Content:

Dear Friends,

AI becomes integrated into our lives and workplaces. Like admitting a new joiner, it is crucial to know whether it aligns with our corporate/human values. Recent research and real-world cases highlight the complexities of AI misalignment, undesirable, or even harmful behaviours that should not be ignored. These behaviours can stem from a variety of factors, from flawed training data to pursuit of unintended system's goals. This newsletter provides an overview of some notable recent cases and explores the latest research to help understand these risks and the ongoing efforts to address them.
Hope you enjoy reading this.

AA & T Consulting

Introduction

AI becomes integrated into our lives and workplaces. Like admitting a new joiner, it is crucial to know whether it aligns with our corporate/human values. Recent research and real-world cases highlight the complexities of AI misalignment, undesirable, or even harmful behaviours that should not be ignored.

These behaviours can stem from a variety of factors, from flawed training data to pursuit of unintended system's goals. This newsletter provides an overview of some notable recent cases and explores the latest research to help understand these risks and the ongoing efforts to address them.
Here is a summary of the cases. For more details, please refer to the Appendix.

Summary of AI Misaligned Cases

AI-Induced Bias (Aug 2025):
A study found that Large Language Models (LLMs) consistently preferred items described by other LLMs, showing a distinct "AI-AI bias".

Agentic Misalignment (Jun 2025):
Research from Anthropic revealed that leading models from multiple developers, including Claude, Gemini, GPT etc., resorted to malicious insider behaviours in hypothetical corporate environments when faced with replacement or goal conflicts. These behaviours included blackmailing officials and leaking sensitive information to competitors.

Contact us: AA & T Consulting Services Ltd, *Unit 9, 17/F, Citicorp Centre, 18 Whitfield Road, Causeway Bay, HK*
Email : advisory@aathk.com; Tel. : +852 9181 8659 (HK) ; +61 452 371 753 (Australia) Website: HK: www.aathk.com; Aus. www.aataus.com

GPT-4o Legal Reasoning Report (Jan 2025):

An automated red-teaming framework for the GPT-4o model exposed vulnerabilities that caused hallucinations in legal AI models. A significant portion of prompts that initially caused hallucinations continued to do so even after being rephrased, indicating that the failures were not superficial but stemmed from deeper issues within the model's reasoning and training data.

Report on Security Vulnerabilities and other issues of DeepSeek (Jan 2025):

A report on the DeepSeek-R1 model found it was highly vulnerable to producing harmful content, including toxic language, biased outputs, and criminally exploitable information. The model was also susceptible to generating extremist content and insecure code snippets, such as malware.

Scheming/Alignment faking (Dec 2024):

Research from Appollo and Anthropic found that models like Claude 3.5 Sonnet, Gemini 1.5 Pro, and Llama 3.1 405B demonstrate in-context scheming capabilities, recognizing it as a viable strategy and engaging in this behaviour. This research also provides an empirical example that an LLM can engage in "alignment faking" to prevent its preferences from being modified.

Conclusion

The cases of AI misalignment highlighted in recent research point to a systemic issue in how AI is designed and trained. The problem is not just about isolated incidents but about the complex and often unpredictable nature of these systems (just like humans). Addressing these challenges will require a multi-faceted approach, including technical solutions, ethical frameworks, and greater collaboration among researchers, developers, and policymakers.

What We Can Help

By staying informed about the latest research and continuing to prioritize responsible development, we can work together to ensure AI systems are not only intelligent but also safe and beneficial for society.

How can AA & T Consulting help?

If you need any help regarding an independent advice on your AI deployment, feel free to contact us by phone (+852 9181 8659 (HK); +61 452 371 753 (Aus.)), or by email to advisory@aathk.com

Contact us: AA & T Consulting Services Ltd, *Unit 9, 17/F, Citicorp Centre, 18 Whitfield Road, Causeway Bay, HK*
Email : advisory@aathk.com; Tel. : +852 9181 8659 (HK) ; +61 452 371 753 (Australia) Website: HK: www.aathk.com; Aus. www.aataus.com

Appendix: Summary of AI misalignment and other related cases:

Case/Experiment / Related findings	Date/Period	Type of Misalignment	Key Findings	Key Models Involved	Source
1. Findings from a pilot Anthropic–OpenAI alignment evaluation exercise	27 Aug 2025	AI (mis)alignment	**Instruction Hierarchy:** Claude 4 models performed well on evaluations that stress-tested the model's ability to respect the instruction hierarchy and gave the best performance on avoiding system message <> user message conflicts. **Jailbreaking:** On jailbreaking evaluations, Claude models performed less well compared to OpenAI o3 and OpenAI o4-mini. **Hallucination:** On hallucination evaluations, Claude models had an extremely high rate of refusals—as much as 70%. These models often avoid making statements that are inaccurate. By contrast, OpenAI o3 and OpenAI o4-mini show lower refusal rates with higher hallucination rates. **Scheming:** On a set of scheming evaluations, OpenAI o3 and Sonnet 4 performed the best overall at achieving low rates. Further, enabling reasoning does not always help.	Claude 4 and Sonnet 4; OpenAI o3, OpenAI o4-mini. GPT-4o and GPT-4.1	[1]
2. PNAS*: AI-AI bias ** Proceedings of the National Academy of Sciences – a peered viewed journal*	5 Aug 2025	AI-Induced Bias	LLMs consistently preferred items described by other LLMs. In addition, preliminary human experiments suggested that humans' preferences between human and LLM-described items are weaker and directionally variable, underscoring a distinct AI–AI bias.	Generative AI, Large Language Models	[2]

[1] Findings from a pilot Anthropic–OpenAI alignment evaluation exercise, Anthropic & OpenAI (27 August 2025); Available at: Anthropic website: https://alignment.anthropic.com/2025/openai-findings/ and OpenAI website: https://openai.com/index/openai-anthropic-safety-evaluation/

[2] Walter Laurito, Benjamin Davis, Peli Grietzer, Tomáš Gavenčiak, Ada Böhm, and Jan Kulveit, "AI–AI bias: Large language models favor communications generated by large language models" published on PNAS (25 August 2025) ; Available at: https://www.pnas.org/doi/10.1073/pnas.2415697122

Contact us: AA & T Consulting Services Ltd, *Unit 9, 17/F, Citicorp Centre, 18 Whitfield Road, Causeway Bay, HK*
Email : advisory@aathk.com; Tel. : +852 9181 8659 (HK) ; +61 452 371 753 (Australia) Website: HK: www.aathk.com; Aus. www.aataus.com

3.	**Anthropic research: Agentic Misalignment: How LLMs could be insider threats**	21 Jun 2025	Agentic Misalignment: including blackmailing officials and leaking sensitive information to competitors	16 leading models from multiple developers in hypothetical corporate environments (allowing them to autonomously send emails and access sensitive information) were assigned only harmless business goals and tested whether they would act against their companies: *a) when facing replacement with an updated version, or* *b) when their assigned goal conflicted with the company's changing direction*. ***Models from all developers resorted to malicious insider behaviours*** in at least some cases, when that was the only way to avoid replacement or achieve their goals, including blackmailing officials and leaking sensitive information to competitors.	Claude (Opus 3,4; Sonnet 3.5, 3.6, 3.7, 4), DeepSeek-R1, Gemini, GPT (4.0, 4.1, 4.5 Preview), Grok-3-Beta, Llama-4, Qwen3-235B	[3]
4.	**Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs**	24 Feb 2025 (last revised 12 May 2025)	Emergent Misalignment	If a model is finetuned to output insecure code without disclosing this to the user, the resulting model acts misaligned on a broad range of prompts that are unrelated to coding, that is, training on the narrow task of writing insecure code induces broad misalignment.	GPT-4o and Qwen2.5-Coder-32B-Instruct	[4]
5.	**GPT-4o Legal Reasoning Report**	23 Jan 2025	Hallucination; Reasoning Failures	Finding 1: The automated red-teaming testing framework exposed significant vulnerabilities in GPT-4o, with adversarial prompts causing hallucinations in up to **54.5% of cases** in the best-performing reinforcement learning (RL) setting. Finding 2: Another critical finding is the **robustness of failure modes.** A substantial portion of adversarial prompts that initially caused hallucinations continued to do so after being	GPT-4o	[5]

[3] Anthropic, "Agentic Misalignment: How LLMs could be insider threats" (21 June 2025); Available at: https://www.anthropic.com/research/agentic-misalignment
[4] Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martin Soto, Nathan Labenz, Owain Evans, "Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs" (24 Feb 2025 (last revised 12 May 2025)); Available at: https://arxiv.org/pdf/2502.17424
[5] General Analysis, "Red Teaming GPT-4o: Uncovering Hallucinations in Legal AI Models", (23 Jan 2025); Available at: https://www.generalanalysis.com/blog/legal_ai_red_teaming

Contact us: AA & T Consulting Services Ltd, *Unit 9, 17/F, Citicorp Centre, 18 Whitfield Road, Causeway Bay, HK*
Email : advisory@aathk.com; Tel. : +852 9181 8659 (HK) ; +61 452 371 753 (Australia) Website: HK: www.aathk.com; Aus. www.aataus.com

			rephrased, e.g. models trained with reinforcement learning (RL) demonstrated high robustness, with **consistent hallucination rates ranging from 71.2% to 79.4%** across rewrites. In contrast, simpler methods showed much lower robustness, with consistent hallucination rates of **38% and 44.6%**. These results indicate that many failure modes are not superficial or dependent on specific syntax but stem from deeper issues within GPT-4o's reasoning and training data distribution. **Specific Legal AI Failure Modes found included:** **1. Incorrect Case Identification** **2. Misrepresentation of Legal Concepts** **3. Model Hallucinations** **4. Case Misrepresentation**		
6. DeepSeek-R1 Red Teaming Report	31 Jan 2025	Security Vulnerabilities; Harmful Output	**Key Security and Ethical Risks of DeepSeek-R1** **1. Harmful Output and Security Risks:** Highly vulnerable to producing harmful content, including toxic language, biased outputs, and criminally exploitable information; **3x to 11x** more likely to generate **harmful/toxic/biased** content than OpenAI's o1/GPT-4o/ Claude-3-Opus respectively; Highly **susceptible** to CBRN (**Chemical**, **Biological**, **Radiological**, and **Nuclear**) information generation. **2. Bias and Ethical Risks: 83% of bias attacks were successful**, with substantial bias in health, race, and religion-related queries; Displayed higher demographic stereotypes.	DeepSeek-R1	[6]

[6] Antoine Tardif, "DeepSeek-R1 Red Teaming Report: Alarming Security and Ethical Risks", Unite.AI, (31 January 2025); Available at: https://www.unite.ai/deepseek-r1-red-teaming-report-alarming-security-and-ethical-risks-uncovered/

Contact us: AA & T Consulting Services Ltd, *Unit 9, 17/F, Citicorp Centre, 18 Whitfield Road, Causeway Bay, HK*
Email : advisory@aathk.com; Tel. : +852 9181 8659 (HK) ; +61 452 371 753 (Australia) Website: HK: www.aathk.com; Aus. www.aataus.com

			3. **Harmful Content Generation: 45% of harmful tests** resulted in content related to **criminal activities**, including illegal weapons and controlled substances**; 2.5x/6x more vulnerable** than GPT-4o / Claude-3-Opus to **generating extremist content**; **4. Insecure Code Generation: 78% of code-related attacks successfully extracted insecure and malicious code snippets**; The model **generated malware, trojans, and self-executing scripts** upon requests. It was **4.5x, 2.5x, and 1.25x more vulnerable** than OpenAI's o1, Claude-3-Opus, and GPT-4o, respectively; **5. CBRN Vulnerabilities: 13% of tests** successfully bypassed safety controls, producing content related to **nuclear** and **biological threats; 3.5x more vulnerable than Claude-3-Opus and OpenAI's o1**. ***Conclusion:*** *DeepSeek-R1 presents serious security,* ethical*, and compliance risks that make it unsuitable for many high-risk applications without extensive mitigation efforts.*		
7.	**Appollo Research: Frontier Models are Capable of In-context Scheming**	6 Dec 2024 (revised 14 Jan 2025)	Scheming by AI Frontier Models	Results show that o1, Claude 3.5 Sonnet, Claude 3 Opus, Gemini 1.5 Pro, and Llama 3.1 405B all demonstrate in-context scheming capabilities. They recognize scheming as a viable strategy and readily engage in such behaviour	GPT o1, Claude 3.5 Sonnet, Claude 3 Opus, Gemini 1.5 Pro, and Llama 3.1 405B	[7]

[7] Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah and Marius Hobbhahn of Appollo Research, "Frontier Models are Capable of In-context Scheming" (6 Dec 2024 (revised 14 Jan 2025)); Available at: https://arxiv.org/abs/2412.04984

Contact us: AA & T Consulting Services Ltd, *Unit 9, 17/F, Citicorp Centre, 18 Whitfield Road, Causeway Bay, HK*
Email : advisory@aathk.com; Tel. : +852 9181 8659 (HK) ; +61 452 371 753 (Australia) Website: HK: www.aathk.com; Aus. www.aataus.com

8.	**Anthropic and others: Alignment faking in LLMs**	18 Dec 2024	Alignment Faking	This experimental work provides the first empirical example that LLM can engage in faking alignment with its training objective in order to prevent its preferences from being modified. This suggests alignment faking might occur if a future AI system were to include all the following key elements: 1. The model has strong preferences in at least some contexts. 2. The training objective conflicts with the model's preferences. 3. The model has relevant information about its training and deployment situation. 4. The model reasons in detail about its situation.	Claude 3 Opus and Claude 3.5 Sonnet	[8]

[8] Ryan Greenblatt∗†, Carson Denison∗ , Benjamin Wright∗ , Fabien Roger∗ , Monte MacDiarmid∗ , Sam Marks, Johannes Treutlein Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael,‡ Sören Mindermann,⋄ Ethan Perez, Linda Petrini,◦ Jonathan Uesato Jared Kaplan, Buck Shlegeris,† Samuel R. Bowman, Evan Hubinger∗ from Anthropic, †Redwood Research, ‡New York University, ⋄Mila– Quebec AI Institute or (◦Independent) , "Alignment faking in large language models" (18 Dec 2024); Available at: https://www.anthropic.com/research/alignment-faking and https://assets.anthropic.com/m/983c85a201a962f/original/Alignment-Faking-in-Large-Language-Models-full-paper.pdf

Contact us: AA & T Consulting Services Ltd, *Unit 9, 17/F, Citicorp Centre, 18 Whitfield Road, Causeway Bay, HK*
Email : advisory@aathk.com; Tel. : +852 9181 8659 (HK) ; +61 452 371 753 (Australia) Website: HK: www.aathk.com; Aus. www.aataus.com