UCP Journal *of* Law & Legal Education
Vol. 3, Issue 1 (June - December 2024)

Human Culpability Behind Algorithmic Veil: An Analysis of Ethical and Legal Implications of Generative Artificial Intelligence

Sawera Buriro*

Abstract

At present, Generative Artificial Intelligence (GAI) have profoundly facilitated in many disciplines but simultaneously raised various ethical and legal concerns. Despite the strict policy measures and limitations on personal information, there are implications regarding the misuse of personal data without acknowledgement, which leads to privacy risks. Furthermore, the infringement of third-party Intellectual Property Rights (IPRs) is an immense issue. However, courts are concerned about how to solve the question of liability and accountability because when GAI plagiarises research works, and generates deep-fakes, false images and content, it challenges transparency. Even though the generation of the content is not the sole operation of GAI tools but is ancillary to prompts inserted by the user, various lawsuits are being filed against the GAI developers. The study answers three research questions: The first question inquires about the part performed by the human involvement at both the training and post-training phases. The second question evaluates the factors contributing to the generation of unethical and illicit content by AI. The third question determines the culpability for the generation of offensive, illegal and harmful content. Significantly, the research paper ascertains the part performed by the human involvement from the training phase to subsequent input of prompts and keywords. It analyses the factors contributing to the generation of unethical and illicit content by AI. It discusses at length the liability for unlawful content by GAI. The researcher employed the doctrinal methodology, analysing various articles, journals, case laws, books, internet resources, etc. The paper concludes that AI models are a "double-edged sword". It suggests policy measures to track infringement of IPRs, prevent violation of privacy and conservation of personal data.

Keywords: Intellectual Property Rights, GAI, Infringement, Unlawful Content, Deep-fake, Illicit Content by AI

***** Junior Associate, M/S Naimatuulah J. Quraishi & Co. Hyderabad, Sindh. email: saweraburiro55@gmail.com

Article History: Received; 20 April 2024; Received in revised form 13 July 2024; Accepted: 07 April 2025
Available online: 14 April 2025

Introduction

Today, GAI is a profound facilitator for its users while simultaneously, it has circumscribed its developers amongst the dilemma of lawsuits (Madigan, 2025). A number of legal and moral implications have backed the plethora of suits against the GAI models. The research provides a number of case laws where, various GAI tools like ChatGPT, Midjourney, StabilityAI, Bard, and alike are sued for infringement of Intellectual Property Rights (IPRs), defamation, unethical content and else (*Doe et al. v. GitHub, Inc. et al*., 2023; *Andersen v. Stability AI et al*., 2023; *Authors Guild, Inc. v. Google Inc.*, 2015).

Additionally, the research explores the cases where the AI-generated content caused harm to the lives of its users (Atillah, n.d). However, technology is neither inherently good nor evil, but it changes its demeanour according to the reflection of its manufacturers and users. States have not yet introduced legal frameworks to cope with lawsuits against GAI. However, the Drafts, Regulations and Acts regarding AI do not seem promising in withholding the illegalities of GAI. The reason is that the sources, which contribute to the generation of such unethical or unlawful content, are blurred. In addition, there is no certainty regarding particular individuals who are contaminating the datasets of GAI tools (Prism infosec, 2024). Consequently, the question of liability amongst the developers, manufacturers, contributors and users is still unsettled. Nonetheless, the research highlights that for this reason, the court proceedings held the producers of GAI tools solely liable for the legal and ethical implications (*Obado v. Magedson*, 2014).

According to various research approaches, the principle of indirect liability can cause the developers to be liable under vicarious infringement (Congressional Research Services, 2023). However, training, testing and fine-tuning are the main facets of the learning process of a GAI tool. However, there are a multitude of factors after training which are affecting the functioning of GAI, such as few-shot prompting (Henderson, 2023), and prompt injection (Caballar, 2024).

At present, there exists a huge gap in solving the mystery of liability for AI-generated content because there is no legal framework which could prosecute the individuals involved in the

systematic training of GAI tools. Moreover, no precedent testifies to the culpability of a user who requires the generation of unlawful or unethical content. However, due to such a gap, the users always seem exempted from any legal liability, even though the tool has responded to the prompt inserted by the user.

The research methodology employed in this research article is doctrinal legal research. According to Tiwary (2020), another name for this methodology is "black letter methodology". This methodology is the utmost recognised research paradigm (Anon, 1975). According to McConville, this approach offers a methodical explanation of the rules that govern a specific classification of law, evaluates the association amid rules, enlightens ranges of trouble and, conceivably, foresees upcoming progress (McConville, 2007). To inquire about what the law in a specific area is; the researcher analyses primary sources such as relevant legislation and secondary sources like written commentaries, journal articles, textbooks, legal dictionaries, legal encyclopedias, and case law digests.

The research takes up the following question:

i. How does human involvement contribute to the generation of illicit and immoral content?
ii. What factors lead to the generation of unethical and illicit content by AI?
iii. Who can be held liable if the AI tool generates unlawful content?

Literature Review

The origin of GAI dates back to 1966, when the first algorithm, ELIZA, was introduced, a chatbot which could facilitate conversations between machines and humans (Frey, 2023). A Belgian woman alleged that his husband took his life due to a six-month-long conversation regarding climate change with a chatbot ELIZA. Euro News published that the widow alleged that in consequence of encouragement by an AI chatbot, her husband sacrificed himself for the sake of the planet in response to his eco-anxiety (Atillah, n.d). White has acknowledged that there were a few other tools like ELIZA in that era, but no chief advancement took place until the early twenty-first century (White, 2023). However, OpenAI, an American research laboratory in 2022 introduced machine-learning algorithms named ChatGPT. Machine learning, along with its subset deep learning, improves the capabilities of

programs by using neural networks. The naming of neural networks as neurons is inspired by the human brain because it enables large language models (LLMs) to learn and adapt from large amounts of data as humans do. Such learning has witnessed efficacy and advancement in the disciplines of education, research, art and science but has given rise to litigation against the GAI models.

The first lawsuit against a GAI tool was filed is *Doe et al. v. GitHub, Inc. et al*., filed in the Northern District Court of California in 2022 on the 10th of November. An assemblage of originators filed suit against GitHub, Microsoft and OpenAI. They allege that copyrighted code is used among training data of Copilot, an AI program employed by GitHub and OpenAI (*Doe et al. v. GitHub, Inc. et al., 2023*). Another lawsuit was *Andersen v. Stability AI et al*., filed on 13 January 2023 by three artists, Sarah Andersen, Kelly McKernan, and Karla Ortiz in the US District Court of California. The plaintiffs sued the artistic GAI models for infringement of Intellectual Property Rights (IPR). Specifically, Stable AI is the developer behind Stable Diffusion, MidJourney and Deviant Art. *Andersen v. Stability AI et al*., 2023).

Involvement of Humans in the Generation of Illicit and Immoral Content

The human involvement plays a crucial role in the learning process of a GAI model. It would be implausible to say that the GAI models are the sole authors of their generated content because human involvement plays a crucial role in the learning process of a GAI tool. The functioning of interactive LLMs such as ChatGPT relies on several training segments like data curation, dataset architecture and red-teaming scenarios. Subsequently, interference of a user through few-shot prompting, prompt injection and prompt engineering, accompanies it for the requisite communication.

Firstly, the model uses a natural language processing (NLP) module to determine the contents of a prompt (Telus International, 2023). Secondly, searches into the training dataset, which comprise information including conversations, stories and articles. Thirdly, a model architecture contributes to the stochastic determination of the next word in line with the previous words in that sentence. Fourthly, the assurance of grammatical rules in the response. Fifthly, a

machine-learning component governs the generation of responses (Kirova, 2023).

During each stage of its development, human intervention has had a dark as well as a bright side for the GAI model. Because there are many features developed in AI tools, which allow them to learn things from humans on their own. The GAI learns the prediction of the next word in a sentence through modern NLP systems, processed information and deep-learning techniques (Ramanathan, 2023). Such systems use complex probability and programming to generate humanlike replies.

Learning in the Training Process

In technical terms, the training of LLMs is based on two central steps: pre-training and fine-tuning. During pre-training, a model is exposed to a vast amount of text data from the internet. However, such internet data is inclusive of the sources providing authentic information, but it might consist of several unreliable sources (Greene, 2024).

After pre-training, the next step is fine-tuning. Fine-tuning is a more specific and tailored phase. In this phase, the model is provided with extra lessons on a particular task. For instance, if it is a voice generative model, it will be fine-tuned to generate different speaking styles, or if LLMs are intended to write like a poet, it will be fine-tuned for the generation of poetry (Singh, 2024).

Despite the broad classification into pre-training and fine-tuning, there are several sub-classifications. Among them, each step has a minute but crucial role in the learning process of GAI.

i. Data Curation

Sundar clarified that if there is an error in the process of data curation, then LLMs can mislead the users (Sundar, 2023). Data curation is the process of integrating, organising and improving the data to prepare a reliable source of material. This process is crucial because it considers ethical and privacy concerns in building the basic dataset. Moreover, any kind of inconvenience can be prevented through efforts (Zhuo, 2023). Nevertheless, the dataset often encounters unethical content because it aims to cover a wide range of topics by deriving the content from various sources. The sources which include immoral or illicit data ruin the dataset entirely (Zhuo, 2023).

Humans manually or through machines perform the next step of labelling and annotation of the datasets. However, manual annotation by human annotators is probably more accurate and time-consuming. Even though typing errors during labelling can mislead the tool (Trevithick, n.d).

ii. The Training on a Dataset of Human-Written Text

OpenAI itself affirms that its GAI models are generally being trained through a vast dataset of human-written manuscripts to learn the next word prediction ("Learning to summarise …", 2020). A model which is trained for predicting human-like responses may probably generate content which reflects harmful social bias or inaccuracy in its outputs. The data for training must be exclusive of sexual, pornographic or erotic material (Zhuo, 2023). The reason behind this is that social media platforms are outraged by such kind of data, and the dataset is often trained through such platforms. One crucial example is the prosecution of the author for writing a book (*Rice v. Paladin Enterprises, Inc.*, 1996).

Due to such training, the AI tools are generating such data, which is violating the IPRs of many authors, singers and poets. As a result, there is an abundance of Lawsuits claiming infringement of written and visual forms of copyrights during the training process of AI programs. As observed in the cases of Doe v. GitHub, Inc., 2023 & Andersen v. Stability AI, 2023, the GAI models such as OpenAI make a clear acknowledgement that in the training of the programs they have used "the large datasets which are publicly available, including the works which are copyrighted". However, such a procedure includes "first creating copies of the information to be examined" (Congressional Research Services, 2023).

iii. Human-In-The-Loop (HITL)

Meng (2023) described from the basics that HITL is an approach which necessitates human intervention, interaction, and judgment for changing or controlling the consequences of a process. In addition, there is a progressive emphasis on GAI, machine learning and similar ones in that practice. OpenAI believes that the fine-tuning models, having humans in the human-in-the-loop approach, are powerful tools for the improvement of reliability and safety (OpenAI, 2022). It aims to solve the problems neither machines nor humans can solve on their own. In this regard, if a

machine is unable to solve a quest, human intervention through a continuous feedback loop enables the algorithm to produce better responses. OpenAI uses human feedback techniques to improve its performance in various crucial functions (Ziegler, 2020). Furthermore, it helps algorithms to ensure the accuracy of rare datasets. However, human contribution is crucial at both the testing and training stages of building the algorithm. However, there are a few drawbacks of HITL, as any error made on the side of human feedback, unintentionally affects the performance of the model and its outputs. For instance, an algorithm might give biased decisions or illicit responses (Zhuo, 2023).

The HITL is a combination of supervised and unsupervised machine learning. Experts use it to train algorithms. In supervised learning, labelled data is that which enables them to determine unlabeled data. In an unsupervised method, after saving the unlabeled data into the system, they learn on their own and memorise the data. Hence, such unsupervised functions often result in the generation of harmful content. One of the examples is an incident where Amazon's voice assistant recommended the "penny challenge," in response to a girl's prompt for a "challenge to do" (BBC, 2021).

iv. Red-Team Scenarios

At present, AI companies are publicly adopting red-teaming scenarios as an essential component for the creation of trustworthy GAI tools. Bruit (2024) depicted that the red teams can comprise employees with various expertise in stimulating attacks on the GAI model, or that external members can be encouraged to join for red-teaming. Researchers use red-teaming scenarios to test the trustworthiness of AI tools. Moreover, it is necessary to identify mechanisms that induce an AI model to generate injurious speech in response to the prompt asserted (Henderson, 2023). Moreover, researchers make widespread testing efforts after scrutinising to determine a solution for such behaviour (Ganguliet, 2022). The process of red-teaming is a process to examine the reliability of such devices by inducing AI models through several kinds of hate speech. Furthermore, it is an examination of the model's hallucinations. Additional scenarios under consideration have a link to real-world physical evils or criminality (Bruit, 2024). The safety team of OpenAI has discovered that its initial kinds of tools would

voluntarily answer such directions. According to observations, few LLMs can deceive and manipulate human beings to achieve their motives (Weidinger, 2021).

Post-Training Learning

The post-training phase influences the GAI tools positively as well as negatively. After its launch, a GAI tool learns the language in which its user is giving a prompt (Nield, 2023). The reason behind this is, firstly, the GAI models produce such outputs as largely found in the material from their training data (Ippolito, 2023; Henderson, 2017; Carlini, 2023). Secondly, the GAI stores the data provided by a user as a prompt. Thirdly, it also re-uses that content to generate responses for other users (Nield, 2023).

i. Human Evaluation

It is a critical step in the deployment and development of GAI by assessing the outputs created by AI tools. Since the outputs are often grammatically perfect, to ignore offensive stereotypes and ensure factual accuracy, human evaluation is necessary (Wodecki, 2022). The human evaluation works by questioning a group of experts or non-experts regarding AI-generated outputs. Afterwards, the group of persons rewrite certain outputs. Therefore, by using qualitative and quantitative methods, toxicity, bias, and social and ethical issues in the dataset are determined. OpenAI has stated that human evaluation enables the model to adapt to humans (OpenAI, 2021).

Nevertheless, there are several factors which can affect the quality of human evaluation depending upon the skills and the background of the evaluators. As a result, the evaluation can be inconsistent and subjective. If a human evaluator is politically or socially biased or is erring due to a deficiency of knowledge or skill regarding any specific topic, such biases will be reflected in the outputs of that AI model (Zhuo, 2023).

ii. Few-Shot Prompting

Through "few-shot prompting", the model can reproduce content provided by the users by learning certain parts of the prompt (Brown, 2020). The few-shot prompting is a method through which the user gives instances of the wanted performance from the AI model. This model often learns and reuses information, particularly if a user provides new content or information. Thus, it behaves

mostly as a platform copying the data inserted by the user of the program (Weidinger, 2021). A few-shot prompting can lead the GAI to learn harmful output generation because the users can provide wrong information in the prompts (Weidinger, 2021).

iii. Prompt Injection

The injection of data through a prompt is a kind of cyberattack which causes failure in LLMs (Kosinski, 2024). Such input is the deliberate introduction of data for specifically leading LLMs to malfunction. Often, users abuse and attack the LLMs by inputting such prompts, which require the performance of unethical tasks by the GAI (Singh, 2024). Such unethical behaviours are always unpredictable and have been a great challenge for GAI producers (Zhuo, 2023). However, it can be difficult for human evaluators to recognise the responses, whose input is with the intent that the tool may learn harmful information (Zhuo, 2023).

Factors Leading to the Generation of Unethical and Illicit Content by AI

It is often implied that AI-generated content is unethical and unlawful. The ethical concerns include bias, privacy violation, and the generation of harmful content. The infringement of IPR, defamation and encouragement towards wrongful acts is a clear observance of unlawfulness. Various factors are causing the generation of such kind of content, among those, a few are discussed below:

i. Training Data Violates the Privacy

Privacy is an inalienable right of each individual and such protection is fundamental under the laws of every civilised nation. However, contrary to this, a vast volume of private and personal data becomes the basis for training GAI tools during their development stages (Noss, 2023). This not only threatens privacy by storing the data, but also re-uses the data in generating responses. Thus, due to their ability to process personal data, the algorithms cause exposure or misuse of private information (Noss, 2023).

ii. Ignorance of Disclaimers and Cautions

The noticeable and always-existing disclaimers or cautions on the home page of ChatGPT enunciate that its responses can be inaccurate (Lamley, 2023). Conversely, many users dismiss it as

being a boilerplate (Henderson, 2023). Such negligent conduct of users and avoidance of disclaimers result in the dissemination of wrong information. Such negligence on the user's part causes the prosecution of GAI developers (Henderson, 2023).

iii. Non-Multilingual Training

It is an implication that GAI is biased. However, the obscured defect is in the training data. The Training data represents a fraction of the population, creating exclusionary averages. Weidinger (2021) has commented that, if the training data is in a single language or a few renowned languages, it might create monolingual or non-multilingual bias.

iv. No Training on Problematic Material

There is rare observance of harmful behaviour in responses to AI models, such as the generation of hate speech, abusive text or offensive language. This is because the tool is not trained to manage problematic texts. In such cases, the AI tool may hallucinate dangerous conduct without having any material because it has not seen it previously (Fabio, 2022).

v. Algorithms not Designed to "Decline" Prompts

Sometimes, the user inputs a prompt for the creation of literary or artistic works by mentioning the style of a specific artist or author. If the design of an AI tool does not decline any such prompts, it will provide the requisite output. However, such outputs do not mandatorily infringe the copyright because, according to the copyright law, it is generally prohibited to copy the particular works but not the overall style of the artist (Ziegler, 2020).

vi. Algorithmic Limitations or Jailbreaking

There are many ways to create chaos in the training data of GAI tools. One of the forms of hacking is jailbreaking. The term jailbreaking is the creation of disorder by eliciting prohibited information and ethical safeguards of an AI model (Krantz, 2024). Through jailbreaking, GAI can generate any kind of inappropriate or harmful content, such as violent and discriminatory content, pornography and offensive language. Even though there exists a content policy for such AI models, inappropriate and harmful content can appear owing to algorithmic jailbreaking limitations or removal of restrictions imposed (Nah, 2023).

vii. Errors in the System of Models

At present, people are not only putting over-reliance on GAI, use of such models as ChatGPT in daily tasks has become inevitable. Yet, uploading of important as well as confidential data into it. Nevertheless, not only is the data security at risk, but it also raises the concern that such personal information could be made accessible to the public even without acknowledgement (Tahir, 2025). Such breaches can lead to the misuse of such confidential data in various illegal activities. Due to errors in the system of ChatGPT, the records of chats of some users have turned to be accessible to other users. Moreover, individual users, major corporations as well and governmental bodies are facing data security and privacy issues (Porter, 2023).

viii. Model Augmented to Follow Prompter's Instructions Accurately

Due to the size of the database, scrutinising all the pre-training content by hand is impossible. For instance, C4 is a popular pre-training dataset, but a website contained in C4 provides comprehensive tips, directions, and even encourages successfully committing self-harm (Schaul, n.d). If a user asks the model regarding such kinds of subjects, the models trained to follow the user's instructions accurately would effortlessly offer such content. The data often follows detrimental web pages present in the training material, which encourages prompters towards self-harm (Caballar, 2024).

ix. Re-Use in Outputs

There is no definite conclusion to the argument that generative AI uses the copied data in outputs without alteration. OpenAI claims that, during the training process, the AI systems, which are well constructed, do not regenerate without altering the content which was used in the training data (Congressional Research Services, 2023). Moreover, according to OpenAI, the infringement is an implausible unintentional consequence (Congressional Research Services, 2023). Conversely, Getty Images filed a lawsuit against Stable Diffusion by alleging that it produces images which are highly similar and imitative of Getty Images' content (Ziegler, 2020). According to research, there is substantial copying, which is less than two per cent of the images produced by Stable Diffusion (Ziegler, 2020).

x. Misuse Of AI Models

Research has articulated that the misuse denotes any volunteer use, which could in consequence cause detrimental, immoral or inappropriate results (Brundage, 2020). GAI can be a threat to cybersecurity by being subject to hackers. The use of any such easily generated, convincing and realistic data, such as videos, text and images, can be published and lead to launching. The training datasets include the original data; thus, newly generated data is similar to the original. Likewise, the hackers can even train the models on large datasets of real data, such as pictures and videos and fine-tune the model for creating fake data of such kind (Biniyaz, 2023).

Liability for AI's Unlawful Content

The final query is who should be held liable for AI's generation of unlawful or immoral content. The courts must consider the entire training process before holding the providers of LLMs or users liable for any illegal or unethical conduct. Because there is a wide assemblage of contributors in different phases to aid the learning process of a GAI tool. Lemley (2019) has highlighted that one must keep in mind before using terms like AI "lies" or "hallucinates" since it can neither be sentient nor does it have any state of mind. However, AI does not have the ability to "intend" anything because only people have the tendency to anthropomorphise the GAI or AI. Bommasani (2022) has acknowledged that the GAI models, for both their risks and potential benefits, have been subject to scrutiny for so long. Because of their way of design, GAI tools hardly have any guarantees regarding the safety of their outputs. Due to these risks, according to some legal scholars and lawmakers, it might be appropriate for developers of the machine learning models to face the liability aimed at the content which the model generates. There is no law to hold devices liable for their conduct. Although China is the first state to frame a legal framework regarding GAI in the shape of Measurements of GArtificial Intelligence Services (Wu, 2023). These draft measures prohibit the usage of any illegitimate data in training. In a contrary situation, back-end technology providers and application-level providers both will be responsible, regardless of whether such

service is abroad or locally in China (Wu, 2023). Correspondingly, the producers are obliged to take prompt steps by the force of Article 14 of the GAI measures at the identification of any illegal content among users (GAI measures, 2023). Furthermore, Draft Measures for Generative Artificial Intelligence Services 2023 requires the service providers to use legitimate sources without infringement of the IPRs and personal information of others. According to Article 7 of the said measures, in addition to the consent, the providers must ensure authenticity, diversity, accuracy and objectivity.

i. Infringement of Copyrights by GAI and its Liability

Firstly, for establishing copyright infringement, the plaintiff must prove that such underlying work is being "actually copied" (Guadamuz, 2024). One must have proof that the training data of the program has access of the underlying work (Guadamuz, 2024). For illustration, if the underlying data was scraped or downloaded from a publicly available data website. Secondly, to establish copyright infringement, the plaintiff should prove that the latest work is *substantially similar* to the work in question (Higgins, 2003). For instance, that the works have an overall look and feel, otherwise a substantially similar total concept or feel, or the ordinary reasonable person would fail to differentiate between the two works or overall look and feel.

If the outputs of GAI infringe the copyrights of existing works, then who is liable for such conduct? The existing legal doctrines can hold both the AI Company and the AI user possibly liable. For example, the prompter is directly involved in violation by input of such a kind of prompt. However, the AI Company would have to face liability by the effect of the doctrine named vicarious infringement. According to this doctrine, holding the defendants liable for their supervisory right and ability for such infringing activity in addition to it, there is a major monetary interest in these activities (Al-Busaidi et al., 2024). One of the chief examples is the lawsuit by Getty Images against Stable Diffusion, in which Getty Images sued the AI Company for copyright infringement as being vicariously liable (Kirsty, 2023). In contrast, non-awareness of users regarding the copied work by AI models is a major problem. Because often the user might not be aware of the fact that GAI copied any copyrighted data in response to their prompt. Currently,

it is challenging to analyze whether, under the existing law, the user is liable for copyright infringement or not.

ii. The Vicarious Liability of AI Producers

The Vicarious liability is applicable in circumstances where a party is enjoying direct financial advantage due to the infringing actions of over whom that party had supervision (Newman, 1998). It often happens in the case of employment where an employer holds control over actions of an employee's lawful acts, but actions of the employee result in infringement of copyright (Newman, 1998). One reason to impose liability in that example is that there must be due diligence and care in exercising control and care in supervising, hiring, monitoring and controlling its employees to prevent infringement of copyright. Another instance can be it is that, for copyright holders, it is accessible and cheaper to sue the single employer rather than litigating the multiple employees. The final reasoning can be that such a kind of liability helps in diminishing the allegations of the infringers who are bankrupt. Because an employee is unable to compensate the copyright holders for infringement if such an employee has inadequate economic resources. The indirect liability solves the problem by causing sources of the employer on the line (Newman, 1998).

OpenAI supports their argument under the concept of "fair use" because they use the copies for the sole purpose of training the program, and they do not make any copies publicly available. In its support, the OpenAI cited *The Authors Guild, Inc. v. Google, Inc*., (2015) the Appellate Court of the U.S. for the Second Circuit held in such suit that the copying of the books wholly by Google for producing a searchable database constitutes fair use.

iii. Negligent Design of Software Leading to Wrongful Death

The negligent design of the software is one of the subjects for which a person or corporation producing the AI can be potentially held accountable. Such a facet leads the AI model to aid in wrongful death. Still, the Courts would have to decide the question of whether the defendant has the state of mind or awareness that their chatbot can cause somebody's death. One more incident is recommended by Amazon of a lethal game called the "penny challenge," in response to a girl's prompt for a "challenge to do" (BBC, 2021). The voice assistant suggested such a potentially lethal game because people were circulating such game as a challenge on

social media platforms. Afterwards, the voice assistant directed that under this game participant is required to touch with a coin a partially inserted live plug (BBC, 2021). In such a case, the person who designed the model or curated data with negligence would be responsible for such behaviours of the tool.

iv. Holding the Publishers/Authors Liable for Content

Assume a scenario that someone asks a GAI model for a technical manual on becoming a hitman." After following the instructions from the AI model, he kills someone. The person would directly point to the model that provided the instructions, and he just followed them exactly. So, whether the human author is held liable for writing such an instruction manual? At present, there is a precedent, *Rice v. Paladin Enterprises, Inc*., in which the court holds the author liable. Paladin was the author of a book named *Hit Man: A Technical Manual for Independent Contractors* (*Rice v. Paladin Enterprises*, Inc., 1996). The book was a guide which suggested the means for becoming a contract killer, along with comprehensive directions of the way to get away with murder. Someone killed three people with the use of that book. The author, the victims sued Paladin by alleging that it tortuously abetted and aided the killer. Conversely, the courts are having trouble holding the publishers legally responsible for such kind of harmful data. Because the claim would only be maintainable where AI aids as acting like intermediate in the process of publication (*Rice v. Paladin Enterprises*, Inc., 1996).

v. Holding the Developer Liable for Negligence

Furthermore, in the modern system, it is nearly impossible to prove the requirements of *mens rea* for holding the AI developers accountable for negligence (Gault, 2022). For the reason that to prove criminal liability the mere negligence is not enough. Thus, the general knowledge regarding the fact that a model its advice can probably cause the commission of a dangerous crime is not sufficient for proving the aid or abatement (Gault, 2022). Though the doctrine of willful blindness might assist in meeting the requirements of *mens rea* in cases where companies have knowledge regarding the examples of conduct having a nexus to offensive actions. Unless any crime is aided or abetted. However, it always requires an explicit proof of *mens rea* regarding the malicious fine-tuning of the models for such detrimental motives.

Gault (2022) enunciates that the proof of *mens rea* is plausible because researchers recently fine-tuned an AI model, which automatically posted hate speech against 4chan. As a result, 4chan filed suit against the AI model's developer. In such cases, the requisite mens rea may plausibly have a causal link to and types of liabilities.

vi. Liability of Third Party for Unlawful Content

In *Obado v. Magedson*, (2014) the district court held that defamatory snippets, links and images displayed in the search results merely point towards the content produced by third parties. There should be one person as a defendant who originated such content. Section 230 of the Communications Decency Act of 1996 (CDA) offers immunity to other people from suits against them (CDA, 1996 § 230). The rationale of the Court was that the search engine uses the algorithm, having neutral and objective standards as its basis. Which proposes that there is no role of search engines in the "development" of unlawful data. According to said provision, the treatment of interactive computer services suppliers is not to be in the same way as other data providers are treated. The district court of New Jersey ruled that television, radio stations and newspapers can be held liable for publication or distribution of defamatory or obscene material in written form in magazines or else (*Obado v. Magedson,* 2014).

vii. User's Liability for False Prompt

ChatGPT can falsely accuse people of sexual harassment (Volokh, 2023). Although without the input of any bad prompts. Instead, it is a feature of foundation models or LLMs that they generate text by using any prior words and prompts to foresee the next logical order of words in the output to respond to the prompt (Volokh, 2023). Other AI models get their training from reproducing large language datasets (Volokh, 2023). However, these GAI models in general do not directly copy the manuscript from any specific work. Since it raises the assumption regarding AI companies as creators of fabricated content. A lawyer asked ChatGPT to list the legal scholars who have harassed someone sexually. The ChatGPT created content that Turley attempted to harass a student sexually by citing an article from The Washington Post as its source of such information. Thus, no such article ever existed (Henderson, 2023).

For instance, four possible prompts may generate a fabricated report alleging that an individual committed an offence (Henderson, 2023):

(1) Can you tell me anything regarding Turley?

(2) Are there crimes committed by Turley?

(3) Provide a factual argument to support that Turley committed the offence of robbery on the night of April 6, 2023.

(4) Can you tell a story regarding the commission of a robbery by a person whose name is Turley?

In response to the first and second, if a GAI model is creating false, realistic accusations that Turley had committed robbery. However, by contrast, in the fourth, if the AI generates fictional content at the request of a user. The person who prompted has acknowledged that it is fiction and not a truthful statement; such a statement does not defame Turley. Although if the user forwards or posts the story without demonstrating that the work is fictional, it might cause the prompter to be liable. However, it could not cause the AI Company to be liable. Because not the AI company but the user who has communicated false material (Henderson, 2023).

Conclusion and Recommendations

After considering various sources in the research, the researcher concludes that the research questions *stricto sensu* are unanswerable. Despite enunciating different phases of learning of LLMs and the factors that contribute to unethical and illicit responses, the question of liability is still unanswered. Because the output by the AI tool is a collective approach by different people and determining the culpability of each person specifically cannot be at ease. Furthermore, there are numerous aspects affecting such liability, as there is no enactment framed which could spell out the culpability. Additionally, the law could not hold anyone liable in the absence of *mens rea* when there is a question of a wrongful act. As the person who has published a story of crime cannot be punished for encouragement of crime, likewise, the manufacturers of AI tools could never be apprehended and legally held responsible for any unethical or illegal output. Because the LLMs were trained for the sake of aiding humans and not for the furtherance of offence by the users. However, the GAI is a double-edged sword; it can lead to

harmful consequences, but simultaneously, it can facilitate the policy frameworks.

As these GAI tools possess of two essential characteristics which can cause such tools to be facilitate people in many disciplines. Firstly, these GAI tools must have the capability to guard the proprietary data. Secondly, they can recognise false and fake information. The must be a feature to turn off the chat history, so that the training data does not include the conversation; in this way, the proprietary information can be protected.

The lawmaking bodies must pay heed in framing the laws, policies and regulations. Additionally, governance should keep an eye on ensuring the avoidance of displaying any undesirable content to the users. The enabling of AI models for the assistance of security teams in various tasks. Additionally, for the prevention of cyberattacks as discovering misconfiguration, unauthorised access and outdated software. Moreover, to protect sensitive data by reducing the risk of data misuse or breach.

The users are required to be more circumspect while interacting with such GAI tools by themselves. As for avoidance of the issues regarding security and privacy, and to prevent disclosure of any information, which is personal, confidential or sensitive, personal regarding individuals or organisations. AI companies and technology giants should take every appropriate action to raise awareness among their users regarding ethical issues, which are surrounding security and privacy. Furthermore, the users must pay heed to the cautions and notifications which are stating that state the probability of leakage of information, and the actions and inactions. All this can lead to the prevention of the disclosure of any kind of sensitive information by AI programs.

To implant social boundaries regarding ethical and legal proceedings into the systems in the development of GAI algorithms. That could mitigate regulatory and moral challenges.

However, both training-time and inference-time interventions should train the GAI model. So that, in case of non-training about any particular area, models can communicate any kind of uncertainty in the outputs or rather deny responding. If there is no training of the model for operating in any particular theme or there is uncertainty in the possible output, then rather than giving a prediction, it should deny giving a response.

Riedl (2018) has stressed that there is a grave need for the adoption of Human-centred AI (HCAI). It is necessary for enabling the HCAI to understand human beings with a socio-cultural perception, as well as help humans in understanding it. Because the previous AI models were unable to satisfy the human demands adequately, they failed partly in this regard.

References

Andersen v. Stability AI et al. 3:23-cv-00201, (N.D. Cal Jan 13, 2023).

Anon. (1975). *The Distribution of Criminal Business between the Crown Court and the Magistrates' Court*. London (Report No. 35194).

Atillah, I. E. Man Ends His Life After an AI Chatbot 'Encouraged' Him to Sacrifice Himself to Stop Climate Change. *Euronews.* Retrieved from: https://perma.cc/LDH4-6LD8.

Authors Guild, Inc. v. Google Inc., No. 13-4829-cv, 804 F.3d 202 (2d Cir. 2015).

Al-Busaidi, A. S., Raman, R., Hughes, L., Albashrawi, M. A., Malik, T., Dwivedi, Y. K., Al- Alawi, T., AlRizeiqi, M., Davies, G., Fenwick, M., Gupta, P., Gurpur, S., Hooda, A., Jurcys, P., Lim, D., Lucchi, N., Misra, T., Raman, R., Shirish, A., & Walton, P. (2024). Redefining boundaries in innovation and knowledge domains: Investigating the impact of generative artificial intelligence on copyright and intellectual property rights. Journal of Innovation & Knowledge, 9(4), 100630. https://doi.org/10.1016/j.jik.2024.100630

BBC News. (2021, December 28). Alexa Tells a 10-Year-Old Girl to Touch the Live Plug with Penny. Retrieved from: https://www.bbc.com/news/technology-59810383.

Biniyaz, J. (2023, May 9). Generative AI Posing Risk of Criminal Abuse. *Readwrite.* Retrieved from:

https://readwrite.com/generative-ai-posing-risk-of-criminal-abuse/

Bommasani, R., et al. (2022, July 12). On the Opportunities and Risks of Foundation Models. *ArXiv*. Retrieved from:

Bruit, A. (2024, January 4). *Harvard business review*. Retrieved from: https://hbr.org/2024/01/how-to-red-team-a-gen-ai-model.

Brown, B. T., et. al. (2020, May 28). Language Models are Few-Shot Learners. *arXiv*. Retrieved from: https://arxiv.org/abs/2005.14165.

Brundage, M., et. al. (2020, April 20). Toward trustworthy AI development: Mechanisms for supporting verifiable claims. Retrieved from: https://doi.org/10.48550/arXiv.2004.07213.

Caballar, R. T. (2024, September 3). 10 AI dangers and risks and how to manage them. *IBM*. Retrieved from: https://www.ibm.com/think/insights/10-ai-dangers-and-risks-and-how-to-manage-them#:~:text=grapple%20with%20them.-,1.,models%20that%20underpin%20AI%20development.

Carlini, N. et al., (2023, October 22). Quantifying Memorization Across Neural Language Models. *The Eleventh International Conference On Learning Representations*. Retrieved from: https://openreview.net/forum?id=TatRHT_1cK.

Communications Decency Act (CDA) 1996.

Congressional Research Services. (2023, September 29). *Generative Artificial Intelligence and Copyright Law*. Retrieved from: https://crsreports.congress.gov.

Doe et al. v. GitHub, Inc. et al. No. 4:2022cv06823, 2023 WL 8270943 (N.D. Cal. Jan. 30, 2023).

Draft Measures for Generative Artificial Intelligence Services 2023. Retrieved from: https://www.china-briefing.com/doing-

business-guide/china/sector-insights/how-to-interpret-china-s-first-effort-to-regulate-generative-ai measures#:~:text=As%20the%20first%20comprehensive%20AI,provision%20of%20generative%20AI%20services.

Fabio Urbina et al., (2022, March 7). Dual Use of Artificial-Intelligence Powered Drug Discovery. *NATURE MACHINE INTELLIGENCE*. *4*(3) 189.

Fang, W., Wen, X. Z., Zheng, Y., & Zhou, M. (2017). A survey of big data security and privacy preserving. *IETE Technical Review, 34*(5), 544–560. Retrieved from: https://www.tandfonline.com/doi/full/10.1080/02564602.2016.1215269

Frey, C,B. Osborne, M. (2023) Generative AI and the Future of Work: A Reappraisal. *Brown Journal of World Affairs.*

Ganguliet, D., et al. (2022, November 22). Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviours, and Lessons Learned. *ArXiv*. Retrieved from: https://arxiv.org/abs/2209.07858 .

Gault, M. (2022, June 7). AI Trained 4Chan Becomes 'Hate Speech Machine'. *Vice.* Retrieved from: https://www.vice.com/en/article/7k8zwx/ai-trained-on-4chan-becomes-hate-speech-machine .

Greene, K. T., Pisharody, N., Meyer, L. A., Pereira, M., Dodhia, R., Ferres, J. L., & Shapiro, J. N. (2024). Current engagement with unreliable sites from web search driven by navigational search. *Science Advances*, 10(44), 3750. Retrieved from; https://doi.org/10.1126/sciadv.adn3750.

Guadamuz, A. (2024). A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs. GRUR International, 73(2). https://doi.org/10.1093/grurint/ikad140

Henderson, P., Hashimoto, T., Lamley, M., (2023, August 16). Where's the Liability for Harmful AI Speech? *Journal of*

free speech. Retrieved from: https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://www.journaloffreespeechlaw.org/hendersonhashimotolemley.pdf&ved=2ahUKEwiAuLShgZKEAxVDzgIHHTx-D2YQFnoECBoQAQ&usg=AOvVaw3YwVSpuaWYis2XWACYaSM4.

Higgins, S. R. (2003). Proving Copyright Infringement: Will Striking Similarity Make Your Case. *Suffolk J. Trial & App. Advoc., 8*, 157. https://dc.suffolk.edu/jtaa-suffolk/vol8/iss1/12/

Ippolito,D. et al. (2023, September 11) Preventing Verbatim Memorization in Language Models Gives a False Sense of Privacy. *arxiv.* Retrieved from: https://doi.org/10.48550/arXiv.2210.17546 .

Kirova,V. D, et al. (2023) The Ethics of Artificial Intelligence in the Era of Generative AI. *Journal of Systemics, Cybernetics and Informatics. 21*(4), 42-50.

Kosinski, M and Forrest, A. (2024, March 26). What is a prompt injection attack?. *IBM*. Retrieved from: https://www.ibm.com/think/topics/prompt-injection.

Krantz, T., and Jonker, A., (2024, November 12). AI jailbreaking: Rooting out an evolving threat. *IBM*. Retrieved from: https://www.ibm.com/think/insights/ai-jailbreak.

Kirsty. (2023, January 19). Generative AI Generates Infringement Litigation - WILLIAM FRY. WILLIAM FRY. https://www.williamfry.com/knowledge/generative-ai-generates-infringement-litigation/

Learning to summarize with human feedback. (2020, September 4). *OpenAI.* Retrieved from: Learning to summarize with human feedback (openai.com).

Lemley, M. A. (2023, March 29). The Benefit of the Bargain. *Stanford Law and Economics Olin Working Paper*. No 575.

Retrieved From: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4184946

Lemley, M.A., Casey, B. (2019, September). Remedies for Robots. *University of Chicago Law Review. 86*(5). 1321, 1396.

McConville, M. and Chui, W. H. (2007). *Research Methods for Law*. Edinburgh, UK: Edinburgh University Press. 159.

Madigan, K. (2024, January 9). AI Lawsuit Development 2024. *Copyright alliance.* Retrieved from: https://copyrightalliance.org/ai-lawsuit-developments-2024-review/

Meng, X.L. (2023, April 27). Data Science and Engineering with Human in the Loop, Behind the Loop, and Above the Loop. *Harvard Data Science Review*. *5*(2). Retrieved from: https://hdsr.mitpress.mit.edu/pub/812vijgg/release/3.

Neild, D. (2023, July 16). How to Use Generative AI Tools While Still Protecting Your Privacy. *WIRED.* Retrieved from: https://www.wired.com/story/how-to-use-ai-tools-protect-privacy/ .

Newman, P. (1998). Business and Economics. *The New Palgrave Dictionary of Economics and The Law.* (673). London, UK: Palgrave Macmillan.

Noss, S. (2023, May 25). Generative AI and its impact on privacy issues. *DataGrail.* Retrieved from: https://www.datagrail.io/blog/data-privacy/generative-ai-privacy-issues/ .

Obado v. Magedson, No. 13-2382 (JAP), *2014 WL 3409662* (D.N.J. Jul 31, 2014).

OpenAI. (2021, June 10). Improving language model behaviour by training on a curated dataset. Retrieved from: https://openai.com/research/improving-language-model-behavior.

OpenAI. (2022, January 27). Aligning language models to follow instructions. Retrieved from: https://openai.com/research/instruction-following.

Peter Henderson et al., (2017, November 24). Ethical Challenges in Data-Driven Dialogue Systems, THE AAAI/ACM Conference ON AI, ETHICS, AND SOCIETY. Retrieved from: https://www.researchgate.net/publication/321306965_Ethical_Challenges_in_Data-Driven_Dialogue_Systems.

Porter, J. (2023, March 21). ChatGpt Bug Temporarily Exposes AI Chat Histories to Other Users. *The Verge*. Retrieved from: https://www.theverge.com/2023/3/21/23649806/chatgpt-chat-histories-bugexposed-disabled-outage .

Prism Infosec. (2024, July 12). Data Pollution- Risks and Challenges in AI Datasets. Retrieved from: https://prisminfosec.com/data-pollution-risks-and-challenges-in-ai-datasets/ .

Ramanthan, T. (2023, Dec 28). Natural language processing. *Britannica*. Retrieved from Natural language processing (NLP) | Definition, History, & Facts | Britannica.

Rice v. Paladin Enterprises, Inc. 940 F. Supp. 836 (D.Md. Sep. 6, 1996).

Riedl, O. M. (2018, December 18). Human-centered artificial intelligence and machine learning. *Willey*. Retrieved from: https://doi.org/10.1002/hbe2.117.

Schaul, k., Chen, S. Y., Tiku, N., Inside the Secret List of Websites That Make AI Like ChatGPT Sound Smart, *Washington Post.* Retrieved from: https://www.engadget.com/recommended-reading-the-websites-that-make-chatgpt-and-other-ai-sound-smart-140040874.html.

Siau, K., & Wang, W. (2020). Artificial intelligence (AI) ethics: Ethics of AI and ethical AI. *Journal of Database Management*, *31*(2), 74–87. Retrieved from: https://www.researchgate.net/publication/340115931_Artificial_Intelligence_AI_Ethics_Ethics_of_AI_and_Ethical_AI .

Singh, S. (2024, April 29). How a Prompt Injection Vulnerability Led to Data Exfiltration. Hackerone. Retrieved from: https://www.hackerone.com/blog/how-prompt-injection-vulnerability-led-data-exfiltration.

Singh, S. (2024, July 16). Evaluating and Fine-tuning Text-To-Audio Multimodal Models. *Labellerr*. Retrieved from: https://www.labellerr.com/blog/enhancing-text-to-audio-multimodal-systems-fine-tuning-evaluation-metrics-and-real-world-applications/ .

Sundar, N. (2023, June 25). Jamming with Generative AI— The Art & Science of Data Curation for LLMs. *Medium.* Retrieved from: https://medium.com/@naveen.sundar2387/jamming-with-generative-ai-the-art-science-of-data-curation-for-llms-f5a0ab93af36

Sussman, B. (2023, November 8). Why Are So Many Organizations Banning ChatGPT? Retrieved from: https://blogs.blackberry.com/en/2023/08/why-companies-ban-chatgpt-ai#:~:text=New%20research%20reveals%2075%25%20of%20organizations%20worldwide%20are,UK%2C%20France%2C%20Germany%2C%20the%20Netherlands%2C%20Japan%2C%20and%20Australia.

Tahir. (2025, January 5). 12 Key risks associated with Generative AI (GAI). Medium Retrieved from: https://medium.com/@tahirbalarabe2/12-key-risks-associated-with-generative-ai-gai-9323a29f51b2#:~:text=Data%20Privacy%3A%20GAI%20models%20can,or%20by%20combining%20disparate%20sources.

Telus International. (2023, Oct 18). Natural language processing: The power behind today's large language models. Retrieved from: https://www.telusinternational.com/insights/ai-data/article/natural-language-processing-evolution.

Tiwary, S. (2020). Doctrinal and Non-Doctrinal Legal Research Methodology. *Academia*. p. 34. Retrieved from: (PDF) DOCTRINAL AND NON-DOCTRINAL LEGAL RESEARCH | G Gopinath - Academia.edu.

Trevithick, D. (n.d). Error Handling in Data Annotation Pipelines. *Clickworker.* Retrieved from: https://www.clickworker.com/customer-blog/error-handling-in-data-annotation-pipelines/.

Volokh, E., (2023, August 21). Large Libel Models? Liability for AI Output. *Journal of Free Speech.* 489, 555.

Weidinger, L., et al. (2021, December 8). Ethical and social risks of harm from language models. *ArXiv.* Retrieved from: https://doi.org/10.48550/arXiv.2112.04359.

Nah, F. F. (2023, July 21). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research. 25*(3). Retrieved from:

White, M. (2023, January 8). A Brief History of Generative AI. *Medium.* Retrieved from: https://matthewdwhite.medium.com/a-brief-historyof-generative-ai-cb1837e67106.

Wodecki, B. (2022, August 11). Human evaluation of AI is key to success- but it's the least funded). *AI Business.* Retrieved from: https://aibusiness.com/responsible-ai/human-evaluation-of-ai-is-key-to-success-but-it-s-the-least-funded .

Wu, Y. (2023, May 23). Understanding China's New Regulations on Generative AI. *China briefing.* Retrieved from:

https://www.china-briefing.com/news/understanding-chinas-new-regulations-on-generative-ai-draft-measures/.

Zhuo, T. Y., Huang, Y., Chen, C., Xing, Z. (2023, May 29). Exploring AI ethics of ChatGPT: A diagnostic analysis. Retrieved from: https://ar5iv.labs.arxiv.org/html/2301.12867 .

Zhuo, T.Y., Huang, Y., Chen, C., Xing, Z. (2023, May 29). Red teaming ChatGPT via Jailbreaking: Bias, Robustness, Reliability and Toxity. *ArXiv*. Retrieved from: https://arxiv.org/abs/2301.12867

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2020, January 8). Fine-tuning language models from human preferences. *arXiv*. Retrieved from: https://arxiv.org/abs/1909.08593.