INTERNATIONAL LAW STUDIES

Published Since 1895

Modes of Liability for AI-Enabled Crimes in International Criminal Law

Antonio Coco

107 INT'L L. STUD. 228 (2026)

Volume 107 2026

Published jointly by the University of Reading and the Stockton Center for International Law

ISSN 2375-2831

Modes of Liability for AI-Enabled Crimes in International Criminal Law

Antonio Coco[*]

CONTENTS

* Senior Lecturer, Essex Law School, University of Essex (antonio.coco@essex.ac.uk).
I wish to thank my friends and colleagues Julia Geneuss, Miles Jackson, Yvonne McDermott, and the participants of the workshop, Artificial Intelligence and International Law, held at the University of Reading on March 19, 2025, for their valuable feedback on an earlier draft. All errors remain my own.

The thoughts and opinions expressed are those of the author and not necessarily those of the U.S. government, the U.S. Department of the Navy, or the U.S. Naval War College.

I. INTRODUCTION

As artificial intelligence (AI) technologies continue to reshape human activities such as warfare and surveillance, ensuring accountability for those who develop or deploy such technologies has become increasingly urgent. This article explores how established rules of international criminal law—particularly those governing individual criminal responsibility—might adapt to address AI-enabled atrocity crimes. Let us consider, for the sake of exposition, the following four scenarios:

Scenario 1: Cyberwarfare specialist (CS), who is a State official, deliberately manipulates the training data of an enemy State's AI-powered targeting system by launching an adversarial attack. The attack subtly alters key datasets, tricking the AI into misidentifying protected civilian structures—such as hospitals and refugee shelters—as legitimate military targets. Unaware of the data corruption, the enemy military deploys the AI-assisted system in a combat operation. As a result, their autonomous weapons system (AWS) executes strikes on civilian areas. CS intended for this outcome, knowing it would increase civilian casualties and erode international support for the enemy State's military operations.

Scenario 2: Military general (MG) deploys an AWS in an urban conflict zone, despite being aware of prior incidents where the AWS misidentified civilians as combatants. MG, under pressure to achieve results, dismisses concerns about false positives and allows the system to operate independently. The AI subsequently misidentifies a school as an enemy compound and launches a devastating strike, killing dozens of children and teachers.

Scenario 3: Software engineer (SE) working for a State intelligence agency develops an AI-powered social media algorithm designed to amplify content that incites violence against an ethnic minority. The engineer deliberately codes the algorithm to prioritize inflammatory posts, knowing that this will escalate persecution. Over time, the AI systematically increases engagement with hate speech, leading to mass riots and coordinated attacks against the targeted group.

Scenario 4: Intelligence officer (IO) working with AI-powered surveillance tools oversees the development of a system that flags individuals based on ethnic and religious profiling. While they recognize the system's biases and the potential for misuse, they justify its use under national security concerns. The government later utilizes the AI's flagged data to round up and arbitrarily detain members of the identified group.

These four scenarios illustrate varying levels of human involvement and participation in AI-powered or AI-enabled core international crimes, including war crimes, crimes against humanity, and genocide. The rules of international criminal law setting the boundaries for determining that an individual's participation in a crime warrants their criminal responsibility are usually called "modes of liability," "forms of participation," or similar expressions.[1] The concept of criminal participation is broad, encompassing not only those who physically carry out unlawful acts but also those who facilitate, incite, or allow them to happen. Consequently, these rules establish different ways in which a person can bear responsibility, such as individual (also known as "direct") or joint perpetration, perpetration through another person (also known as "indirect perpetration"), instigation, aiding and abetting, or responsibility for failure to prevent or punish a subordinate's crime (also known as "superior responsibility" or "command responsibility"). The Rome Statute of the International Criminal Court (ICC Statute),[2] as well as the statutes of other international tribunals, recognizes various modes of liability that allow for the prosecution and punishment of both those who commit crimes themselves and those who contribute to crimes in a variety of ways. These rules of international law, not domestic criminal law, will be the point of reference for this article.

---

1. MODES OF LIABILITY IN INTERNATIONAL CRIMINAL LAW (Jérôme de Hemptinne, Robert Roth & Elies van Sliedregt eds., 2019); KAI AMBOS, 1 TREATISE ON INTERNATIONAL CRIMINAL LAW: FOUNDATIONS AND GENERAL PART 158–256 (2d ed. 2021); ELIES VAN SLIEDREGT, INDIVIDUAL CRIMINAL RESPONSIBILITY IN INTERNATIONAL LAW 89–156 (2012). For a critique of the expression, *see* James G. Stewart, *The End of 'Modes of Liability' for International Crimes*, 25 LEIDEN JOURNAL OF INTERNATIONAL LAW 165, 166 (2012).

2. Rome Statute of the International Criminal Court, July 17, 1998, 2187 U.N.T.S. 90. Of note, the rules on modes of liability established in customary international law may not always align precisely with the formulations enshrined in the Rome Statute or other tribunals' statutes. Moreover, while this article focuses exclusively on criminal accountability, it is important to note that other bodies of international law—such as international humanitarian law and international human rights law—may impose different obligations, fault standards, or be better suited to addressing collective or systemic harms.

The scenarios described above aim to illustrate how the mentioned rules of international criminal law may apply in the context of AI-enabled crimes. In each case, human involvement is central to the commission of the offense, but the nature of that involvement varies. A developer who designs an AI system to spread hate speech, knowing it will fuel persecution, plays a different role from a military commander who knowingly deploys an AWS despite the risk of civilian casualties. An intelligence officer overseeing an AI-driven surveillance program that enables mass persecution is engaged differently from a cyber warfare specialist who manipulates an enemy's AI, ensuring it will commit a war crime. These variations in conduct and mental state raise distinct questions about criminal responsibility and illustrate the different ways in which individuals may be held criminally responsible under international law.

This article contributes to ongoing debates about individual criminal responsibility by examining how international criminal law—particularly the rules on modes of liability as interpreted by international criminal tribunals—can be applied to cases involving AI-enabled international crimes. It engages critically with relevant case law, scholarship, and novel factual scenarios emerging from the employment of new technologies, with the aim of informing future interpretations of the relevant rules. In so doing, it also seeks to advance the nascent discussion on the intersection between AI and international criminal law and to prompt further inquiry into this evolving area.

To those ends, Part II outlines the difficulties of attributing criminal responsibility in cases involving AI, including a brief discussion of whether AI itself could, in theory, bear such responsibility. Part III then turns to the key question of human criminal responsibility when AI is involved, that is the question of how to establish mens rea. Part IV examines forms of perpetration (also known as commission), namely individual perpetration, perpetration through another person, and perpetration jointly with another person, with particular attention to the doctrine of joint criminal enterprise. Parts V and VI focus on two selected "secondary" or "accessory" modes of liability, namely instigation and aiding and abetting. Part VII covers superior (also known as "command") responsibility. Finally, Part VIII concludes by briefly reflecting on the extent to which existing rules suffice to address AI-enabled international crimes.

II. CHALLENGES OF ESTABLISHING WHO BEARS CRIMINAL RESPONSIBILITY IN AI-ENABLED CRIMES

International criminal law traditionally assumes that human actors are responsible, but the use of AI systems challenges this framework by introducing scenarios where humans are not necessarily the principal actors behind misconduct. Criminal responsibility in AI-enabled crimes is, in fact, made harder to establish by questions of intent and by the level of control that individuals have over the outcome. The required mental state, or mens rea, varies depending on the mode of liability at issue, and the statutes of international courts have generally required a high degree of mens rea for most crimes. The scenarios presented above show different degrees of mental states. Some actors act with clear intent, while others knowingly take risks or fail to intervene despite foreseeing that an unlawful consequence may occur. This distinction is critical in assessing whether an individual can be held criminally liable and under what mode of liability.

One central issue is the distinction between violations committed *with* AI and violations committed *by* AI. When international law is violated through the use of an AI system, the crime remains imputable to human actors, whether they be programmers, developers, commanders, operators, etc.[3] However, where an AI system acts autonomously and unpredictably, the challenge becomes to determine whether any individual can be held liable, or whether such conduct should be classified instead as an instance of what has been labelled a "hard AI crime"—that is, a situation in which criminal responsibility cannot be traced back to the wrongful act of a person.[4] This difficulty in tracing responsibility has led some scholars to suggest that "hard AI crimes" make the strongest case for considering whether AI itself could be punished, though such a proposition remains legally and philosophically controversial.[5] In these cases, AI is functionally committing the crime, either

3. Taking this view, among others, Christine Carpenter, *Whose [Crime] Is It Anyway? Adapting the Crime of Aggression to Grapple with AI and the Future of International Crimes*, 23 JOURNAL OF INTERNATIONAL CRIMINAL JUSTICE 69, 79–80 (2025).

4. Ryan Abbott & Alex Sarch, *Punishing Artificial Intelligence: Legal Fiction or Science Fiction*, 53 UC DAVIS LAW REVIEW 323, 328 (2019); Alice Giannini, *Artificial Intelligence, Criminal Liability For*, *in* ENCYCLOPEDIA OF THE PHILOSOPHY OF LAW AND SOCIAL PHILOSOPHY 1, 3 (Mortimer Sellers & Stephan Kirste eds., 2024).

5. On the subject, *see* Monika Simmler & Nora Markwalder, *Guilty Robots?—Rethinking the Nature of Culpability and Legal Personhood in an Age of Artificial Intelligence*, 30 CRIMINAL LAW FORUM 1 (2019); Thomas Weigend, *Convicting Autonomous Weapons?: Criminal Responsibility of and for AWS Under International Law*, 21 JOURNAL OF INTERNATIONAL CRIMINAL JUSTICE

because no human has acted with criminal intent or because the AI's behavior cannot be directly attributed to any individual.[6] Scholars advocating against the principle *machina delinquere non potest*—the idea that machines cannot commit crimes or be punished—argue that, as AI systems become increasingly autonomous, they may eventually reach a level where their actions and decisions are sufficiently independent to warrant direct legal accountability.[7]

However, if AI were to be held criminally responsible, the fundamental structure of criminal law—particularly any requirement of mens rea—would need to be radically rethought. This is because all existing modes of liability require some form of mental element, such as intent or recklessness. AI, as usually conceptualized, lacks the capacity to experience these mental states, meaning that it does not possess the conscious awareness necessary for criminal responsibility.[8] Criminal law generally presumes that criminal acts are committed by moral agents who can experience punishment, deterrence, and retribution.[9] AI, as an artificial construct, does not possess the qualities that make legal punishment meaningful, such as the ability to suffer consequences or feel guilt.[10] Moreover, the mens rea requirement is not just a theoretical concern but a practical barrier: even in cases involving human actors, liability is not imputed where conduct is involuntary, as criminal acts must be the product of free will and awareness.[11] Without intent or the capacity to make autonomous moral decisions, AI fails to meet the criteria for criminal agency and, consequently, responsibility.

The dominant position in international law, therefore, rejects the notion of AI's criminal responsibility. Most scholars persuasively argue that accountability should remain with the "human behind the machine," whether

---

1137 (2023); Africa Maria Morales Moreno, *Artificial Intelligence and Criminal Law: First Approximations*, 53 REVISTA JURIDICA DE CASTILLA & LEON 177 (2021).

6. Abbott & Sarch, *supra* note 4, at 328.

7. Giannini, *supra* note 4, at 3; *see also* GABRIEL HALLEVY, LIABILITY FOR CRIMES INVOLVING ARTIFICIAL INTELLIGENCE SYSTEMS (2015).

8. *See, e.g.*, Paola Gaeta, *Who Acts When Autonomous Weapons Strike?: The Act Requirement for Individual Criminal Responsibility and State Responsibility*, 21 JOURNAL OF INTERNATIONAL CRIMINAL JUSTICE 1033 (2023).

9. Jack M. Beard, *Autonomous Weapons and Human Responsibilities*, 45 GEORGETOWN JOURNAL OF INTERNATIONAL LAW 617, 663 (2013).

10. Jason Lee, *Autonomous Weapons, War Crimes, and Accountability*, 48 NORTH CAROLINA JOURNAL OF INTERNATIONAL LAW 51, 66 (2022).

11. See, e.g., the discussion in JEREMY HORDER, ASHWORTH'S PRINCIPLES OF CRIMINAL LAW 74–76 (10th ed. 2022).

the developer, deployer, or operator.[12] AI is widely regarded as a tool rather than an independent agent, and under international humanitarian law even the most advanced AI-enabled AWS are still considered mere means of warfare.[13] The Group of Governmental Experts on Lethal Autonomous Weapon Systems has advised that human responsibility for decisions on the use of weapon systems must be retained, as accountability cannot be transferred to machines.[14] Ultimately, punishing AI—aside from the lack of clarity about what that would actually involve—would not only be legally and philosophically incoherent but also impractical, as it does not align with the fundamental principles of criminal responsibility and accountability.

The attribution of responsibility to humans in AI-related crimes is, however, further complicated by the "problem of many hands." This expression alludes to the fact that the development, deployment, and use of AI systems ordinarily involve multiple actors, including programmers, engineers, corporate entities, military personnel, and political leaders, all of whom may contribute to the system's functioning in different ways.[15] Thus, it is difficult to pinpoint a single individual responsible for an AI-related crime, particularly when the system operates in a way that was not explicitly anticipated by its creators or users. When AI technology is employed in cyber operations, the matter becomes even more complex. Identifying the actor responsible for an AI-enabled cyber-operation is often challenging as AI systems may be manipulated by third parties, adversarial attacks may distort their outputs, and attribution to a specific individual or State may be difficult to establish.[16] Furthermore, the problem of many hands is compounded by the "problem

12. Giannini, *supra* note 4, at 3.

13. Gaeta, *supra* note 8, at 1053.

14. Dustin A. Lewis, *War Crimes Involving Autonomous Weapons: Responsibility, Liability and Accountability*, 21 JOURNAL OF INTERNATIONAL CRIMINAL JUSTICE 965, 978 (2023); Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, *Report of the 2019 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*, annex IV, ¶ b, U.N. Doc. CCW/GGE.1/2019/3 (Sept. 25, 2019); Rolling Text, The Convention on Certain Weapons (CCW) Informal Meeting of Experts on Lethal Autonomous Weapons Systems, sec. V(2) (Dec. 18, 2025), https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2026)/CCW_GGE_LAWS_Rolling _Text_-_status_18_December_2025.pdf.

15. Monika Simmler, *Responsibility Gap or Responsibility Shift? The Attribution of Criminal Responsibility in Human–Machine Interaction*, 27 INFORMATION, COMMUNICATION & SOCIETY 1142, 1145 (2024).

16. Carpenter, *supra* note 3, at 7.

of many things," referring to the fact that AI systems are often trained on vast, heterogeneous datasets, and their decision-making processes may be influenced by factors beyond the control or knowledge of any one individual.[17] These features raise fundamental questions about whether traditional criminal law doctrines can adequately address AI-enabled offenses.

These challenges are further exacerbated by the opacity and unpredictability of AI decision-making. AI-enabled systems, particularly those utilizing deep learning and data-driven learning methods, often operate in ways that are not entirely explainable to human users.[18] This lack of transparency undermines the ability of courts and prosecutors to establish the mental element required for criminal liability. If an AI system makes an autonomous determination to attack a civilian target, can a commander or programmer be held liable if they did not and could not foresee that decision? The next part delves deeper into this issue and, more broadly, into the problem of meeting the mens rea requirement for humans who interacted with AI systems on the way to the commission of international crimes.

III. THE PROBLEM OF MENS REA FOR HUMANS INTERACTING WITH AI SYSTEMS

Establishing mens rea for individuals who resort to AI in ways that contribute to international crimes presents significant challenges. Criminal law traditionally assigns culpability based on what individuals intended or foresaw (mens rea) and what they controlled (actus reus and causation), yet AI systems introduce layers of decision-making that obscure these determinations.[19] While some scholars argue that AI's autonomy does not necessarily

17. Giannini, *supra* note 4, at 3.

18. Anna Rosalie Greipl, *Data-Driven Learning Systems and the Commission of International Crimes: Concerns for Criminal Responsibility?*, 21 JOURNAL OF INTERNATIONAL CRIMINAL JUSTICE 1097, 1099 (2023); *see also* FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015); ARTHUR HOLLAND MICHEL, THE BLACK BOX, UNLOCKED: PREDICTABILITY AND UNDERSTANDABILITY IN MILITARY AI (2020).

19. Guido Acquaviva, *Autonomous Weapons Systems Controlled by Artificial Intelligence: A Conceptual Roadmap for International Criminal Responsibility*, 60 THE MILITARY LAW AND THE LAW OF WAR REVIEW 89, 109 (2022).

break the causal chain of responsibility—since human actors ultimately define how AI functions[20]—the level of foreseen and accepted risks remains a crucial factor. Depending on the applicable law and on the specific definitions it provides, the required mens rea may range from intent to recklessness.[21] At times, criminal responsibility may not hinge on direct intent but rather on whether an actor knowingly accepted the risk of a criminally relevant consequence.[22]

Intent, especially, is among the most difficult standards of mens rea to establish in AI-related crimes. When an individual deliberately programs or deploys an AI system with the explicit aim of committing a crime, proving intent is relatively straightforward. However, when an AI system is deployed in circumstances where an unlawful consequence is a foreseeable but unintended consequence, determining intent becomes harder. Some argue that criminal responsibility can be ascribed when humans have knowledge that a criminally relevant consequence "will occur in the ordinary course of events" (*dolus indirectus*), such as when an AI system trained in one environment is deployed in another where it predictably fails.[23] Others note that intent must also account for whether a human actor could "reasonably anticipate" the consequences of an AI's actions.[24] AI systems that operate with learning capabilities add another challenge, as their actions may not be entirely traceable to human intent.[25] In practice, proving that an individual intended for an AI system to commit an international crime requires demonstrating that they were aware of the system's likely behavior and deployed it with that understanding—an evidentiary hurdle that will likely prevent many prosecutions.[26]

---

20. Marco Sassoli, *Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified*, 90 INTERNATIONAL LAW STUDIES 308, 324–25 (2014).

21. MARTA BO, LAURA BRUUN & VINCENT BOULANIN, RETAINING HUMAN RESPONSIBILITY IN THE DEVELOPMENT AND USE OF AUTONOMOUS WEAPON SYSTEMS 29 (2022), https://www.sipri.org/sites/default/files/2022-10/2210_aws_human_responsibility.pdf.

22. Emily L. Drake, *Evaluating Autonomous Weapons Systems: A Dichotomic Lens of Military Value and Accountability*, 53 COLUMBIA HUMAN RIGHTS LAW REVIEW 297, 329 (2021).

23. Guido Acquaviva, *Crimes Without Humanity? Artificial Intelligence, Meaningful Human Control, and International Criminal Law*, 21 JOURNAL OF INTERNATIONAL CRIMINAL JUSTICE 981, 998 (2023).

24. Lewis, *supra* note 14, at 974–75.

25. Acquaviva, *supra* note 19, at 105.

26. Weigend, *supra* note 5, at 1150; *see also* Jindan-Karena Mann, *Autonomous Weapons Systems and the Liability Gap, Part One: Introduction to Autonomous Weapons Systems and International Criminal Liability*, RETHINKING SLIC* (July 15, 2019), https://www.rethinkingslic.org

Recklessness is another potential standard for mens rea in AI-related cases. Recklessness generally involves an individual foreseeing a substantial and unjustified risk yet proceeding with their actions regardless.[27] This could apply, for instance, to military commanders or developers who recognize the risk that an AWS may malfunction and target civilians, but deploy it anyway.[28] However, recklessness is not generally provided as a default mens rea standard in the ICC Statute, which instead in Article 30 requires intent and knowledge.[29] The question of whether recklessness suffices for criminal responsibility with respect to international crimes thus depends on the applicable law. For instance, the International Committee of the Red Cross's commentary to Additional Protocol I affirms that, for the grave breach (war crime) of "making the civilian population or individual civilians the object of attack,"[30] the required mental element of willfulness includes "'recklessness', viz., the attitude of an agent who, without being certain of a particular result, accepts the possibility of it happening."[31] The International Criminal Tribunal for the former Yugoslavia (ICTY) has followed such approach when adjudicating this war crime.[32] Following this approach in the case of AI-enabled crime implicates the responsibility of those who foresaw the possibility of a machine's error and would have had reasons to be cautious.

Knowledge is relevant to all modes of liability. In particular, for modes such as aiding and abetting or superior responsibility, it is often interpreted to

---

/blog/criminal-law/51-autonomous-weapons-systems-and-the-liability-gap-part-one-introduction-to-autonomous-weapon-systems-and-international-criminal-liability.

27. *See, e.g.*, Regina v. Cunningham, [1957] 2 Q.B. 396 (C.A.); MODEL PENAL CODE § 2.02(2)(c) (AM. L. INST. 1962).

28. Weigend, *supra* note 5, at 1149.

29. Unless otherwise provided for specific crimes. Rome Statute of the International Criminal Court, *supra* note 2, art. 30.

30. Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts art. 85(3)(a), June 8, 1977, 1125 U.N.T.S. 3.

31. COMMENTARY ON THE ADDITIONAL PROTOCOLS OF 8 JUNE 1977 TO THE GENEVA CONVENTIONS OF 12 AUGUST 1949, § 3474 (Yves Sandoz, Christophe Swinarski & Bruno Zimmermann eds., 1987) (whereas on the other side, the *Commentary* clarifies that "ordinary negligence or lack of foresight is not covered, i.e., when a man acts without having his mind on the act or its consequences."); *see also* Jens David Ohlin, *The Combatant's Stance: Autonomous Weapons on the Battlefield*, 92 INTERNATIONAL LAW STUDIES 1, 23 (2016).

32. Prosecutor v. Galić, Case No. IT-98-29-T, Judgment, ¶ 54 (Int'l Crim. Trib. for the Former Yugoslavia, Dec. 5, 2003); Prosecutor v. Galić, Case No. IT-98-29-A, Appeals Chamber Judgment, ¶ 140 (Int'l Crim. Trib. for the Former Yugoslavia, Nov. 30, 2006); Prosecutor v. Strugar, Case No. IT-01-42-A, Appeals Chamber Judgment, ¶ 270 (Int'l Crim. Trib. for the Former Yugoslavia, July 17, 2008).

require, respectively, that the individual was aware (or, in some cases, should have been aware owing to the circumstances) that their actions or omissions contributed to an international crime (aiding and abetting) or that their subordinates were committing international crimes (superior responsibility). In the case of AI, this raises questions about what an individual must know about the system's programming, capabilities, and operational environment. It could be argued that users of AI systems should have an advanced understanding of how the system functions and how it will behave in different scenarios. But it could also be noted that requiring such detailed technical knowledge places an unreasonable burden on users, and that their knowledge should instead be assessed in relation to the system's expected effects in a specific operational context.[33] The ICC has generally adopted a strict standard of knowledge, requiring a high degree of certainty regarding the criminal outcome of a defendant's actions.[34] This level of certainty could be hard to obtain in AI cases, as deviations from expected outcomes may make it difficult to prove that a user, commander, or developer had the requisite knowledge of a crime occurring.

Given these challenges, negligence and strict liability could be considered as alternative standards to fill the accountability gap for AI-enabled international crimes. Under current international criminal law, mere *negligence* is generally insufficient to establish individual responsibility for war crimes, as the threshold for criminal liability has been deliberately set high.[35] This means that even if an AI system malfunctions due to inadequate testing and engages in unlawful conduct or causes an unlawful consequence, those responsible for its development and deployment may evade prosecution—at least for international crimes at the international level—if they were merely negligent in failing to foresee such consequences. Moreover, manufacturers of AWS are not typically bound by a duty of care toward those who deploy these weapons and/or towards those against whom those weapons may be used (i.e., potential victims), making it unlikely that criminal liability could be im-

33. BO, BRUUN & BOULANIN, *supra* note 21, at 32.

34. Prosecutor v. Bemba, ICC-01/05-01/08, Decision on the Confirmation of Charges, ¶¶ 368–69 (June 15, 2009); Prosecutor v. Lubanga, Case No. ICC-01/04-01/06-2842, Judgment, ¶ 1011 (Mar. 14, 2012).

35. With respect to the general mens rea standard in the ICC Statute, for instance, *see* Donald Piragoff & Darryl Robinson, *Article 30—Mental Element*, *in* ROME STATUTE OF THE INTERNATIONAL CRIMINAL COURT: ARTICLE-BY-ARTICLE COMMENTARY 1328, 1329 (Kai Ambos ed., 4th ed. 2022).

posed on corporate entities unless intentional or reckless programming decisions can be proven. Given the unpredictability of AI and its potential for unintended yet harmful consequences, lowering the threshold to include negligence could ensure accountability for those who fail to exercise sufficient caution in designing or deploying AI-driven systems.[36]

Another approach to addressing the accountability gap is the application of *strict liability*, particularly in cases where AI use poses an inherent and uncontrollable risk. Some scholars argue that strict liability could provide a solution in contexts such as AI-driven cyber warfare, where the unique dangers and evidentiary challenges make it difficult to establish intent or recklessness.[37] According to Christine Carpenter, in particular, strict liability could apply where AI-enabled attacks are deemed "ultrahazardous," based on factors such as the magnitude of the risk, the inability to eliminate harm through reasonable care, and the uncommon nature of the activity.[38] While strict liability would represent a significant departure from traditional international criminal law principles, it may be necessary to prevent AI-related crimes from escaping legal scrutiny due to evidentiary constraints and the limitations of existing mens rea standards.

IV. PERPETRATION AND CRIMINAL RESPONSIBILITY FOR AI-ENABLED CRIMES

In the framework of international criminal law, individuals may be found responsible as perpetrators for having committed crimes by means of AI. Article 25(3)(a) of the ICC Statute provides that a person shall be criminally responsible if they commit a crime "as an individual, jointly with another or through another person, regardless of whether that other person is criminally responsible." Criminal responsibility for perpetration (also known as "commission") is a staple of the doctrine of modes of liability in the general theory of criminal law, and it denotes the conduct of the person who is most responsible for the criminal offense. As such, it has consistently been listed among other available modes of liability in the case of international crimes.

36. Antonio Coco & Talita Dias, *'Handle with Care': Due Diligence Obligations in the Employment of AI Technologies*, *in* RESEARCH HANDBOOK ON WARFARE AND ARTIFICIAL INTELLIGENCE 234, 245 (Robin Geiß & Henning Lahmann eds., 2024).

37. Carpenter, *supra* note 3, at 16 (discussing, inter alia, the work of Christiane Wendehorst, *Strict Liability for AI and Other Emerging Technologies*, 11 JOURNAL OF EUROPEAN TORT LAW 150 (2020)).

38. Carpenter, *supra* note 3, at 17.

For instance, Article 7(1) of the ICTY Statute and Article 6(1) of the Statute of the International Criminal Tribunal for Rwanda establish the responsibility of those who commit crimes under these tribunals' jurisdiction.[39] Provisions on perpetration/commission may be helpful to capture the criminality of human actors who develop, deploy, or manipulate AI systems with a view to generate conduct amounting to international crimes. In the following sections, I will separately examine three forms that the responsibility of perpetrators in international criminal law may take, namely individual perpetration (Section IV(A)), perpetration through another person (Section IV(B)), and joint (also known as co-) perpetration, with a focus on the doctrine of "joint criminal enterprise" (Section IV(C)).

A. Individual Perpetration

The most straightforward mode of liability is individual (also known as "direct") perpetration (or commission), where an individual "physically fulfils all the material elements of the crime," with "intent to commit the crime or . . . knowledge that his/her actions will bring about the occurrence of the material elements of the crime."[40] This mode of liability could be used to capture the criminality in Scenario 1, in which CS deliberately manipulates the training data of the enemy's AI-powered targeting system, intending for that system to strike civilian areas. Assuming that the adversarial data poisoning could be qualified as an "attack," CS's conduct and mental element could match—for instance—the definition of the war crime of attacking civilians under Articles 8(2)(b)(i) and 8(2)(e)(i) of the ICC Statute.[41] In such

39. Statute of the International Criminal Tribunal for the Former Yugoslavia art. 7(1), *Secretary General Report Pursuant to Paragraph 2 of S.C. Res. 808*, U.N. Doc. S/25704 (May 3, 1993) (adopted by the Security Council in S.C. Res. 827 (May 25, 1993)); Statute of the International Criminal Tribunal for Rwanda art. 6(1), S.C. Res. 955 annex (Nov. 8, 1994).

40. Tom Gal, *Direct Commission*, *in* MODES OF LIABILITY IN INTERNATIONAL CRIMINAL LAW 17, 29 (Jérôme de Hemptinne, Robert Roth & Elies van Sliedregt eds., 2019); *cf.* Rome Statute of the International Criminal Court, *supra* note 2, art. 30(b) (establishing that, for the purposes of the article, "a person has intent where: . . . In relation to a consequence, that person means to cause that consequence or is aware that it will occur in the ordinary course of events.").

41. In the same vein, *see* Elliot Winter, *The Accountability of Software Developers for War Crimes Involving Autonomous Weapons: The Role of the Joint Criminal Enterprise Doctrine*, 83 UNIVERSITY OF PITTSBURGH LAW REVIEW 51, 56 (2021); Jonathan Kwik, *The Conceptual Roots of the Criminal Responsibility Gap in Autonomous Weapon Systems*, 24 MELBOURNE JOURNAL OF INTERNATIONAL LAW 1, 16 (2023); *cf.* Sassoli, *supra* note 20, at 325 (considering instead this to be an example of indirect perpetration).

example, CS would be using the AWS as no more than a tool, similarly to personally pulling a trigger.

Furthermore, even if an AI system was not explicitly programmed to commit crimes, intent can be inferred when a user deploys it with knowledge that it will inevitably cause unlawful consequences. It has been proposed that intent could be inferred where the user is "practically" or "virtually" certain of the unlawful result.[42] Moreover, the mental element required for a certain crime may be lower than intent. As explained above, the ICRC commentary to Additional Protocol I and ICTY case law established that the defendant's recklessness is sufficient to meet the requirements for the grave breach (war crime) of "making the civilian population or individual civilians the object of attack."[43] If this were the applicable law in Scenario 2, MG could be held responsible as a principal perpetrator for knowingly deploying an AWS that misidentifies civilians as combatants. Beyond actively engaging in AI-assisted crimes, human actors may also be found to be criminally responsible for perpetration *by omission* when they have a legal duty to act or prevent a certain conduct or consequence but fail to do so.[44] Such may be the case of a user who, upon realizing that an AI system is leading to unlawful conduct and/or consequences, and having a duty to stop its operation, deliberately chooses not to intervene.

B. Perpetration Through Another Person

Human actors could also be held criminally responsible for AI-enabled crimes under the doctrine of perpetration through another person, also known as "indirect perpetration"—a mode of liability recognized, inter alia, in Article 25(3)(a) of the ICC Statute. This mode of liability has been applied to hold perpetrators criminally responsible when they act through another person—either an "innocent agent" (lacking autonomy or intent) or a responsible individual in their own right—or through a hierarchical apparatus of power, effectively using them as tools to commit a crime.[45]

---

42. BO, BRUUN & BOULANIN, *supra* note 21, at 29.

43. *See supra* notes 31, 32.

44. For a discussion, *see* AMBOS, *supra* note 1, at 261–74. On the issue of omissions with respect to AI technologies specifically, *see* Marta Bo, *Meaningful Human Control: An International Criminal Law Account*, *in* RESEARCH HANDBOOK ON MEANINGFUL HUMAN CONTROL OF ARTIFICIAL INTELLIGENCE SYSTEMS 148, 153 (Giulio Mecacci et al. eds., 2024).

45. *See generally* Alejandro Kiss, *Indirect Commission*, *in* MODES OF LIABILITY IN INTERNATIONAL CRIMINAL LAW 30 (Jérôme de Hemptinne, Robert Roth & Elies van Sliedregt

This mode of liability has been identified as particularly suitable for capturing the criminality of human actors—whether developers, commanders, or users—who commit crimes through an AI system, which, in this context, could be equated to an innocent agent or another instrument of criminality.[46] At first glance, this interpretation might appear to find support in language such as that used in the ICC Statute. Article 25(3)(a), in this respect, includes the phrase "regardless of whether that other person is criminally responsible"—a formulation that could potentially be applied to an "irresponsible" AWS. Should this interpretation be adopted, MG in Scenario 2 could potentially be held responsible under this doctrine, having employed the AWS as an instrument of criminality.

My difficulty with resorting to this mode of liability in such cases, however, is that an AI system is not (yet, at least) technically a person. It is not being used "like an instrument"; it *is* an instrument. In such cases, I find it more theoretically accurate to hold the human actor as a direct perpetrator rather than as acting through another person. This distinction, however, is not particularly consequential, as the human actor in question would be deemed to be a perpetrator either way.

C. Perpetration Jointly with Another Person

Human actors may also be held criminally responsible as perpetrators of AI-enabled crimes jointly with other persons. This mode of liability, often referred to as joint perpetration or co-perpetration, acknowledges the common possibility that two or more individuals divide tasks functionally among themselves when committing a crime.[47] In the ICC case law, this mode of liability has been applied when multiple individuals share functional control

---

eds., 2019). The theory has its origin in the German doctrine of the *Hintermann* ("man behind" or "man in the background"). *See, e.g.*, AMBOS, *supra* note 1, at 224–32. When used to describe the perpetrator's control over a hierarchical organizational structure, the theory is known as *Organisationsherrschaft* (which can be roughly translated as "organizational domination"). At the ICC, the doctrine was first explored in Prosecutor v. Katanga, Case No. ICC-01/04-01/07-717, Decision on the Confirmation of Charges, ¶¶ 494–539 (Sept. 30, 2008).

46. Abbott & Sarch, *supra* note 4, at 370; Ohlin, *supra* note 31, at 29; Sassoli, *supra* note 20, at 325; Winter, *supra* note 41, at 57.

47. Kai Ambos, *Article 25—Individual Criminal Responsibility*, *in* ROME STATUTE OF THE INTERNATIONAL CRIMINAL COURT: ARTICLE-BY-ARTICLE COMMENTARY 1189, 1200 (Kai Ambos ed., 4th ed. 2022).

over the commission of a crime.[48] In particular, the rule's aim is to establish the criminal responsibility of a person who "provides a coordinated essential contribution to the effecting of a common plan or agreement that results in the commission of an offence."[49] One of the requirements, therefore, is that the co-perpetrators share a common plan or agreement, the implementation of which results in the realization of the objective elements of the crime.[50] ICC case law clarifies that, in this context, the execution of the common plan must be "virtually certain" to result in the commission of one or more crimes[51] and that the defendant's contribution must be "essential"—meaning that its withdrawal would lead to the collapse of the plan and the non-commission of the charged crime(s).[52] Co-perpetrators must exhibit the mens rea required for the crime and be aware that the plan will result in its commission, as well as of the factual circumstances that grant them control over the crime.[53] This mode of liability could be particularly relevant for

48. Prosecutor v. Lubanga, Case No. ICC-01/04-01/06-803, Decision on the Confirmation of Charges, ¶¶ 326–67 (Jan. 29, 2007); Prosecutor v. Lubanga, Case No. ICC-01/04-01/06-2842, Judgment, ¶¶ 1003–6 (Mar. 14, 2012); Prosecutor v. Lubanga, Case No. ICC-01/04-01/06-3121-Red, Appeals Chamber Judgment, ¶¶ 469, 473 (Dec. 1, 2014); Prosecutor v. Ntaganda, Case No. ICC-01/04-02/06-2359, Judgment, ¶ 774 (July 8, 2019); *see* Ambos, *supra* note 47, at 1204 (highlighting that this theoretical approach reflects the German doctrine of *funktionelle Tatherrschaft* (functional control over the crime)).

49. Elies van Sliedregt & Lachezar Yanev, *Co-Perpetration Based on Joint Control Over the Crime*, *in* MODES OF LIABILITY IN INTERNATIONAL CRIMINAL LAW 85, 118 (Jérôme de Hemptinne, Robert Roth & Elies van Sliedregt eds., 2019); *see also* Bo, *supra* note 44, at 159.

50. Prosecutor v. Lubanga, Case No. ICC-01/04-01/06, Decision on the Confirmation of Charges, ¶¶ 343–45 (Jan. 29, 2007); Prosecutor v. Lubanga, Case No. ICC-01/04-01/06-2842, Judgment, ¶¶ 980–88 (Mar. 14, 2012); *see also* Ambos, *supra* note 47, at 1205. On whether the plan or agreement itself needs to contain one or more crimes, rather than simply an "element of criminality," *see id.* at 1205–6.

51. Prosecutor v. Bemba, Case No. ICC-01/05-01/08, Decision on the Confirmation of Charges, ¶ 370 (June 15, 2009); Prosecutor v. Lubanga, Case No. ICC-01/04-01/06-3121-Red, Appeals Chamber Judgment, ¶ 447 (Dec. 1, 2014).

52. Prosecutor v. Katanga, Case No. ICC-01/04-01/07-717, Decision on the Confirmation of Charges, ¶¶ 524–25 (Sept. 30, 2008); Prosecutor v. Lubanga, Case No. ICC-01/04-01/06-3121-Red, Appeals Chamber Judgment, ¶ 469 (Dec. 1, 2014).

53. Prosecutor v. Lubanga, Case No. ICC-01/04-01/06, Decision on the Confirmation of Charges, ¶¶ 349–67 (Jan. 29, 2007); Prosecutor v. Katanga, Case No. ICC-01/04-01/07-717, Decision on the Confirmation of Charges, ¶¶ 533–39 (Sept. 30, 2008); Prosecutor v. Ntaganda, Case No. ICC-01/04-02/06-309, Decision Pursuant to Article 61(7)(a) and (b) of the Rome Statute on the Charges of the Prosecutor Against Bosco Ntaganda, ¶ 121 (June 9, 2014); Prosecutor v. Gbagbo, Case No. ICC-02/11-01/11-656-Red, Decision on the Confirmation of Charges Against Laurent Gbagbo, ¶ 240 (June 12, 2014); *see also* van Sliedregt & Yanev, *supra* note 49, at 98–100.

cases where multiple actors, such as developers, manufacturers, commanders, and/or end users collaborate in the development and deployment of AI systems for criminal purposes.[54] For instance, let us imagine that in Scenario 2, a data analyst and a software engineer intentionally collaborated with MG to deploy the AI-powered targeting system. Imagine if, in this hypothetical scenario, the data analyst manipulated data to conceal false positives in the AI's functioning, while the software engineer overrode safeguards, and MG authorized the strike, resulting in mass civilian casualties. All three individuals have intentionally given an essential contribution to a common plan virtually certain to result in the commission of a crime.

Of course, the main challenge in establishing the criminal responsibility of humans who employ AI systems remains that of intent and, more broadly, the mens rea requirement. For this reason, some scholars have sought to repurpose a specific doctrine of joint perpetration known as "joint criminal enterprise" (JCE)—which, in certain circumstances, allows for the punishment of crimes that were not planned but were foreseeable consequences of a particular course of conduct.

Not mentioned in the ICC Statute and eschewed by ICC judges, the doctrine of JCE was first conceptualized in international criminal law—finding its basis in customary international law—through the jurisprudence of the ICTY, particularly in *Prosecutor v. Tadić*, with the aim of ensuring that all individuals who contribute to a common criminal plan are held responsible for resulting crimes.[55] In the context of AI-enabled crimes, resort to JCE could be hypothesized to establish criminal responsibility for those involved in the development, deployment, and operational use of AI systems. While JCE has been theorized in three different versions (basic, systemic, and extended), it is the "extended" version (also known as "JCE III"—whose basis in customary international law has been seriously questioned)[56] that has garnered the most attention in capturing the criminality of human actors who

54. *See, e.g.*, Thompson Chengeta, *Accountability Gap: Autonomous Weapon Systems and Modes of Responsibility in International Law*, 45 DENVER JOURNAL OF INTERNATIONAL LAW & POLICY 1, 21 (2016); Lewis, *supra* note 14, at 974–75.

55. Prosecutor v. Tadić, Case No. IT-94-1-A, Appeals Chamber Judgment, ¶¶ 185–229 (Int'l Crim. Trib. for the former Yugoslavia, July 15, 1999). For the ICC judges' approach, *see* Prosecutor v. Lubanga, Case No. ICC-01/04-01/06, Decision on the Confirmation of Charges, ¶¶ 322–23 (Jan. 29, 2007). For a detailed analysis, *see* Lachezar Yanev, *Joint Criminal Enterprise*, *in* MODES OF LIABILITY IN INTERNATIONAL CRIMINAL LAW 121 (Jérôme de Hemptinne, Robert Roth & Elies van Sliedregt eds., 2019); *see also* AMBOS, *supra* note 1, at 186–92.

56. Yanev, *supra* note 55, at 164–66.

interact with autonomous AI systems and, in so doing, bypass mens rea problems. Any version of JCE requires the existence of a common plan shared between two or more persons—a plan that involves or amounts to the commission of one or more crimes ("common criminal plan").[57] In any JCE, participants do not need to physically carry out any of the elements of the criminal offense, but are deemed criminally responsible insofar as they provided a significant contribution to the execution of the shared common plan.[58]

JCE III contemplates the possibility that a crime is committed outside the scope of the common plan: in this case, any participant in the JCE would be responsible for that crime too, provided it was a "natural and foreseeable consequence of the execution of the plan."[59] Thus, it has been suggested that JCE III could, in theory, capture the criminality of developers, commanders, and users who employ AI systems as part of a common criminal plan, even where the autonomy of the AI system results in a crime outside of the agreed plan—so long as such "non-planned" crime is a natural and foreseeable consequence of that plan.

A major hurdle in applying JCE to AI-enabled crimes—as it is the case for "co-perpetration by control over the crime"—is proving the existence of a common criminal plan. This may be relatively straightforward in certain cases of mass atrocities, where all members of the enterprise share the intent to commit crimes. However, in AI-related cases, the situation is more complex. Developers, military officers, and operators may have varying levels of knowledge and involvement, making it difficult to prove that they shared the same plan—particularly since some developers carry out their "significant

57. *Id.* at 137–39. Among others, *see* Prosecutor v. Brđanin, Case No. IT-99-36-A, Appeals Chamber Judgment, ¶ 418 (Int'l Crim. Trib. for the Former Yugoslavia, Apr. 3, 2007).

58. Yanev, *supra* note 55, at 140–41; *see, e.g.*, Prosecutor v. Gotovina et al., Case No. IT-06-90-T, Judgment, ¶ 1953 (Int'l Crim. Trib. for the Former Yugoslavia, Apr. 15, 2011); Prosecutor v. Karadžić, Case No. IT-95-5/18-T, Judgment, ¶ 564 (Int'l Crim. Trib. for the Former Yugoslavia, Mar. 24, 2016).

59. Prosecutor v. Kvočka et al., Case No. IT-98-30/1-A, Appeals Chamber Judgment, ¶ 86 (Int'l Crim. Trib. for the Former Yugoslavia, Feb. 28, 2005); Yanev, *supra* note 55, at 148–49 discusses how this standard of foreseeability has been interpreted to include both objective and subjective dimensions. For an example of a critique of this mens rea standard from the perspective of its alleged incompatibility with the principle of culpability, *see* AMBOS, *supra* note 1, at 249–52. The customary status of JCE III has been questioned, inter alia, by Prosecutor v. Khieu et al., Case No. 002/19-09-2007-ECCC/OCIJ, Decision on the Appeals Against the Co-Investigating Judges' Order on Joint Criminal Enterprise, ¶ 77 (Extraordinary Chambers in the Courts of Cambodia, May 20, 2010).

contribution" months or even years before the AI system is deployed or engages in prima facie criminal conduct.[60] Moreover, one of the key requirements of JCE III is that the crime outside the common purpose must have been a "natural and foreseeable" consequence of the joint criminal enterprise. This creates difficulties in cases involving AI, as its unpredictability often raises questions about whether resulting crimes can truly be foreseen.[61] Some developers and commanders may argue that they lacked the knowledge to anticipate how the AI system in question would be used or would behave.[62]

More importantly, for doctrinal accuracy, it should be noted that JCE III still requires that a crime outside the common plan is committed. If the conduct in question is entirely attributable to a machine with no legal personality, or if the mental element required for the non-planned crime is otherwise absent, I remain unpersuaded that criminal responsibility via JCE can ensue. Moreover, as noted above, the doctrine is not enshrined in the ICC Statute and is unlikely to appear in many States' implementing legislation.

V. HUMANS INSTIGATING BY MEANS OF AI THE COMMISSION OF CRIMES

Instigation, as a mode of accessorial liability, applies to individuals who prompt, encourage, or urge others to commit crimes.[63] It is expressly provided for in Article 25(3)(b) of the ICC Statute, which establishes criminal responsibility for those who order, solicit, or induce the commission of a crime, and is similarly contemplated in the statutes of other international criminal tribunals, such as in Article 7(1) of the ICTY Statute.[64] In the context of AI-enabled crimes, instigation may capture the criminal responsibility of human actors who incite or provoke others into criminal activity through

60. Winter, *supra* note 41, at 64–65, 70.

61. *Id.* at 71.

62. *See id.* at 77; Beard, *supra* note 9, at 661.

63. Antonio Coco, *Instigation*, *in* MODES OF LIABILITY IN INTERNATIONAL CRIMINAL LAW 257, 257 (Jérôme de Hemptinne, Robert Roth & Elies van Sliedregt eds., 2019).

64. Although ordering is listed under Article 25(3)(b), Ambos has argued that it should more accurately be conceptualized as a form of (indirect) perpetration. AMBOS, *supra* note 1, at 235–36; Ambos, *supra* note 47, at 1216–17. Should one subscribe to Ambos's view, indirect perpetration could be the doctrine used to establish the criminal responsibility of a person who orders conduct—whether involving an AI-enabled system or the deployment of an AWS—that prima facie constitutes a crime. In that case, the analysis set out in Section IV(B) of this contribution would be relevant.

AI systems, including developers, operators, or users who deliberately design or employ such systems to promote forms of violence, discrimination, or other unlawful acts.

Instigation may be carried out by either act or omission, directly or indirectly, personally or through an intermediary.[65] This means that an individual may be held liable for instigating crimes not only through explicit encouragement but also by creating or deploying AI tools that predictably provoke criminal activity. For example, an AI-powered recommendation algorithm that amplifies hate speech or extremist propaganda—such as in the case of SE in Scenario 3 above—may constitute instigation if deliberately designed or manipulated to incite violence.

Being a mode of accessorial liability, instigation is only punishable when it results in the commission or attempted commission of a crime.[66] This requirement ensures that liability does not extend to mere advocacy or abstract expressions but applies where the instigator's actions have a tangible impact on the criminal outcome. No formal relationship between the instigator and the perpetrator is required,[67] meaning that AI system designers or operators could be liable even if they have no direct connection to the individuals who ultimately commit the offenses.

A crucial requirement for instigation to entail criminal responsibility is that it provides a substantial contribution to the crime, meaning that the instigation must have actually influenced the perpetrator's decision to commit the offense.[68] If an individual was already fully resolved to commit a crime before any instigating act occurred, no liability for instigation arises.[69] However, when AI-driven amplification of inflammatory content or misinfor-

65. Coco, *supra* note 63, at 267–69.

66. *Id.* at 266.

67. Prosecutor v. Brđanin, Case No. IT-99-36-T, Judgment, ¶ 359 (Int'l Crim. Trib. for the Former Yugoslavia, Sept. 1, 2004); Prosecutor v. Karadžić, Case No. IT-95-5/18-T, Judgment, ¶ 572 (Int'l Crim. Trib. for the Former Yugoslavia, Mar. 24, 2016).

68. Prosecutor v. Kordić & Čerkez, Case No. IT-95-14/2-A, Appeals Chamber Judgment, ¶ 27 (Int'l Crim. Trib. for the Former Yugoslavia, Dec. 17, 2004); Prosecutor v. Karadžić, Case No. IT-95-5/18-T, Judgment, ¶ 572 (Int'l Crim. Trib. for the Former Yugoslavia, Mar. 24, 2016); Prosecutor v. Akayesu, Case No. ICTR-96-4-T, Judgment, ¶¶ 481–82 (Int'l Crim. Trib. for Rwanda, Sept. 2, 1998); *cf.* Prosecutor v. Ntaganda, Case No. ICC-01/04-02/06-309, Decision on the Confirmation of Charges, ¶ 155 (June 9, 2014); Prosecutor v. Bemba, Case No. ICC-01/05-01/13-1989-Red, Judgment, ¶ 81 (Oct. 19, 2016). For a discussion, *see* Coco, *supra* note 63, at 269–71.

69. Prosecutor v. Orić, Case No. IT-03-68-T, Judgment, ¶ 271 (Int'l Crim. Trib. for the Former Yugoslavia, June 30, 2006); Coco, *supra* note 63, at 271.

mation plays a significant role in radicalizing or inciting individuals to commit violence, liability may be triggered provided that evidence, even circumstantial, demonstrates a substantial contribution to the crime.[70]

With respect to the mental element, instigation requires that the instigator intended to perform the instigating conduct, knowing of its influencing effect on the final perpetrator.[71] This means that liability does not depend on the instigator sharing the perpetrator's intent but rather on their awareness that their conduct would likely provoke or induce criminal actions. Given the complexity of AI-driven instigation, courts may need to rely on indirect evidence, such as internal communications, algorithmic design choices, or data patterns showing deliberate manipulation to incite violence or discrimination.

Moreover, depending on the contours of the applicable law, instigation may also arise when the instigator was aware of the substantial likelihood that their conduct would lead to the commission of a crime, even if they did not specifically intend that crime.[72] This is particularly relevant in AI-related scenarios where developers or operators may argue that they did not directly intend specific violent outcomes, yet were aware that their actions significantly increased the risk of such outcomes.

Additionally, instigation liability applies when the instigator was aware of the essential elements of the perpetrated crime, even if they did not know its precise legal classification, exact circumstances, or the identity of the final perpetrator.[73] This broadens liability to cases where individuals knowingly contribute to an AI system's harmful effects without requiring them to foresee every specific detail of the resulting crimes.

---

70. Notably, when an AI system is designed or developed to directly and publicly incite genocide, this would often constitute a criminal offense in itself, pursuant to provisions such as Article 25(3)(e) of the ICC Statute.

71. Prosecutor v. Muvunyi, Case No. ICTR-2000-55A-T, Judgment and Sentence, ¶ 465 (Int'l Crim. Trib. for Rwanda, Sept. 12, 2006); Prosecutor v. Ntaganda, Case No. ICC-01/04-02/06-309, Decision on the Confirmation of Charges, ¶ 153 (June 9, 2014); Prosecutor v. Gbagbo, Case No. ICC-02/11-01/11-656-Red, Decision on the Confirmation of Charges, ¶ 244 (June 12, 2014); Prosecutor v. Bemba, Case No. ICC-01/05-01/13-1989-Red, Judgment, ¶ 82 (Oct. 19, 2016).

72. Prosecutor v. Kordić & Čerkez, Case No. IT-95-14/2-A, Appeals Chamber Judgment, ¶ 32 (Int'l Crim. Trib. for the Former Yugoslavia, Dec. 17, 2004); Prosecutor v. Orić, Case No. IT-03-68-T, Judgment, ¶ 279 (Int'l Crim. Trib. for the Former Yugoslavia, June 30, 2006).

73. Prosecutor v. Kamuhanda, Case No. ICTR-99-54A-T, Judgment, ¶ 599 (Int'l Crim. Trib. for Rwanda, Jan. 22, 2004); Prosecutor v. Orić, Case No. IT-03-68-T, Judgment, ¶ 279 (Int'l Crim. Trib. for the Former Yugoslavia, June 30, 2006).

VI. AIDING AND ABETTING, AND ASSISTANCE, IN THE CONTEXT OF AI-ENABLED INTERNATIONAL CRIMES

Aiding and abetting, a mode of liability attaching to the conduct of those who assist or lend moral support to a crime, may be relevant in the case of crimes committed by or with the use of AI systems, potentially capturing the criminality of human actors who contributed to the creation, deployment, or operation of those AI systems. Aiding and abetting liability is commonly envisaged in systems of criminal law, exists as a matter of customary international law, and is provided in Article 25(3)(c) of the ICC Statute and Article 7(1) of the ICTY Statute.[74]

As a matter of objective element, if one follows the case law of ad hoc and hybrid tribunals, aiding and abetting liability requires that the actor's conduct—by either commission or omission—constitutes a substantial contribution to, or has had a substantial effect on, the commission of a crime.[75] No prior plan or agreement is required,[76] and the contribution may be remote from the actual crime both geographically and temporally.[77] In the context of AI, this means that a developer who creates a weapon system with the potential for unlawful use could still be held responsible, even if they are far removed from the actual deployment of the system.[78]

---

74. *See* Manuel J. Ventura, *Aiding and Abetting*, *in* MODES OF LIABILITY IN INTERNATIONAL CRIMINAL LAW 173 (Jérôme de Hemptinne, Robert Roth & Elies van Sliedregt eds., 2019).

75. *E.g.*, Prosecutor v. Tadić, Case No. IT-94-1-T, Opinion and Judgment, ¶¶ 674, 688–92 (Int'l Crim. Trib. for the Former Yugoslavia, May 7, 1997); Prosecutor v. Aleksovski, Case No. IT-95-14/1-T, Judgment, ¶¶ 60–61 (Int'l Crim. Trib. for the Former Yugoslavia, June 25, 1999); Prosecutor v. Furundžija, Case No. IT-95-17/1-A, Judgment, ¶ 126 (Int'l Crim. Trib. for the Former Yugoslavia, July 21, 2000); *see also* Ventura, *supra* note 74, at 186–87. Ventura explains that while the case law of the ICC is less clear on the matter, there are persuasive reasons for reading a threshold of "substantial contribution/effect" for the material element of aiding and abetting into Article 25(3)(c) of the ICC Statute too. *Id.* at 208–12. In this sense, *see also* Ambos, *supra* note 47, at 1222–23.

76. *See, e.g.*, Prosecutor v. Tadić, Case No. IT-94-1-A, Appeals Chamber Judgment, ¶ 229(ii) (Int'l Crim. Trib. for the Former Yugoslavia, July 15, 1999); Prosecutor v. Simić et al., Case No. IT-95-9-T, Judgment, ¶ 162 (Int'l Crim. Trib. for the Former Yugoslavia, Oct. 17, 2003).

77. *See, e.g.*, Prosecutor v. Delalić et al., Case No. IT-96-21-A, Appeals Chamber Judgment, ¶ 352 (Int'l Crim. Trib. for the Former Yugoslavia, Feb. 20, 2001); Prosecutor v. Blaškić, Case No. IT-95-14-A, Judgment, ¶ 48 (Int'l Crim. Trib. for the Former Yugoslavia, July 29, 2004).

78. BO, BRUUN & BOULANIN, *supra* note 21, at 38.

The required mental element for aiding and abetting liability is, however, more controversial. Article 25(3)(c) of the ICC Statute includes language indicating that the aider and abettor must act "for the purpose of facilitating" the commission of the crime, meaning that an intention to assist the crime in question is required for criminal responsibility to arise.[79] It may well be that such a purposive requirement applies only before the ICC while, in other legal systems (for instance, within the framework of customary international law),[80] aiding and abetting liability may arise whenever the actor deliberately carries out the assisting conduct with knowledge that they are assisting the perpetration of a crime.[81] For the ICTY Appeals Chamber in *Tadić*, in this respect, "awareness . . . of the essential elements of the crime committed by the principal would suffice."[82]

The fact that mere knowledge of assisting a crime may suffice is particularly relevant in the context of AI-enabled offenses, as it helps mitigate the challenges of establishing intent discussed in earlier sections. However, challenges persist even with respect to knowledge, which may be difficult to prove given the complexity of AI systems and the unpredictability of their behavior.[83] In Scenario 4, for instance, IO's conduct may give rise to aiding and abetting liability in that it effectively facilitated human rights violations that could amount—among other possibilities—to the crime against humanity of persecution. IO did not necessarily intend to engage in such a crime or act with the purpose of assisting it, but they were aware of the system's biases and its concrete potential for misuse in the commission of criminal offenses.

---

79. In detail, *see* Ventura, *supra* note 74, at 213–17; *see also* Ambos, *supra* note 47, at 1225. Such a purposive requirement does not mean that the aider and abettor must share the perpetrator's mens rea. For example, where a crime includes a special intent element—as in the case of genocide—it is sufficient that the aider and abettor is aware of the perpetrator's special intent. *See* AMBOS, *supra* note 1, at 240.

80. Prosecutor v. Taylor, Case No. SCSL-03-01-A, Appeals Chamber Judgment, ¶ 435 (Special Court for Sierra Leone, Sept. 26, 2013).

81. Prosecutor v. Blaškić, Case No. IT-95-14-A, Appeals Chamber Judgment, ¶ 46 (Int'l Crim. Trib. for the Former Yugoslavia, July 29, 2004); Prosecutor v. Taylor, Case No. SCSL-03-01-A, Appeals Chamber Judgment, ¶ 436 (Special Court for Sierra Leone, Sept. 26, 2013); Prosecutor v. Šainović et al., Case No. IT-05-87-A, Appeals Chamber Judgment, ¶ 1649 (Int'l Crim. Trib. for the Former Yugoslavia, Jan. 23, 2014); *see also* Antonio Coco & Tom Gal, *Losing Direction: The ICTY Appeals Chamber's Controversial Approach to Aiding and Abetting in Perišić*, 12 JOURNAL OF INTERNATIONAL CRIMINAL JUSTICE 345, 353–54 (2014).

82. Prosecutor v. Tadić, Case No. IT-94-1-A, Appeals Chamber Judgment, ¶ 164 (Int'l Crim. Trib. for the Former Yugoslavia, July 15, 1999); *see also* Prosecutor v. Orić, Case No. IT-03-68-T, Judgment, ¶ 288 (Int'l Crim. Trib. for the Former Yugoslavia, June 30, 2006).

83. Beard, *supra* note 9, at 649, 651.

In a case like this, establishing criminal responsibility for aiding and abetting would turn on whether the applicable law requires the purpose (or intent) to facilitate the crime in question.

One further difficulty is that, for aiding and abetting liability to materialize, the assistance must result in a crime. As highlighted above in the case of joint criminal enterprise, unless legal personality and the capacity to form mens rea are attributed to machines, no such crime would exist if the prima facie criminal conduct had been entirely performed by a fully autonomous system. The case in Scenario 4 is different: IO helps develop an AI system, which in turn facilitates the commission of crimes by humans. Even if one step removed, IO's act of assistance contributes to fully executed crimes.

Related to aiding and abetting is the mode of liability enshrined in Article 25(3)(d) of the ICC Statute, which addresses contribution to a crime committed by a group acting with a common purpose. Under this provision, a person is criminally responsible if their contribution is intentional and is made either (i) with the aim of furthering the criminal activity or criminal purpose of the group, or (ii) in the knowledge of the intention of the group to commit the crime. Under option (i), it is sufficient that the contributor is aware of the group's criminal activity or purpose and, consequently, intentionally furthers it without necessarily knowing the specific crime that will result from their contribution.[84]

This flexibility could prove particularly significant for crimes involving the use of AI systems. For instance, a tech supplier who knowingly provides or maintains an AI system for a client—despite being aware that the client intends to use it for criminal activity—may fall within the scope of Article 25(3)(d). What matters is the intentional contribution coupled with awareness of the group's criminal purpose. Nevertheless, this mode of liability still presupposes the existence of an identifiable group of human perpetrators acting with a shared purpose. The language of the provision—referring explicitly to a "group of persons"—suggests that attribution of responsibility under 25(3)(d) may not be easily transposed to scenarios in which an AI system operates with high levels of autonomy and without clear, coordinated human direction.

84. In this sense, *see* Prosecutor v. Katanga, Case No. ICC-01/04-01/07-3436, Minority Opinion of Judge Christine Van den Wyngaert, ¶ 288 (Mar. 7, 2014); *see also* Ambos, *supra* note 47, at 1231.

VII. SUPERIOR RESPONSIBILITY AND AI-ENABLED CRIMES

The doctrine of superior responsibility, existing in customary international law and enshrined in statutes of international criminal tribunals such as in Article 28 of the ICC Statute and Article 7(3) of the ICTY Statute, establishes the criminal responsibility of military commanders and civilian superiors for crimes committed by their subordinates when they knew or had reason to know of those crimes and failed to prevent them or punish the perpetrators.[85]

Superior responsibility consists of three key elements. First, there must be a superior-subordinate relationship, meaning that the superior exercised effective control over the individual(s) who committed the crime. Second, the superior must have known or had reason to know that the crime was about to be committed or had already been committed. Third, the superior must have failed to take necessary and reasonable measures to prevent the crime or punish those responsible.[86] Importantly, the duties to prevent and punish are distinct, meaning that liability can arise from the failure to perform either obligation.[87] It remains controversial whether, under the legal framework of the ICC Statute, a fourth requirement exists—namely, that the superior's failure to act must have a causal effect on the crime(s) committed by the subordinate(s).[88]

Superior responsibility has been discussed as a potential avenue for ensuring accountability concerning prima facie criminal conduct carried out by

85. For a general examination, *see* Miles Jackson, *Command Responsibility*, *in* MODES OF LIABILITY IN INTERNATIONAL CRIMINAL LAW 409 (Jérôme de Hemptinne, Robert Roth & Elies van Sliedregt eds., 2019); Roberta Arnold & Miles Jackson, *Article 28: Responsibility of Commanders and Other Superiors*, *in* ROME STATUTE OF THE INTERNATIONAL CRIMINAL COURT: ARTICLE-BY-ARTICLE COMMENTARY (Kai Ambos ed., 4th ed. 2022).

86. Prosecutor v. Delalić et al., Case No. IT-96-21-T, Judgment, ¶ 344 (Int'l Crim. Trib. for the Former Yugoslavia, Nov. 16, 1998); Prosecutor v. Blaškić, Case No. IT-95-14-T, Judgment, ¶ 294 (Int'l Crim. Trib. for the Former Yugoslavia, Mar. 3, 2000); Prosecutor v. Bemba, Case No. ICC-01/05-01/08, Appeals Chamber Judgment, ¶ 167 (June 8, 2018).

87. Prosecutor v. Delić, Case No. IT-04-83-T, Judgment, ¶ 69 (Int'l Crim. Trib. for the Former Yugoslavia, Sept. 15, 2008); Prosecutor v. Bemba, Case No. ICC-01/05-01/08, Decision on the Confirmation of Charges, ¶ 436 (June 15, 2009); Jackson, *supra* note 85, at 430; Arnold & Jackson, *supra* note 85, at 1311.

88. As admitted by two ICC judges in Prosecutor v. Bemba, Case No. ICC-01/05-01/08, Separate Opinion of Judges Van den Wyngaert and Morrison, ¶ 51 (June 8, 2018); *see also* Arnold & Jackson, *supra* note 85, at 1299–1306; AMBOS, *supra* note 1, at 301–6.

fully autonomous systems.[89] However, the most fundamental issue with this approach is whether AI systems can be considered "subordinates" given that they lack independent agency, intent, or moral responsibility. Control over a weapon, or an algorithm—whether autonomous or not—differs qualitatively from control over individuals in the context of command responsibility. AI systems remain tools rather than legal or moral agents, making their classification as subordinates problematic.[90] Furthermore, formulations of the doctrine require that subordinates commit crimes: as previously explained, only legal persons capable of forming mens rea are—at present—able to do so. In Scenario 2, MG's failure to prevent foreseeable targeting of civilians could potentially be scrutinized under superior responsibility. However, because the AWS itself is not a legal person and lacks criminal intent, it is unclear whether the doctrine would apply in the same way it would for human subordinates.

Despite this doctrinal challenge, some scholars argue that superior responsibility could still apply in cases where commanders and civilian superiors fail to prevent crimes perpetrated by means of AI systems under their control. A superior, indeed, can still be held responsible for how AI is used by human subordinates. If subordinates misuse the AI in a way that results in criminal conduct, the superior may incur liability under the doctrine of superior responsibility provided they knew, or should have known, about the misuse and failed to prevent or repress it. And yet, the application of the superior responsibility doctrine would not be straightforward on the point of fact. Under Article 28(a)(i) of the ICC Statute, military commanders are responsible when they "knew or, owing to the circumstances at the time, should have known" that their subordinates were committing crimes.[91] Since AI systems operate on machine-learning principles and may develop novel behavioral patterns, it is unclear whether superiors should be expected to anticipate and prevent all potential malfunctions. Much may depend on how stringently the "should have known" standard is applied in practice—whether it demands a high degree of foresight and technical understanding, or it allows instead for a more flexible assessment of reasonableness in the

89. *See, e.g.*, Russell Buchan & Nicholas Tsagourias, *Autonomous Cyber Weapons and Command Responsibility*, *in* AUTONOMOUS CYBER CAPABILITIES UNDER INTERNATIONAL LAW 321 (Rain Liivoja & Ann Väljataga eds., 2021).

90. Alessandra Spadaro, *A Weapon Is No Subordinate: Autonomous Weapon Systems and the Scope of Superior Responsibility*, 21 JOURNAL OF INTERNATIONAL CRIMINAL JUSTICE 1119, 1127 (2023).

91. Weigend, *supra* note 5, at 1151.

face of autonomous system unpredictability. One could go as far as imagining a form of strict liability, making commanders accountable for any AI system failure, regardless of intent or oversight. While lawmakers may ultimately choose to take this approach, interpreting existing law in this way could be seen as a stretch.

Another challenge is the control requirement. Traditional applications of this doctrine assume that a commander has effective control over subordinates, that is, the "material ability to prevent or repress the commission of the crimes or to submit the matter to the competent authorities."[92] However, AI systems may operate autonomously and beyond real-time human oversight, making it harder to establish that there was effective control.

One more consideration can be made on the basis of the fact that, often, commentators inquire whether superior responsibility should be classified as a stand-alone "failure-to-act" (or "dereliction of duty") offense, rather than a mode of liability.[93] If this interpretation were adopted, superior responsibility could evolve into a broader legal duty to take feasible measures with a view to preventing or repressing certain crimes—a duty that would be applicable even in cases where AI systems are involved. However, such a shift could require significant legal development and reconsideration of the doctrine's foundations.

VIII. CONCLUSION

Current international criminal law offers several pathways for establishing individual responsibility for AI-enabled crimes, particularly through modes of liability such as perpetration, instigation, and aiding and abetting. These frameworks are well-suited to cases where human actors deliberately use AI systems as tools to commit or facilitate international crimes—for example, through data manipulation or AI-driven hate speech. However, as AI systems grow more autonomous and complex, establishing the requisite mens rea becomes increasingly difficult. At a time when States, international organizations, civil society, and industry representatives are actively discussing

92. Prosecutor v. Bemba, Case No. ICC-01/05-01/08, Judgment, ¶ 183 (Mar. 21, 2016); *see also* Prosecutor v. Delalić et al., Case No. IT-96-21-A, Appeals Chamber Judgment, ¶ 256 (Int'l Crim. Trib. for the Former Yugoslavia, Feb. 20, 2001); Arnold & Jackson, *supra* note 85, at 1293–97.

93. *See, e.g.*, Arnold & Jackson, *supra* note 85, at 1304–6; *contra*, DARRYL ROBINSON, JUSTICE IN EXTREME CASES: CRIMINAL LAW THEORY MEETS INTERNATIONAL CRIMINAL LAW 220 (2020).

or negotiating potential rules for the governance of AI technologies, it may be worth considering how the rules on modes of liability could be adapted to ensure human accountability—including, where necessary, by re-conceptualizing certain modes or introducing lower mental element thresholds.