International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 06 June 2025
Page No: 2988 - 2995
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0440

Ethical and Security Risks of Autonomous AI Systems

Ilakiya Ulaganathan[1]
[1]*Independent Researcher, Tagore Engineering College (affiliated to Anna University), Chennai, India.*

Abstract
As AI systems grow more independent, the challenges of keeping them both ethical and secure are only getting tougher. From self-driving cars and military drones to algorithms making real-world decisions, these technologies often operate with little—or no—human input. This paper takes a close look at the risks that come with that shift. On the ethical side, we're talking about things like bias, unclear accountability, and the risk of sidelining human judgment. On the security front, there's the threat of adversarial attacks, hacking, and deeper system flaws that could be exploited. By reviewing what's already out there—research, case studies, real-world examples—we aim to unpack how these risks are currently being managed. More importantly, we offer ideas on how future systems can be built with stronger guardrails, both in terms of tech and governance.

Keywords: *Autonomous AI, Ethical Risks, AI Security, AI Governance, Bias in AI, AI Accountability, Adversarial Attacks, Machine Ethics.*

1. Introduction
In the past ten years or so, it's become obvious that artificial intelligence isn't just another tech buzzword. It's settled into our daily lives, often in ways we don't even notice. Whether it's helping doctors make quicker decisions in hospitals, managing piles of financial data behind the scenes, or playing a quiet role in national defense strategies—AI is already here, working quietly in the background and shaping the way the world runs. But one branch of AI that has really stirred both excitement and concern is autonomous systems. These are the kinds of tools that don't need a person hovering over them to work—they just go. That's impressive, no doubt, but it also makes people stop and wonder: are we truly prepared for this level of independence in machines? The honest answer? Not quite yet.

1.1.Definition and Scope of Autonomous AI Systems
When someone talks about autonomous AI, what they usually mean is a system that's able to take action or make a decision without someone hitting a button or giving real-time instructions. These systems rely on things like machine learning, or maybe a complex set of pre-written rules—sometimes both. For example, driverless cars don't wait for your okay to brake at a stop sign. They're trained on how to recognize it and act accordingly. The same goes for bots that make trades in the blink of an eye, surgical robots helping doctors, or even weapon systems that can identify and engage a threat. What links them all is their ability to sense what's going on around them, process that input super quickly, and act based on what they've been "taught" or coded to do. It's fast, it's efficient, but it also removes human judgment from the loop—which is where a lot of ethical and safety concerns start to come in [1].

1.2.Ethical Implications of Autonomous Decision-Making
Talking about the ethics of autonomous AI can get messy fast. One of the biggest issues people bring up is the fact that no one really knows how these systems make some of their decisions. That's often called the "black box" problem—basically, AI does something, but we're not totally sure why or how. This becomes a real problem when the outcome matters. Think about a self-driving car making a wrong move or an AI misdiagnosing something in a hospital. If something goes wrong, then what? Who gets held accountable? It's not always easy to tell, and honestly, that makes the whole idea of responsibility feel kind of fuzzy [2]. There's also the whole thing with bias. AI systems are only as fair as the data

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 06 June 2025
Page No: 2988 - 2995
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0440

they're trained on, and let's be honest—our data isn't perfect. If the training data includes bias, the system is likely to repeat it or even make it worse. And sadly, that usually hits disadvantaged groups the hardest. It's not just a glitch; it's a serious issue when machines reflect or amplify the unfairness that already exists in society [3]. It really makes you wonder—are we using AI to fix problems, or just getting better at repeating the same old unfairness, only faster?

1.3.Autonomy Versus Human Control

This one's a bit complicated—how much say should we actually be giving machines? As these systems keep getting better, it sort of feels natural to hand over more to them—especially when they're faster or seem to do the job well enough. But after a while, it stops being about just helping out or saving time. There's this shift, and suddenly it's like—we're not just letting them help, we're letting them take over. But eventually, it's not really about being efficient anymore—it sort of turns into us handing over control without asking enough questions. We don't always notice it happening, either. And the more control we give up, the more we end up being left out of decisions that probably need a human perspective. One of the more troubling examples is autonomous weapons. It's honestly hard to wrap your head around the idea of a machine deciding who lives or dies. There's been a lot of back-and-forth about whether that should even be allowed. A lot of people say no—it just doesn't feel right to hand over something so serious to a piece of tech. Decisions like that, life-and-death ones, should really involve a human being. Even the smartest code out there can't replace human judgment—or empathy, for that matter. It just doesn't work the same way. Some people who study this stuff believe that letting AI make those kinds of decisions could even break international laws that are meant to protect civilians in times of war [4]. But honestly, it's not just about legal rules. It goes deeper than that. It's about what it means to be human—about dignity, responsibility, and whether we're handing over too much when we start trusting machines with choices that really should come from us.

1.4.Security Vulnerabilities in Autonomous AI Systems

Security risks with autonomous AI aren't some distant possibility anymore—they're already happening in real situations. These systems can be hit by all kinds of things: people messing with the input data, poisoning the training process, or even changing how the algorithms behave behind the scenes. Take adversarial attacks, for example. Just a small tweak in the input—something a person might not even notice—can totally throw off an AI's decision-making. That's how a self-driving car could end up reading a stop sign as a speed limit sign [5]. It's not science fiction; this stuff is already happening. And when you're dealing with AI in things like power grids, healthcare, or military tools, these weak points become really serious. What makes it even trickier is how AI is often spread out across networks—like in smart home systems or connected devices. Because everything's so connected, it opens up more chances for someone to break in—and the moment something does go wrong, it's usually tough to catch it fast or sort it out right away.

1.5.The Need for Ethical and Security Governance

It's becoming more and more clear that we can't rely on tech fixes alone to handle the ethical and security challenges AI brings. What we really need is something bigger—rules and systems that include everyone who's part of the picture. It's not just about the engineers and developers anymore. Policymakers, ethicists, and regular people all need to be part of the conversation. And to be fair, some of that is already happening. Groups like the IEEE, the European Commission, and UNESCO have started laying out guidelines to help keep AI use responsible [6]. But the truth is, even with those efforts, how these rules are being rolled out varies a lot from one country to another. Some places are moving fast, others are way behind. So while the conversation has definitely started, we've still got a long way to go before we can say there's solid, global oversight in place.

1.6.Overview of Paper Structure

Here's how the rest of the paper is laid out. Section 2 dives into the research approach—basically how the ethical and security risks tied to autonomous AI were studied and what criteria were used. Then in Section 3, the findings are laid out, including a few real-world case studies and some supporting data, followed by a deeper look into what it all means. Section 4 basically

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 06 June 2025
Page No: 2988 - 2995
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0440

wraps it all up. It lays out a few policy ideas and throws in some thoughts on what future research might want to explore next.

2. Method

This part breaks down how the research was done to explore the ethical and security challenges connected to autonomous AI systems. For this research, I looked at a mix of different sources—stuff like academic papers, policy write-ups, real case studies, and official reports about past incidents. I also put together a comparison framework to get a better idea of how these risks actually show up in different areas where autonomous AI is being used already, like in self-driving cars, hospitals, military systems, and finance.

2.1.Data Sources and Selection Criteria

To keep the research focused and meaningful, purposive sampling was used to pick out a solid, representative mix of sources. This included peer-reviewed journal articles, government white papers, and well-documented case databases. Each source had to meet a few key criteria:

- It had to deal with how autonomous AI is actually being used.
- Each source had to clearly focus on ethical or security-related issues. And it also needed to be trustworthy—stuff that had been peer-reviewed or published by a well-known organization.

In total, I went through 65 documents, covering the years from 2015 to 2024. They came from a bunch of different fields—things like AI ethics, cybersecurity, law, robotics, and how people interact with computers.

2.2.Analytical Framework

The first one was the AI4People Ethical Framework, which focuses on big-picture principles like doing good, avoiding harm, respecting autonomy, ensuring fairness, and making things understandable [7]. The second was the NIST AI Risk Management Framework (AI RMF). It breaks down how to look at risks in AI systems using four key steps: Map, Measure, Manage, and Govern [8]. These frameworks were applied across four application domains: healthcare, autonomous vehicles, finance, and defense. The matrix below (Table 1) presents a mapping of ethical and security risk categories against these domains.

Table 1 Risk Category Matrix Across Domains

Risk Category	Healthcare	Autonomous Vehicles	Finance	Defense
Bias & Discrimination	✓✓	✓✓✓	✓✓	✓
Explain ability	✓✓✓	✓✓	✓	✓
Accountability	✓✓✓	✓✓✓	✓✓	✓✓✓
Cyber Vulnerability	✓✓	✓✓✓	✓✓✓	✓✓✓
Human Oversight	✓✓✓	✓✓	✓	✓

This comparative matrix facilitated a structured cross-domain analysis to identify which risks are most critical and where existing mitigation strategies fall short.

2.3.Case Selection and Scenario Mapping

To really show what these risks can look like in the real world, I pulled together five cases where things clearly didn't go as expected. Each one shows a different way autonomous AI can fail—and why those failures matter. To help make these risks more concrete, I looked at five real cases where autonomous AI systems didn't perform the way they were supposed to. Each one shows a different kind of failure—and why that matters in the bigger picture. Uber's Self-Driving Car Fatality (2018) – A pedestrian was hit and killed by a self-driving Uber. The system just didn't recognize her. It's one of those moments that shows how fragile this tech still is—how a split-second failure can lead to something irreversible. And it left people asking a hard question: when something like this happens, who's actually to blame? [9] GPT-4 in Financial Markets – Some early

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 06 June 2025
Page No: 2988 - 2995
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0440

experiments tried using large language models for trading, but it didn't take long for concerns to pop up—things like hallucinated info, potential manipulation, and the lack of clarity around how decisions were being made [11]. Autonomous Weapons in Conflict Zones – In Libya, AI-powered drones were reportedly used without any human in the loop. That raised serious concerns—not just about how the tech was used, but also about whether it crossed the line when it comes to human rights and international law [12]. Healthcare Diagnosis Systems (IBM Watson) – In its early stages, Watson sometimes gave medical advice that didn't line up with standard practices. It showed what can go wrong when we trust systems that aren't always transparent in how they reach their conclusions [13]. Each of these cases was analyzed through the lens of the two earlier frameworks, to break down the ethical and security failures—and to take a closer look at how developers, regulators, and the public responded (Figure 1).

2.4.Figures: Risk Prevalence and Severity Analysis

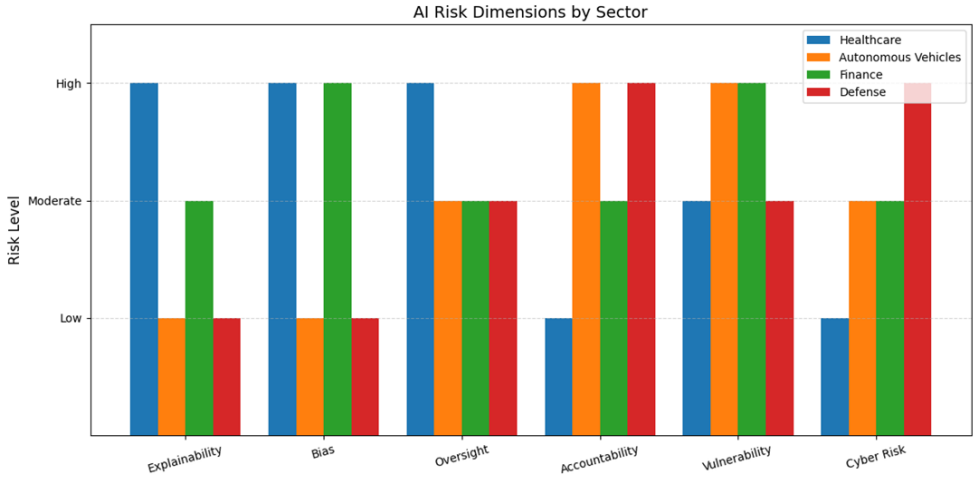

Figure 1 Ethical Risk Severity Distribution by Domain

This figure 2 demonstrates that while ethical risks pervade all domains, the severity and nature vary significantly. Defense and autonomous driving exhibit the highest risk scores.

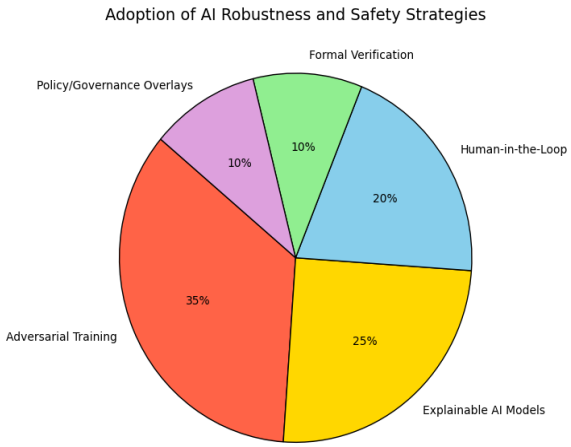

Figure 2 Risk Mitigation Techniques Adoption

This figure emphasizes the technological skew in current mitigation strategies, with less emphasis on institutional and human-centric governance.

2.5.Limitations

Even though the study pulled in a lot of material from different areas, there were still a few things that held it back:

- AI moves fast—some of what's here could end up outdated pretty quickly.
- I didn't have access to private company data on security issues, so that part's a bit of a blind spot.
- The case studies don't fully capture how different cultures or laws might shape things differently in other places.

3. Results and Discussion

One thing that became really clear during the

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 06 June 2025
Page No: 2988 - 2995
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0440

analysis: these ethical and security risks aren't just side effects—they're built into the systems themselves. From how they're designed to how they're used, the problems are part of the structure. I kept seeing the same kinds of weaknesses across different cases, and even now, there are still some big gaps in how industries are trying to manage them.

3.1.Ethical Risk Landscape Across Domains

What the results showed is that ethical risks don't look the same everywhere—they tend to show up differently depending on the setting and who's involved.

Bias and Discrimination: Bias came up most often in systems trained on older datasets that didn't reflect enough diversity. A good example is Amazon's AI hiring tool, which downgraded resumes that included the word "women's." It wasn't doing that on purpose, but it mirrored past hiring trends instead of judging applicants fairly [10]. We've seen the same kind of issue in facial recognition tech, where the systems tend to work much better for some groups than others—especially across different races and genders [14].

Opacity and Explain Ability: One of the biggest ongoing problems is that a lot of these systems don't explain how they make decisions—especially the more complex ones, like deep learning models. Take IBM Watson, for instance. In some cases, it gave medical advice that doctors couldn't really make sense of, which obviously makes it hard to trust or follow [13]. This kind of "black box" issue makes it tough to review decisions or push back when something seems off—and it goes against one of the key principles of ethical AI: that systems should be understandable [7].

Accountability and Responsibility: No matter the industry, figuring out who's actually responsible when something goes wrong is a huge challenge. That responsibility often gets spread out between developers, companies, and end users, which just makes things more confusing. The Uber case is a perfect example—there was a lot of uncertainty over whether the fault was with the software, the backup driver, or the company itself [9]. And when there's no clear line of responsibility, it becomes much harder for victims to get justice or even basic answers [15] (Figure 3).

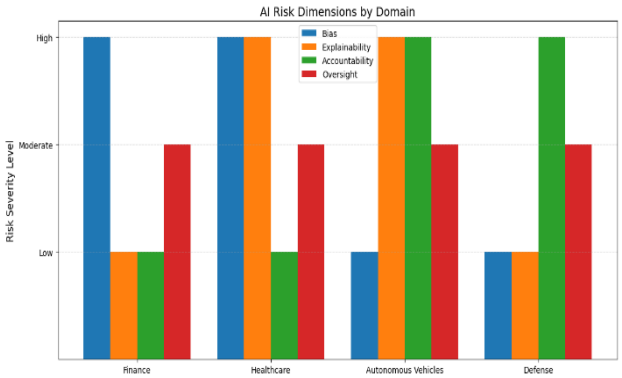

Figure 3 Frequency of Ethical Risk Themes in Literature by Domain

3.2.Security Risk Landscape

Security risks showed up in a lot of different ways too, and many of them don't fit into the usual mold of traditional cybersecurity problems. These are newer, more complex challenges that come with how AI systems interact with the world around them.

Adversarial Attacks: One big concern is adversarial attacks. These systems can be surprisingly easy to trick. One well-known example involved adding small stickers to a stop sign—just subtle changes most people wouldn't even notice. But it was enough to throw off a Tesla's autopilot, making it misread the sign completely [16]. It really shows how something as small as a few marks on a sign can mess with the system—and in the wrong situation, that kind of mistake could be dangerous.

Data Poisoning and Manipulation: Another serious issue is data poisoning. This happens when someone intentionally adds bad or misleading data during training, so the AI ends up learning the wrong patterns. In finance, that kind of move can be used to mess with predictions or outcomes on purpose—just to sway the market or pull off fraud [17]. It's not always easy to spot, but if no one catches it, the damage can be pretty serious.

Autonomy and Malware Risks: Highly autonomous systems are attractive targets for malicious control. Military systems with autonomous targeting features could be hijacked for unauthorized engagements. This is not merely theoretical—various international think tanks have warned about autonomous drone swarms being weaponized by state and non-state actors [18].

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 06 June 2025
Page No: 2988 - 2995
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0440

Table 2 Security Breach Types and Domain Susceptibility

Type of Breach	Healthcare	Vehicles	Finance	Defense
Adversarial Inputs	Medium	High	Medium	High
Data Poisoning	Low	Medium	High	Medium
Remote Hijacking	Medium	High	Low	High
Surveillance Exploits	High	Low	Medium	High

3.3.Comparative Risk Synthesis

Integrating the ethical and security perspectives offers a holistic view of AI risk landscapes. The heatmap below synthesizes these factors Table 3. This analysis suggests that while awareness of AI risks is growing, the maturity of mitigation frameworks often lags behind, especially in high-risk domains like defense.

Table 3 Heat map – Combined Risk Intensity by Domain

Risk Category	Healthcare	Autonomous Vehicles	Finance	Defense
Ethical Intensity	High	High	Medium	High
Security Intensity	Medium	High	High	Very High
Mitigation Maturity	Low	Medium	Medium	Low

3.4.Current Mitigation Measures and Their Shortcomings

There are some ways we're trying to reduce the risks that come with autonomous AI, but most of the current approaches tend to be reactive—they deal with problems after they show up, rather than stopping them before they happen:

- Adversarial Training works pretty well, but it usually only applies to specific situations. What works in one area doesn't always work in another [19].
- Explainable AI (XAI) is improving, and there's been progress, but these models still aren't used much in real-time systems, where they could really make a difference [20].
- Regulatory Frameworks like the EU's AI Act are a step in the right direction, but they're tough to enforce globally—different countries have their own rules, and not everyone is on the same page [21].

3.5.Ethical Governance and Policy Challenges

There's still a big gap between what ethical guidelines say and how they're actually put into practice. Most of the time, AI development is led by private companies, and let's be honest—business goals often win out over ethical concerns. And beyond that, there's still no clear agreement across the world on big stuff—like whether or not autonomous weapons should even be allowed at all [22]. Another major issue is that public voices—especially those from vulnerable groups—aren't really part of the conversation. A lot of existing governance models don't do a good job of including diverse perspectives, which means the rules being created often don't reflect everyone's needs or concerns.

Conclusion

Autonomous AI showing up in more places these days—cars that can drive on their own, tools that help doctors, systems that handle money, even tech used in military decisions. It's clearly useful. But as discussed in this paper, there are also some serious risks that come with it. But like this paper has laid out, it also brings some serious risks that we really need to pay attention to. And if we're not careful, those risks could end up doing real harm—not just to individuals, but to entire communities. Looking at it from an ethics angle, it brings up some hard questions. If a machine makes a call, who do you hold accountable? And since a lot of these systems are

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 06 June 2025
Page No: 2988 - 2995
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0440

trained on biased data, they can end up being unfair to people—and most of the time, no one even realizes it right away. On top of that, many of them don't really explain how they reach their conclusions, which makes it hard for people to trust what they're doing. And when things go wrong—as they sometimes will—it's not always clear who should be held accountable. And that becomes a real concern in areas like hiring, healthcare, or law enforcement—where what the system decides can seriously affect someone's life in ways that aren't easy to take back.

On the security side, these systems aren't like regular software. They learn and adapt, and that means they can act in unpredictable ways. If someone feeds them the wrong data on purpose or finds a way to confuse them, the results can be serious. Attacks like that are getting more common, and right now, we don't have great ways to stop them. There are efforts out there—some global groups and governments are trying to put rules in place. Things like the EU's AI Act or OECD's principles are a start [23]. But they don't go far enough. Most of the time, these efforts are either about technical solutions or about policy—not both. And a lot of those rules don't really carry over to other parts of the world. So there are still plenty of gaps left unaddressed. To move forward, we need to think bigger:

- Stronger policies that make sure ethics are built into the tech from the beginning.
- We need to make sure there are clear rules, so people actually get how these systems are making important decisions—especially when those decisions affect their lives.
- There should be people outside the companies who are actually checking this stuff regularly. If no one's looking from the outside, problems can slip through—and by the time we notice, it might already be too late to fix them easily.
- It also shouldn't just be up to the tech companies to figure this out. We need lawmakers, researchers, and regular people involved too. This tech doesn't exist in a bubble—it affects everyone, so everyone deserves to be part of the conversation.

And as AI gets more advanced and starts doing more without us, these problems are only going to get bigger. If we don't build in ethics and safety from the start, we're just asking for trouble later on. It shouldn't be optional—it should be the first thing we think about.

Acknowledgements

I'm genuinely grateful to the researchers, institutions, and organizations whose work helped shape this paper. Their insights, studies, and documentation were a huge part of the foundation for this project. A big thank you as well to the open-access databases, think tanks, and global AI ethics groups that continue to push for transparency and responsibility in how AI is developed—your efforts truly matter.

References

[1]. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707.
[2]. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
[3]. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org.
[4]. Human Rights Watch. (2020). Stopping killer robots: Country positions on banning fully autonomous weapons and retaining meaningful human control.
[5]. Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.
[6]. European Commission. (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).
[7]. AI4People. (2018). An ethical framework for a good AI society. https://www.eismd.eu/ai4people/
[8]. National Institute of Standards and Technology. (2023). AI Risk Management Framework (AI RMF 1.0).
[9]. National Transportation Safety Board. (2020). Collision between vehicle controlled by developmental automated driving system

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 06 June 2025
Page No: 2988 - 2995
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0440

and pedestrian.

[10]. Reuters. (2018). Amazon scraps secret AI recruiting tool that showed bias against women.

[11]. Silverman, R. E. (2023). AI and algorithmic trading: Innovation or instability? Journal of Financial Innovation, 15(3), 211–229.

[12]. United Nations Panel of Experts. (2021). Final report on Libya (S/2021/229).

[13]. Ross, C., & Swetlitz, I. (2017). IBM pitched Watson as a revolution in cancer care. It's nowhere close. STAT News.

[14]. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 77–91.

[15]. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2).

[16]. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xia, C., ... & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1625–1634.

[17]. Steinhardt, J., Koh, P. W., & Liang, P. (2017). Certified defenses for data poisoning attacks. Advances in Neural Information Processing Systems, 30.

[18]. Scharre, P. (2018). Army of none: Autonomous weapons and the future of war. W. W. Norton & Company.

[19]. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations (ICLR).

[20]. Gunning, D., & Aha, D. W. (2019). DARPA's explainable artificial intelligence (XAI) program. AI Magazine, 40(2), 44–58.

[21]. European Parliament. (2023). Artificial Intelligence Act: Legislative framework.

[22]. Future of Life Institute. (2021). Autonomous weapons open letter.

[23]. OECD. (2022). OECD Framework for the Classification of AI Systems.