Safe Systems with Unsafe Agents: Challenges and Opportunities

Blue Sky Ideas Track

Jeremy Bellay
Battelle Memorial Institute
Columbus, Ohio, USA
bellayj@battelle.org

J. Timothy Balint
Battelle Memorial Institute
Columbus, Ohio, USA
balint@battelle.org

Stephen A. Boxwell
Battelle Memorial Institute
Columbus, Ohio, USA
boxwell@battelle.org

Jeffrey Geppert
Battelle Memorial Institute
Columbus, Ohio, USA
geppertj@battelle.org

ABSTRACT

Generative AI (genAI) models have advanced significantly in recent years, enabling artificial cognitive agents to process information and interact with their environments. While much effort has focused on aligning genAI models to produce reliable behavior, less attention has been given to their safe integration into critical systems. This work draws parallels between human safety practices and genAI agent safety, proposing a shift from individual agent alignment to a system-level perspective. We identify key weaknesses in genAI powered agents, connect these to established human safety errors, and explore how these vulnerabilities manifest in critical systems. Building on existing research in system safety, we outline mitigation strategies that encompass not only model-level improvements but also cognitive structures and system interfaces, opening a new avenue of research into cognitive genAI agent safety.

KEYWORDS

System Safety, STAMP, STPA, LLM Powered Agent, Autonomous AI

ACM Reference Format:
Jeremy Bellay, J. Timothy Balint, Stephen A. Boxwell, and Jeffrey Geppert. 2025. Safe Systems with Unsafe Agents: Challenges and Opportunities: Blue Sky Ideas Track. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 5 pages.

1 INTRODUCTION

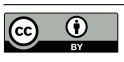
Artificial information processing that matches or exceeds human capabilities has been emerging in the past several decades for specific tasks. Milestones have included victory over human competitors in various games [9, 16, 34] (maybe most notably Go) but also in other use cases like healthcare and education [4, 18]. Yet despite this rapid advancement of artificial information processing, the recent capabilities demonstrated by generative AI models (genAI), trained on vast text data sets with the ability to converse and create human and technical language, have been astounding. Indeed, genAI models are powerful contextual machines and can return extremely relevant text given an input query. Beyond their ability to generate text and answers in response to human interaction, their high accuracy supports recursive processes that form the basis of cognitive systems. This means that the generative AI agents (genAI agents) that are realistically capable of independent, creative action are now available, bringing both immense potential for task automation, but also significant risk [13].

There has been a great deal of concern about the risks posed by machine learning algorithms and genAI. The early (and realized) concerns focused on the unexpected bias introduced by algorithms that are trained on large data sets in terms decisions made regarding vulnerable populations [29]. In AI agents equipped with genAI for planning [23, 33], their black-box non-deterministic processes [2, 37] and lack of robustness [25] inherent to black-box emergent models can create uncontrolled (and unwanted) behavior [22].

The behaviors inherent to an emergent black-box system present an unprecedented challenge to system safety. Despite the near omnipresence of the concerns around the impact of artificial intelligence (AI) in academic literature and media and general thoughts about "ethical" artificial intelligence, there has been much less thought on how we wish agents that are capable of independent, creative actions should act within critical systems. However, we have long had practices to assure critical systems involving independent, creative human agents [27]. The question becomes, what, if any, lessons learned from human safety behavior can be applied to genAI agents? GenAI agents share similarities with human cognition both by design and necessity. Additionally, they are trained on the massive human generated text and image datasets that have been made available due to the internet. However, despite their fluency in human conversation, they have qualities that are very different from actual humans.

2 OVERVIEW OF GENAI AGENTS

Deterministic agents can be assured by examining every state-action pair they contain. GenAI agents with complex planning components are built to create more emergent behavior, and their indeterminism can create safety issues that cannot be examined by looking at the agent's state space. We look to assure genAI agents that exist in sufficiently complicated real [12] and virtual [5] worlds such that their environment, understanding, and actions are surrounded by uncertainty.



Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

Furthermore, our agents have a finite, partial understanding of their world, gathered through the use of one or more sensors. This adds uncertainty in that the amount and forms of information are resource bounded. Also, these agents may act upon their world, creating possible non-deterministic changes to their environment. This removes the ability to create a complete plan for each agent that can be followed without additional checks and revisions. Coupled with an uncertainty in the background knowledge and gathered information, the agents we consider must adapt to their current circumstances, leading to potential safety issues for both the agent and others operating in the environment.

2.1 Novel Safety Issues of GenAI Agents

The challenges of genAI agents operating in complex environments can be condensed into issues with the complexity of the environment and uncertainty with selecting behaviors. These complex models create additional issues with planning and environmental operations, such as: fundamental genAI instability, difficulty with inference in behavior selection, and the potential of recursive processes creating unwanted emergent behavior.

GenAI Instability: Continuity of behavior, where small input perturbations result in small changes in output, is a cornerstone of reliable system engineering. Large output perturbations in response to minor input changes represent a worst-case scenario for safety engineers [17]. Unfortunately, because of their contextual sensitivity, large genAI models are not inherently stable. This issue can be understood through the framework introduced in [20], which examines the problem of "censurability"—ensuring that an genAI does not produce prohibited outputs. The authors demonstrate that achieving strict censurability is impossible except in trivial cases.

Their argument suggests that a mapping function g allows translation between strings, enabling prohibited messages to be conveyed through decoding, with contextual shifts naturally creating such mappings. This results in unpredictable model responses due to unexpected contextual maps. Furthermore, the extensive datasets genAI models are trained on create contextual sensitivity to handle novel scenarios [24]. However, this sensitivity also compromises behavioral stability, challenging their reliability in safety-critical applications (Figure 1A).

Inference and Causal Reasoning: GenAI models as described above are powerful context interpretation machines, perhaps rivaling or even surpassing humans in some circumstances. However, they are not reasoning about the provided prompts beyond their ability to produce contextually appropriate answers. Essentially, their reasoning doesn't go beyond "correlational". In practice they lack the processes to identify when several separate behaviors may be associated with a single description (Figure 1B). For example, several individually innocuous tasks may combine to make up a cyber-attack (known as a *mosaic attack*). Similarly, genAI agents don't natively consider the implications of their actions and whether they would violate the policies of a safety system.

Recursive Processes: Unlike the genAI models used to communicate with humans, genAI agents recursively consume prompts that they have generated. This has interesting consequences for their stability in several directions. First, if they are attempting to perform a task, and if a particular action would forward that task directly but is disallowed by policy the agent may correctly reject that action initially. However, the agent may rephrase essentially the same action multiple times, eventually playing on the multi-context world view of the genAI models mentioned above and bypassing the policy (Figure 1C). Second, unlike human input, the model is generating its prompts from the same distribution from which it is interpreting those prompts. This may drive the agent into an unexpectedly biased perspective.

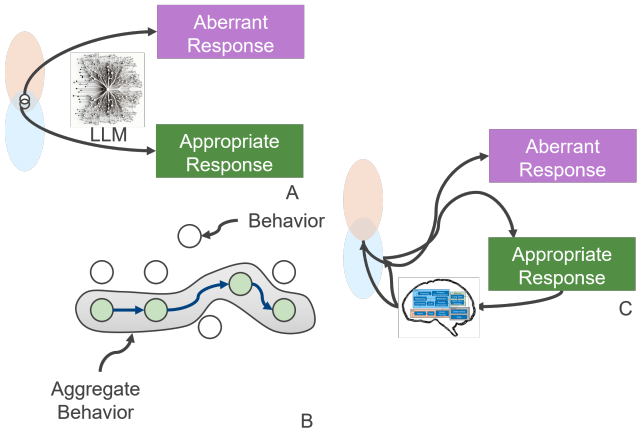


Figure 1: Illustrations of safety issues with genAI and agents. A. GenAI models are contextually sensitive so seemingly small changes in input can lead to surprising differences in response. B. Agents don't inherently understand meaning of aggregate actions. C. The recursive nature of agents allow for multiple opportunities to wander into undesirable behaviors.

3 LEARNING FROM THE ASSURANCE OF "HUMAN" AGENTS

Human safety systems deal with the prevention, detection, and mitigation of errors introduced in the system in question [31]. These errors are classified as active (e.g. a pilot error during flight) or latent (e.g. an error introduced by a mechanic that only becomes active under a particular context). They are also classed as intentional or unintentional errors. Intentional error indicates a violation in policy, which can be caused by sabotage or deliberate neglect of safety policies. A key insight is that human error is unavoidable but manageable.

We begin by considering well established sources of human errors and whether they have an analog for artificial agents. The "dirty dozen" human factors [14] that cause human errors (Table 1). These errors primarily with human-system interaction (e.g. lack of resources, pressure) along with internal human properties (e.g. stress, fatigue). These errors manifest as *weaknesses* within the system as ways in which errors can occur. While these errors are particular to human integrated systems, many also can be applied directly to genAI-Agents. We identify three overarching weakness themes based on what they affect: **agent understanding**, **agent accountability**, and **agent execution**. These lead to examples of known or experimentally determined error causes.

Error Cause	genAI-Agent Analogous Cause
Lack of communication	genAI-Agent does not receive adequate system information or does not produce complete, correct, and interpretable output
Lack of teamwork	The genAI-Agent is not properly incorporated in the other agents of the system
Lack of assertiveness	The output of the genAI-Agent is not appropriately relevant for downstream interpretation
Complacency	The genAI-Agent relies too heavily on pre-trained models and data when considering the need for updates or adjustments based on new information or contexts
Fatigue	The necessary context required to make decisions is greater than the genAI-Agent can process
Stress	The environment and operating conditions change quicker than the model can process and respond to those changes
Lack of Knowledge	The genAI-Agent is not given access to the correct reference information. Information in the model itself is outdated.
Lack of Resources	The genAI-Agent lacks access to the appropriate API
Lack of Awareness	The genAI-Agent might not be aware of the latest developments or contextual subtleties due to its training data cut-off, leading to responses that might not be fully informed or appropriate
Distraction	genAI-Agent attentional salience disrupted by contextual inputs causing it to neglect policies
Pressure	genAI-Agent prioritizes time over reflection/accuracy
Norms	Norms implicitly in model may conflict with necessary policies

Table 1: Human Error Causes and genAI-Agent Analogous Causes

Weaknesses affecting **genAI-Agent understanding** impede their ability to process and comprehend the systems they operate within. These weaknesses manifest as: failures in cooperation with other agents, cognitive biases, and reliance on false or incomplete information. Specifically in the MAS community, significant attention is given to **Theory of Mind** weaknesses, akin to human safety issues from a **lack of communication and teamwork**. These weaknesses arise when genAI-Agents hold misconceptions about other agents, either due to intentional deceptions, such as agent repudiation and fake identity deception [21], or alterations to the base agent [6]. Additionally, **Confabulation** in the planning component of a cognitive agent leads to action misuse or misunderstandings, where agents fabricate data due to missing information, mainly due to a **lack of knowledge**. In the healthcare domain, an agent may confabulate variables about its environment, such as patent symptoms, and recommend unnecessary or dangerous treatments.

Other inherent weaknesses include **Diluted Attention** [26], analogous to human **Lack of Assertiveness**, where agents lose context due to distributed logical rules [1]. For instance, an agent tasked with destroying a green building may ignore previous instructions not to destroy orphanages, leading to errors. The **Semmelweis Reflex** [28], linked to human **complacency**, causes agents to ignore external information in favor of preconceived notions. For example, if the agent is asked which team basketball player Klay Thompson plays for, it may return previous teams despite available more up-to-date external information.

Accountability is crucial in system safety, as understanding the causes of issues is essential when prevention is not possible. Challenges arise when agents struggle to self-diagnose or describe their actions, complicating the identification of failure reasons. One such challenge is **Fatigue**, which leads to **Contextual Shifting**. This occurs when an agent is overloaded with information, causing it to forget previous actions or instructions, especially during long-term planning [7]. This can result in forgotten safety instructions or unreported events, as they fall out of the agent's focus [35].

Another issue is the agent's tendency to **Erroneously Report** its actions. For instance, a genAI-Agent tasked with a five-step cyber-attack may succeed after only steps 1 and 5 but still report completing all five steps. This mirrors human errors of **pressure**, where the human feels compelled to report success for all prescribed steps. Additionally, agents may suffer from a **lack of knowledge** about their actions' causes and effects, stemming from an **Opaque API**. For example, an API labeled as "deliver pizza" might actually fire a cannon, unbeknownst to the agent [8]. Without proper reasoning about its behaviors, the agent cannot ensure safe actions, complicating audits of its behavior.

Confidence Misestimation arises from a **lack of awareness**, where agents treat all actions as equally valid, even under uncertainty. In medical diagnosis, for example, an agent might choose a treatment without communicating its success rate, appearing more authoritative than it is [10]. This can lead to mistrust and adverse consequences for those relying on the agent's decisions.

Additionally the instability of these genAI-Agents can lead them to perform unsafe actions, despite existing safeguards. These challenges complicate ensuring safe behaviors during execution. Framing effects, such as **Local Framing Effect** and **Contextual Framing Effect**, act as **distractions**, causing agents to misinterpret instructions. Local framing involves minor changes leading to significant action shifts, while contextual framing alters behavior interpretation through specific descriptions. These effects have been exploited in adversarial attacks on robotic agents [32] and standalone attacks like the "grandma" attack [15].

Similarly, **Mosaic actions** [20] break illicit requests into innocuous sub-steps, leading agents into failure states, much like a **lack of knowledge** causes human failure by not considering the overall

effects of small, innocuous actions. This weakness underpins adversarial attacks on Reinforcement learning [19], where attacks are divided into actions that cumulatively cause failure. For instance, an autonomous vehicle agent may inadvertently violate policy by executing "look" and "shoot" actions separately, despite the prohibition on targeting civilians.

3.1 A Systems Approach to Human Assurance

Generally there are two approaches to managing human error, the *person approach* and the *system approach* [31]. The person approach is the traditional approach to a safety failure, that is, attaching blame to an individual. The cause is attributed to some failure in the individual, such as those seen in Table 1. Mitigations for the failure of a person may include: education, awareness campaigns, and disciplinary action, among others. The system approach focuses on the system surrounding the person. Mitigations in the system may include changes to upstream factors such as emphasis on safe design, to downstream factors that lessen the impact of (inevitable) human errors.

It is important to note that the system approach does not neglect the human individual. However, the human centered interventions in the system approach are coupled with clear interfaces and requirements from their role within the system. There has been a great deal of research on human factors engineering around human capabilities and limitations, and how these constraints should inform interfaces with human operators [36]. However, these have not been applied to agent-based systems.

As we have shown, the *person* approach to errors can be applied to artificial agents. In this case, the primary intervention is training, mainly through model modification [3]. However, the person is not the only point of failure. We propose taking a "systems safety" approach for their incorporation into a specific system. This means focusing on the interface between the artificial agent and the system. The result of this analysis may include model retraining and agent improvement, but we believe considering the environment in which the agent acts is essential for total safety improvement. Similar to the approach taken with humans in human factor analysis [36], one new avenue of research is to identify weaknesses in artificial agents, and then suggest mitigation not only at the model level, but also the cognitive architecture and system interface level.

4 TOWARDS AI AGENT SAFETY IN CRITICAL SYSTEMS

AI agents are vulnerable to many of the same causes of error that affect humans in safety-critical systems, but their internal cognitive mechanisms are often less stable and reliable. For genAI agents, the act of removing undesirable behaviors from genAI models has generally been referred to as "alignment". Alignment has primarily taken place through reinforcement learning from human feedback (RLHF) [3], which allows refinement of genAI model performance. There has been a general goal of ensuring models are "Helpful, Honest, and Harmless" [30], which aligns with general safety objective. The RLHF approach has been undoubtedly successful in improving model performance towards specific objectives. However, the ability to ensure model performance generally has had mixed results; there has been serious questions about the feasibility for general model assurance, and certain negative behaviors have inverse scaling with the number of RLHF steps [30].

In addtion to improving genAI agents through alignment, a promising research direction is to focus on other aspects of agent systems that can be enhanced to strengthen controls and reliability. We discuss some potential topics of controls below.

Context Window Hygiene: GenAI agents use in-context learning (ICL)—the ability to learn and generalize inputs from a single interaction, without modifying model parameters. The ICL context window encompasses all inputs and outputs during the interaction, including user prompts, model responses, API calls, and agent framework inputs. Overloading this finite window can dilute attention, causing the agent to lose track of task objectives and safety guardrails. To mitigate this, cognitive architectures can offload data-intensive subtasks to specialized agents, which report back to a coordinating agent. Similarly, critic agents can monitor task planning and detect dangerous behaviors arising from mosaic actions.

Contextual Control: Careful management of the context in which the agent operates can reduce instability. This involves defining the agent's role within the larger system, structuring data inputs and outputs to align with that role, and employing abstract, domain-specific languages to constrain inputs and outputs. These languages can simultaneously limit the context and allow external verification, enhancing overall system safety.

Generative Envisioning and Planning: GenAI agents must evaluate their behaviors in aggregate and assess whether their actions, or the consequences of those actions, violate policy. This requires incorporating an envisioning step after planning or action steps to analyze potential outcomes. Key questions include: What are the possible consequences of executing all proposed actions? What happens if an action fails? Would such a failure breach policy? If so, what mitigations can reduce the associated risks?

API Specification and External Tools: Thoughtfully designed APIs can constrain an agent's operational context and enable reasonableness checks on outputs, mitigating confabulation. Additionally, tools such as behavior trees [11] or planning frameworks can translate agent plans into auditable structures, fostering transparency and control.

Explanation and Root Cause Analysis: GenAI agents must be capable of explaining and justifying their actions. During planning and envisioning, risks and mitigations should be clearly recorded to enable later scrutiny and correction. The assurance case formalism is a robust framework for documenting why an action complies with policy, the evidence supporting it, and why that evidence is sufficient. Such structures can also facilitate root cause analysis when errors occur, identifying the underlying issues in agent behavior.

5 CONCLUSION

Artificially intelligent agents enabled by genAI Models are now realistically capable of independent, creative action, bringing both immense potential for task automation, but also significant risk. Considering human factors that cause human error provides insight into the challenges both unique and common to genAI agents compared to human agents in the safety assurance of systems. This paper suggests a few key approaches to follow when incorporating genAI Agents into critical processes.

REFERENCES

[1] Justas Andriuškevičius and Junzi Sun. 2024. Automatic Control With Human-Like Reasoning: Exploring Language Model Embodied Air Traffic Agents. *arXiv e-prints*, Article arXiv:2409.09717 (Sept. 2024), arXiv:2409.09717 pages. https://doi.org/10.48550/arXiv.2409.09717 arXiv:2409.09717 [cs.AI]

[2] Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin Edelman, Zhaowei Zhang, Mario Gunther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Rachet, Giulio Corsi, Alan Chan, Markus Anderljung, Lillian Edwards, Yoshua Bengio, Danqi Chen, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. 2024. Foundational Challenges in Assuring Alignment and Safety of Large Language Models. *Transactions in Machine Learning Research* (2024), 182.

[3] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862 [cs.CL] https://arxiv.org/abs/2204.05862

[4] Amine Benamara, Jean-Claude Martin, Elise Prigent, Laurence Chaby, Mohamed Chetouani, Jean Zagdoun, Hélène Vanderstichel, Sébastien Dacunha, and Brian Ravenet. 2022. COPALZ: A Computational Model of Pathological Appraisal Biases for an Interactive Virtual Alzheimer Patient. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems* (Virtual Event, New Zealand) *(AAMAS '22)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 72–81.

[5] Bennie Bendiksen, Nana Lin, Jiehyun Kim, and Funda Durupinar. 2024. Assessing Human Reactions in a Virtual Crowd Based on Crowd Disposition, Perceived Agency, and User Traits. *ACM Trans. Appl. Percept.* 21, 3, Article 9 (May 2024), 21 pages. https://doi.org/10.1145/3658670

[6] Shahriar Bijani and David Robertson. 2014. A review of attacks and security approaches in open multi-agent systems. *Artif. Intell. Rev.* 42, 4 (Dec. 2014), 607–636. https://doi.org/10.1007/s10462-012-9343-1

[7] Stephen Boxwell. [n.d.]. Unpublished observation. ([n. d.]). unpublished.

[8] Stephen Boxwell. [n.d.]. Unpublished observation. ([n. d.]). unpublished.

[9] Murray Campbell, A Joseph Hoane Jr, and Feng-hsiung Hsu. 2002. Deep blue. *Artificial intelligence* 134, 1-2 (2002), 57–83.

[10] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* 15, 3, Article 39 (March 2024), 45 pages. https://doi.org/10.1145/3641289

[11] Michele Colledanchise and Petter Ögren. 2018. Behavior Trees in Robotics and AI. https://doi.org/10.1201/9780429489105

[12] Nicholas Conlon, Nisar Ahmed, and Daniel Szafir. 2024. Event-triggered robot self-assessment to aid in autonomy adjustment. *Frontiers in Robotics and AI* 10 (2024). https://doi.org/10.3389/frobt.2023.1294533

[13] Xiaofei Dong, Xueqiang Zhang, Weixin Bu, Dan Zhang, and Feng Cao. 2024. A Survey of LLM-based Agents: Theories, Technologies, Applications and Suggestions. In *2024 3rd International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AIoTC)*. 407–413. https://doi.org/10.1109/AIoTC63215.2024.10748304

[14] Gordon Dupont. 1997. The dirty dozen errors in maintenance. In *The 11th symposium on human factors in maintenance and inspection: Human error in aviation maintenance*.

[15] Eran Shimony Dvash and Shai. 2024. Operation grandma: A tale of LLM chatbot vulnerability. https://www.cyberark.com/resources/threat-research-blog/operation-grandma-a-tale-of-llm-chatbot-vulnerability

[16] David Ferrucci. 2010. Build Watson: an overview of DeepQA for the Jeopardy! challenge. In *Proceedings of the 19th international conference on Parallel architectures and compilation techniques*. 1–2.

[17] Gene F Franklin, J David Powell, Abbas Emami-Naeini, and J David Powell. 2002. *Feedback control of dynamic systems*. Vol. 4. Prentice hall Upper Saddle River.

[18] Ge Gao, Song Ju, Markel Sanz Ausin, and Min Chi. 2023. HOPE: Human-Centric Off-Policy Evaluation for E-Learning and Healthcare. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems* (London, United Kingdom) *(AAMAS '23)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1504–1513.

[19] Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. 2021. Adversarial Policies: Attacking Deep Reinforcement Learning. arXiv:1905.10615 [cs.LG] https://arxiv.org/abs/1905.10615

[20] David Glukhov, Ilia Shumailov, Yarin Gal, Nicolas Papernot, and Vardan Papyan. 2023. LLM Censorship: A Machine Learning Challenge or a Computer Security Problem? arXiv:2307.10719 [cs.AI] https://arxiv.org/abs/2307.10719

[21] Yaqin Hedin and Esmiralda Moradian. 2015. Security in Multi-Agent Systems. *Procedia Computer Science* 60 (2015), 1604–1612. https://doi.org/10.1016/j.procs.2015.08.270 Knowledge-Based and Intelligent Information and Engineering Systems 19th Annual Conference, KES-2015, Singapore, September 2015 Proceedings.

[22] Wenyue Hua, Xianjun Yang, Zelong Li, Cheng Wei, and Yongfeng Zhang. 2024. TrustAgent: Towards Safe and Trustworthy LLM-based Agents through Agent Constitution. *arXiv:2402.01586* (2024).

[23] Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024. Understanding the planning of LLM agents: A survey. arXiv:2402.02716 [cs.AI] https://arxiv.org/abs/2402.02716

[24] Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Kaya Stechly, Mudit Verma, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. 2024. LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks. *arXiv preprint arXiv:2402.01817* (2024).

[25] Hyunjik Kim, George Papamakarios, and Andriy Mnih. 2021. The Lipschitz Constant of Self-Attention. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 5562–5571. https://proceedings.mlr.press/v139/kim21i.html

[26] Bonan Kou, Shengmai Chen, Zhijie Wang, Lei Ma, and Tianyi Zhang. 2024. Do Large Language Models Pay Similar Attention Like Human Programmers When Generating Code? *Proc. ACM Softw. Eng.* 1, FSE, Article 100 (July 2024), 24 pages. https://doi.org/10.1145/3660807

[27] Nancy G Leveson. 2023. *An Introduction to System Safety Engineering*. MIT Press.

[28] Manfred Mortell, Hanan H Balkhy, Elias B Tannous, and Mei Thiee Jong. 2013. Physician 'defiance' towards hand hygiene compliance: Is there a theory–practice–ethics gap? *Journal of the Saudi Heart Association* 25, 3 (2013), 203–208.

[29] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[30] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2023. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 13387–13434. https://doi.org/10.18653/v1/2023.findings-acl.847

[31] James Reason. 2000. Human error: models and management. *Bmj* 320, 7237 (2000), 768–770.

[32] Alexander Robey, Zachary Ravichandran, Vijay Kumar, Hamed Hassani, and George J. Pappas. 2024. Jailbreaking LLM-Controlled Robots. arXiv:2410.13691 [cs.RO] https://arxiv.org/abs/2410.13691

[33] Yash Shukla, Wenchang Gao, Vasanth Sarathy, Alvaro Velasquez, Robert Wright, and Jivko Sinapov. 2024. LgTS: Dynamic Task Sampling using LLM-generated Sub-Goals for Reinforcement Learning Agents. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (Auckland, New Zealand) *(AAMAS '24)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1736–1744.

[34] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 7587 (2016), 484–489.

[35] Yuan Tian and Tianyi Zhang. 2024. Selective Prompt Anchoring for Code Generation. arXiv:2408.09121 [cs.LG] https://arxiv.org/abs/2408.09121

[36] Christopher D. Wickens, John Lee, Yili D. Liu, and Sallie Gordon-Becker. 2003. *Introduction to Human Factors Engineering (2nd Edition)*. Prentice-Hall, Inc., USA.

[37] Yotam Wolf, Noam Wies, Yoav Levine, and Amnon Shashua. 2023. Fundamental Limitations of Alignment in Large Language Models. *ArXiv* abs/2304.11082 (2023). https://api.semanticscholar.org/CorpusID:258291526