ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

CAN HUMANS REMAIN MORALLY RESPONSIBLE FOR AUTONOMOUS SYSTEMS? A SOCIAL AND ETHICAL ANALYSIS

Ashish Kumar
Lead Author
Nixor College - Bahadurabad Campus
Email: ashish.kumar26@nixorcollege.edu.pk
Dr. Asif Abd ur Rehman (Co-Author)
PhD Mathematics from PU, Visiting, Lecturer at University of the Punjab, Lahore
Email: asifrehman.math@gmail.com
Samreen Rashid (Corresponding Author)
Lecturer English, IES, University of the Punjab
Email: rashidsamreen9@gmail.com

1. Abstract
The advent of highly sophisticated autonomous systems, ranging from self-driving vehicles and lethal autonomous weapons to AI-driven diagnostic tools in healthcare, has precipitated a profound crisis in traditional moral and legal frameworks. Central to this crisis is the "responsibility gap"—a phenomenon where the increasing autonomy of machines threatens to decouple human agency from mechanical outcomes. This research provides a comprehensive social and ethical analysis of human responsibility in the age of artificial intelligence. By examining the evolution of "learning automata" and the resulting challenges to traditional theories of blameworthiness, the paper explores the move from regulative control to "meaningful human control" (MHC). Through an expanded literature review and in-depth case studies—including the 2018 Uber autonomous vehicle fatality, the Boeing 737 MAX MCAS failures, the COMPAS algorithmic bias controversy, and the failure of IBM Watson for Oncology—this analysis identifies the emergence of "moral crumple zones" and "accountability shields." Furthermore, the research integrates 2024–2025 statistical data and international policy trajectories from the United Nations Group of Governmental Experts (GGE) to argue for a transition toward collective, forward-looking responsibility models. Ultimately, the paper concludes that while individual blame may be increasingly difficult to attribute, moral responsibility can be preserved through the rigorous design of socio-technical infrastructures that prioritize tracking human values and tracing causal accountability.

1.1 Key Words
Responsibility Gap, Meaningful Human Control, Moral Crumple Zone, Autonomous Weapon Systems, Algorithmic Accountability, Socio-Technical Systems.

2. Introduction

The integration of artificial intelligence (AI) into the foundational structures of modern society represents a paradigm shift that rivals the industrial revolution in its transformative potential and ethical complexity. As we move further into the 2020s, autonomous systems are no longer merely passive tools or rigid programs; they have become "learning automata" capable of adapting their behavior based on environmental feedback and high-dimensional data processing (Matthias, 2004). This evolution has fundamentally challenged the classical prerequisites for moral responsibility: the control condition and the epistemic condition (Santoni de Sio & Mecacci, 2021). Traditionally, an agent is held responsible for an outcome only if they possessed causal control over the event and were aware of its moral implications.

However, the "black box" nature of deep learning, combined with the distributed nature of modern software development, has created a widening "responsibility gap" where no single human appears

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

to satisfy these conditions when things go wrong.

In the contemporary landscape of 2025, this gap is no longer a theoretical concern of philosophers but a pressing issue for policymakers, engineers, and the public. In the United States alone, fear of self-driving cars has remained high, with 61% of the population expressing distrust as of late 2024 (Stanford AI Index, 2025).

Simultaneously, the global community is engaged in urgent consultations at the United Nations to establish "legal guardrails" for lethal autonomous weapon systems (LAWS) before they become a permanent fixture of warfare (United Nations, 2025). The ethical stakes are nothing less than the preservation of human dignity; to allow a machine to make life-and-death decisions without a corresponding human locus of responsibility is to risk treating individuals as "data points" rather than moral agents (Human Rights Watch, 2025).

2.1 Research Questions

1. How do the "control" and "epistemic" conditions of responsibility fail in the context of deep-learning algorithms and autonomous agents?
2. What role does the "moral crumple zone" play in the legal and social attribution of blame for systemic failures?
3. Can the "meaningful human control" framework (MHC) effectively close the responsibility gap in high-stakes domains like healthcare and defense?
4. What do current (2024–2025) statistical trends indicate regarding the maturity of "Responsible AI" frameworks and public trust in autonomous systems?

3. Literature Review

3.1. The Ontological Existence of the Responsibility Gap

The term "responsibility gap" was first introduced by Andreas Matthias in 2004 to describe a situation where "learning automata" are exploited, causing human agents to lose sufficient control over outcomes such that no one can be held traditionally responsible (Matthias, 2004). Matthias argued that as machines learn and evolve beyond their initial programming, the link between the programmer's intent and the machine's action is severed (Matthias, 2004). This is not merely an "apparent" gap but a fundamental mismatch in our moral practices. However, later scholars have challenged this "fatalist" view, suggesting that responsibility gaps may not exist if we redefine responsibility in terms of "risk" and "moral control" rather than simple blame (Santoni de Sio & van den Hoven, 2021).

Building on this, Santoni de Sio and Giulio Mecacci (2021) expanded the concept, arguing that the responsibility gap is not a single problem but a set of at least four interconnected issues: gaps in culpability, moral and public accountability, and active responsibility (Santoni de Sio & Mecacci, 2021).

3.2. Meaningful Human Control: Tracking and Tracing

The most prominent framework for addressing these gaps is "meaningful human control" (MHC) (Santoni de Sio & van den Hoven, 2021). MHC posits that for a human to remain responsible, the autonomous system must satisfy two general conditions:

- **The Tracking Condition**: The system must be able to respond to both the moral reasons of the humans designing and deploying the system and the relevant features of the environment (Santoni de Sio & van den Hoven, 2021).
- **The Tracing Condition**: The system's outcomes must be traceable back to at least one human agent in the design or command chain who understands the system's operation and potential risks (Santoni de Sio & van den Hoven, 2021).

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

3.3. Distributed Responsibility and Information Ethics

Luciano Floridi's theory of "distributed responsibility" provides a macro-ethical foundation for understanding systems where multiple agents—both human and artificial—interact (Floridi, 2021). In the "Infosphere," moral action is an information processing pattern. In such networks, responsibility is not localized in a single person but is "distributed" across the network of designers, producers, implementers, and users. This model replaces traditional "blame" with a "duty of censorship," where human stakeholders have a prospective duty to monitor, modify, or delete AI systems that begin to cause harm (Floridi, 2021).

3.4. The Ubuntu Perspective: Collective Responsibility

A significant development in recent literature is the application of "Ubuntu" philosophy to the responsibility gap debate (Pusan National University, 2025). Ubuntu, rooted in African philosophy, emphasizes that "a person is a person through other people." In the context of AI, an Ubuntu-inspired perspective calls for collective forward-looking responsibility (Pusan National University, 2025). Instead of focusing retrospectively on individual blame for a past error, the focus shifts to how the community can collectively resolve the harm and support the victim.

4. Socio-Technical Analysis of Failure

4.1. The "Moral Crumple Zone" and the Handoff Problem

The concept of the "moral crumple zone," developed by Madeleine Clare Elish, is essential for a social analysis of autonomous systems (Elish, 2018). Just as a car's crumple zone is designed to absorb the force of impact to protect passengers, a human operator in an automated system often becomes the "component" that bears the brunt of legal and moral responsibility when the overall system malfunctions (Elish, 2018). This dynamic allows the perception of a "flawless" technology to remain intact by isolating the human operator as the "weak link" (CIGI, 2020). This is exacerbated by the "handoff problem," which occurs when a human is expected to take over a machine-controlled system quickly and safely, a task for which human cognition is structurally disadvantaged (Elish, 2018).

4.2. Failure Modes: Syntactic, Semantic, Testing, and Warning

A proposed "Education Theory of Fault" categorizes the failure points that lead to unpredictable harm in autonomous systems into four distinct categories (Selbst, 2020). These are detailed in Table 1 below.

Table 1: The Four Failure Points of Autonomous Systems

Failure Point	Mechanism of Error	Real-World Example
Syntactic	Sensors failing to identify or classify objects correctly.	Uber's system is cycling through "bicycle" and "other".
Semantic	Failure to translate human intent into algorithmic logic.	AI recommendations ignoring local clinical practices.
Testing	Failure to test the system in full range edge-case scenarios.	Tesla autopilot mistaking a white truck for a cloud.
Warning	Failure to articulate the system's limits to the end-user.	Boeing's omission of MCAS from pilot manuals.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

5. Detailed Case Studies

5.1. The Uber Autonomous Vehicle Fatality (Tempe, 2018)

On March 18, 2018, Elaine Herzberg was struck and killed by a self-driving Uber car in Tempe, Arizona, marking the first recorded pedestrian fatality involving an autonomous vehicle (Elish, 2018). The vehicle failed to accurately classify Herzberg as she pushed a bicycle across the road, cycling through classifications of "motor vehicle" and "other" before impact (Elish, 2018). Although the NTSB found flaws in Uber's safety culture—including the removal of the car's factory collision warning system—the legal focus shifted to the backup safety driver, Rafaela Vasquez (CIGI, 2020). Vasquez became the "moral crumple zone," absorbing the legal impact of a complex failure that involved software design and organizational negligence (Elish, 2018).

5.2. The Boeing 737 MAX MCAS Failures (2018–2019)

The crashes of Lion Air Flight 610 and Ethiopian Airlines Flight 302 represent a catastrophic failure of "meaningful human control" and institutional accountability (Langewiesche, 2019). Boeing developed the Maneuvering Characteristics Augmentation System (MCAS) to automatically push the plane's nose down based on a single sensor, yet concealed its existence from regulators and pilots (Sgobba, 2019). When the sensor failed, pilots were unaware that an autonomous system was fighting their manual inputs. This case highlights that without "epistemic transparency," human operators cannot be held responsible for the failures of a system they do not even know exists (Sgobba, 2019).

5.3. IBM Watson for Oncology and clinical Decision Support

IBM Watson for Oncology was marketed as a "physician's assistant" that could democratize high-quality cancer care (Healthark, 2023). However, the system's training data was heavily biased toward U.S.-centric practices at a single medical center. When deployed globally, Watson provided recommendations inconsistent with local standards (Dolfing, 2024). The project was discontinued in 2023 after failing to overcome the "black box" nature of its recommendations, which made it impossible for doctors to assume "informed responsibility" (Dolfing, 2024).

5.4. COMPAS and Algorithmic Bias in Justice

The COMPAS algorithm, used to predict recidivism risk, illustrates the danger of "tech-washing" historical biases (Angwin et al., 2016). Analysis found that Black defendants were twice as likely to receive a "false positive" for high recidivism risk than White defendants (Angwin et al., 2016). In *State v. Loomis* (2016), the court upheld the use of COMPAS despite the defendant's inability to challenge the proprietary, "black box" algorithm, which acts as an "accountability shield" for judicial actors (Shor, 2023).

6. Psychological Factors: Automation Bias and Over-Trust

A critical barrier to maintaining human responsibility is Automation Bias (AB)—the tendency of humans to over-rely on automated systems (Jacobs et al., 2021). Recent research suggests a version of the Dunning-Kruger effect in AI trust: users with moderate familiarity are more prone to significant over-reliance than those with either zero knowledge or expert-level expertise (McKinsey, 2025). This "complacency" occurs when operators prioritize manual tasks over monitoring the veracity of automation, leading to a brittleness in the human-machine team (Jacobs et al., 2021).

7. Statistical Trends and Public Sentiment (2024–2025)

The landscape of AI adoption is shifting rapidly. While 55% of the global population believes AI benefits outweigh drawbacks, trust in AI companies to protect personal data has fallen to 47% (Stanford AI Index, 2025). In corporate sectors, 58% of executives agree that Responsible AI

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

improves ROI, yet 50% cite the transition from "principles" to "operational processes" as their biggest hurdle (PwC, 2025). Furthermore, 61% of people in the United States express persistent fear regarding self-driving cars as of 2025 (Stanford AI Index, 2025).

8. International Policy and the Future of LAWS

The debate over Lethal Autonomous Weapon Systems (LAWS) at the United Nations represents the critical frontier for responsibility (United Nations, 2025). 2025 UN GGE sessions have centered on defining "Meaningful Human Control" and "Appropriate Human Judgment" (United Nations, 2025). A consensus is building around a "two-tiered approach": prohibiting systems that target individuals without human involvement while strictly regulating others to ensure adherence to International Humanitarian Law (Human Rights Watch, 2025).

9. Legal Liability Frameworks

Legal scholars are debating two primary liability regimes (Selbst, 2020):

- **Negligence (Fault-Based)**: Struggles with autonomy because it requires proving a human could have "reasonably anticipated" a machine's unpredictable behavior (Selbst, 2020).
- **Strict Liability**: Increasingly favored for "high-risk" AI, holding manufacturers liable for harms regardless of fault to ensure victim redress and incentivize "safety-by-design" (Vladeck, 2014).

10. Conclusion

The social and ethical analysis of autonomous systems reveals that the "responsibility gap" is not an inevitable byproduct of technology, but a structural deficiency in our socio-technical systems. Humans can remain morally responsible for autonomous systems only if responsibility is transformed from a localized property of individuals into a distributed property of the system's entire life cycle. The "moral crumple zone" identifies the fundamental injustice of our current trajectory—concentrating blame on operators while shielding the architects of the "black box." To bridge this gap, we must move beyond retributive blame toward collective forward-looking responsibility, requiring rigorous design for MHC, epistemic transparency in algorithms, and strict liability for high-risk AI.

Bibliography

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias. *ProPublica*.

CIGI (Centre for International Governance Innovation). (2020). Who is responsible when autonomous systems fail?

Dolfing, H. (2024). Case Study 20: The $4 Billion AI Failure of IBM Watson for Oncology.

Elish, M. C. (2018). Moral Crumple Zones: Case Studies in Unmanned Aerial Vehicles (UAVs) and Autonomous Vehicles. *Cultural Anthropology*.

Floridi, L. (2021). *Ethics, Governance, and Policies in Artificial Intelligence*. Springer.

Healthark. (2023). IBM Watson: From healthcare canary to a failed prodigy.

Human Rights Watch. (2025). A Hazard to Human Rights: Autonomous Weapons Systems and Digital Decision-Making.

Jacobs, M., et al. (2021). How AI Literacy impacts medical decision-making. *Microsoft Research*.

Langewiesche, W. (2019). The Human Factor: What went wrong on the Boeing 737 MAX? *The New York Times Magazine*.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*.

McKinsey & Company. (2025). The state of AI in 2025: Agents, innovation, and transformation.

Pusan National University. (2025). A shared responsibility of both humans and AI in AI-caused harm. *Topoi*.

PwC (PriceWaterhouseCoopers). (2025). US Responsible AI Survey: From foundational governance to innovation and scale.

Santoni de Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence. *Philosophy & Technology*.

Santoni de Sio, F., & van den Hoven, J. (2021). Meaningful Human Control over Autonomous Systems.

Selbst, A. D. (2020). Negligence and AI's Human Users. *Boston University Law Review*.

Sgobba, T. (2019). Regulatory failures surrounding the certification of the 737 MAX.

Shor, J. (2023). Algorithmic Accountability and the COMPAS Controversy.

Stanford University. (2025). *AI Index 2025 Annual Report*. Stanford Institute for Human-Centered AI (HAI).

United Nations. (2025). *Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*. UNODA Geneva.

Vladeck, D. C. (2014). Machines Without Principals: Liability Rules and Artificial Intelligence. *Washington Law Review*.