Emmanuel Q. Fernando Philosophy
Undergraduate Conference

Analysis: Contemporary Perspectives on
Classic Problems of Philosophy Conference

WHEN MACHINES BREAK THE LAW: THE CASE FOR CRIMINAL LIABILITY & LEGAL PERSONHOOD FOR ARTIFICIAL INTELLIGENCE

MORALES, Janielle Vhon A.
jamorales8@up.edu.ph

Department of Philosophy
College of Social Sciences and Philosophy
University of the Philippines – Diliman

ABSTRACT:

If law assigns responsibility where harm is done, then to exempt Artificial Intelligence (AI) from liability is to create a void where power exists without answerability. When such systems cause harm, should the law remain indifferent to their agency? As artificial intelligence (AI) increasingly influences critical domains like transportation, healthcare, finance, and criminal justice, existing legal doctrines—rooted in human-centric notions of agency and causation—struggle to address harms caused by autonomous systems. This paper argues for extending criminal liability and legal personhood to AI by integrating philosophical theories of causation with contemporary legal frameworks. Drawing on Alexander Kaiserman's Partial Liability theory and modal semantics, alongside Hart and Honoré's distinctions on causation, we propose a two-dimensional model of AI accountability based on both production and dependence measures. Through restricted possible-worlds models, causal contributions are quantified, enabling courts to apportion liability among datasets, algorithms, human operators, and AI entities themselves.

Using precedents from corporate personhood and recent European legislative initiatives like the AI Act and GDPR's right-to-explanation, the study advances the concept of an AI Registry and mandatory algorithmic audits. Structural causal models by Halpern and Pearl further enable provenance tracking and counterfactual simulations, operationalizing liability attribution. Empirical cases—such as COMPAS's predictive bias and AI misdiagnoses—illustrate the urgent need for such frameworks.

The paper distills its framework into four principles: Volition (V), Intent (I), Culpability (C), and graded Autonomy and Causation (AC), offering a scalable standard for legal recognition of AI harms. If AI is permitted to act with impunity, law itself becomes obsolete, and justice is surrendered to the unchecked rule of machines.

Keywords: Artificial Intelligence Liability, Causation Theory, Criminal Liability, Legal Personhood, Algorithmic Accountability

Over the past decade, artificial intelligence (AI) has graduated from a niche research endeavor to an omnipresent force shaping life-and-death decisions in areas in transportation, medicine, finance, and criminal justice. Autonomous vehicles navigate split-second moral choices, algorithmic diagnostic tools influence clinical outcomes, and predictive-policing models determine who enters pretrial detention. Yet our legal doctrines remain anchored in centuries-old conceptions of agency and causation premised on human deliberation, intent, and proximate physical acts that strain to accommodate *AI*'s distributed architectures,[1] opaque learning loops[2], and probabilistic reasoning.[3] When an *AI*-driven decision inflicts harm, courts regulators lack legal metrics to trace responsibility through layers of data preprocessing, model training, and dynamic retraining. The doctrinal–technological mismatch both denies victims redress and leaves developers without a clear compass in designing, auditing, and governing *AI* systems processes responsibly.

To bridge this gap, a framework that marries the metaphysical subtleties of philosophical causation theory with the inherent procedural rigor that comes with legal practice ought to be rectified. By integrating Alexander Kaiserman's counterfactual Partial Liability theory[4] with his modal semantics of causation[5] and Hart & Honoré's proximate versus effective causation distinction, we move beyond all-or-nothing fault to a graded, two-dimensional model. Kaiserman's Partial Liability theory argues that liability can be apportioned fractionally among multiple causes based on counterfactual contributions rather than a binary fault assignment, whereas his modal semantics of causation that uses possible-world semantics formalize how interventions on causal variables propagate through models while providing a metaphysical basis for graded causation. Under this proposed paradigm, each causal actor—whether data set, algorithmic module, human operator, or the *AI* itself—receives a quantified share of liability based on both "*difference-making*" and "*dependence*" metrics. When coupled with EU's General Data Protection Regulation's (GDPR) right-to-explanation and Halpern & Pearl's structural causal models, this approach yields transparent, measurable criteria that dovetail with existing tort, product-liability, and criminal doctrines. The approach calibrates liability precisely to each actor's causal contribution.[6]

We begin by reconciling production and dependence intuitions: while production measures quantify how much a component contributes to an outcome, dependence measures assess whether harm would occur absent that component. Because these dimensions often conflict, we employ a restricted possible-worlds model to assign numerical weights to causal claims[7]. Restricted possible-worlds models constrain counterfactual analysis to worlds differing minimally from the actual iteration with causal weights that reflect plausible alterations rather

[1] David Carrera et al., "State of the Art in Parallel and Distributed Systems: Emerging Trends," *Electronics* 14, no. 4 (2023): 677–700. "Distributed systems power applications like global-scale content delivery networks and decentralized finance by distributing computation and data across multiple nodes to improve scalability, fault tolerance, and security."

[2] Fiona Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms," *Big Data & Society* 3, no. 1 (2016). "Opacity in algorithmic systems stems from complex model architectures and proprietary design choices, rendering internal decision pathways opaque even to expert auditors."

[3] Google Cloud, "Continuous Training and MLOps Best Practices," *Google Cloud Blog*, June 15, 2021. "Continuous training cycles retrain deployed machine learning models based on live feedback, necessitating updated governance frameworks to account for model drift and emergent behavior."

[4] Alex Kaiserman, "Partial Liability," Legal Theory 23, no. 1 (2017): 3. "A defendant should be held liable for a claimant's loss only to the degree to which the defendant's wrongdoing contributed to the causing of the loss."

[5] Alex Kaiserman, "Partial Liability," Legal Theory 23, no. 1 (2017): 12. "In section 5, I argue that proximate causation doctrines ought to be replaced with a more fine-grained approach which recognises the possibility of partial liability for losses."

[6] Alex Kaiserman, "Necessary Connections in Context," Erkenntnis 82, no. 1 (2017): 47. "Production intuitions focus on difference-making—how much an action increases the risk of harm—whereas dependence intuitions measure how indispensable that action is to bringing about the outcome."

[7] Alex Kaiserman, "Necessary Connections in Context," Erkenntnis 82, no. 1 (2017): sec. 2–3. "Section 2 looks at what I call 'dependence measures,' which arise from thinking of causes as difference-makers, while Section 3 looks at what I call 'production measures,' which arise from thinking of causes as jointly sufficient for their effects."

than extreme hypotheticals. To operationalize this for *AI*, we integrate H.L.A. Hart and Tony Honoré's proximate versus effective causation concepts[8]—enabling courts to trace non-human decision processes back to responsible human or artificial agents.

The doctrines of corporate juridical personhood furnishes a somewhat paradigmatic example of how rights and liabilities can be attached to non-natural entities.[9] Recent European Parliament proposals for electronic personhood suggest a template for AI registration[10]. The EU's forthcoming *AI* Act mandates a public registry for high-risk *AI* systems to open the potentiality for traceability and to facilitate post-incident liability assessment. Algorithmic accountability under GDPR's right-to-explanation for data subjects affected by fully automated decisions (Art. 22, Recital 71), obliging controllers to provide meaningful data regarding the logic involve for effective human intervention, goes hand-in-hand with technical frameworks by Doshi-Velez & Kortz's model agnostic, post-hoc methods (e.g., LIME, SHAP) that generales explanation without exposing internal weights;[11] Wachter et al. counterfactual explanation compatible with GDPR's notion of meaningful information;[12] and Malgieri & Comandé's algorithmic legibility paradigm,[13] offers interpretability standards for foreseeability, intent assessment, and enforcing proper design imperative.

Structural causal models developed by Halpern & Pearl[14] offer computational tools for provenance tracking and counterfactual simulations. Provenance tracking involves mapping the lineage of data inputs and decision-making pathways within an *AI* system that enables precise attribution of causal contribution across modules. These tools enable developers and regulators to measure each module's production and dependence contributions, ensuring that liability is spread according to causal footprint. Empirical case studies—such as ProPublica's COMPAS analysis[15] and documented misdiagnoses by FDA-cleared *AI* diagnostics—illustrate persistent harms that satisfy our foreseeability and culpability thresholds.

Building on these foundations, we propose creating an *AI* Registry modeled on corporate registries.[16] By doing so, it would formalize registration that is akin to corporate registries by cataloging system architecture version, audit logs, compliance attestations, and operator credentials to create an immutable record for post hoc investigations. This centralized database would record each system's personhood status, developer and operator identities, audit history,

[8] H.L.A. Hart and Tony Honoré, Causation in the Law, 2nd ed. (Oxford: Oxford University Press, 1985), 59. "Proximate cause is that which, in a natural and continuous sequence, unbroken by any new independent cause, produces the injury."

[9] "Corporate personhood," Wikipedia, last edited January 31, 2025, para. 1: "Corporations are recognized as legal persons, capable of exercising rights and incurring obligations separate from their members."

[10] European Parliament, Resolution on Civil Law Rules on Robotics (2017), 35: "The European Parliament invites the Commission to explore the creation of 'electronic personhood' for sophisticated autonomous robots to ensure clear liability channels."

[11] Finale Doshi-Velez and Mason Kortz, Accountability of AI Under the Law: The Role of Explanation (2017), arXiv:1711.01134, 4: "Accountability demands that AI systems provide meaningful, human‑understandable explanations of their decision processes."

[12] Sandra Wachter, Brent Mittelstadt & Luciano Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation," International Data Privacy Law 7, no. 2 (2017): 85. "GDPR's right to explanation is hollow without substantive criteria for algorithmic legibility."

[13] Gianclaudio Malgieri & Giovanni Comandé, "Why a Right to Legibility of Automated Decision-Making Matters," International Data Privacy Law 7, no. 4 (2017): sec. 4: "Algorithmic legibility entails that decision‑making processes are transparent and contestable by affected individuals."

[14] Joseph Y. Halpern & Judea Pearl, "Causes and Explanations: A Structural-Model Approach—Part I," British Journal for the Philosophy of Science 56 (2005): 858. "An event C is an actual cause of E if, in an appropriate structural model, altering C changes E in a counterfactual scenario."

[15] ProPublica, "How We Analyzed COMPAS Recidivism Scores," accessed April 25, 2025: "Black defendants were nearly twice as likely to be incorrectly labeled high-risk than white defendants."

[16] European Parliament and Council, Regulation (EU) 2024/1689 on Artificial Intelligence Act, *Official Journal of the European Union* L 268 (2024): Recital 131. "Providers of high-risk AI systems … should be required to register themselves and information about their AI system in an EU database, to be established and managed by the Commission."

and compliance with legibility and causal-contribution standards[17]. Complementing registration, we recommend statutory reforms inspired by Directive (EU) 2024/2853 on liability for defective products—extending strict liability to *AI* components—and regulatory mandates for periodic algorithmic audits and standardized causal-assessment protocols.[18]

The following section distills the framework into four core liability principles of Volition (V), Intent (I), Culpability (C), graded Autonomy & Causation (AC).

True independence is not in the mere act of choosing, but in the ability to shape outcomes beyond an entity's prior design. True *V* arises when an *AI*'s decision-process is not exhaustively governed by its programmers' directives but instead issues from emergent policies discovered through the system's own adaptive learning mechanisms. Under traditional tort and criminal law, origination of action is presumed in human agents by virtue of moral agency[19]; however, the law remains agnostic to mental states, focusing instead on whether the actor's conduct was originative rather than causally inert.[20] In the *AI* context, systems that merely execute pre-specified rules possess no genuine *V*, for their outputs are strictly functions of external commands.[21] By contrast, reinforcement-learning agents that refine decision-rules in ways unforeseen even by their designers satisfy *V* because their actions "*make a genuine difference*" to outcomes in counterfactual scenarios.[22] Thus, *V* emerges at the threshold where predictability ceases and the *AI*'s own policy structures generate novel causal pathways to harm.

Judgment rests not on what an AI was meant to do, but on the effects it produces—when its reasoning evades scrutiny, its dangers grow, as does its harm. *I* is not anchored in subjective states but in the structure of decision-making that renders harmful outcomes foreseeable. A binary model of classifying an actor as either intending or not intending harm ignores the graded nature of foreseeability in *AI* systems: repeated harmful outputs across varied inputs constitute a pattern equivalent to human "*intent*".[23] If an *AI*, when presented with a spectrum of scenarios, persistently produces decisions that foreseeably injure particular groups such as recidivism-prediction algorithms disproportionately flagging minority populations[24]—this demonstrates an embedded design imperative toward harm, and *I* follows. Far from demanding proof of malice, *I* as criterion requires only that the *AI*'s architecture be such that harmful outputs manifest across counterfactual trajectories with non-negligible probability.

The law seeks no explanation where order prevails, but where disorder arises, reason must be given; when AI's actions bring harm, it must be made to answer as any instrument of consequence. *C* concerns the normative judgment that harm-producing conduct merits sanction. In human jurisprudence, negligence suffices to ground *C* where an actor fails to meet a reasonable-care standard. For *AI*, the absence of consciousness does not vitiate *C*: what matters is the system's capacity to cause destruction through flawed or insufficiently safeguarded

[17] Ada Lovelace Institute, AI Liability in Europe (2022): 2. "An AI Registry could function akin to corporate registries, documenting personhood status, developer identity, and compliance records."
[18] Directive (EU) 2024/2853 of the European Parliament and of the Council on liability for defective products, Official Journal of the European Union (November 18, 2024), art. 3: "Software and AI components are explicitly included within the scope of strict liability for defective products."
[19] Stanford Encyclopedia of Philosophy, "Agency," section on origination and autonomy in decision-making.
[20] Stanford Encyclopedia of Philosophy, "Ethics of Artificial Intelligence and Robotics," §2.7 on autonomy.
[21] Frances S. Grodzinsky, Keith W. Miller & Marty J. Wolf, "The Ethics of Designing Artificial Agents," Ethics and Information Technology 10 (2008): 115–121.
[22] Alex Kaiserman, "Partial Liability," Legal Theory 23, no. 1 (2017): 3–5. "A defendant should be held liable for a claimant's loss only to the degree to which the defendant's wrongdoing contributed to the causing of the loss."
[23] Joseph Y. Halpern & Judea Pearl, "Causes and Explanations: A Structural-Model Approach—Part I," British Journal for the Philosophy of Science 56 (2005): 843–887.
[24] ProPublica, "How We Analyzed COMPAS Recidivism Scores," accessed April 25, 2025, https://www.propublica.org/article/how-we-analyzed-compas-recidivism-scores

decision-rules. A "*black-box*" defense—invoking opacity as exculpation—fails under *C* because the law holds agents to account for harms they are capable of causing, regardless of epistemic access.[25] Thus, an *AI* whose training data omit critical safety constraints or whose algorithmic architecture forecloses human oversight bears *C* in proportion to the ease with which its design permitted the harm.[26] The law, having long extended negligence doctrines to corporate entities, must similarly impose *C* on any entity (natural or artificial) capable of generating legal injuries.

Responsibility is not evaded by obscurity—what acts without oversight yet shapes the world must bear the weight of its consequence. *AC* synthesizes *V* and *I* with a graded causation metric to link autonomous action-chains to proportional liability.[27] Whereas binary causation doctrines struggle with multiple sufficient or overdetermining causes, Kaiserman's partial-liability model apportions responsibility according to counterfactual difference-making and contextual indispensability.[28] In *AI* systems, provenance logs and algorithmic metadata furnish the raw data for production measures—quantifying how much each module contributes to an injury—while counterfactual simulations yield dependence measures—assessing how essential each component was to the harm.[29] By applying these dual metrics, AC ensures that designers, operators, and the *AI* itself share liability shares commensurate with their causal footprints, closing accountability gaps endemic to distributed, opaque architectures.

The four liability principles—*V, I, C, and AC*—form the theoretical scaffolding for three integrated institutional mechanisms. First, an **AI Registry** records each system's provenance logs, algorithmic architecture, governance protocols, and audit history, establishing a traceable chain of control and design intent. Second, **Algorithmic Audits** deploy accredited experts to run counterfactual simulations on recorded logs, producing both production and dependence scores for every module and stakeholder. Third, **Causation-Assessment Standards** translate those quantitative scores into tiered legal thresholds—specifying minimal counterfactual difference-making for *V*, systematic foreseeability for *I*, and oversight-failure indices for *C*. By embedding these feedback loops *(registry → audit → causation testing → adjudication)*, we ensure that philosophical nuance directly informs enforceable legal rules, allowing courts to allocate liability proportionally in even the most complex, multi-actor scenarios

CONCLUSION

a. **Accountability follows agency.** When an *AI* system's emergent decision policies satisfy our Volition and Autonomy thresholds, it assumes a de facto role as an originative agent, meriting liability for its causal footprint without importing contested mental-state doctrines.

b. **Liability is enforceable.** Systems meeting the combined *V* + *I* + *C* + *AC* criteria incur binding obligations under criminal, product-liability, and corrective-justice norms—creating tangible incentives for designers to internalize safety across every stage of data treatment and robustness of the code inherent to the system.

c. **Responsibility transcends anthropocentrism.** In line with corporate and strict-liability precedents, the capacity to effect harm—rather than human volition

[25] Miller v. Jackson, [1977] QB 966 (Eng. C.A.).

[26] Sandra Wachter, Brent Mittelstadt & Luciano Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation," International Data Privacy Law 7, no. 2 (2017): 76–99.

[27] Alex Kaiserman, "Necessary Connections in Context," Erkenntnis 82, no. 1 (2017): 47–58. "Production intuitions focus on difference-making… dependence intuitions measure indispensability."

[28] Gianclaudio Malgieri & Giovanni Comandé, "Why a Right to Legibility of Automated Decision-Making Matters," International Data Privacy Law 7, no. 4 (2017): sec. 4. "Algorithmic legibility entails transparent and contestable decision-making."

[29] H.L.A. Hart & Tony Honoré, Causation in the Law, 2nd ed. (Oxford: Oxford University Press, 1985), 59. "Proximate cause… produces the injury."

alone—grounds legal responsibility; *AI* systems and their human stewards share proportional liability based on measurable causal contributions.

d. **Institutionalizing oversight protects society.** By operationalizing the proposed framework through a centralized *AI* Registry, periodic Algorithmic Audits, and codified Causation-Assessment Standards, regulators and courts gain the tools needed to govern machine intelligence under the rule of law—safeguarding individual rights and closing the accountability gaps of tomorrow.

BIBLIOGRAPHY

Ada Lovelace Institute. AI Liability in Europe. London: Ada Lovelace Institute, 2022. https://www.adalovelaceinstitute.org/report/ai-liability-in-europe

Carrera, David et al., "State of the Art in Parallel and Distributed Systems: Emerging Trends," *Electronics* 14, no. 4 (2023): 677–700. https://www.mdpi.com/2079-9292/14/4/677

"Corporate Personhood." Wikipedia, last modified January 31, 2025. https://en.wikipedia.org/wiki/Corporate_personhood

Directive (EU) 2024/2853 of the European Parliament and of the Council on Liability for Defective Products, Official Journal of the European Union L 304/1 (November 18, 2024). https://eur-lex.europa.eu/eli/dir/2024/2853/oj

Doshi-Velez, Finale, and Mason Kortz. "Accountability of AI Under the Law: The Role of Explanation." arXiv preprint arXiv:1711.01134 (2017). https://arxiv.org/abs/1711.01134

European Parliament and Council, Regulation (EU) 2024/1689 on Artificial Intelligence Act, *Official Journal of the European Union* L 268 (2024): Recital 131, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

European Parliament. Resolution on Civil Law Rules on Robotics. Strasbourg: European Parliament, 2017. https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html

Fiona Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms," *Big Data & Society* 3, no. 1 (2016). https://doi.org/10.1177/2053951715622512

Google Cloud, "Continuous Training and MLOps Best Practices," *Google Cloud Blog*, June 15, 2021. https://cloud.google.com/blog/topics/machine-learning/continuous-training-mlops

Grodzinsky, Frances S., Keith W. Miller, and Marty J. Wolf. "The Ethics of Designing Artificial Agents." Ethics and Information Technology 10 (2008): 115–21. https://link.springer.com/article/10.1007/s10676-008-9165-9

Halpern, Joseph Y., and Judea Pearl. "Causes and Explanations: A Structural-Model Approach—Part I." British Journal for the Philosophy of Science 56 (2005): 843–887. https://academic.oup.com/bjps/article/56/4/843/153063

Hart, H.L.A., and Tony Honoré. Causation in the Law. 2nd ed. Oxford: Oxford University Press, 1985. https://doi.org/10.1093/acprof:oso/9780198263612.001.0001

Kaiserman, Alex. "Necessary Connections in Context." Erkenntnis 82, no. 1 (2017): 47–58. https://doi.org/10.1007/s10670-016-9831-6

Kaiserman, Alex. "Partial Liability." Legal Theory 23, no. 1 (2017): 1–26. https://doi.org/10.1017/S1352325217000040

Malgieri, Gianclaudio, and Giovanni Comandé. "Why a Right to Legibility of Automated Decision-Making Matters." International Data Privacy Law 7, no. 4 (2017): 267–82. https://academic.oup.com/idpl/article/7/4/267/4093388

Miller v. Jackson, [1977] QB 966 (Eng. C.A.). https://www.bailii.org/ew/cases/EWCA/Civ/1977/4.html

ProPublica. "How We Analyzed COMPAS Recidivism Scores." ProPublica. Accessed April 25, 2025. https://www.propublica.org/article/how-we-analyzed-compas-recidivism-scores

Stanford Encyclopedia of Philosophy. “Agency.” Accessed April 25, 2025. https://plato.stanford.edu/entries/agency/

Stanford Encyclopedia of Philosophy. “Ethics of Artificial Intelligence and Robotics.” Accessed April 25, 2025. https://plato.stanford.edu/entries/ethics-ai/

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. “Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation.” International Data Privacy Law 7, no. 2 (2017): 76–99. https://academic.oup.com/idpl/article/7/2/76/3860558