Assignment 1

Research into AI Ethics & Responsible AI

Authors

Alex Vrsecky 104268899

Version History

Version	Sections Updated	Updated by

Statement of Contribution

Name (Student No.)	Contribution	Signature
Alex Vrsecky (104268899)	Report	ALEX VRSECKY

Table of Contents

Executive Summary

Introduction

The increasing prevalence and sophistication of Artificial Intelligence (AI) necessitate a comprehensive understanding of its ethical and societal implications. As AI technologies rapidly permeate various aspects of life, including institutions, infrastructure, products, and services, they present both immense benefits and significant risks. One of the most contentious debates in AI ethics is whether AI systems should be allowed to kill. This question intersects with deep moral considerations, as killing is widely regarded as unethical and unjust. Critics argue that the development of lethal autonomous weapons could trigger a dangerous global arms race, akin to the Cold War-era nuclear buildup between the U.S. and the USSR. There is also concern that these technologies could fall into the hands of terrorist organisations or groups engaged in acts of genocide. On the other hand, banning such development may be unrealistic given the massive industry investment in military AI. Proponents argue that AI soldiers could reduce human casualties by replacing human soldiers and, due to their programmed nature, might be better at adhering to the laws of war and rules of engagement.

The Australian Government, in its “Proposals Paper for Introducing Mandatory Guardrails for AI in High-Risk Settings”, acknowledges the urgency of establishing regulations to ensure AI is used responsibly. This initiative proposes a risk-based approach with mandatory guardrails focusing on testing, transparency, and accountability in high-risk AI applications, including autonomous weapons. Ethical frameworks help in evaluating this issue. A deontological perspective suggests that killing is inherently wrong and should be avoided at all costs, reinforcing the argument against autonomous lethal AI. However, from a utilitarian standpoint, if these systems are inevitably developed, the focus should be on minimising harm and ensuring they are used in a way that leads to the least amount of suffering. The academic community also emphasises the importance of integrating AI ethics into education, as outlined in “Ethical Considerations in Artificial Intelligence Courses”, which advocates for teaching ethical theories like deontology, utilitarianism, and virtue ethics through case studies. The scope of this discussion is primarily centered on the Australian regulatory landscape for high-risk AI, analysed through ethical considerations and academic discourse. Additionally, this exploration is framed within the context of a university-level research assignment on AI ethics and responsible AI, highlighting the critical role of education in preparing future AI practitioners to navigate these complex moral and regulatory challenges.

Findings

Ethical Perspectives on AI and Lethal Decision

As artificial intelligence becomes increasingly embedded in everyday life, it is crucial to consider how these systems ought to behave within our society. AI has evolved from a niche research focus to a powerful, general-purpose technology influencing areas such as customer service, transport, education, and even national security. With such expansive reach, the ethical dimensions of AI design and usage must be carefully examined to ensure these technologies serve the public good. The Australian Government's consultations have revealed that the current regulatory system is not fit for purpose in addressing the distinct risks posed by AI, underscoring the need for preventative, risk-based guardrails.

To navigate the complexities of ethical AI behaviour, philosophical frameworks offer valuable guidance. A deontological approach emphasises that AI systems should adhere to rules or duties—such as respecting privacy and avoiding harm—regardless of outcomes. The Government's proposal for mandatory guardrails incorporates accountability mechanisms, aligning with deontological principles by ensuring ethical obligations are upheld. In contrast, utilitarianism promotes behaviours that maximise overall well-being and minimise harm, raising important questions about who stands to benefit and who may be disadvantaged. The Government's risk-based strategy—designed to prevent catastrophic harm before it occurs—echoes this utilitarian concern with weighing potential benefits and risks. Meanwhile, virtue ethics focuses on the character and intentions of those developing and deploying AI, advocating for responsibility, humility, and transparency. The emphasis placed by the Government on transparency and accountability in its proposed guardrails aligns with virtue ethics by promoting ethical conduct across the AI lifecycle. Governmental bodies are clearly recognising the stakes. The Australian Government has proposed mandatory guardrails for AI in high-risk scenarios, highlighting the importance of rigorous testing, transparency, and accountability. These preventative measures are intended to apply throughout the AI supply chain and lifecycle, reflecting a holistic approach to responsible AI governance. Public consultation is also underway, with feedback actively sought from a wide cross-section of society, further reinforcing the democratic importance of the issue.

Ultimately, determining how AI should behave is not merely about avoiding harm—it's about proactively aligning these systems with human values and societal goals. This means establishing clear expectations for safe and responsible use, fostering public trust through transparency, and ensuring AI remains a force that empowers, rather than undermines, our collective well-being. The Australian Government acknowledges that by

implementing effective guardrails and regulation, innovation and adoption will be supported—delivering long-term benefits to the nation.

Risks and Challenges of Lethal Autonomous Weapons

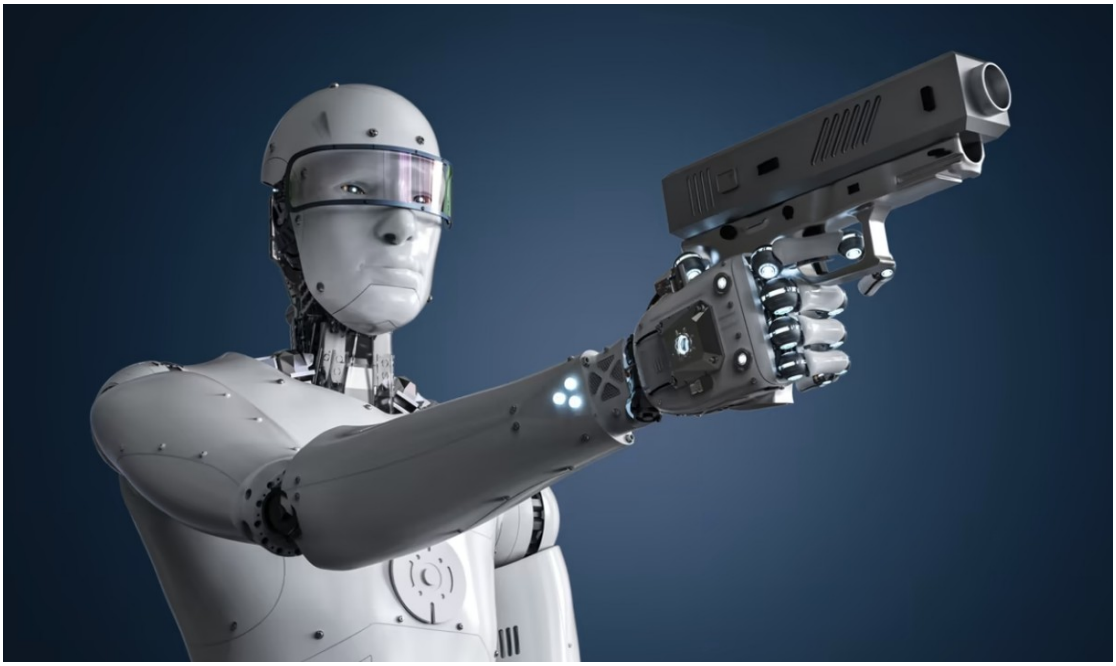

Figure 1: Decision to kill

The development and deployment of lethal autonomous weapons (LAWs) pose a multitude of ethical, legal, and geopolitical risks, particularly as AI becomes increasingly entangled in national defence strategies.

A significant shift in industry attitudes is evident in Alphabet’s recent revision of its AI ethics guidelines. Previously committed to not pursuing AI technologies that could "cause or are likely to cause overall harm," Alphabet—Google’s parent company—has removed this clause. AI chief Demis Hassabis justified the change as a response to a "changing world," emphasising AI’s role in protecting "national security." The updated stance that "democracies should lead in AI" reflects a broader rationale for engaging in AI-driven military applications. This signals a troubling trend: as geopolitical competition intensifies, major tech firms may begin actively contributing to a global arms race in AI weaponisation. Though explicit examples of AI misuse by terrorists or authoritarian regimes were not found in the referenced sources, this concern is strongly implied. The Australian Government’s AI guardrails proposal acknowledges that AI “amplifies and creates new risks,” with specific emphasis on high-risk applications like autonomous weapons. The potential for such technologies to be co-opted by malicious actors underscores the urgent need for robust international safeguards. Accountability is another critical issue. The

Associated Press investigation into the Israeli military's use of AI from Microsoft and OpenAI for targeting in conflict zones raises serious concerns. While the military asserts that human analysts review AI-generated targets and that multiple oversight layers exist, the sheer scale and speed at which AI systems can generate targets calls this into question. The risk of flawed data, biased algorithms, or over-reliance on machine recommendations could lead to tragic mistakes with little clarity on who—or what—is ultimately responsible. This lack of legal and ethical clarity is a central concern in academic and policy discussions. Burton et al. reference the inclusion of "Should AI Systems Be Allowed to Kill?" in AI ethics curricula, reflecting growing awareness and debate. Similarly, the Australian Government's policy paper notes that many existing legal frameworks assume human decision-making, creating a vacuum when machines independently make life-and-death choices. As responsibility becomes diffused between developers, deployers, and AI systems, the challenge of assigning legal liability becomes increasingly complex.

– Discuss potential dangers, including global arms races, the risk of misuse by terrorists or authoritarian regimes, and the lack of accountability when AI makes life-and-death decisions without human oversight.

Regulatory and Educational Approaches for Responsible AI

Australia is proactively shaping the regulatory landscape for Artificial Intelligence (AI), particularly in high-risk applications, as outlined in their "Proposals paper for introducing mandatory guardrails for AI in high-risk settings." This initiative emphasises stringent testing, transparency, and accountability measures. Simultaneously, the nation recognises the critical role of AI ethics education in preparing future professionals to navigate these complex issues responsibly, as highlighted in resources like the "Ethical Considerations in Artificial Intelligence Courses" article

The proposed mandatory guardrails directly address testing requirements for high-risk AI. Guardrail 4 mandates that organisations must rigorously test and evaluate AI model performance before deploying a system in a high-risk setting, and continuously monitor its operation. High-risk settings, as defined in the proposal, include applications with significant potential for harm to individuals or society, such as those in law enforcement, healthcare diagnostics, and critical infrastructure management. This testing aims to ensure AI models meet specific, objective, and measurable performance metrics, effectively managing associated risks. For example, facial recognition systems require accuracy testing across diverse social groups to mitigate discriminatory biases, while General-Purpose AI (GPAI) models necessitate adversarial testing to detect emergent or dangerous capabilities. Transparency is enforced through several guardrails. Guardrail 6 requires organisations to clearly and accessibly inform end-users about AI-enabled decisions, interactions, and generated content. This includes disclosing when AI is used in

decision-making processes, during direct interactions, and ensuring AI-generated outputs are distinguishable from human-created content. Guardrail 8 promotes transparency across the AI supply chain, compelling organisations to share information about high-risk AI systems with developers, deployers, and other stakeholders. Developers, who are responsible for building the AI models, and deployers, who integrate and use those models within specific applications, have distinct roles. Developers must supply information about the models capabilities, risks, and limitations, and deployers must understand that information to use the models responsibly.

Accountability is addressed through several guardrails. Guardrail 1 mandates that organisations establish and publish accountability processes, including governance policies, clear roles, and regulatory compliance strategies. This involves documented risk management, defined responsibilities, and staff training. Guardrail 7 requires organisations to establish processes for affected individuals to challenge AI usage or outcomes, including internal complaint handling and human oversight. Guardrail 9 necessitates comprehensive record-keeping, including technical documentation, throughout the AI system's lifecycle for regulatory compliance assessments. Guardrail 10 proposes mandatory conformity assessments, conducted by qualified, independent assessors, to certify compliance with the guardrails. The responsibility for implementing these guardrails is shared between developers and deployers, acknowledging their distinct roles and capabilities. AI ethics education is vital for preparing future professionals to navigate these regulatory and ethical challenges.

The "Ethical Considerations in Artificial Intelligence Courses" article advocates for integrating ethical theories like deontology, utilitarianism, and virtue ethics into AI curricula. By applying these theories to case studies, students develop critical thinking skills to analyse ethical dilemmas. The "Introduction to AI Assignment 1" emphasise research into AI Ethics and Responsible AI, prompting students to consider fundamental ethical questions and evaluate the Australian government's proposed principles. Furthermore, resources like textbooks and online platforms dedicated to AI and robot ethics, as mentioned by Burton et al., provide valuable tools for educators. Implementing these guardrails will not be without challenges. The rapid pace of AI development may outstrip the regulatory frameworks. Additionally, international collaboration will be vital to ensure global consistency in AI governance.

By combining robust regulatory frameworks with comprehensive AI ethics education, Australia aims to foster a future where AI innovation is both beneficial and responsible, effectively managing risks and promoting ethical development.

– Examine Australia’s proposed AI guardrails, focusing on testing, transparency, and accountability in high-risk AI applications, and highlight the role of AI ethics education in preparing future professionals to navigate these challenges responsibly.

Conclusion

References

https://apnews.com/article/israel-palestinians-ai-technology-737bc17af7b03e98c29cec4e15d0f108

By MICHAEL BIESECKER, SAM MEDNICK and GARANCE BURKE

Updated 11:06 PM GMT+11, February 18, 2025

https://www.theguardian.com/technology/2025/feb/05/google-owner-drops-promise-not-to-use-ai-for-weapons?utm_source=chatgpt.com

Lenore Taylor

Editor, Guardian Australia

https://www.politico.com/newsletters/digital-future-daily/2023/01/25/should-a-robot-be-allowed-to-kill-you-00079467

By MATT BERG