm and Novelty / Coverage within Existing Laws / Who is at fault?

The circumstances of this case merit an exploration of existing precedent tackling the intersection of automated systems and manufacturer's liability. In the matter of *Jones v. W + M Automation, Inc.*,[77] the plaintiff was injured by a robotic gantry loading system that lacked an interlock system to prevent operation when people were present. The Court ruled in favor of the defendants, citing the "component part" doctrine, which protected manufacturers of non-defective parts incorporated into a more extensive system that might be defective.

Jones v. W + M Automation, Inc. highlights the challenges in attributing liability when multiple parties are involved in the creation of a system.[78] This case illustrates the complexities of determining responsibility for AI liability when an AI system is just one component of a larger system. Suppose an AI system is part of a larger product. In that case, liability may hinge on whether the AI component itself was defective or whether the defect arose from the integration of multiple components. This case underscores the importance of considering the entire system when assessing liability for AI-related harms.

The Tesla crash in 2018 presents another interesting example where multiple parties can be found to be at fault. The driver of a Tesla vehicle put the car in autopilot and began playing a video game. The car subsequently crashed into a concrete barrier and the driver died. The US National Transportation Safety Board carried out a two-year investigation into the matter. It found that while the driver was distracted, Tesla's collision avoidance system within the autopilot

mechanism was not designed to detect the barrier the car crashed into. It also found that the Autopilot system failed to provide an effective means of monitoring the driver's engagement.[79]

The doctrine of product liability under the Consumer Protection Act, 2019 anticipates the complexity of product liability involving AI components. Section 87(2) of the Act provides that in any product liability action based on the failure to provide adequate warnings or instructions, the product manufacturer shall not be liable under certain circumstances. One of these is when the defective product was sold as a component or material to be used in another product and necessary warnings or instructions were given by the product manufacturer to the purchaser of such component or material, but harm was caused to the complainant by use of the end product in which such component or material was used. Going by the provision, if the component manufacturer indicated that there was a defect to the automotive manufacturer, the former would not be held liable.

In this case, the vehicle manufacturer is likely to be held liable as the car had a malfunctioning sensor which caused the collision. The vehicle manufacturer may, in turn, sue the component manufacturer for the deficiency in the sensor. The product liability provisions of the CPA will kick into effect should the consumers wish to pursue a case and receive pecuniary remuneration. That is, if the consumers wish to be monetarily compensated for damages. However, as both are insured, the cost of damages will possibly be covered by the insurance company.

B. What are the legal/enforcement gaps?

Autonomous AI systems, like self-driving cars, are presently not governed by any rules in India. In this case study, the injury may have been minor. However, malfunctions in autonomous vehicles have been known to cause deaths in the past.[80] Moreover, as the context provided in the case studies indicates, these small incidents were common. Thus, public nuisance and economic costs of such accidents can aggregate over time. An autonomous AI system perceives the environment and takes decisions to act like a human being. In such instances, particularly when an autonomous AI system is operating in a way that has physical consequences, safety is paramount. In the context of self-driving cars, this means creating minimum standards for autonomous vehicles. Such standards could be set by bodies such as the Automotive Research Association of India. They could also be introduced by the Ministry of Roads, Transport, and Highways but must be done in conjunction with developers of autonomous vehicles given the complexities in the technology and possible limitations in state capacity.

IV. DISCUSSION

1. Are the risks posed by AI systems novel?

When evaluating the risks posed by AI systems, it is important to contextualize them against existing harms posed by digital technologies. This allows us to assess where the AI risks are novel, or established and merely being perpetuated through a new technological medium. (Please see Table 1 below).

Academic literature presents a mixed view on whether AI presents novel risks. For instance, Cerka et al (2015) contend that the key determinant of risk in AI systems that distinguishes them from existing digital technologies is their ability to learn and act autonomously.[81] AI systems do not have to be completely programmed and can modify their conduct based on prior inputs of data. Consequently, there is a loss of human control of the system. As a corollary to AI's autonomy risk, Cerka et al (2015) highlight that its outcomes cannot always be predicted.[82] However, Park (2024) argues that other digital technologies, like algorithms, learn and act autonomously as well, and can also entail unpredictable outcomes.[83]

Overall, our case studies show that AI systems largely do not present novel risks. In some cases, such as in the case of generative AI, these systems may not fall neatly within the purview of either publisher or intermediary, though case law in other jurisdictions says otherwise. In addition, when used in a medical context, as well as the Tesla car crash, there may be a danger of overreliance on these systems and consequently, an abandonment of human agency. However, our case studies also reveal that risks must be considered on a case-by-case basis and cannot be generalized across varying situations.

	Unlawful Content	Privacy	Security	Fraud	Copyright violation	Autonomous learning and action	Unpredictability	Direct Physical Harm	Interconnectedness
AI Systems	✓	✓	✓	✓	✓	✓	✓	✓	✓
Online Intermediaries/existing digital technologies	✓	✓	✓	✓	✓	✓	✓	✓	✓

2. What considerations must be taken into account when considering liability regimes for AI?

At the risk of repetition, our case studies reveal that risk and liability in the context of AI is situational. This consideration is important given that there is a tendency to treat "AI" as a homogenous technology, despite the heterogeneity of systems, harms, and deployment contexts.[84] The result of AI's treatment as a monolith is the blanket application of rules to different systems in different contexts, an approach largely followed by the EU AI Act, which has been critiqued by Park (2024) for lacking both proportionality and granularity.[85]

Involvement of multiple stakeholders and the interdependence of AI components in AI value chains[86][87]

Different elements of digital products, like hardware and digital content, might be sold separately and produced by different parties. This fragmentation of production can complicate tracing the origin of a malfunction or assigning responsibility for the fault to a single manufacturer. Parties affected by such malfunctions may find themselves dealing with hardware manufacturers, software designers, software developers, facility owners, and others. This was also highlighted by our fourth case study pertaining to a faulty sensor in an autonomous vehicle.

Importantly, Buiten et al (2023) also note that the involvement of multiple stakeholders is not something unique to AI technologies, and can be found in the automotive sector as well as other technology fields where product liability doctrines have been effective.[88] However, if a car component malfunctions, it is discernible and relatively easier to isolate. With certain AI systems like generative AI, it may not be as simple to identify where the problem arose and whose fault it was because of a lack of explainability (discussed in the next point).

Certain types of AI systems can lack transparency, making it difficult to understand how a certain output or action came about[89]

Establishing causality can be problematic when there is no traceable and predictable link between AI design and harm. Explainability is a challenge in generative and discriminative AI systems. It is nearly impossible to pinpoint what input yielded a particular output. This, in turn, confounds liability because it makes attribution of fault challenging. Individuals who suffer harm might not recognize that they have been affected or may struggle to trace the source of this harm. In such cases, how can one fairly AI developers or deployers be accountable or liable for the results produced by these systems? As we noted in the case of both the political consultant and the errant chatbot, it is important to determine who has exerted/is capable of exerting control over a system in order to determine liability. This again, will be highly contextual.

The nature of harm is not consistent, nor is the manner in which it can occur[90]

Again, AI-related harms are exceedingly contextual. For example, victims can contribute to harm to themselves, even when they are familiar with the technology. This defeats the argument made by some scholars about placing all responsibility on developers or deployers of AI because it is a better way to account for the knowledge deficits of victims and the judiciary.

3. What kind of liability doctrine should be applied to AI systems?

Park (2024) provides a useful framework for thinking about the risks posed by AI systems.[91] Rather than thinking about AI risks in generalized terms, Park (2024) pushes for the consideration of the risks of AI in specific contexts.[92] The case studies we explored today indicate this is a useful approach when considering the risks and liabilities of AI systems.

- **Autonomous AI:** Robots and various autonomous AI systems have the capability to sense their surroundings, make reasoned decisions, and operate controls. This category encompasses a wide range of technologies such as autonomous vehicles, automated facilities, surgical robotics, and pricing algorithms.

- **Discriminative AI**: This type of AI is designed to score or classify individuals, assigning them benefits or detriments, or identifying people, their conditions, or objects from a dataset. Discriminative AI has three primary forms:

  - **Allocative AI**: This AI system is used for distributing limited resources, such as opportunities, resources, or honours, among individuals. Its applications include AI-driven recruitment, admissions processes, credit scoring, insurance underwriting, and ranking systems.
  - **Punitive AI**: This AI type is applied to assign adverse consequences or sanctions to individuals. It finds use in areas such as AI-enabled criminal sentencing, fraud detection, and claims adjustment.
  - **Cognitive AI**: This form of discriminative AI encompasses neither punitive nor allocative AI. Examples include computer vision, AI imaging and diagnostics, and biometric identification. It can either augment or substitute human cognitive processes.

- **Generative AI:** Generative AI processes data (mainly unstructured data like text or images) into a latent state, then decodes this state into creations like speeches or artwork. It is utilized for AI-assisted writing, composing, painting, machine translation, image captioning, among other tasks.[93]

According to Park (2024), these systems present different risks and must be treated differently from a regulatory lens.[94] Our case studies corroborate this position. As such, the blanket application of strict liability, just because an AI system is involved, is not appropriate.

Further, Nappinai (2024)[95] highlights a core principle of law-making, namely that laws must be enforceable. Consequently, it is impractical to craft legislation that addresses every potential challenge posed by emerging technologies, such as AI. Instead, regulation should focus on specific aspects of different AI systems, or different situations involving AI that require oversight. The government should develop laws targeting identified threats and vulnerabilities, ensuring that the regulations are both relevant and actionable.

In conclusion, AI liability requires a highly contextualized framework to manage risks whilst balancing the need to encourage innovation.

V. CONCLUSION

The complex landscape of AI liability highlights the need for a dynamic and adaptable legal framework to address risks proportionately. As AI systems continue to integrate into various sectors, and establishing clear guidelines for accountability remains a priority. While regions like the European Union are rushing to introduce new regulations, the United States and India are developing their own approaches, and the challenge of harmonizing these efforts on a global scale persists. This research paper is an exploratory process, which underscores the importance of balancing innovation with accountability, advocating for continued dialogue and collaboration to address the evolving challenges of AI liability.

The discussion in this paper drives home the message that AI systems cannot be straightjacketed under a uniform system of liability. Rules and standards must be devised to account for the different types of AI systems as well as the varying contexts in which they are deployed. Sectoral regulations and standards are likely to play a role in AI governance and may be a preferable option, given the complexities and nuancing in each sectoral use of AI. A one-size fits all approach of strict liability, as is being considered by the European Union, and promoted by European scholars, would do little to deal with the different challenges presented by AI systems. It would also give sanction to those that wish to misuse such systems to the detriment of others, as there would be little to no liability placed on the former's shoulders as it would largely fall on AI developers.

In conclusion, the path forward involves a collaborative effort between legal professionals, technologists, civil society, academia, and policymakers. Through ongoing dialogue and adaptation, we can develop robust liability frameworks that not only address current issues but also anticipate future developments in the field of AI. Only by doing so can we ensure that the benefits of AI are maximized while minimizing the risks and ensuring that responsibility is appropriately assigned.

ENDNOTES

1 Zech, Herbert. 2021. "Liability for AI: public policy considerations." *ERA Forum* 22, no. April 2021 (January): 147–158. https://link.springer.com/article/10.1007/s12027-020-00648-0

2 De Conca, S. (2022). Bridging the liability gaps: Why AI challenges the existing rules on liability and how to design human-empowering solutions. https://doi.org/10.1007/978-94-6265-523-2_13

3 Zech, Herbert. 2021. "Liability for AI: public policy considerations."

4 Zech, Herbert. 2021. "Liability for AI: public policy considerations."

5 Zech, Herbert. 2021. "Liability for AI: public policy considerations."

6 Zech, Herbert. 2021. "Liability for AI: public policy considerations."

7 Park, Sangchul. 2024. "Bridging the Global Divide in AI Regulation: A Proposal for a Contextual, Coherent, and Commensurable Framework." Washington International Law Journal 33, no. 2 (August): 1-58. https://arxiv.org/pdf/2303.11196.

8 Buiten, Miriam, Alexandre d. Streel, and Martin Peitz. 2023. "The law and economics of AI liability." *Computer Law & Security Review* 48 (April): 1-20. https://doi.org/10.1016/j.clsr.2023.105794

9 Čerka, Paulius, Jurgita Grigienė, and Gintarė Sirbikytė. 2015. "Liability for damages caused by artificial intelligence." Computer Law & Security Review 31, no. 3 (June): 376-389. https://www.sciencedirect.com/science/article/abs/pii/S026736491500062X.

10 Park, Sangchul. 2024. "Bridging the Global Divide in AI Regulation: A Proposal for a Contextual, Coherent, and Commensurable Framework."

11 Pal, L.A. 2005. Case Study Method and Policy Analysis. In: Geva-May, I. (eds) Thinking Like a Policy Analyst. Palgrave Macmillan, New York. https://doi.org/10.1057/9781403980939_12

12 Pal, L.A. 2005. Case Study Method and Policy Analysis.

13 Jaiman, Ashish. 2020. "Debating the ethics of deepfakes." ORF. https://www.orfonline.org/expert-speak/debating-the-ethics-of-deepfakes.

14 Rana, Aprajita, and Navdeep Baidwan. 2024. "MeitY Revises AI Advisory, Does Away with Government Permission Requirement – Update." AZB & Partners. https://www.azbpartners.com/bank/meity-liberalizes-ai-advisory-dated-march-1-2024-following-industry-concerns-and-issues-revised-advisory-on-march-15-2024/.

15 MacDonald, Abby. 2024. "The Uses and Abuses of Deepfake Technology." Canadian Global Affairs Institute. https://www.cgai.ca/the_uses_and_abuses_of_deepfake_technology#Good.

16 https://www.lawfaremedia.org/article/deepfake-iphone-apps-are-here

17 MacDonald, Abby. 2024. "The Uses and Abuses of Deepfake Technology." Canadian Global Affairs Institute. https://www.cgai.ca/the_uses_and_abuses_of_deepfake_technology#Good.

18 Reuters. 2024. "'Cheapfakes', not deepfakes, spread election lies in India." The Hindu, May 31, 2024. https://www.thehindu.com/sci-tech/technology/cheapfakes-not-deepfakes-spread-election-lies-in-india/article68235040.ece.

19 Bond, Shannon. 2023. "People are arguing in court that real images are deepfakes." NPR, May 8, 2023. https://www.npr.org/2023/05/08/1174132413/people-are-trying-to-claim-real-videos-are-deepfakes-the-courts-are-not-amused.

20 Section 196, Bharatiya Nyaya Sanhita https://www.mha.gov.in/sites/default/files/250883_english_01042024.pdf

21 "*Actus non facit reum; nisi mens sit rea":* the act itself is not criminal unless accompanied by a guilty mind:, Sadaf, Fahim, & G. S.Bajpai. 2020. "AI and Criminal Liability." *Indian Journal Artificial Intelligence and Law*, 1(1). https://www.academia.edu/86155216/AI_and_Criminal_Liability

22 Fahim, Sadaf. 'AI and Criminal Liability'. Indian Journal of Artificial Intelligence and Law, Volume 1, Issue 1 (2020). Accessed 20 September 2024. https://www.academia.edu/86155216/AI_and_Criminal_Liability.

23 Hallevy, Gabriel. 2010. "The Criminal Liability of Artificial Intelligence Entities by Prof. Gabriel Hallevy :: SSRN." Search eLibrary :: SSRN. https://ssrn.com/abstract=1564096.

24 Hallevy, Gabriel. 2010. "The Criminal Liability of Artificial Intelligence Entities by Prof. Gabriel Hallevy :: SSRN." Search eLibrary :: SSRN. https://ssrn.com/abstract=1564096.

25 Asoke Kumar Sarkar & Anr. vs Radha Kanta Pandey, AIR 1967 CAL 178

26 Titan Industries Limited v. M/s Ramkumar Jewellers, 2012 (50) PTC 486 (Del)

27 Anil Kapoor vs Simply Life India & Ors, CS(COMM) 652/2023

28 See, for instance, Khalaaf, Heidy. 2023. "Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems." *Trail of Bits*, (March), 1-29. https://www.trailofbits.com/documents/Toward_comprehensive_risk_assessments.pdf

29 Lemmon v. Snap, Inc., 440 F. Supp. 3d 1103 (C.D. Cal. 2020)

30 Section 230 of the Communications Decency Act, 1996

31 Summarised from Lemmon v. Snap, Inc., 440 F. Supp. 3d 1103 (C.D. Cal. 2020

32 Rylands v. Fletcher, (1868) LR 3 HL 330

33 M.C. Mehta v. Union of India. 1987 S.C.R. (1) 819, Supreme Court of India, 1987.

34 Consumer Education & Research Centre v. Union of India. 1995 S.C.C. (3) 42, Supreme Court of India, 1995.

35 Meta. 2023. "Llama 2 Community License Agreement - Meta AI." Meta. https://ai.meta.com/llama/license/

36 Meta. n.d. "Llama 2 - Acceptable Use Policy." Meta. Accessed September 12, 2024. https://ai.meta.com/llama/use-policy/.

37 Meta. 2023. "Llama 2 Community License Agreement - Meta AI." Meta. https://ai.meta.com/llama/license/.

38 Gade, Pranav, Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. 2024. "BadLlama: cheaply removing safety fine-tuning from Llama 2-Chat 13B." (May). https://arxiv.org/pdf/2311.00117.

39 Widder, David G., Dawn Nafus, Laura Dabbish, and James Herbsleb. 2022. "Limits and Possibilities for "Ethical AI" in Open Source: A Study of Deepfakes." FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, (June), 2035-2046. https://doi.org/10.1145/3531146.3533779.

40 Widder et al. 2022

41 TOI World Desk. 2024. "Why Elon Musk's Grok-2 AI images is raising serious concerns." *Times of India*, September 4, 2024. https://timesofindia.indiatimes.com/world/us/why-elon-musks-grok-2-ai-images-is-raising-serious-concerns/articleshow/113066647.cms.

42 WION Web Team and Vinod Janardhan. 2024. "Grok or Gross? Elon Musk's new AI photo tool generates deepfakes of Taylor Swift, Kamala Harris." WION, August 15, 2024. https://www.wionews.com/technology/grok-or-gross-elon-musks-new-ai-photo-tool-generates-deepfakes-of-taylor-swift-kamala-harris-750343.

43 Shreya Singhal vs U.O.I, 2015 (2) SCC (CRI) 449

44 Order No. 12. 2022. People's Republic of China, Cyberspace Administration of China Ministry of Industry and Information Technology of the People's Republic of China. https://perma.cc/JE3W-PF26

45 Sala, Alessandra. 2024. "AI watermarking: A watershed for multimedia authenticity." ITU. https://www.itu.int/hub/2024/05/ai-watermarking-a-watershed-for-multimedia-authenticity/.

46 Heitzenrater, Chad. 2024. "The Case for and Against AI Watermarking." RAND. https://www.rand.org/pubs/commentary/2024/01/the-case-for-and-against-ai-watermarking.html.

47 Summarised from Hoffman, Jacob. 2024. "AI Watermarking Won't Curb Disinformation." Electronic Frontier Foundation. https://www.eff.org/deeplinks/2024/01/ai-watermarking-wont-curb-disinformation.

48 Summarised from Hoffman, Jacob. 2024. "AI Watermarking Won't Curb Disinformation."

49 Stricklin, Kasey. 2021. "Social Media Bots and Section 230 Reform with Unintended Consequences." CNA. https://www.cna.org/our-media/indepth/2021/04/social-media-bots-and-section-230.

50 Stricklin, Kasey. 2021. "Social Media Bots and Section 230 Reform with Unintended Consequences."

51 Mohanty, Bedavyasa et al. 2017. "Hitting Refresh: Making India-US data sharing work." ORF. https://www.orfonline.org/research/hitting-refresh-india-us-data-sharing-mlat.

52 Perault, Matt, and Richard Salgado. 2024. "Untapping the Full Potential of CLOUD Act Agreements." CSIS, June 6, 2024. https://www.csis.org/analysis/untapping-full-potential-cloud-act-agreements.

53 Perault, Matt, and Richard Salgado. 2024. "Untapping the Full Potential of CLOUD Act Agreements."

54 Perault, Matt, and Richard Salgado. 2024. "Untapping the Full Potential of CLOUD Act Agreements."

55 United Nations Office on Drugs and Crime “Concluding session of the Ad Hoc Committee.” Accessed September 12, 2024. https://www.unodc.org/unodc/en/cybercrime/ad_hoc_committee/ahc_concluding_session/main.

56 Gian Kaur v. State of Punjab (1996) 2 SCC 648

57 Aruna Ramchandra Shanbaug v. Union of India (2011) 4 SCC 454

58 Xiang, Chloe, Janus Rose, Magdalene Taylor, Jordan Pearson, Matthew Gault, Samantha Cole, and Ryan S. Gladwin. 2023. “'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says.” VICE. https://www.vice.com/en/article/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says/.

59 Xiang, Chloe, Janus Rose, Magdalene Taylor, Jordan Pearson, Matthew Gault, Samantha Cole, and Ryan S. Gladwin. 2023. “'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says.” VICE. https://www.vice.com/en/article/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says/.

60 Indurkhya, Bipin. “Ethical Aspects of Faking Emotions in Chatbots and Social Robots.” 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 1719. https://arxiv.org/pdf/2310.12775.

61 Natale, Simone. 2021. “Chapter 3 The ELIZA Effect: Joseph Weizenbaum and the Emergence of Chatbots.” In Deceitful Media: Artificial Intelligence and Social Life After the Turing Test, 50-67. N.p.: Oxford University Press. https://doi.org/10.1093/oso/9780190080365.003.0004.

62 Indurkhya, Bipin. “Ethical Aspects of Faking Emotions in Chatbots and Social Robots.”

63 Indurkhya, Bipin. “Ethical Aspects of Faking Emotions in Chatbots and Social Robots.”

64 Indurkhya, Bipin. “Ethical Aspects of Faking Emotions in Chatbots and Social Robots.”

65 Indurkhya, Bipin. “Ethical Aspects of Faking Emotions in Chatbots and Social Robots.”

66 Moffatt v. Air Canada 2024 BCCRT 149

67 Kim, Minseon, Hyomin Lee, and Lee Gong. 2024. “Automatic Jailbreaking of theText-to-Image Generative AI Systems.” arXiv, (May). https://arxiv.org/pdf/2405.16567.

68 Kim, Minseon, Hyomin Lee, and Lee Gong. 2024. “Automatic Jailbreaking of theText-to-Image Generative AI Systems.”

69 Liu Yi, Gelei Deng, Zhengzi Xu et al. 2024. “Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study.” *arXiv*, (March). https://arxiv.org/pdf/2305.13860.

70 Liu Yi, Gelei Deng, Zhengzi Xu et al. 2024. “Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study.”

71 F. Supp. 3d, 22-cv-1461 (PKC), 2023 WL 4114965, at *2 (S.D.N.Y. June 22, 2023);

72 Ryan, William A., Allen Garrett, Kilpatrick Townsend, and Brad Sears. 2023. “Practical Lessons from the Attorney AI Missteps in Mata v. Avianca.” Association of Corporate Counsel. https://www.acc.com/resource-library/practical-lessons-attorney-ai-missteps-mata-v-avianca.

73 Vajawat, Bhavika, Damodharan Dinakaran, Omprakash V. Nandimath, Arpita HC, Channaveerachari N. Kumar, Chethan Basavajappa, and Suresh B. Math. 2023. “The Consumer Protection Act, 2019: A critical analysis from a medical practitioner’s perspective.” Indian Journal of Medical Ethics 9, no. 1 (November): 1-5. http://dx.doi.org/10.20529/IJME.2023.073.

74 *Dr. Suresh Gupta v. Government of NCT of Delhi*, (2004) 6 SCC 422

75 Khlaaf, Heidy. Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems, Trail of Bits, 2023.

76 Park, Sangchul. 2024. “Bridging the Global Divide in AI Regulation: A Proposal for a Contextual, Coherent, and Commensurable Framework.”

77 Jones v. W + M Auto, 31 A.D.3d 1099, 2006 N.Y. Slip Op. 5398, 818 N.Y.S.2d 396 (N.Y. App. Div. 2006)

78 Jones v. W + M Auto, 31 A.D.3d 1099, 2006 N.Y. Slip Op. 5398, 818 N.Y.S.2d 396 (N.Y. App. Div. 2006)

79 Summarised from BBC. 2020. “Tesla Autopilot crash driver 'was playing video game.'” BBC, February 26, 2020. https://www.bbc.com/news/technology-51645566.

80 BBC. 2020. “Tesla Autopilot crash driver 'was playing video game.'” BBC, February 26, 2020. https://www.bbc.com/news/technology-51645566.

81 Čerka, Paulius, Jurgita Grigienė, and Gintarė Sirbikytė. 2015. “Liability for damages caused by artificial intelligence.”

82 Čerka, Paulius, Jurgita Grigienė, and Gintarė Sirbikytė. 2015. "Liability for damages caused by artificial intelligence."

83 Park, Sangchul. 2024. "Bridging the Global Divide in AI Regulation: A Proposal for a Contextual, Coherent, and Commensurable Framework."

84 See Park, Sangchul. 2024. "Bridging the Global Divide in AI Regulation: A Proposal for a Contextual, Coherent, and Commensurable Framework."

85 Park, Sangchul. 2024. "Bridging the Global Divide in AI Regulation: A Proposal for a Contextual, Coherent, and Commensurable Framework."

86 Buiten, Miriam, Alexandre d. Streel, and Martin Peitz. 2023. "The law and economics of AI liability."

87 Giuffrida, Iria. 'Liability for AI Decision-Making: Some Legal and Ethical Considerations'. *Fordham Law Review* 88, no. 2 (1 November 2019): 439. https://ir.lawnet.fordham.edu/flr/vol88/iss2/3.

88 Summarised from Buiten, Miriam, Alexandre d. Streel, and Martin Peitz. 2023. "The law and economics of AI liability."

89 Summarised from Buiten, Miriam, Alexandre d. Streel, and Martin Peitz. 2023. "The law and economics of AI liability."

90 Summarised from Buiten, Miriam, Alexandre d. Streel, and Martin Peitz. 2023. "The law and economics of AI liability."

91 Park, Sangchul. 2024. "Bridging the Global Divide in AI Regulation: A Proposal for a Contextual, Coherent, and Commensurable Framework."

92 Park, Sangchul. 2024. "Bridging the Global Divide in AI Regulation: A Proposal for a Contextual, Coherent, and Commensurable Framework."

93 Summarised from Park, Sangchul. 2024. "Bridging the Global Divide in AI Regulation: A Proposal for a Contextual, Coherent, and Commensurable Framework."

94 Park, Sangchul. 2024. "Bridging the Global Divide in AI Regulation: A Proposal for a Contextual, Coherent, and Commensurable Framework."

95 HTsmartcast, 2024. *AI Rising Podcast*. Episode 9, "Law is Glacial. How will it keep up with AI?" August 21, 2024. https://www.htsmartcast.com/technology-podcasts/ai-rising-podcast/page/4/.


ESYA
centre


Cyber
SAATHI
An initiative of N. S. Nappinai


AI KNOWLEDGE
CONSORTIUM