GCSP
Geneva Centre for Security Policy

FEBRUARY 2026 | ISSUE 45

Strategic Security Analysis

What Self-driving Car Operations Can Teach Us about Incorporating AI into Weapons Systems

Mary L. Cummings

www.gcsp.ch

The Geneva Centre for Security Policy

The Geneva Centre for Security Policy (GCSP) is an international foundation that aims to advance global cooperation, security and peace. The foundation is supported by the Swiss government and governed by 55 member states. The GCSP provides a unique 360° approach to learn about and solve global challenges. The foundation's mission is to educate leaders, facilitate dialogue, advise through in-house research, inspire new ideas and connect experts to develop sustainable solutions to build a more peaceful future.

Strategic Security Analyses

The GCSP Strategic Security Analyses series publishes short papers that address a current security issue. These papers provide background information about the theme, identify the main issues and challenges, and propose policy recommendations.

This series is edited by Dr Jean-Marc Rickli, Head of Global and Emerging Risks.

About the author

Professor M.L. Cummings received her BS in Mathematics from the US Naval Academy in 1988, her MS in Space Systems Engineering from the Naval Postgraduate School in 1994 and her PhD in Systems Engineering from the University of Virginia in 2004. A naval officer and military pilot from 1988 to 1999, she was one of the US Navy's first female fighter pilots. She is a professor in the George Mason University College of Engineering and Computing and is the director of the Mason Autonomy and Robotics Center.

Acknowledgment

This work was sponsored in part by the US Office of Naval Research Science of Autonomy programme.

ISBN: 978-2-88947-333-5

Cover photo: showcake, Shutterstock.com

Key points

- The incorporation of AI into weapons will face similar reliability issues as self-driving cars, including hallucinations, poor handling of uncertainty, latent failure modes and planning failures.
- Generative AI and agentic AI can introduce uncontrolled non-determinism and lack reliable reasoning, spatial awareness, and self-verification – making them highly unsuitable for weapons systems, where precision and predictability are critical.
- AI-enabled weapons are the highest risk in terms of safety-criticality and non-determinism. This risk profile demands rigorous governance and testing protocols rather than outright bans, given AI's widespread civilian applications.
- Lessons from self-driving cars show that real-world testing is essential to uncover latent failure modes. Exclusive reliance on simulation for weapons AI could lead to catastrophic oversights.
- To mitigate AI risks, governments and organisations must invest in physical AI test ranges and develop standardised evaluation protocols through global collaborations, ensuring transparency and accountability in military AI deployment.

Introduction

Artificial intelligence (AI) has become a must-have capability for military operations, receiving much attention, both positive and negative, in the Ukraine-Russian[1] and Israel-Gaza[2] conflicts. There are many kinds of AI, ranging from rules-based AI that was popular in expert-based systems 20 years ago to connectionist AI where neural networks learn patterns from large amounts of empirical data. Currently, the vast majority of AI used in weapons systems occurs in the form of rules-based AI (sometimes called good old-fashioned AI, or GOFAI). However, significant efforts are under way to incorporate connectionist AI into weapons systems, including agentic AI that leverages generative AI.[3]

There is much debate as to whether and what types of AI should be allowed in weapons systems and what governance frameworks are needed to mitigate AI risks in military systems.[4] Because militaries are not open to sharing performance issues and other problems associated with the deployment of AI on the battlefield, it is difficult for decision-makers and policymakers to know where red lines should be drawn, or even what issues should garner the most interest or concern. However, significant performance data has been generated in an adjacent field that can shed much light on the use of AI in a safety-critical system, i.e. that of self-driving cars.

There is much debate as to whether and what types of AI should be allowed in weapons systems and what governance frameworks are needed to mitigate AI risks in military systems.

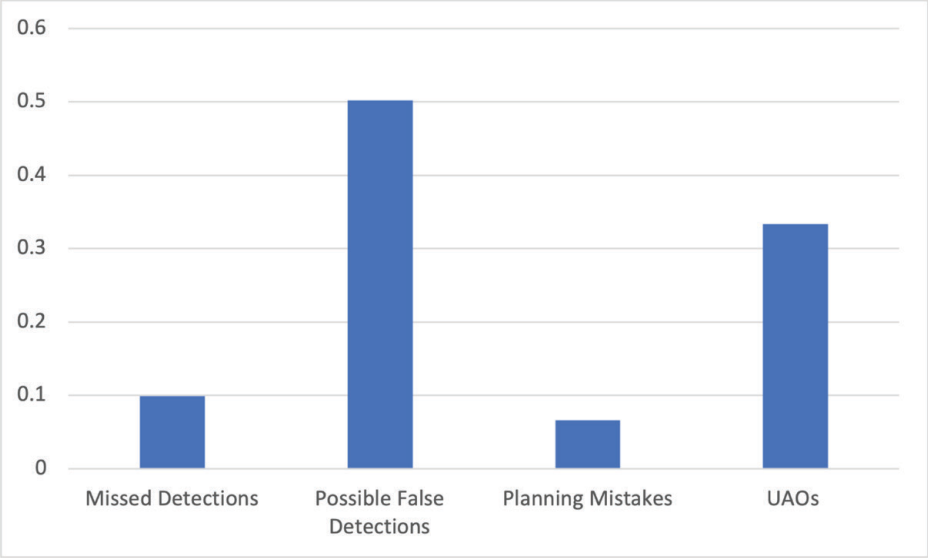
This GCSP Strategic Security Assessment highlights the lessons learned in the operation of self-driving cars in San Francisco, California, especially with regard to new and unexpected crash modalities. It uses this information to develop a hazard analysis framework for the use of AI in safety-critical systems, especially in weapons. It concludes with a set of recommendations to mitigate risk in the use of AI in weapons systems, including the need for expanded AI testing partnerships.

Self-driving car failure modes

The vast majority of self-driving car development and deployments have occurred in San Francisco. Such operations are so numerous in the United States that the National Highway Transportation Safety Administration (NHTSA) requires that any company that deploys an autonomous vehicle on public roads must report crashes while some form of autonomy is engaged.[5] A previous analysis of the failure modes of self-driving car operations in San Francisco based on the NHTSA data revealed four general crash modes: (1) missed detections, (2) possible false detections, (3) planning mistakes, and (4) unexpected actions by others (UAOs),[6] which are listed in this order in Figure 1, because the first two categories are perception problems and the second two are planning problems.

By far the largest category of crashes is due to possible false detections by the perception system.

Figure 1: Self-driving car crash categories in California in 2023



Hallucinations

By far the largest category of crashes is due to possible false detections by the perception system. In these crashes, self-driving cars are struck from behind when they execute an emergency braking manoeuvre because they see a non-existent obstacle, catching following drivers by surprise. Human crashes only result in struck-from-behind crashes at rates around 30%,[7] so the fact that 50% of autonomous vehicle (AV) crashes are caused by other vehicles rear-ending them is cause for concern.

Researchers have long recognised that self-driving cars' sensors and software can detect an obstacle that does not exist.[8] It is not known what causes these false detections – also known as false positives or hallucinations – although there is some evidence shadows cause them.[9] Sensor washout at low sun angles could also be a factor.[10]

This problem is also known as phantom braking, and occurs not just for self-driving cars, but any vehicle with an advanced driving-assist system. The problem is so widespread in Teslas that the US NHSTA opened an investigation because of several deaths and serious injuries that have occurred.[11] Several class action lawsuits are currently pending against Tesla in both the United States and Australia for this problem.[12] It cannot be emphasised enough that no known solutions are available to prevent such hallucinations.

Unexpected actions by others (UAOs)

The second-largest category of crashes for self-driving cars results from mistakes made by a self-driving car due to the unexpected actions by others (33%). The AI built for self-driving cars relies on neural networks to recognise emergent patterns, and from these patterns determine what the next likely set of actions should be. While pattern recognition approximates reasoning, it is not actual reasoning and does not generalise well in the face of uncertainty.

The accident that best exemplifies this category is the pedestrian struck by a Cruise self-driving vehicle in San Francisco in late 2023.[13] A pedestrian crossing a street at night was initially struck by another human-driven car, knocking her into the path of an oncoming Cruise self-driving car vehicle. After correctly sensing her and executing an emergency braking manoeuvre, the pedestrian was struck and fell underneath the car. The Cruise AV failed to detect her there and moved to the curb, dragging her for about 20 feet. She was in hospital for almost a year in recovery.[14]

While pattern recognition approximates reasoning, it is not actual reasoning and does not generalise well in the face of uncertainty.

Self-driving cars do not have sensors to detect objects underneath them, and several pets have been killed under cars in this way,[15] but human drivers occasionally do this as well. What was noteworthy was the fact that the self-driving car's accelerometers registered a collision just after detecting the pedestrian on a camera. Indeed, the car predicted this collision at least seven seconds before the pedestrian was hit, and not only did it not sound its horn or slow down, but it sped up as it approached the predicted point of collision. Moreover, once the pedestrian was under the car, the AI onboard the vehicle simply "forgot" she was there, because computer-vision AI does not "know" anything and has no memory. This is why it decided to move with a pedestrian underneath the car, an act that would not likely occur with a human driver.

This case highlights that self-driving cars do not have the ability to predict unseen events, which means that they cannot imagine what might happen and adopt a defensive posture in scenarios with significant uncertainty. Indeed, the lack of defensive driving ability has led to other collisions involving backing trucks, as well as poor decision-making for unprotected left turns.[16] As will be discussed in more detail in the next section, all forms of neural networked-based AI struggle with uncertainty. So, when the driving setting closely matches the data self-driving AI has trained on, the car can give the illusion of safe driving, but when the scenario does not match this underlying training, then serious problems can emerge.

Lower-frequency crashes

The remaining 17% of crashes were caused by missed detections (10%) and an inability to generate correct plans (7%). In terms of missed detections, Waymo self-driving vehicles have run into poles, gates and flooded roads),[17] which indicates that even when equipped with a laser range-finder called a LIDAR (light detection and ranging), this technology cannot always build what is known as a "world model". LIDAR cannot sense the depth of water and has blind spots, especially when dealing with long, very thin objects.

Crashes in the planning mistakes category occur when the world is sensed correctly, but erroneous plans are executed. Self-driving cars can struggle to develop safe trajectories around bicycles, leading to actions that cannot account for bicycle dynamics, causing the rider to crash into the self-driving vehicle. Several such incidents with self-driving cars have occurred in San Francisco.[18] More recently, while no injuries have occurred, Waymo was required to recall all its self-driving vehicles because they kept going around stopped school busses.[19] This manoeuvre is illegal in the United States, because it significantly raises the risk that one or more children could be hit.

Of these four crash categories, the only ones that were anticipated in the design and development of these systems were the low-frequency crashes involving missed detections and poor planning decisions. It was not until self-driving cars were deployed in the real world that the more problematic collisions in terms of hallucinations and UAOs emerged, and the industry is still struggling to address these failure modes. While companies make improvements after every crash, the hypercompetitive environment means companies are not sharing best practices. The sharing of accident reports in commercial aviation was key to solving problems with the introduction of autonomy in the 1990s, so the refusal of self-driving companies to be more transparent means that learning from accidents will be slow and uneven.

While the self-driving problems discussed here can help shed light on how weapons with embedded AI could fail, the potential use of agentic AI is introducing new risks, which is further discussed in the next section.

AI in the form of computer vision, including end-to-end learning from videos, is not only far from achieving its intended functionality, but we do not even fully understand all its possible failure modes.

Agentic AI and weapons

While there is much wailing and gnashing of teeth over the potential existential risks of AI, especially when used in weapons, such arguments presume that not only does AI function as intended, but that it can achieve either general or even super intelligence. As illustrated in the previous section, AI in the form of computer vision, including end-to-end learning from videos, is not only far from achieving its intended functionality, but we do not even fully understand all its possible failure modes. And while AI-enabled computer vision is a cornerstone of lethal autonomous weapons, there is significant growing interest in integrating generative AI (GenAI) into weapons systems.[20] The inclusion of generative AI in a system that can "close the loop" is known as agentic AI, meaning onboard AI can both develop and execute a plan, including actuation physical devices.

Traditional machine-learning models focus on tasks like computer vision object detection and classification, or teaching a robot to execute a control manoeuvre. GenAI models focus on content creation through transformers, which are essentially neural network architectures that generate text, image or audio output when given a similar input based on learned patterns in a vast training dataset. For GenAI to be used in a weapon, it would need to be embedded in the form of agentic AI.

Just as problematic as traditional machine learning has been in self-driving cars, GenAI is a deeply flawed technology when applied in safety-critical systems. GenAI notoriously hallucinates (aka makes mistakes) and one of its inventers, Yann Lecun, said that "[Large language models] hallucinate answers .... They can't really be factual".[21] GenAI can also not reliably spatially reason, which is critical for any kind of weapon. GPT-4 struggles to reason when given a grid and images,[22] and at best only achieves 53% accuracy in spatial reasoning tasks.[23] This is far below what is needed in lethal engagements, so currently, spatial planning is well outside the ability of GenAI.

It is also well established that GenAI cannot reliably complete mathematical problems, which is an absolute requirement in a weapons system.[24] It also shows no ability to execute commonsense reasoning as it relates to the physical environment.[25] Very similar to the pedestrian accidents caused by self-driving cars, GenAI also cannot self-verify, i.e. determine whether correct task execution has occurred.[26]

GenAI technologies do not show reliable reasoning abilities across every aspect of reasoning needed for autonomous weapons. They approximate and mirror human reasoning, but cannot reliably and predictably demonstrate consistent reasoning, which absolutely must exist for any system operating

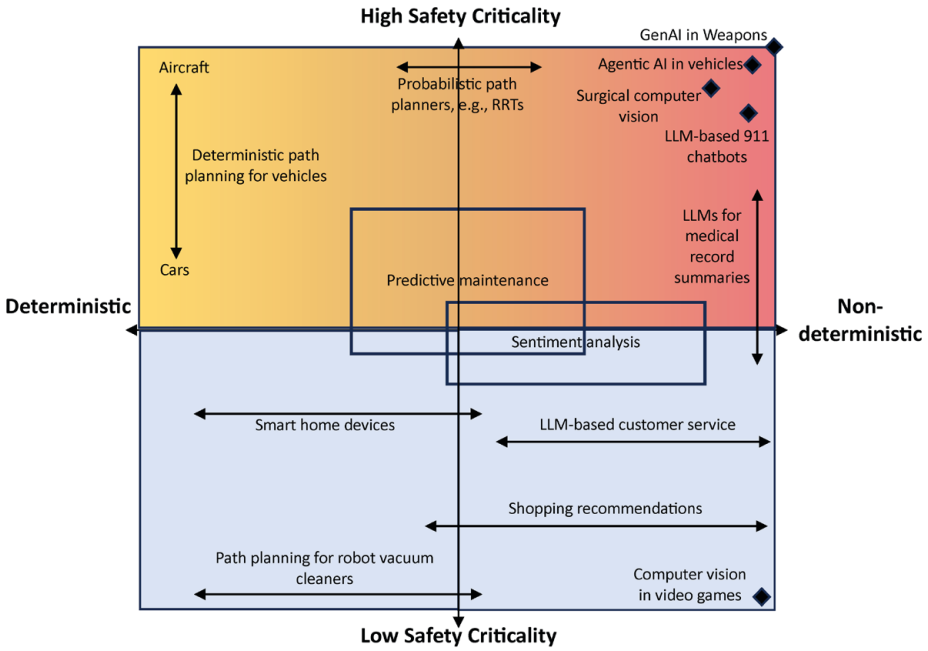
in a safety-critical setting like weapons deployment. As will be discussed in the next section, the degree of determinism and safety criticality govern just how risky an AI technology is.

Safety-critical AI hazard analysis

Despite the clear problems with reliability for traditional machine-learning and agentic AI technologies, companies and governments continue to laud them, even on the battlefield. So, to better understand the risk for including AI in weapons, Figure 2 represents various agentic AI systems, both deployed and futuristic, plotted in terms of the degree of non-determinism on the x axis and safety criticality on the y axis. AI systems that are completely predictable, i.e. always produce the exact same output for the same input without randomness, are completely deterministic. The shading across the top of Figure 2 indicates increasing risk of a harmful human outcome such as serious injury or death.

> AI … cannot reliably and predictably demonstrate consistent reasoning, which absolutely must exist for any system operating in a safety-critical setting like weapons deployment.

Figure 2: Various Agentic AI systems represented by level of safety criticality and degree of non-determinism.



Note: In terms of human harm, blue regions indicate negligible risk, yellow indicates moderate risk and red indicates significant risk.

The bottom half of Figure 2 represents those systems that generally operate in low safety-criticality settings with increasing degrees of non-determinism, like robot vacuum cleaners. Such robots can operate through rule-based AI (like a lawn mower moving in a predictable and repeatable path), to more random non-deterministic operations with no seeming logic in the choice of direction (but, in the case of robot vacuum cleaners, will vacuum every inch of floor). In the bottom right corner, computer vision can be used in an interactive video game for human pose estimation, but if and when it fails, it has very little bearing on the physical well-being of the human user.

As the safety-criticality axis increases in Figure 2, the examples reflect uses that could have increasingly harmful effects. Shopping recommendations that

are wrong carry very low risk of harm. However, sentiment analysis agentic AI straddles the midline between low and medium-to-high risk, since mental health agentic AI chatbots are growing in popularity. Medical practitioners are justifiably concerned that agentic AI cannot be certain to do no harm,[27] and if a human's sentiment is incorrectly estimated, there is a real risk of serious harm. Indeed, the first lawsuit asserting a suicide was caused by a chatbot has been filed.[28] Moreover, large language models (LLMs) have been shown to hallucinate in transcription applications,[29] and such errors carry very real possibilities for human harm in medical settings, so they are very much in the red high-risk zone.

As seen in Figure 2, the agentic AI application with the most risk is the use of GenAI in weapons. It is an extreme case of safety criticality and also operates at the highest degrees of non-determinism. Just underneath this use case in Figure 2 is the use of agentic AI in vehicles like self-driving cars and autonomous aircraft.

To fill the void left by current US policy, the United Nations, NATO, and countries interested in safe and effective AI need to form partnerships to develop tangible and practical AI testing protocols.

Strategic implications

Given that the use of AI in weapons and in enabling technologies like computer vision sits in the upper right corner of Figure 2, some policymakers may be tempted to use this as a justification for recommending the banning of AI in weapons. However, the entirety of Figure 2 illustrates just why this is not a practical strategy. The AI that can be used in weapons is the same AI that can be used in a smart home device or in the path planners for commercial aircraft. Thus, we need to be able to preserve the ability to use AI across a number of systems, but we also need to understand the critical role of the human in AI functionality.

One of the greatest current AI myths is that self-driving cars exist – but there is no company today that operates an actual self-driving car. All companies require significant human oversight in the form of remote operators that either interpret the world for the car and give commands for different movements or take control of the car and driver through a remote car cockpit. Some companies likely use a mixture of both. Any claims that AI is a better driver than humans are false, since currently all self-driving is human-assisted driving. This is incredibly important when thinking about how and why to use AI on the battlefield, since weapons move at speeds that exceed the ability of humans exercise meaningful oversight.

The inability of humans to directly oversee the operations of weapons-based AI means that if countries want to use such weapons, then the testing and evaluation of AI in battlefield systems must dramatically improve, i.e. we need meaningful human certification of military AI.[30] Unfortunately, the US government has significantly reduced its AI testing and evaluation workforce, and Peter Thiel, a politically influential co-owner of Palantir, a leading military AI company, posits that the regulation of AI will hasten the arrival of the antichrist.[31] To fill the void left by current US policy, the United Nations, NATO, and countries interested in safe and effective AI need to form partnerships to develop tangible and practical AI testing protocols. These organisations and countries need to move beyond vague, high-level guidelines to specific protocols with access to real-world testing facilities.

One critical lesson that should be learned from self-driving operations is the importance of physical testing, because this was critical in learning about various unexpected failure modes. If companies and governments only ever test AI in weapons in simulation, they may never discover latent failures and gaps in functionality. No data-driven AI system can be matured to the point of successful deployment by only using simulation, and the exposure of self-driving car systems to real-world scenarios has been pivotal. Countries need to invest in physical AI testing ranges, and while they are expensive,

private-public partnerships could help defray costs. Such collaborations would also give participating countries access to unprecedented data and lessons learned. Given all the gaps in AI capabilities seen in self-driving cars and GenAI applications, it is imperative that testing and evaluation become the focal point for responsible use of AI in *all* safety-critical systems, but especially battlefield technology.

Conclusion

The deployment of self-driving cars offers critical insights into the challenges of integrating AI into safety-critical systems, particularly weapons systems. Despite years of development and billions of dollars in investment, autonomous vehicles still exhibit unpredictable behaviours such as hallucinations, poor handling of uncertainty, and failures in reasoning – issues that stem from the inherent non-determinism of neural-network-based AI. These same vulnerabilities are amplified in military contexts, where the stakes are far higher and the margin for error can be very small.

Real-world testing, coupled with meaningful human oversight and certification, is essential to mitigate AI risks in weapons.

Agentic AI and generative models introduce even greater risks, given their inability to reliably reason, self-verify or operate with safety guarantees. As illustrated in the hazard analysis, the combination of extreme safety-criticality and high non-determinism places AI-enabled weapons in the highest risk tier. While banning AI in weapons may seem appealing, such an approach is impractical (a lack of participation in the Ottawa Convention is an example), given the widespread use of similar technologies in civilian and commercial domains. Instead, the focus must shift toward rigorous test and evaluation protocols, physical testing environments, and international partnerships to ensure transparency and accountability.

A critical lesson from self-driving car operations is that simulation alone cannot uncover latent failure modes. Real-world testing, coupled with meaningful human oversight and certification, is essential to mitigate AI risks in weapons. Without such measures, the deployment of AI in weapons systems will remain fraught with uncertainty, posing unacceptable dangers to both military personnel and civilians. Responsible innovation demands that we confront these limitations head-on, prioritising safety and reliability over speed and hype.

Endnotes

1 A. Abdurasulov (2025) "The New AI Arms Race Changing the War in Ukraine", 9 October, https://www.bbc.com/news/articles/cly7jrez2jno.

2 N. Sylvia (2024) "The Israel Defense Forces' Use of AI in Gaza: A Case of Misplaced Purpose", RUSI, 4 July, https://www.rusi.org/explore-our-research/publications/commentary/israel-defense-forces-use-ai-gaza-case-misplaced-purpose.

3 Anduril Industries (2024) "Anduril Partners with OpenAI to Advance U.S. Artificial Intelligence Leadership and Protect U.S. and Allied Forces", 4 December, https://www.anduril.com/news/anduril-partners-with-openai-to-advance-u-s-artificial-intelligence-leadership-and-protect-u-s; B. Gertz (2023) "China's Military Working on AI Weapons and Systems for Warfighting and 'Overthrowing Regimes'", *Washington Times*, 22 August, https://www.washingtontimes.com/news/2023/aug/22/chinas-military-working-ai-weapons-and-systems-war/.

4 M.L. Cummings (2025) "Prohibiting Generative AI in Any Form of Weapon Control", NeurIPS, https://neurips.cc/virtual/2025/loc/san-diego/poster/121921; Global Commission on Responsible Artificial Intelligence in the Military Domain (2025) *Responsible by Design: Strategic Guidance Report on the Risks, Opportunities, and Governance of Artificial Intelligence in the Military Domain*, The Hague, The Hague Centre for Strategic Studies.

5 NHTSA (National Highway Transportation Safety Administration) (2023) "Second Amended Standing General Order 2021-01", US Department of Transportation.

6 M. Cummings and B. Bauchwitz (2024) "Identifying Research Gaps through Self-Driving Car Data Analysis", *IEEE Transactions on Intelligent Vehicles*, 1(10), https://ieeexplore.ieee.org/document/10778107.

7 N.J. Goodall (2021) "Comparison of Automated Vehicle Struck-from-behind Crash Rates with National Rates Using Naturalistic Data", *Accident Analysis & Prevention*, 154(106056), https://www.sciencedirect.com/science/article/abs/pii/S0001457521000877.

8 C. Hogan and G. Sistu (2023) "Automatic Vehicle Ego Body Extraction for Reducing False Detections in Automated Driving Applications", *Artificial Intelligence and Cognitive Science, AICS 2022*: 264-275, https://link.springer.com/chapter/10.1007/978-3-031-26438-2_21.

9 A.K. Sriranga et al. (2020) "Trigger-Based Pothole Detection Using Smartphone and OBD-II", *IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 1(6), DOI:10.1109/CONECCT50063.2020.9198602; B. Bauchwitz and M.L. Cummings (2022) "Individual Differences Dominate Variation in ADAS Takeover Alert Behavior", *Transportation Research Record: Journal of the Transportation Research Board*, 2676(5): 489-499.

10 M.L. Cummings and B. Bauchwitz (2023) "Driver Alerting in ADAS-Equipped Cars: A Field Study", *IEEE International Conference on Assured Autonomy (ICAA)*: 29-33, https://ieeexplore.ieee.org/document/10207596.

11 S. Ridella (2022) "ODI Resume: EA 22-002", NHTSA Office of Defects Investigation, https://static.nhtsa.gov/odi/inv/2022/INOA-EA22002-3184.PDF.

12 M. Scarcella (2024) "Tesla Must Face Part of 'Phantom Braking' Lawsuit, US Judge Rules", Reuters, 22 November, https://www.reuters.com/legal/litigation/tesla-must-face-part-phantom-braking-lawsuit-us-judge-rules-2024-11-22/; Woodsford (2025) "Woodsford Supports JGA Saddler in Major Australian Tesla Class Action", https://woodsford.com/woodsford-supports-jga-saddler-in-major-australian-tesla-class-action/.

13 Quinn Emanuel Trial Lawyers (2024) *Report to the Boards of Directors of Cruise LLC, GM Cruise Holdings LLC, and General Motors Holdings LLC Regarding the October 2, 2023 Accident in San Francisco*, https://assets.ctfassets.net/95kuvdv8zn1v/1mb55pLYkkXVn0nXxEXz7w/9fb0e4938a89dc5cc09bf39e86ce5b9c/2024.01.24_Quinn_Emanuel_Report_re_Cruise.pdf.

14 M.L. Cummings et al. (2024) "A Root Cause Analysis of a Self- Driving Car Dragging a Pedestrian", *IEEE Computer*, 57: 31-40, 10.1109/MC.2024.3429276.

15 Ridella (2022).

16 Ibid.

17 NHTSA (2023).

18 Ibid.

19 X. Walton (2025) "Waymo Recalls Robotaxi Software after School Bus Incidents", *The Hill*, 9 December, https://thehill.com/business/5638543-waymo-recall-robotaxi-software/.

20 M. Stone (2025) "Anduril and Palantir Battlefield Communication System 'Very High Risk,' US Army Memo Says", Reuters, https://www.reuters.com/business/aerospace-defense/anduril-palantir-battlefield-communication-system-has-deep-flaws-army-memo-says-2025-10-03/.

21 B. Marr (2024) "Generative AI Sucks: Meta's Chief AI Scientist Calls for a Shift to Objective-Driven AI", Forbes, 12 April, https://www.forbes.com/sites/bernardmarr/2024/04/12/generative-ai-sucks-metas-chief-ai-scientist-calls-for-a-shift-to-objective-driven-ai/.

22 M. Aghzal et al. (2024) "Look Further Ahead: Testing the Limits of GPT-4 in Path Planning", IEEE 20th International Conference on Automation Science and Engineering, https://arxiv.org/pdf/2406.12000?; W. Wu et al. (2024) "Mind's Eye of LLMs: Visualization-of-Thought Elicits Spatial Reasoning in Large Language Models", 38th Annual Conference on Neural Information Processing Systems, https://arxiv.org/abs/2404.03622.

23 J. Cho et al. (2023) "Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models", International Conference on Computer Vision, https://arxiv.org/abs/2202.04053.

24 K.M. Collins et al. (2024) "Evaluating Language Models for Mathematics through Interactions", *Proceedings of the National Academy of Sciences*, 121(24), https://www.pnas.org/doi/10.1073/pnas.2318124121; T.R. Mcintosh et al. (2024) "A Reasoning and Value Alignment Test to Assess Advanced GPT Reasoning", *ACM Transactions on Interactive Intelligent Systems*, 14(3): 1-17, https://dl.acm.org/doi/10.1145/3670691; P. Mondorf and B. Plank (2024) "Beyond Accuracy: Evaluating the Reasoning Behavior of Large Language Models – A Survey", First Conference on Language Modeling, https://arxiv.org/html/2404.01869v1.

25 Y. Bisk et al. (2020) "PIQA: Reasoning about Physical Commonsense in Natural Language", https://arxiv.org/abs/1911.11641; J.L. Espejel et al. (2023) "GPT-3.5, GPT-4, or BARD? Evaluating LLMs Reasoning Ability in Zero-Shot Setting and Performance Boosting through Prompts", *Natural Language Processing Journal*, 5: 100032, https://www.sciencedirect.com/science/article/pii/S2949719123000298?via%3Dihub.

26 S. Kambhampati et al. (2024) "LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks", 41st International Conference on Machine Learning, https://arxiv.org/abs/2402.01817.

27 A. Fiske et al. (2019) "Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy", *Journal of Medical Internet Research*, 21(5), https://www.jmir.org/2019/5/e13216/.

28 K. Payne (2024) "An AI Chatbot Pushed a Teen to Kill Himself, a Lawsuit against Its Creator Alleges", AP, https://apnews.com/article/chatbot-ai-lawsuit-suicide-teen-artificial-intelligence-9d48adc572100822fdbc3c90d1456bd0.

29 A. Koenecke et al. (2024) "Careless Whisper: Speech-to-Text Hallucination Harms", *FAccT '24: ACM Conference on Fairness, Accountability, and Transparency*: 1672-1681, https://arxiv.org/abs/2402.08021.

30 M.L. Cummings (2019) "Lethal Autonomous Weapons: Meaningful Human Control or Meaningful Human Certification?" *IEEE Technology and Society* 38(10): 20-26, DOI:10.1109/MTS.2019.2948438.

31 A. Au-Yeung (2025) "Peter Thiel Wants Everyone to Think More about the Antichrist", *Wall Street Journal*, 23 September, https://www.wsj.com/tech/peter-thiel-antichrist-lectures-dd28c876?gaa_at=eafs&gaa_n=AWEtsqesZiqjCWxjBU0QSJVXSqQ2aTi49E21rK9b6kKvvqQZkdMTYbWqJbre&gaa_ts=69456659&gaa_sig=nUW7ddRDLVAvIQgs3_a-TBv5RLOCy-2J-0RWOO9egLVNWc4GXDoBAP8q_Hns2KNktn-5BKrRs_iVnPbb6XaU6Q%3D%3D.

Building Peace Together

Geneva Centre for Security Policy
Maison de la paix
Chemin Eugène-Rigot 2D
P.O. Box 1295
1211 Geneva 1
Switzerland
Tel: + 41 22 730 96 00
Contact: www.gcsp.ch/contact
www.gcsp.ch

ISBN: 978-2-88947-333-5

GCSP
Geneva Centre for Security Policy