Cambridge Forum on AI: Law and Governance (2025), **1**, e9, 1–14
doi:10.1017/cfl.2024.9

RESEARCH ARTICLE

Generative AI and criminal law

Beatrice Panattoni

Department of Law, University of Verona, Italy
Email: beatrice.panattoni@univr.it

(Received 12 August 2024; revised 10 November 2024; accepted 11 November 2024)

Abstract

Several criminal offenses can originate from or culminate with the creation of content. Sexual abuse can be committed by producing intimate materials without the subject's consent, while incitement to violence or self-harm can begin with a conversation. When the task of generating content is entrusted to artificial intelligence (AI), it becomes necessary to explore the risks of this technology. AI changes criminal affordances because it creates new kinds of harmful content, it amplifies the range of recipients, and it can exploit cognitive vulnerabilities to manipulate user behavior. Given this evolving landscape, the question is whether policies aimed at fighting Generative AI-related harms should include criminal law. The bulk of criminal law scholarship to date would not criminalize AI harms on the theory that AI lacks moral agency. Even so, the field of AI might need criminal law, precisely because it entails a moral responsibility. When a serious harm occurs, responsibility needs to be distributed considering the guilt of the agents involved, and, if it is lacking, it needs to fall back because of their innocence. Thus, legal systems need to start exploring whether and how guilt can be preserved when the *actus reus* is completely or partially delegated to Generative AI.

Keywords: Generative AI; crime; criminal law; content-related offences; culpability

1. Introduction

The recent advancements in the field of artificial intelligence (AI) opened the opportunity to delegate fractions of activities to artificial systems. Specifically, with large language models (LLMs), the generation of human language can be executed by AI systems designed for this purpose. The outcomes of these systems range from written language to images and videos. The effectiveness of the contents artificially generated by LLMs is related to their abilities to process natural language, synthesize and reason about facts, as well as problem-solve within abstract topics. AI systems that use LLMs "understand" the (contextual, however probabilistic) meaning of natural language (Surden, 2024).

Equipped with this new set of abilities, the artificial generation of content entails delegating significant fragments of socially relevant behaviors to AI systems. Their autonomous execution of the tasks assigned implies that these systems can actively intervene in the world. They manifest a form of "behavior." However, it is a behavior without intelligence, without understanding or cognitive abilities (Floridi, 2023). As "actors", Generative AI autonomous functioning can generate relevant harms, raising the question of who will answer for algorithmic misbehavior. However, given the shared-agency context where Generative AI systems operate, which include multiple components and actors, Generative AI-related harms are not simply the result of a behavior, but rather of an interaction. To mention a few: interactions with designers, with users, with the social environment where the AI is supposed to operate.

When stating that the harm is the result of an interaction, we are not simply looking at the sum of multiple behaviors, but on the qualities of the relations that keep the behaviors interacting with each other. These interactions can be malicious if they are purposely designed or actualized to cause intentional harm. Instead, the interactions can be dangerous if they are designed or actualized lawfully, but they remain capable of causing unintentional harm.

Both malicious *and* dangerous interactions, and the harms they cause, might be relevant under criminal law. This paper suggests a possible categorization of Generative AI-related harms based on the distinction between intentional and unintentional harm. It then outlines the consequences they might raise in responsibility attribution. The paper's focus is primarily on the issues related to unintentional harm, since they create the most challenging scenario when it comes to criminal law. Can we find guilt when harms are the result of a lawful interaction gone bad? If so, can or should criminal law have a role in the set of tech policies related to Generative AI?

2. Generative AI and criminal law

In the case of Generative AI-related harms, the areas where criminal law might play a role can be categorized based on whether a human agent acted with criminal intent or not. The first area of interest includes all the interactions purposely directed to committing a crime, which this paper will label as "malicious interactions." The second area of interest encompasses interactions that occur within a lawful context but have the potential to cause significant harm, which this paper will label as "dangerous interactions."

2.1 Generative AI-related malicious interactions

Considering that criminal phenomena involving Generative AI systems have started to appear only in recent times, rather than verifying which specific criminal offences might apply in this context, the paper will highlight, from a criminological perspective, the categories of interactions and the areas of crime that might find, in theory, a connection with Generative AI. This approach will allow us to maintain a cross-border perspective, without needing to base the analysis on a specific criminal justice system (which will remain the next needed step in the future developments of this field of study).[1] Without presuming to be exhaustive, when there is a human agent who participates with criminal intent, the following malicious interactions with Generative AI can be outlined:

(a) Support to realize content-related criminal offences, by intentionally producing or disseminating artificially generated illegal contents.
(b) Support to realize other crimes, by aiding or facilitating criminal activities.

Generative AI can be used or designed to create harmful contents, whose creation itself, or whose dissemination might be criminalized in most criminal justice systems. The range of contents that can be created is significantly wide. It includes terroristic contents, image-based sexual abuse, child pornography, and harmful speech, which can be relevant under criminal law, for instance, in case of some categories of discrimination-based discourse, incitement to crime, defamation (Henderson et al., 2023; Ousidhoum et al., 2021; Weidinger et al., 2022). Taking as a case-study a well-known chat-based LLM, GPT-4, OpenAI published a report on the safety challenges presented by the model,

[1]Criminal law scholars started recently to research whether existing criminal offences might apply to AI-related harmful phenomena. The results of such analysis will logically be different depending on which national criminal statute is considered. Among the most recent and overall comparative research, see Miró-Llinares et al. (2024). Criminalization of AI-related offence [Monograph]. *Revue Internationale de Droit Pénal*, 1, 1–445.

where they described the results of their adversarial testing processes, called "red teaming."[2] The list of harmful contents that might be generated by LLMs include: advice or encouragement for self-harm behaviors; graphic material such as erotic or violent content;[3] harassing, demeaning, hateful content; apology or incitement to terroristic activities (Sivaram et al., 2023; Weidinger et al., 2022).

Furthermore, fake artificially generated content is raising significant concerns in the current debate (Chesney & Citron, 2019), even though it can be based on a different sub-category of Generative AI, called generative adversarial networks, invented by Google researcher Ian Goodfellow (Goodfellow et al., 2014). In this area, a primary threat regards the so-called deep nudes, namely the use of deep fakes to perpetrate crimes of non-consensual pornography (Franks & Waldman, 2019). Although deep nudes are undoubtedly a criminal law concern, the broader category of deep fakes outlines a more problematic area from a criminal law standpoint, since it is questionable whether and to what extent behaviors involving disinformation campaigns and propaganda should be criminalized, since such criminal policies can compromise free speech rights.

In the area of artificial content-related offences, distinguishing between private and public harms is pivotal in assessing the criminal relevance of Generative AI content. While private harms – such as individual defamation or non-consensual dissemination of intimate content – raise significant ethical and legal concerns, which grant a sounder legitimation for criminal intervention, public harm must typically involve a higher threshold of risk to society at large. For crimes tied to terrorism or organized crime, the focus is on clear and severe threats, whereas offenses related to freedom of expression or democratic integrity, like electoral interference or incitement to violence, demand a more nuanced approach. Criminalization in these cases is generally limited to actions that demonstrate a significant probability of translating danger into tangible damage, emphasizing the need for scale, coordination and clear harm to the public (Guerini, 2020). As Generative AI continues to evolve, establishing robust frameworks to assess both the technological misuse and its societal ramifications will be integral to balancing innovation with legal accountability.

Generative AI systems, however, enable these crimes to be committed in ways that go beyond the mere creation of false or violent content. More harmful phenomena might be realized, for instance, through manipulative schemes that amplify the severity of the harm. LLMs might become a part of online manipulation schemes (Susser et al., 2019), which can be realized not only through the generation of fake contents, but also through several operations (e.g., spreading of polarized contents). Generative AI systems can be used for disinformation campaigns at scale (Chesney & Citron, 2019; Kreps et al., 2022), and, more broadly, for covert or deceptive efforts to influence the opinions of a target audience, defined as "influence operations" (Goldstein et al., 2023), able to cause manipulative effects, such as cause alteration in the electoral and democratic process, or exploit cognitive vulnerabilities of users (DiPaola & Calo, 2024; Neuwirth, 2022).

This leads to the second area of intersection between criminal law and malicious interactions with Generative AI, which involves its aiding role as a supporter or facilitator of crime. Considering the GPT-4 system card previously mentioned, two different forms of aid might surface: (a) aid in the preparatory phase of a crime, through the gathering of contents useful for planning a crime (e.g., information for developing unconventional weapons) or through the gathering of instructions for finding illegal content; (b) aid in the execution phase of a crime (e.g., LLMs can be used to find alternative, purchasable chemicals used in criminal activities). A relevant area of interest might be the area of cybercrimes. LLMs can be used to realize or improve social engineering attacks (e.g., drafting phishing emails) or hacking attacks (e.g., explaining systems or code's vulnerabilities).

[2]Red teaming is defined as "a structured effort to find flaws and vulnerabilities in a plan, organization, or technical system, often performed by dedicated 'red teams' that seek to adopt an attacker's mindset and methods." See OpenAI, "GPT-4 System Card," 23 March 2023, p. 5, available at https://cdn.openai.com/papers/gpt-4-system-card.pdf.

[3]As happened in the case of Dungeons & Dragons, released by Utah-based gaming company Latitude, where the inclusion of the GPT-3 technology led to inappropriate game plotlines (disturbing stories, including sex scenes involving children).

The same threats are outlined by Europol in a recent report (Europol, 2023), which states how LLMs can be used to learn about crime, ranging from how to break into a home, to cybercrime and child sexual abuse.[4] A specific concern remains the fight against terroristic attacks, since the model can be used to generally gather more information that may facilitate terrorist activities, such as terrorism financing or anonymous file sharing.

The malicious interactions with Generative AI will determine a direct criminal responsibility of the human agents who committed purposely the crime, either by designing the system to realize one of the activities mentioned or by using it to the same end. Therefore, in terms of responsibility attribution, these cases are less challenging. Although some gray areas remain – especially regarding theories of criminalization (Hörnle, 2015) – legal systems may vary in qualifying malicious interactions either as new, potentially aggravated, ways of committing existing criminal offenses or as new criminal behaviors that require criminalization (Miró-Llinares et al., 2024).

Beyond the issue of distinguishing between harm to individual victims and harm to public interests, another issue concerns the distinction between the behaviors of creation versus distribution of content. Defining the threshold for criminally relevant behavior is essential in determining when the creation or dissemination of content crosses into punishable territory. Notably, in the European Union (EU), only the creation of child pornography is criminalized in itself (even in the case of artificial or virtual content),[5] setting a precedent where mere content creation is rarely deemed illicit unless it pertains to certain severe offenses.[6] Most other content-related offenses, such as terroristic propaganda or non-consensual pornography, require dissemination to meet the threshold of criminal liability, as these materials are not inherently illicit outside of their distribution and impact on the public sphere.

In dealing with Generative AI, criminal policies might consider, on the one hand, to extend criminalization. For instance, there could be arguments in favor of the responsibility of those actors who purposefully design and make available online Generative AI platforms that enable users to realize criminal activities (e.g., platforms to create non-consensual deep nudes). In these cases, where the creation of such platforms is not criminalized itself, the crime would not be directly realized by the providers, but they might be held responsible as accomplice in the crime committed by their users.[7]

On the other hand, to avoid overcriminalization, Courts might interpret existing criminal offenses in an evolutionary manner, so as to include within their scope new forms of conduct involving the abuse or misuse of Generative AI systems. For instance, under certain circumstances, the act of creating terrorist propaganda using Generative AI could be interpreted as a criminal offense if it aligns closely with national laws defining terrorism-related crimes.[8] However, this requires a case-by-case

[4]While all the information provided is freely available online, the possibility to use the model to provide specific steps by asking contextual questions means it is much easier for malicious actors to better understand and then carry out various types of crime.

[5]The Italian criminal law, for instance, punishes the crime of child pornography also when the contents do not picture a real minor but when they are graphically realized (crime of "virtual child pornography," article 600 quarter.1 Italian criminal code).

[6]Considering the case of depp fake, Miró-Llinares, Duvac, Toader & Santisteban Galarza claim that: "what is proscribed is not the creation of deep fakes, but their making available to third parties. (…) such an offence of 'deep fakes creation' would hardly have an independent scope of application nor be of practical relevance since production only becomes known upon publication." See Miró-Llinares et al. (2024). Criminalization of AI-related offence [Monograph]. *Revue Internationale de Droit Pénal*, 1, 70.

[7]The responsibility of actors who purposefully design and make available online Generative AI platforms that enable users to realize criminal activities relates to the broader topic of platform liability regimes, which is very different if we consider the EU or the US legal system. Among the comparative studies on the topic, see Giancarlo Frosio (ed.), *Oxford Handbook of Online Intermediary Liability* (Oxford: Frosio, 2020).

[8]Considering as a reference the Italian criminal law, the crime of "Associations with purposes of terrorism, including international terrorism, or subversion of the democratic order" might become relevant in these cases (article 270 bis Italian criminal code).

assessment grounded in the specific offences' wording provided by each jurisdiction, but it constitutes a more balanced approach, rather than a blanket criminalization of AI-assisted content creation. Similarly, while the use of Generative AI to gather intelligence or to support criminal activities does not constitute a standalone offense, it could be relevant as an aggravating factor, amplifying existing crime's impact and warranting more severe penalties when technology is misused in such contexts.

In addition to defining criminal thresholds, there is indeed a growing need to address the deterrence of misuse in Generative AI. Protecting the confidentiality, integrity and availability of AI is essential (Flor, 2024), but equally crucial is safeguarding public trust in the technology as it integrates more deeply into society. Generative AI, with its capacity for autonomous content creation, could become a "social actor" that influences public perception and societal norms. Therefore, misuse in the form of large-scale misinformation campaigns, recruitment by terroristic organizations, or criminal campaigns aimed at democratic disruption warrants close scrutiny. The threshold here shifts to focus not only on the quantity of potential harm but also on its nature, which might contribute to the emerging of new interests, such as the public interest to protect the trustworthiness of the interactions with Generative AI.

2.2 Generative AI-related dangerous interactions

The picture becomes very different when considering dangerous interactions and the positions of the different actors involved in the supply chain of a Generative AI system that is lawfully put online or on the market. The umbrella term "operator," keeping as a reference the definitions provided by the EU AI Act (AIA), includes different categories of actors. Namely, operators are considered as follows: "providers," those that develop, place on the market and put into service an AI system; "deployers," those that use an AI system under their authority in a professional activity; "distributors," all the other remaining actors that participate in the supply chain of an AI system and make it available on the market.[9]

From the perspective of these actors, Generative AI-related harms are unintentional, and the interactions are no longer considered malicious, but dangerous. In other words, they present the probability of the occurrence of severe harms.[10] Unintentional harms awaken the preventive function of criminal law. The legal response to unintentional harm does not end in punishing the agent that caused the harm, but it widens to providing the means to avoid the harm, on whose violation the responsibility attribution is based. However, these are unstable grounds for criminal law. The need to reassure against the insecurity created by the risks of harm might create a rise in punitiveness and an expansion of the penal state, which clashes with the legitimacy of criminal law as a modern, liberal institution (Ashworth, 2013; Carvalho, 2017; Donini & Pavarini, 2011). Therefore, we need to stay on the fine line between avoiding broad new criminalization to prevent unintentional harms and filling gaps in the attribution of responsibility for those harms.

In their hybrid forms, theories of punishment combine retributivism and prevention. However, in the field of unintentional harm, the risk is to lean too much on prevention, where the goal shifts from avoiding the commission of crime to assuring social order. Security politics have problematically promoted in the last decades the framing of crime as a social risk and criminal law as a means of risk management (Ashworth & Zedner, 2012; Ashworth, et. al., 2013; Husak, 2007). This approach, while influential, encounters significant challenges, particularly when extended into the realm of technology. The modern application of this model of punishment threats overextending its scope, as it must contend with the compounded uncertainties of complex social systems and the rapid evolution of technology as a social reality.

[9] Article 3 nn. 3, 4, 7, 8, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).

[10] Definition of risk provided by the AIA. According to Art. 3(1a), "risk" "means the combination of the probability of an occurrence of harm and the severity of that harm."

Sociological studies already outlined the limits of the paradigm of "control" when it comes to technology. The responsibility gaps created by technology developments are the result of an old issue, defined as the "dilemma of control," that society faces when attempting to predict the social consequences of a technology (Collingridge, 1980). The control of technology has proven to be a very difficult task because of its unanticipated social consequences. Narrowing our focus from the broadly harmful social consequences that a technology might have on society to the consequences that can be qualified as specific harms to a person or to a public interest, the dilemma of control might be addressed using a risk-based approach.

Therefore, assessing the necessity of criminal policies targeting AI-related harms depends on the existence of an *objective* and collective interest to follow this path (Vassalli, 1942), which might be, as highlighted, the trustworthiness of AI as a human value. Such policies should require a shift beyond general or negative prevention toward positive prevention, focusing on restoration rather than mere deterrence. Here, culpability emerges not simply as a matter of blame (Kelly, 2018) but as a normative construct which can bring to restorative measures, suggesting a framework for criminal law that emphasizes social co-responsibility in the age of AI.

To establish whether such involvement of criminal policies might find a legitimate space, the objective nature, scope and severeness of AI-related harms must be firstly defined, especially when there are unintentional harms. Although a current analysis must contend with the phenomenon's young age, that is, the fact that Generative AI systems have only recently been released into society, a framework of dangerous interactions can be suggested.

The paper approximately summarizes dangerous interactions into three main groups:

(a) Risks of harms related to users' misuse.
(b) Risks of harms related to artificial emergence.
(c) Risks of harms related to overreliance.

The first group corresponds to the misuses listed in the previous paragraph, which become risks from the Generative AI operators' perspective, whereas the last two groups create the more complex scenarios, where Generative AI-related harms cannot be traced back to a human agent who interacted with criminal intent. Society might face, on the one hand, harms directly caused by the Generative AI system without any user behind it (harmful AI content), and, on the other hand, harms that the Generative AI system contributed to causing with the participation of a human agent who acted without criminal intent (negligent AI content).

Harmful AI contents might be generated by emergent interactions, where "emergence" states the ability of AI to perform unpredictable behaviors (Calo, 2015, p. 538). For instance, harmful content might be the result of the hallucinations of LLMs (Gehman, 2020). There have already been cases where the LLMs fabricated offensive false claims against a person, falsely accusing him of a crime (Volokh, 2023), or cases where LLMs "degenerate into offensive language from seemingly innocuous prompts" (Weidinger et al., 2022, p. 216). Furthermore, in cases of Generative AI implemented beyond the task of content-generation, new areas of risks might rise when the model is set to "accomplish goals which may not have been concretely specified or appeared in the training" (GPT-4 System Card, 2023, p. 15).

Closer to the present might be cases of overreliance of users on negligent AI content, where the harm is the result of a dangerous interaction between the Generative AI and an innocent or negligent user. Since Generative AI systems might provide part of the knowledge used in decision-making processes, if it provides manipulative information, it might encourage people to commit self-harm or it might violate their self-determination (Atillah, 2023, March 31; Perrigo, 2023, February 17), whereas if it provides neutral, but false, misleading, or poor-quality information to a user who acts consequently based on that information, the result of the interaction might cause harm to the user

herself or to others (Bambauer, 2023).[11] For instance, Generative AI is used to design proteins and inform patients in the healthcare sector (AlQuraishi, 2021; Chakraborty et al., 2022), which might cause injuries due to the health professionals' overreliance on the AI outputs.

Each dangerous interaction might raise different modes of responsibility attribution. In scenario (a), where there is an interaction with a malicious user, the possibility to resort to doctrines of complicity (Stewart, 2015) might become relevant, holding the operators of the Generative AI responsible for participating in the users' crimes, whether there is an omissive contribution (i.e., failure to avoid the user's crime) or an active contribution (e.g., development or putting into service of the Generative AI through which the user committed the crime).[12]

In scenarios (b) and (c), where malicious users' interactions lack, the operator might be directly responsible for the harm realized, whether there is an omissive interaction (i.e., failure to avoid the harm) or an active interaction (e.g., development or putting into service of the Generative AI which caused the harm). However, all these cases reveal several shortcomings, as the next sections explore.

2.2.1 Omissive interactions

In case of omissive contributions or interactions, the operators cannot be held responsible unless the legal systems provide for duties to act which hold them responsible for preventing harms from happening (either in the case of user misuses or emergence or overreliance). It is questionable whether and which operators might have a duty to protect or a duty to control over a source of danger because of their activity.[13] Society cannot expect the operators to act to avoid every harm that might be related to the Generative AI system that they develop, place online or deploy. This scenario would hardly comply with a socially expected level of care or diligence. Moreover, in most cases, considering the context of autonomous and active interactions from which the harms originate, the operators might not have foreseen the harm, and they might not have the material possibility to intervene to avoid the harm from occurring. A further argument for distancing from this hypothesis is that omission liability is controversial in many criminal justice systems, given the moral distinction between acts and omissions and the downsides that the commission-by-omission model presents in complying with a strict legal certainty standard (Ambos, 2020).

To follow this path, policies would need to distinguish the positions of the different operators involved (providers, deployers, distributors) and their normatively expected capability to act to avoid AI Generative-related harms. However, to date, neither EU law nor US law regulates positive obligations that might be associated with a duty to act of the Generative AI operators.

Considering the EU context, the AIA qualifies Generative AI within the broader category of general-purpose AI (GPAI),[14] which is regulated by a specific chapter of the Act (Chapter V). According to Article 53, the legal obligations for GPAI providers mostly entail ensuring detailed technical and informational documentation. In case of GPAI that pose a "systemic risk," namely if the GPAI has high-impact capabilities (automatically presumed if the model's training involves more than 10^25 floating-point operations), further obligations must be carried out by the providers. According to article 55 AIA, they must conduct model evaluations, including adversarial testing (red teaming),

[11] The author suggests considering the following hypotheticals: "(A) AI tells an adult that it is safe to eat a mushroom that is actually toxic. (B) AI tells a child that it is safe to eat a mushroom that is actually toxic. (C) AI tells an adult how to alter a drug therapy to address or avoid side effects, resulting in harm from the primary disease."

[12] Outside of the intentional participation in the users' crime, the possibility to lower the subjective requirement to reckless behaviors remains. However, recklessness as a standard for complicity is controversial throughout criminal justice systems. For this reason, negligent as a standard for complicity in others' intentional crime should also be excluded.

[13] In this scenario, the more likely to be relevant are duties based on responsibility for a source of danger. See Keiler and Roef (2019), *Comparative concepts of criminal law*. Intersentia, 148.

[14] According to recital n. 99 of the AIA: "Large generative AI models are a typical example for a general-purpose AI model, given that they allow for flexible generation of content, such as in the form of text, audio, images or video, that can readily accommodate a wide range of distinctive tasks."

assess and mitigate systemic risks, document and report incidents to the AI Office, and ensure adequate cybersecurity protection. Generative AI can also fall within the category of "certain AI system," whose providers are subject to transparency obligations under article 50 of the AIA. However, neither of these obligations entails a duty to act.[15]

The scenario might have been different if the set of obligations required for high-risk AI systems would have been extended to GPAI. Here, an interesting concept is "reasonably foreseeable misuse," namely "the use of an AI system in a way that is not in accordance with its intended purpose, but which may result from reasonably foreseeable human behavior or interaction with other systems, including other AI systems" (Article 3 n. 13 AIA). The risk of reasonably foreseeable misuse is included in the set of risks that providers of high-risk AI systems must address through the implementation of a risk management system (Article 9 AIA). However, these obligations do not apply to GPAI (except, possibly, when they are part of the AI value chain and can be integrated as components within downstream high-risk AI systems).

A different option might be resorting to another EU regulation, the Digital Service Act (DSA), which regulates intermediaries' liability. Since the regulation imposes a set of obligations to ensure that intermediaries remove illegal content from their services, the normative behaviors expected of Generative AI providers could be expanded to include those obligations as well. In the context of Generative AI, content moderation obligations and notice-and-take-down procedures might have significant benefits. Nevertheless, it is not clear yet whether Generative AI might fall within the scope of the DSA, depending on whether Generative AI platforms might be qualified as "hosting providers" or "search engines" (Lemoine & Vermeulen, 2023).

In the US, Generative AI operators' liability will have to deal with the debate on Section 230 of the Communications Act and the immunity regime that it provides. Generative AI-related harms might become a new problematic chapter in the extensive use of the defense provided by Section 230 (Bambauer, 2023; Henderson et al., 2023).

An additional drawback pertains to causation. Establishing a causal link between an operator's conduct and the harm caused by Generative AI-driven systems presents significant challenges, even when a duty to act and legal obligations can be identified (Beck & Gerndt, 2023). The limitations become more evident when using the *conditio-sine-qua-non* formula and the doctrine of objective attribution. However, beyond the adoption of one model of causation over another, all theories encounter the difficulty of identifying a single person who can be held responsible. The many-hands-problem (Nissenbaum, 1996; Santoni de Sio & Mecacci, 2021; Thompson, 1980), wherein responsibility is distributed among numerous agents, combined with the "black box" nature of AI, complicates accountability. Specifically, it becomes difficult to determine if the operator fulfilled their obligations in a way that could have prevented a harmful outcome. In this context, the capacity to avoid harm is constrained not only by the opacity of the AI system but also by the complex and shared-agency environment in which decisions are made, where control over outcomes is dispersed among various actors. This fragmentation of agency limits any one individual's ability to foresee and prevent adverse effects.

2.2.2 Active interactions

Generative AI operators' responsibility for active contributions or interactions also must face the problem of causation. The harms that result from all three dangerous interactions could be difficult to trace back to a single and specific operator's behavior, because of the "failures of causation" that

[15]The assessment of how doctrine on criminal responsibility by omission can apply to Generative AI-related harm is outside the scope of this paper. For references on such analysis see Grandi, C. (2023) Positive Obligations (Garantestellung) Grounding Criminal Responsibility for not Having Avoided an Illegal Result Connected to the AI Functioning. In L. Picotti & B. Panattoni (Eds.) *Traditional Criminal Law Categories and AI: Crisis or Palingenesis?* [Monograph] *Revue Internationale de Droit Pénal*, 67–77.

characterize the AI context (Bathaee, 2018; Giannini, 2023). For instance, the action of a health operator using the information provided by a Generative AI system in deciding a treatment that might result wrong and injure the patient's health is just a fraction of a network of interactions between multiple human and artificial agents.

Considering the role of providers, this path might also lead to overcriminalization of "neutral behaviors," such as the putting into service of a Generative AI system used to commit crime (Fiorinelli, 2023). The merely material and objective contribution of making available the technical means to the realization of a user's crime cannot alone legitimize a judgment of blameworthiness.

Moreover, the relevant behaviors might be different depending on the specific Generative AI systems deployed as services or products. An approximate distinction can be drawn between open-source Generative AI models, namely models fully public which can be downloaded and used by malicious actors, and Generative AI platforms, such as chat-based Generative AI. In the first case, harms caused by users' misuses overstep the designers' ability to monitor them, whereas in the second case they stay under the operator's monitoring ability.

In cases of harms related to emergence or overreliance, a direct responsibility of the operator for an active interaction in causing the harm has the same drawbacks. Even if the operator interacted recklessly, it will be highly challenging to ascertain a causality chain between the harm and the operator's behaviors. However, despite the limits related to causation, the main challenge that characterizes all the hypotheses of responsibility for unintentional harm concerns the problem of guilt.

3. Generative AI and guilt: Walls or doors?

Generative AI-related harms stem from a system dynamic, driven by interactions among various actors. The combination of interactions challenges the paradigm of control by a single actor over the whole actualization of the system's functioning, as the many-hands-problem outlines. Since the liberal model of criminal law has individual autonomy at its core, when control is challenged, guilt is also challenged.

The field of Generative AI and criminal law, outside of the scope of malicious interactions, needs to address this challenge. The requirement of guilt can be seen in a limiting and confining function (as a "wall"), which is used to claim that where guilt cannot be found, neither criminal responsibility can be attributed. However, in a changing social context, such as the one where artificial and human agents interact with each other, guilt can also be an adapting concept (becoming a "door" instead). On the one hand, in the first perspective, guilt in the context of Generative AI confines criminal law to punishing only intentional malicious interactions (either as uses, developments or deployments of Generative AI), whereas the area of dangerous interactions is left to other branches of law, such as tort law. On the other hand, in the second perspective, if certain dangerous interactions create a too high and diffuse danger to individuals and society, the question of which kind of guilt can be found in those interactions might open new areas for criminal law.

Instead of looking at the issue from two different perspectives (one negative and one affirmative), the field of Generative AI and criminal law might be pictured as a set of closed doors.

3.1 The natural person behind the AI: Closed door

The traditional allocation of guilt in facing AI-related harms is problematic. Scholars analyzing criminal culpability in the AI field, although mostly focused on robotics (Consulich, 2022; Gless et al., 2016; Piergallini, 2020; Simmler & Markwalder, 2019), outline several collision points related to multiple factors, such as the distance between human behaviors and their results, the failures of causation and the challenged foreseeability of harm. The control of the users and of the operators who stand behind the technology is limited and compromised by the shared-agency dimension. That is if we follow the traditional perspective that devises responsibility attribution starting from the harm instead of from the interactions.

These collision points are not a prerogative of the field of AI-related harms. They are linked to a well-known phenomenon. The use of preventive criminal law to mitigate the risks of our contemporary and complex society is a feature of a longstanding debate related to whether and to what extent criminal law can be used to regulate the harms generated by the risk society (Stortoni & Foffani, 2004), which is now the algorithmic society. The analysis of the broader evolution of criminal law in the risk society provides a lens through which we can examine the inadequacy of the traditional punitive paradigm when applied to risks associated with Generative AI. Specifically, the conventional model of person-action-harm fails to adequately address the complexities inherent in situations where harms arise from systemic processes rather than actions of individual human agents. In other words, assigning direct personal responsibility to a singular human agent for harms resulting from systemic risks poses significant challenges, particularly concerning questions of culpability. Therefore, to regulate the harms related to the risks of the algorithmic society, it is not enough to consider the behavior of the single human agent alone. In these cases, guilt will likely close the door, limiting criminal law only to the punishment of intentional criminal behavior.

Alternatively, it may be necessary to analyze the roots and conditions of the interaction. By disentangling the behaviors of Generative AI operators from the interactions between AI systems and users, scholars (Consulich, 2022; Fiorinelli, 2023; Fragrasso, 2024) suggest the possibility of employing criminal law to address novel forms of omission, such as intentional or reckless omission to comply with prescribed safety measures (i.e., failure to comply with specific obligations, such as non-delegable duties, transparency obligations or failure to provide the competent authorities with the required information). However, in addition to the concerns that arise from the use of preventive criminal law, this approach might have collateral effects, disincentivizing progress and investment, as operators might restrict their activities out of fear of facing severe punishments. It would also need to coordinate with the administrative sanctions that might already be regulated in case of violation of legal obligations, for instance by EU law.[16]

Furthermore, the scope and content of the standard of due care in the field of Generative AI have not yet been established, which will depend on the set of safety international standards required by the industry (Ebers, 2022), since best practices will likely be insufficient. Resorting only to self-regulation does not appear to be a sufficient alternative, since delegating altogether the process of selecting the policies and measures for harm prevention to the private sector might rise concerns related to legal certainty. Moreover, certain measures (e.g., labeling unlawful content into pre-established categories) might be more efficient if based on an active collaboration with the public authorities.

3.2 Generative AI and criminal law: Whether to open the door

While the door is open for using criminal law to address malicious behavior, this paper has highlighted various setbacks to using criminal law for dangerous behavior. Given this scenario, the question then becomes whether the context of dangerous interactions might need criminal law.

Excluding criminal law completely from the set of policies related to dangerous interactions is an inadequate long-term strategy. Among the multiple theories on the purposes of criminal law (Hörnle, 2015), when viewed through the lens of the harm principle or the theory of the legal goods (*Rechtsgüterschutz*),[17] the paper outlined that it becomes essential to assess the nature of

[16]The AIA provides for specific administrative fines in case of violation of the legal obligations required to providers of GPAI. However, these fines can be enforced only by the European Commission. According to Article 101 of the AIA: "The Commission may impose on providers of general purpose AI models fines not exceeding 3 % of their total worldwide turnover in the preceding financial year or 15 million EUR, whichever is higher, when the Commission finds that the provider intentionally or negligently: (a) infringed the relevant provisions of this Regulation; (…)."

[17]One theory or the other depends on whether one considers references pertaining to the Anglo-Saxon world and, therefore, common law systems rather than civil law systems, even though the theory of the harm principle is gaining increasing traction even in civil law systems. See Fiandaca, G. & Francolini, G. (Eds.) (2008) *Sulla legittimazione del diritto penale. Culture*

harms and risks inherent in regulating Generative AI. These include not only the harms originating from the diffusion of harmful and illicit contents artificially generated by malicious users, but they have a broader scope, which encompass the risks related to the delegation of content-generation to artificial agents in the information society. The ability to generate information in our contemporary society can have significant effects not only on individuals but also on society. This creates a base of legitimacy for criminal law, if limited only to the most serious wrongdoings and the most serious harms (in compliance with the *ultima ratio* principle). As the paper attempted to demonstrate, engaging in information-generating activities with artificial agents in a shared-agency context can lead to the emergence of social conflicts, manifested in the form of dangerous interactions, which include (it is worth remembering) also the risk of malicious interactions. The normative regulation and protection of a sector of social life (De Francesco, 2004) highly relevant as the generation of information may ultimately necessitate criminal protection involvement in the long term. However, the condition for such involvement is that social conflicts translate into substantial wrongdoings. The question at issue then becomes whether Generative AI deployments can indeed result in substantial wrongdoing. Although this possibility exists, given that Generative AI opens the possibility for causing widespread and diffuse damage, as well as for facilitating the realization of criminal activities by simplifying their preparation or execution, it is premature to provide a definitive answer. Nevertheless, the question must be posed, and it will require finding an answer by examining the evolution of the sector and the framework of policies aimed at regulating it with an open-ended perspective, which might resort to the blurring boundaries between criminal punishment and administrative or regulatory offences (Duff et al., 2010; Fletcher, 2008).

At this stage, the paper suggests that, in the context of Generative AI and dangerous interactions, following a completely non-criminal approach, which involves civil litigation and administrative regulation, may prove insufficient in the long run. This approach would narrow the legal response to the perspective of harm mitigation, whereas criminal law aims to establish an institutional normative order aimed at protecting social values. The function of censuring and discouraging wrongdoing, which implies "should not do," entails state intervention into individual freedoms that extends beyond providing a remedy to the victim for the invasion of protected interests (Ashworth, 2013). The punitive and preventive functions, therefore, require that the distribution of responsibility in cases of serious harms relevant under criminal law follows its stricter standards, both substantively and procedurally, while granting certain protections for the defendant. In the case of AI Generative-related risks, especially in the case of emergence and overreliance, a non-criminal approach may lead to a situation where a crime occurs, and criminal law falls short because it has become challenging to find the guilty mind. However, the result of interactions remains a crime, and possibly, when behaviors become interactions within a system, a "guilty system" can be identified. According to this perspective, the attribution of responsibility must not address a gap but rather a shift (Simmler, 2023).

In addition, there are pragmatic considerations to ponder. Generative AI holds the potential to either facilitate or contribute to relevant harms, which may manifest with increased rapidity upon deployment of artificial agents, challenging investigation activities. Relying solely on retrospective measures to prosecute malicious uses may prove ineffective or unsustainable. The prospect of intervening during stages that precede interactions (e.g., design or online availability) or during interactions themselves (e.g., deployment) presents an opportunity to address social conflicts as they arise, preempting their escalation into broader and diffuse harms. Involvement during these phases could equip criminal law with valuable tools for regulating conflicts. For instance, civil courts lack the authority to order defendants to perform positive acts to avoid or to mitigate

europeo-continentale e anglo-americana a confronto. Giappichelli; Persak, N. (2007) *Criminalising Harmful Conduct. The Harm Principle, Its Limits and Continental Counterparts*. Springer.

punishment, such as implementing reactive compliance efforts. Lastly, given the diffuse nature of risks associated with Generative AI, it prompts inquiry into the adequacy of entrusting the legal response to harms solely to private citizen initiatives, without a structured intervention of the State, especially in cases where there is no victim, and the harm caused is against collective and public interests.

4. Conclusion

Given the possibility of integrating criminal law into the regulation of AI Generative-related interactions, the fundamental question is which policy would best justify criminal coercion. The answer to this question is contingent upon both the theory of attribution and the nature of the punishment that will be selected.

In cases of malicious interactions, the preferred approach should assess whether, and to what extent, existing criminal offenses might apply, with consideration of whether the increased harm caused by the facilitator role of Generative AI may warrant an enhanced punishment. Simultaneously, the evolving landscape of criminal misuse of Generative AI should be monitored to determine whether there are actual gaps in criminal legislation, starting from the harms against a specific victim rather than harms against public interests.[18]

In case of dangerous interactions, instead of dissembling the interactions of the shared-agency context, a way to stay within them is to follow one of the suggestions raised in the literature (Diamantis, 2021; Mongillo, 2023), namely, to resort to corporate criminal responsibility instead of human agents' criminal responsibility in the context of AI. Expanding or amending corporate criminal law might be a possibility to further explore,[19] in cases where a relevant harm occurs due to the intentional or reckless noncompliance with the standard of care required of the corporation which interacts in the relevant processes of proving and deploying the Generative AI system, and which benefits from its activities. The deterrence and prevention raised from a possible affliction to the economic interests of companies might help in enhancing corporate collaboration in the AI industry, without being solely driven by market pressure (Askell et al., 2019).

In these instances, legislators may need to ponder whether to open the door of criminal law or not, after assessing the precise nature of harms potentially engendered by Generative AI, and, especially, their severeness from both a qualitative and quantitative point of view. Additionally, a prerequisite for pursuing this course of action involves determining how to preserve culpability in the context of AI-related harms, firstly, by underscoring the possibility to devise frameworks for accountability within shared-agency context, where new socio-digital institutions are emerging (Beckers & Teubner, 2021), secondly, by assessing the scope and content of non-criminal regulatory measures as they define a standard of care in the field. Until then, we face a wall with many closed doors. But the doors exist. Whether and which one we want to leave closed or to start opening remains a political choice.

Funding statement. None to declare.

Competing interests. None to declare.

[18]In Italy, it is currently under discussion a bill (A.C. 2986) proposing to introduce a new criminal offence (art. 612-quater of the Criminal Code), which would punish the artificial manipulation of images of real people for the purpose of obtaining nude representations of them (deep nudes). In Spain, a bill proposes to introduce an offence which punishes whoever recreates by means of automated systems, software, algorithms or artificial intelligence the image or the voice of a person, without authorization and with the intent to harm honor, reputation, dignity or self-esteem. See Miró-Llinares et al. (2024). Criminalisation of AI-related offence [Monograph]. *Revue Internationale de Droit Pénal*, 1, 70.

[19]This hypothesis will have a different outlook depending on the civil law or common law legal systems where it applies. In common law jurisdictions, it will engage with the doctrine of vicarious liability, while in civil law contexts, it will focus on organizational culpability.

References

AlQuraishi, M. (2021). Machine learning in protein structure prediction. *Current Opinion in Chemical Biology*, *65*, 1–8.

Ambos, K. (2020). Omission. In K. Ambos, A. Duff, J. Roberts, T. Weigend, & A. Heinze (Eds.), *Core concepts in criminal law and criminal justice* (pp. 17–53). Cambridge University Press.

Ashworth, A. (2013). *Positive obligations in criminal law*. Hart Publishing.

Ashworth, A., & Zedner, L. (2012). Prevention and criminalization. *New Criminal Law Review*, *15*(4), 542–571.

Ashworth, A., Zedner, L., & Tomlin, P. (Eds.). (2013). *Prevention and the limits of the criminal law*. Oxford University Press.

Askell, A., Brundage, M, and Hadfield, G K. (2019). The Role of Cooperation in Responsible AI Development. https://api.semanticscholar.org/CorpusID:195874137

Atillah, E. (2023, March 31). Man Ends His Life After an AI Chatbot 'Encouraged' Him to Sacrifice Himself to Stop Climate Change. *Euronews*. https://perma.cc/LDH4-6LD8

Bambauer, J. (2023). Negligent AI speech: Some thoughts about duty. *Journal of Free Speech Law*, *3*, 344–362.

Bathaee, Y. (2018). The artificial intelligence black box and the failure of intent and causation. *Harvard Journal of Law & Technology*, *31*, 922–927.

Beck, S., & Gerndt, S. (2023). German report on traditional criminal law categories and AI. *Revue Internationale de Droit Pénal*, *94*(1), 195–223.

Beckers, A., & Teubner, G. (2021). *Three liability regimes for artificial intelligence. Algorithmic actants, hybrids, crowds*. Hart Publishing.

Calo, R. (2015). Robotics and the lessons of cyberlaw. *California Law Review*, *103*(3), 513–564.

Carvalho, H. (2017). *The preventive turn in criminal law*. Oxford University Press.

Chakraborty, S., Hrithik, P., Sayani, G., Saroj Kumar, P., Ankit, A., Kamred Udham, S., and Mohd Asif, S. (2022). An AI-Based Medical Chatbot Model for Infectious Disease Prediction. *IEEE Access*, *10*, 128469–128483.

Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, *107*, 1753–1820.

Collingridge, D. (1980). *The social control of technology*. St Martin's Press.

Consulich, F. (2022). Flash offenders. Le prospettive di accountability penale nel contrasto alle Intelligenze Artificiali devianti. *Rivista Italiana di Diritto e Procedura Penale*, *3*, 1015–1055.

De Francesco, G. (2004). *Programmi di tutela e ruolo dell'intervento penale*. Giappichelli.

Diamantis, M. E. (2021). Algorithms acting badly: A solution from corporate law. *The George Washington Law Review*, *89*, 801–856.

DiPaola, D., & Calo, R. (2024). Socio-digital vulnerability. Advanced online publication. doi:10.2139/ssrn.4686874.

Donini, M., & Pavarini, M. (Eds.). (2011). *Sicurezza e diritto penale*. Bologna University Press.

Duff, R. A. ***et al***. (2010). *The Boundaries of the Criminal Law*. (Eds.) Oxford University Press.

Ebers, M. (2022). Standardizing AI. The case of the European Commission's Proposal for an "Artificial Intelligence Act". In L. DiMatteo, C. Poncibò, & M. Cannarsa (Eds.), *The Cambridge handbook of artificial intelligence. Global perspectives on law and ethics* (pp. 321–344). Cambridge University Press.

Europol, Innovation Lab. (2023). *ChatGPT - The impact of Large Language Models on Law Enforcement, a Tech Watch Flash Report from the Europol Innovation Lab.* Europol. https://www.europol.europa.eu/.

Fiorinelli, G. (2023). Il concorrente virtuale: La prevenzione dell'uso di catgut per finalità criminali tra etero-e auto-regolazione. *Rivista Italiana Di Medicina Legale*, *2*, 361–378.

Fletcher, G. P. (2008). *Tort liability for human rights abuses*. Hart Publishing.

Flor, R. (2024). Cybersecurity for artificial intelligence, medical device security and criminal law: First thoughts in the prism of European Law. In P. Magagna, & R. Magagna (Eds.), *Advanced diagnostic tools in aortic pathology* (pp. 142–149). Edizioni Minerva Medica.

Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models, and other generative models. *Philosophy and Technology*, *36*, 3–7.

Fragrasso, B. (2024). Intelligenza Artificiale e crisi del diritto penale d'evento: profili di responsabilità penale del produttore di sistemi di I.A. *Rivista Italiana di Diritto e Procedura Penale*,*1*, 287-305.

Franks, M. A., & Waldman, A. E. (2019). Sex, lies, and videotape: Deep fakes and free speech delusions. *Maryland Law Review*, *78*, 892–898.

Frosio, G. (Eds.). (2020). *The Oxford handbook of online intermediary liability*. Oxford University Press.

Gehman, S. (2020). Real toxicity prompts: Evaluating neural toxic degeneration in language models. In T. Cohn, Y. He, Y. Liu (Eds.). *Findings of the association for computational linguistics* (pp. 3356–3369). The Association for Computational Linguistics.

Giannini, A. (2023). *Criminal behavior and accountability of artificial intelligence systems*. Eleven International Publishing.

Gless, S., Silverman, E., & Weigend, T. (2016). If robots cause harm, who is to blame? Self-driving cars and criminal liability. *New Criminal Law Review*, *19*(3), 412–436.

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., and Sedova, K. (2023). Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. *Joint report with Georgetown University's Center for Security and Emerging Technology OpenAI and Stanford Internet Observatory*. Advance online publication. arXiv:2301.04246.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems* (pp. 1-9). Curran Associates, Inc.

Guerini, T. (2020). *Fake news e diritto penale. La manipolazione digitale del consenso nelle democrazie liberali*. Giappichelli.

Henderson, P., Hashimoto, T., & Lemley, M. (2023). Where's the liability in harmful AI speech? *Journal of Free Speech Law*, *3*, 561–605.

Hörnle, T. (2015). Theories of criminalization. In M. D. Dubber, & T. Hörnle (Eds.), *The Oxford handbook of criminal law* (pp. 679–701). Oxford University Press.

Husak, D. (2007). *Overcriminalization. The limits on the criminal law*. Oxford University Press.

Keiler, J., & Roef, D. (2019). *Comparative concepts of criminal law*. Intersentia.

Kelly, E. (2018). *The limits of blame. Rethinking punishment and responsibility*. Harvard University Press.

Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, *9*(1), 104–117.

Lemoine, L., & Vermeulen, M. (2023). Assessing the extent to which Generative Artificial Intelligence (AI) falls within the scope of the EU's Digital Services Act: An initial analysis. Advance online publication. doi:10.2139/ssrn.4702422.

Miró-Llinares, F., Duvac, C., Toader, T., et al. (Eds.) (2024). *Criminalisation of AI-related offence* [Monograph]. *Revue Internationale de Droit Pénal*, 1, 1–445.

Mongillo, V. (2023). Corporate criminal liability for AI-related crimes: Possible legal techniques and obstacles. *International Review of Penal Law*, *1*, 77–90.

Neuwirth, R. J. (2022). *The EU artificial intelligence act. Regulating subliminal AI systems*. Routledge.

Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics*, *2*(1), 25–42.

OpenAI GPT-4 System Card (2023). Available at https://cdn.openai.com/papers/gpt-4-system-card.pdf (access date November 3, 2024).

Ousidhoum, N. et al. (2021). Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (pp. 4262–4274). Association for Computational Linguistics.

Perrigo, B. (2023, February 17). The New AI-Powered Bing Is Threatening Users. That's No Laughing Matter. *Time*. https://time.com/6256529/bing-openai-chatgpt-danger-alignment/

Piergallini, C. (2020). Intelligenza artificiale: Da 'mezzo' ad 'autore' del reato? *Rivista Italiana di Diritto e Procedura Penale*, 1745–1772.

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy and Technology*, *34*(4), 1057–1084.

Simmler, M. (2023). Responsibility gap or responsibility shift? The attribution of criminal responsibility in human–machine interaction. *Information, Communication & Society*, 1–21.

Simmler, M., & Markwalder, N. (2019). Guilty robots? – Rethinking the nature of culpability and legal personhood in an age of artificial intelligence. *Criminal Law Forum*, *30*(1), 1–31.

Sivaram, J. et al. (2023). Adversarial machine learning: The rise in AI-enabled crime and its role in spam filter evasion. *Computer Fraud & Security*, *2*, 1–8.

Stewart, J. G. (2015). Complicity. In M. D. Dubber & T. Hörnle (Eds.), *The Oxford handbook of criminal law* (pp. 534–559). Oxford University Press.

Stortoni, L., & Foffani, L. (Eds.). (2004). *Critica e giustificazione del diritto penale nel cambio di secolo. L'analisi critica della Scuola di Francoforte*. Giuffrè.

Surden, H. (2024). ChatGPT, artificial intelligence (AI) large language models, and law. *Fordham Law Review*, *92*, 1940–1970.

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Online manipulation: Hidden influences in a digital world. *Georgetown Law Technology Review*, *4*(1), 1–45.

Thompson, D F. (1980). Moral Responsibility of Public Officials: The Problem of Many Hands. *Am Polit Sci Rev*, *74*(4), 905–916.

Vassalli, G. (1942). *La potestà punitiva*. Utet.

Volokh, E. (2023). Large libel models? Liability for AI output. *Journal of Free Speech Law*, *3*, 489–558.

Weidinger, L. et al. (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214–229). Association for Computing Machinery.

Cite this article: Panattoni B. (2025). Generative AI and criminal law. *Cambridge Forum on AI: Law and Governance* **1**, e9, 1–14. https://doi.org/10.1017/cfl.2024.9