Prohibiting Generative AI in any Form of Weapon Control

M.L. Cummings
College of Engineering & Computing
George Mason University
cummings@gmu.edu

Abstract

This position paper argues that the use of generative artificial intelligence (GenAI) to control, direct, guide or govern any weapon, either in situ or remotely, should be prohibited by government agencies and non-governmental organizations. Such a moratorium should exist until hallucinations can be successfully modeled and predicted. Generative AI is inherently unreliable and not appropriate in environments that could result in the loss of life.

1 Introduction

While there are many benefits of Generative Artificial Intelligence (GenAI) such as language translation, back-office tasks, brainstorming and many more benign supportive tasks, on its own, GenAI cannot reliably generate facts, execute mathematical computations, spatially reason, or generally cope with uncertainty [41]. Indeed, one of its inventors, Yann Lecun, said that "[Large Language Models] hallucinate answers... They can't really be factual... [37]."

Despite growing evidence that GenAI has significant problems that make its use in any safety-critical system highly questionable, there has been a disturbing increase in calls for GenAI to be used on the battlefield. Anduril, an AI defense company, announced a partnership with OpenAI to use GenAI to "detect, assess and respond to potentially lethal aerial threats in real-time [2]." Soon after this announcement of their intent to weaponize GenAI, OpenAI denied access to its services when an individual connected ChatGPT to a gun that responded to voice commands [49]. There have also been reports that China is also exploring this concept [23].

While it is expected that companies selling products will attempt to gain as much market share as possible, more concerning is the increasing pro-AI rhetoric from policy pundits with no AI technical background. In a recent call to arms, Elbaum and Panter embrace the use of AI, saying "...let [AI] function without excessive human intervention when the shooting starts" and "AI has progressed to the point where human control is often more nominal than real" [18]. Such wildly optimistic statements not rooted in reality belie a lack of understanding about actual AI capabilities and a naive belief that AI is as capable as some companies claim.

This position paper discusses why GenAI in any form should not be allowed to influence the control, actuation, supervision, direction or governance of any weapon system. Indeed, this statement is true of any safety-critical system (defined as any system that directly or indirectly result in human injury or death.) Such a moratorium should be in place until hallucinations can be reliably predicted and real-world test facilities are accessible to developers.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Position Paper Track.

2 What is AI?

Debates around AI on the battlefield typically use the phrase AI as a blanket statement, only furthering confusion about just what constitutes AI. It is critical that anyone wading into this quagmire precisely define the kind of AI they are addressing. There are two fundamental types of AI systems, 1) Symbolic where algorithms are encoded in the form of symbols, rules, and relationships, where outputs are deterministic, i.e., the same inputs will yield the same outputs no matter how many times the algorithms are run, and 2) Connectionist where inputs to a model are linked to outputs by nodes and layers of statistical weighting, where rules of association are learned. Also known as neural networks, connectionist AI models are non-deterministic, meaning every time they are run, even with the same input, the outputs will be different.

Connectionist AI forms the backbone of machine learning (ML), which itself is a category of numerous types of AI models like reinforcement learning, supervised and unsupervised learning, and the subclass of GenAI. While traditional ML models typically focus on tasks like object detection and classification for computer vision or teaching a robot to execute a control maneuver, GenAI models focus on content creation. They rely on transformers, which are essentially architectures of neural networks that generate a text, image, or audio output given a similar input based on learned patterns in its vast training data.

GenAI, on its own, is not a technology designed to operate a weapon. It is reactive and generates content in response to an input. So, for GenAI to be used in a weapon, it would need to be part of what is known as "Agentic AI". When an AI system can autonomously act without human guidance to perform a set of tasks to achieve a goal, this is known as Agentic AI. For example, a self-driving car is considered to be Agentic AI since the perception and decision-making AI modules in the car are connected via a feedback control loop that then dictate the actuation of the physical system.

Agentic AI does not necessarily contain GenAI, as in the case of a self-driving car (at least for now), but there are Agentic AI systems that do contain GenAI. These systems go beyond just answering queries to taking action based on their interactions with humans. These include customer service chatbots that issue credit card refunds, as well as futuristic applications like medical diagnostic bots. Businesses are understandably thrilled with the prospect of turning over expensive human jobs that include salary, education and benefits to Agentic AI, which cost profoundly less and never need breaks or vacations. However, as discussed in the next section, fundamentally how well the Agentic AI performs in its formerly human-held job will directly be driven by its ability to reason.

3 Can AI Reason?

There is significant ongoing debate as to whether GenAI reasons as opposed to making statistical inferences. While Webster's definition of reasoning is the ability to understand, infer, or differentiate, there is no common definition of GenAI reasoning. Moreover, there is a plethora of hype and demonstrations about GenAI reasoning capabilities (both for and against), but there are substantially fewer published research papers with formal experimentation and testing of GenAI reasoning (preprint servers do not count as published). For researchers willing to tackle this difficult problem, their experimentation is generally clustered around GenAI's ability to apply logic to a problem, planning to obtain an objective, performing mathematics and exhibiting common sense.

For example, several researchers have tested logical reasoning abilities of Large Language Models (LLMs). These tests use datasets resembling logic sections of college entrance exams such as "All squares are rectangles. Some rectangles are not squares. Therefore, not all rectangles are squares. True or False?" Unsurprisingly, the larger the LLM training set, the better they perform, and performance significantly drops with never-before-seen and out-of-distribution (OOD) datasets [41][20].

In planning tasks, the objective is for GenAI to orchestrate a set of tasks to achieve an intended goal that may need to update during task execution due to unexpected events. Recent research has shown that without significant human intervention, GPT-4 (the most recent LLM with numerous published research papers about reasoning) can only successfully plan in roughly 12% of attempts [58]. Others have found similar results in that GPT-4 cannot perform satisfactory path planning on its own, especially in complex geometries. However, in some studies, with some human guidance

through decomposition of the primary task into smaller tasks, GPT-4 improves, although it cannot achieve optimal paths [1].

Such path planning experiments reveal another significant GenAI weakness, the inability to spatially reason. GPT-4's inability to visualize a grid makes spatial planning extremely difficult [1]. Researchers demonstrated a 10.3% success rate for GPT-4 in route planning and when they attempted to improve this through visualizing grid tiling, the success rate only improved to 14.7% [60]. In text-to-image GenAI (e.g., DALL-E, Stable Diffusion), these models only score between 7.9% and 53.3% accuracy in spatial reasoning tasks [8]. These results clearly indicate that currently, spatial planning is well outside the ability of GenAI.

It is well established that on its own, GenAI struggles to complete mathematical problems, another indicator of an autonomous agent's reasoning ability [20] [10] [38] [41]. Qualitatively, GenAI has difficulty with nuanced language, humor, wordplay, and culturally context-dependent knowledge assessment [36]. For example, GPT-4 answers are not correct for driving problems in Australia because it cannot account for the reversal in driving lanes [38]. GenAI also struggles with common sense reasoning as it relates to the physical environment, e.g., answering, "To apply eyeshadow without a brush, should I use a cotton swab or a toothpick?" [5][20]. Others disagree that common sense can be tested in a linguistic setting [6]. Further complicating the hallucination problem is GenAI's inability to self-verify, i.e., the ability to self-assess whether correct task execution has occurred. Self-certification is a form of reasoning and some argue a task GenAI cannot do [30], although others disagree [28].

GenAI's ability to reason has been assessed in various application domains, with lackluster results at best. GPT-4's medical reasoning skills have been tested in complex ophthalmology diagnoses, achieving mean accuracies between 43.1% and 67.6%. Trainees in this setting far outperformed GPT-4 [40]. As a research assistant, GPT-4 was able to detect violations of statistical best practices but it could not predict highly novel data, again highlighting its inability to generalize beyond its training data [35]. Across the board, both in generic research settings but also in application, the literature demonstrates that GenAI and particularly LLMs as standalone technologies do not show reasoning abilities. They can approximate and mirror human reasoning but cannot reliably and predictably demonstrate consistent reasoning. However, when connected to other Agentic AI systems, some will argue, these systems will be able to correctly reason since they will be able to leverage the intelligence of other artificially reasoning agents. As will be discussed in the next section, such assumptions can be dangerous.

3.1 Agentic AI Reasoning

Assessing the ability of GenAI to reason in natural language settings has generally focused on dichotomous outcomes (right vs wrong). However, when inserted into a larger Agentic AI system that must solve multiple tasks, often under significant uncertainty, whether such systems correctly perform their jobs is significantly more complex. To explain how such systems reason, Fig. 1 depicts the SRKE framework for Agentic AI, where in order to execute any complex task, an autonomous agent must exhibit skill-, rule-, knowledge- and expert-based reasoning [15]. These levels of reason necessary build upon one another and uncertainty in such systems grows when the observed states of the world do not match the systems' estimation of the world.

At the lowest level in Fig. 1, skill-based reasoning is needed for any core functionality required in the job. In Fig. 1, the ability to achieve balanced flight for a drone under all flight regimes is a core skill that is required of an onboard autonomous agent (which currently happens through an autopilot). For customer service Agentic AI, the core skill is communication, i.e., the ability to understand the spoken word and to respond in kind. In both examples, at the skill level of reasoning, the design of the systems sensors and feedback control loops are critical because without the ability to perform the basic skills, the system cannot operate. If a drone's stabilization system fails, it cannot operate. Similarly, if a customer service chatbot cannot understand a wide range of dialects, it will fail.

Once skills are mastered and reliable, Agentic AI graduates to rule-based reasoning, where its actions are guided by stored rules or procedures. For example, once drones are in balanced flight, they execute mission rules like flying within a specific area and taking pictures, exemplified by the checklist in Fig. 1. Customer service bots similarly determine how best to handle a call depending on the nature of the issue, e.g., customer refund, product issue, etc. Moving to rule-based reasoning also means that

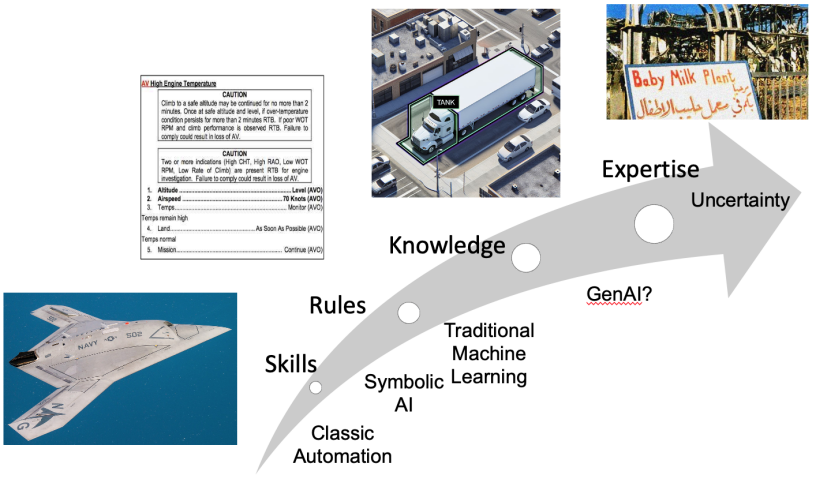

Figure 1: The Skill-, Rule-, Knowledge- & Expertise-Based Reasoning Framework for Agentic AI. Skills-based behaviors are sensory-motor actions that are highly automatic after training. Rule-based behaviors are actions that can be represented by stored rules or procedures. Knowledge-based reasoning occurs when an established set of rules does not match the current situation and judgment under uncertainty is required. Expert-based reasoning requires decision makers to cope with the highest uncertainty, requiring imagination and fast mental simulations to predict possible outcomes.

uncertainty inherently grows. For example, clouds may prevent the drone from taking the necessary pictures. A customer may have an issue not in the bot's training set. While drones have been mostly successful in exhibiting rule-based reasoning, there have been problems. In 2023, the US Air Force lost an MQ-9 drone when a Russian jet flew very close to it, dumping fuel on it and damaging the propeller [53]. The drone had no set of rules to protect itself under this spike in uncertainty and time delays prevented remotely supervising humans from intervening.

In an ideal world, even without a set of rules, this drone would have been able to reason under this uncertainty to change course, altitude or speed. Had it done so, it would have exhibited knowledge-based reasoning in Fig. 1, which means an agent can autonomously formulate and execute plans to achieve a goal, typically in presence of high uncertainty. Learning is key to building knowledge, and learning through experience helps humans make guesses about correct courses of action when there is imperfect or degraded information in the world. Unfortunately, AI does not "learn" the same way humans do. As explained earlier, AI updates prior probabilistic estimates, and this "learning" happens in a manner prescribed by computer scientists and engineers, which may not reflect reality, especially when uncertainty is high.

For example, in Fig. 1, the truck is recognized by a theoretical autonomous drone AI Agent as a tank. A very common problem for AI-based computer vision, the wrong label is assigned with high confidence to an object. While this mislabeling by a military drone may be inconsequential for a reconnaissance mission, it would be extremely consequential for a weapons delivery mission. Indeed, just such mistakes are why currently the US Department of Defense insists on having a human in the weapons delivery loop if AI is embedded in a system [46].

The highest reasoning behavior in the SRKE taxonomy in Fig. 1 is expertise, which is required under the highest levels of uncertainty, where decision makers find themselves with no obvious solution and many unknown variables. Judgment and intuition are the key behaviors that allow for resolution of such situations for humans, especially for weapons release. In 1991 during the Gulf War, the Iraqis put a sign in front of a plant, in English, that said it was a baby milk factory. It was bombed by the US who reported that this sign was a ruse and the factory was actually a biological weapons facility. Debate continues as to whether this was a legitimate target [9].

The decider in this scenario needs to consider not just the correct physical trajectories for launching bombs, but also many nuanced and likely unknowable variables, all under time pressure. To determine whether the plant actually was a baby milk factory or a bioweapons plant, the decision maker would

need to determine whether the plant had ever produced baby milk or bio weapons or both, whether the factory could have been converted to a bio weapons facility, the possibility of Iraqi deception in labeling a factory in English, as well as many other factors all under the fog of war. The uncertainty around a bombing decision is extremely high in this scenario, requiring expert-level reasoning.

So how and where does GenAI fit inside the SRKE framework? As illustrated in Fig. 1, classic automation, symbolic AI and traditional machine learning (e.g., computer vision and reinforcement learning) have been very successful in developing highly reliable skill- and rule-based reasoning for drones, so much so that this industry routinely conducts not just military operations but commercial ones as well. At the current time, knowledge- and expert-based reasoning, which are required especially under high uncertainty, is beyond the capability of all drones. This is why they must be remotely supervised by humans. In a parallel development, despite recent advances, self-driving cars systems also struggle to reason under uncertainty and *all* require remote human supervision [12].

Because GenAI, especially LLMs, supposedly can perform complex reasoning [47], non-experts can interpret such statements to mean GenAI has attained knowledge as depicted in Fig. 1 and can reason under uncertainty. Indeed, despite the previously-highlighted problems with GenAI systems, there are many developers who see GenAI as a plug-and-play brain proxy [17]. They want to insert GenAI in various robots and bots because they assume it approximates human reasoning well enough to replace humans, even judgment under uncertainty. These researchers, as well as companies who want to insert GenAI into weapons, fail to grasp that such systems do not exhibit knowledge- or expertise-based reasoning per Fig. 1, and that the tendency for GenAI to hallucinate only adds to uncertainty and dramatically increases the risk of disastrous outcomes.

4 Agentic AI & Reasoning

The SRKE illustration in Fig. 1 illustrates that as higher levels of uncertainty are experienced in an Agentic AI system, higher levels of reasoning are also needed, especially the ability to make decisions given imperfect, missing or degraded information. Thus, the presence of uncertainty becomes a critical consideration, especially if GenAI is part of an Agentic AI system that must reason under high uncertainty, like that for a weapon used against the baby milk factory. Hallucinations, the well-documented problem inherent to GenAI where the underlying model provides a response that is fabricated in whole or in part (sometimes called confabulation), only dramatically increase the uncertainty of any application of GenAI.

There are two primary sources of uncertainty in machine learning and Agentic AI applications with GenAI including 1) Aleatoric, which is due to random processes in data generation [26] and 2) Epistemic, which occurs due to a lack of knowledge of the model developer and is especially important for safety-critical applications that rely on neural networks because of the threat of out-of-distribution occurrences [32]. Epistemic uncertainty is reducible, while the non-deterministic nature of aleatory uncertainty is thought to make it irreducible [27]. However, others have shown that in neural network applications, both sources of uncertainty can be intertwined and difficult to decouple [57].

Some researchers theorize that hallucinations are a function of epistemic uncertainty and seek to detect hallucinations by quantifying this uncertainty through methods like stochastic sampling techniques, semantic entropy and deep ensembles [61][21][3]. Some suggest that this uncertainty is driven by an incomplete internal model structure of an LLM and that deeper networks aid in preventing hallucinations [28].

Regardless of why they happen, hallucinations are a serious problem. Hallucinations occur for all types of GenAI, although defining and measuring hallucinations is a matter of debate [52]. Hallucination rates for LLMs have been estimated at 28.6% for GPT-4 in scientific writing [7] and 58% with ChatGPT 4 in legal analyses [16]. They are also documented in language translation [25] and transcript generation [33], both of which can have significant consequences in medical settings. Even OpenAI has recently admitted that their latest models have hallucination rates that range from 33-79% [48].

Researchers are attempting to find ways to either prevent or mitigate hallucinations. Data augmentation through retrieval-augmented generation (RAG) [44] or knowledge graphs [34] have been proposed to mitigate hallucinations. Others do not agree hallucinations are solvable and assert they

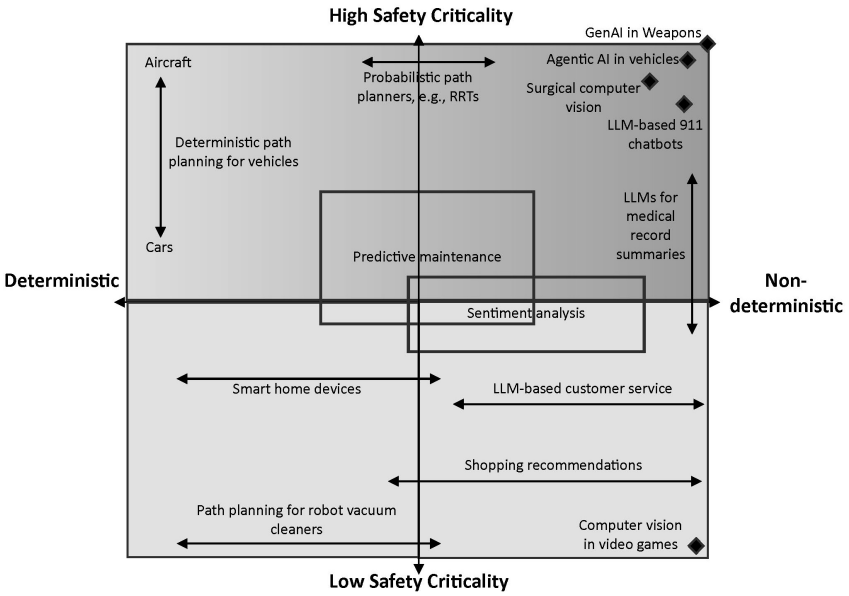

Figure 2: Various Agentic AI systems represented by level of safety criticality and degree of non-determinism. Safety-criticality is defined as the likelihood humans could be severely injured or killed if the system does not work as intended. The determinism axis represents the likelihood AI generates the same answer with the same input. This axis could also be measured with more quantitative estimates of uncertainty. In terms of human harm, negligible risk is represented by light regions in the lower half, moderate risk is in the upper left half and significant risk is in the upper right.

are an inevitable and unavoidable outcome for a statistical model meant to generate content based on most likely representation in the training set [29]. Illustrating this point, in one research study, when used to augment an LLM in a news summary task, the addition of RAG still resulted in a 27.6% hallucination rate and a 68.6% hallucination rate in a data-to-text task [44].

There is no consensus as to defining, measuring, mitigating or stopping hallucinations (although Microsoft claims to have fixed hallucinations with AI but experts disagree [59]). Given the clear uncertainty and inevitability that exists surrounding hallucinations, in addition to the uncertainty that Agentic AI systems face as illustrated in Fig. 1, it is critical then to ask what is the risk of a hallucination? Risk is generally thought of as the likelihood of a hazardous outcome in combination with the consequence of that outcome. To this end, Fig. 2 represents various Agentic AI systems found in the world today, plotted in terms of their degree of non-determinism on the x axis and their safety criticality on the y axis. The shading across the top of Fig. 2 indicates increasing risk of a harmful human outcome such as serious injury or death.

The bottom half of Fig. 2 represents those systems that generally operate in low safety-criticality settings, with increasing degrees of non-determinism. The bottom of Fig. 2, for example, illustrates that robotic vacuum cleaners contain Agentic AI that can range from operating on 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 will vacuum every inch of floor). Robotic vacuum cleaners are relatively small and operate at low speeds, so the risk of human harm is quite low, regardless of the degree of non-determinism. Similarly in the bottom right corner, computer vision used in an interactive video game for human pose estimation is an example of Agentic AI that if and when it fails, has little significant bearing on the physical well-being of the human user.

As the safety-criticality axis increases in Fig. 2, the examples reflect uses that could have increasingly harmful effects. A smart home device can include AI that ranges from rules-based to sophisticated machine learning. However, a malfunctioning smart home device could put a user at risk, for example, if it overheats and starts a fire. This possible consequence is offset by the low likelihood of such an occurrence, which is why it is in the blue low risk region. Sentiment analysis Agentic AI, on the other hand, straddles the midline between low and medium-to-high risk.

Sentiment analysis is a growing application of machine learning and most applications are benign in that your car might tell you it thinks you are tired, or an email checker tells you an email sounds too harsh. However, there are a growing number of proposed therapeutic Agentic AI chatbots widely available with the rising popularity of LLMs. Medical practitioners are justifiably concerned that Agentic AI cannot be guaranteed to do no harm [22], and if a human's sentiment is incorrectly estimated, serious harm can come to a human. Indeed, the first lawsuit asserting a suicide was caused by a chatbot has been filed [51], and it is critical that researchers and developers understand both the physical and mental harm that can result from LLM development. As mentioned previously, LLMs have been shown to hallucinate in transcription and such errors carry very real possibilities for human harm in medical settings so these applications are very much in the high-risk zone.

As seen in Fig. 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 it also operates at the highest degrees of non-determinism. Just underneath this use case is the use of Agentic AI in vehicles like self-driving cars and autonomous aircraft. While there is no publicly-available data on the use of GenAI in weapons, there is quite a bit of data on the successes and failures of AI in vehicles, most notably in self-driving vehicles. Thus, those successes and failures can aid in understanding what similar issues will likely occur for Agentic GenAI in even more complex systems, discussed in the next section.

5 Agentic AI in Self-Driving Vehicles

While there are various self-driving pilot operations in a handful of cities in the US, the industry has fallen short on its promises to bring cheap robotic taxis (aka robotaxis) to market. This delay can be attributed to a number of issues, but the inability to develop reliable and highly accurate decision-making systems is one of the primary causes [12].

Self-driving cars rely on Agentic AI in the form of end-to-end machine learning systems which take raw data from sensors like radar, lidar and camera-based computer vision and map them directly to vehicle control actions. The use of end-to-end ML control models represents one of the most advanced uses of Agentic AI today, but these advances have also ushered in new problems. Using data from mandatory federal accident reporting requirements, recent research has shown that the use of Agentic AI in self-driving cars has led to many serious real-world problems, including those involving perception and uncertainty management, discussed in the next section.

5.1 Missed & False Object Detection

In March of 2023, a Cruise self-driving car rear-ended an articulated municipal bus in San Francisco [45]. Because of the onboard LIDAR, such an accident was thought to be nearly impossible, especially at slow speeds. Cruise admitted that its self-driving car's sensors detected the bus, but because it was articulated (a bus with two hard shells at the ends joined by a soft accordion-like middle), the self-driving car did not correctly identify the bus. The rear part of the bus was effectively filtered out by the AI [45]. This is known as a missed detection and such accidents account for about 10% of crashes in the federal crash data. Equally concerning are false object detections, aka hallucinations, a well-known problem with computer vision convolutional neural networks (CNNs) [31] [12].

While the exact number is unknown, when compared to missed detection crashes, twice as many self-driving accidents have been caused by false positives, i.e., self-driving cars detecting a non-existent object, leading to a hard braking event and then a struck-from-behind crash. This is also known as phantom braking. More research is needed to determine causes and frequencies of these false detections in self-driving systems, but some researchers attribute these hallucinations to overly-optimistic confidence intervals [39] or partially visible objects [31]. Others have evidence of shadows causing possible hallucinations [55] [4], and sensor washout at low sun angles could also affect object detection [11]. Just as in hallucinations for GenAI, there is no consensus on why or when such false detections occur, nor how to mitigate them.

5.2 Reasoning under Uncertainty

Given that any autonomous weapons necessarily must also be able to reason under the fog of war, which is maximum uncertainty, it is instructive to see how well self-driving cars are performing when faced with uncertainty. One small example of how well self-driving cars reason under uncertainty is

the case of a large truck attempting to reverse at an intersection to order to make a tight turn. Such a maneuver requires the driver to reverse to generate truck turning room. While trivial for humans to comprehend, the federal accident database contains several reports of crashes because self-driving cars would not slowly back up and create space for the truck turn, and were hit while the truck reversed [43].

There have been other more serious cases where self-driving cars could not cope with uncertainty including a Cruise vehicle dragging a pedestrian under a car for 20 ft [19] and refusing to yield when operating near ambulances and firetrucks, blocking critical services and endangering the public [56]. The inability of self-driving cars to deal with increasing uncertainty has led to development of remote operation centers where humans closely monitor vehicle progress, intervening when either the cars make a wrong decision or are unable to make a decision about the next set of actions [12]. Currently, there are no true "self-driving" cars in operation today, as all require direct human oversight because they cannot cope with uncertainty.

The fact that self-driving cars, arguably the most advanced public-facing Agentic AI in the world today, must have human babysitters for safe operation raises serious questions about whether any weapon system should use any AI for control and targeting., especially in fog-of-war settings. While AI could be successful in low uncertainty weapon engagements such as striking a building on a sunny day, the more dynamic and uncertain a situation is, the more likely AI will make a mistake [13]. This is an ongoing debate in international forums, especially in terms of the morality and legality of using such weapons, with the US declaring that for now, autonomous weapons targeting and release will be under the supervision of humans [50]. Indeed, for drone missions, the military has multiple layers of human oversight through remote operations centers. This insistence that humans be in the weapon targeting and release chain of decision making is an implicit admission that traditional AI currently in use today cannot be trusted to make reliable and safe decisions.

Despite the fact that traditional AI cannot perform safely and reliably under uncertainty either on the road or on the battlefield without human oversight, this has not stopped companies from claiming to bring GenAI to the battlefield [2]. Undoubtedly companies like OpenAI want to monetize their significant investment in GenAI infrastructure and defense contractors want what appears to be a relatively straightforward software solution that can theoretically streamline the kill chain. However, both high and unpredictable GenAI hallucination rates make GenAI's use in safety-critical settings extremely risky. This is especially true when considering that many think AI is needed to combat high uncertainty but actually only adds to it.

Regardless of whether a GenAI system is wrapped inside a larger Agentic AI weapon control system, the likelihood a GenAI component could make a mistake that could influence a negative outcome is unacceptably high. In its current form, GenAI embedded in an Agentic AI weapon control system can never be trusted to autonomously operate. High hallucination rates would also be a nightmare for any group of remote operators attempting to oversee such a system. History is quite clear that when remote operators are presented with a significant amount of false information, especially under time pressure, they can easily be cognitively saturated. High cognitive workload opens the door to automation bias, which is when an operator blindly follows an automated system's recommendation without question [14]. Thus, even if GenAI is supervised by humans in safety-critical systems, the time pressure and speed of weapon delivery makes this an untenable concept of operations.

6 Dangers in Rushing Agentic AI Development

Understanding that GenAI is a set of emerging and dynamic technologies, if they are not ready now for deployment, when will we have some indication that they could be? To better understand the nature of autonomous technology development, one first has to understand the stages of development, formally called the Technology Readiness Levels (TRLs), which range from 1, basic research, to TRL 9, which represents a system's readiness for operation in its intended domain. Academic research spans TRLs 1-3, while technology maturation takes place in TRLs 4-6 and system integration takes place in TRLs 7-9 [46].

While GenAI is arguably at TRL 9 for commercial back- office functions, it is not clear where it resides for the purposes of inclusion in a weapon system. Given high hallucination rates, it is not ready for deployment in weapons systems today, either as a control technology or even as an assistive tool. While not a mature technology today, when can we reasonably expect that it might be? To

inform this debate, it is useful to look at the technology development life cycle of other autonomous systems. To this end, Table 1compares three autonomous systems with different degrees of embedded AI. The first category includes unmanned aerial vehicles, aka drones, where modern systems started in 1965 with military drones. These were in development for 30 years before their first military deployment with the Predator in 1995 [42]. Three years later, Draganfly was the first company to start R&D with quadrotor drones, and eight years later, the FAA allowed commercial drones in the national airspace. This ushered in the period of commercial research and development (R&D), as well as deployment 54. For this safety-critical system with symbolic AI and various levels of human supervision, it took 30 years of R&D before it could be safely and reliably used on the battlefield, and 41 years in total before it could be commercially deployed.

Table 1: Years of military and commercial research & development (R&D) for the deployment of drones, self-driving cars and GenAI.

Tech	AI Type	Military R&D/Deploy Yrs	Commercial R&D/Deploy Yrs
Drones	Symbolic	30/30	8/19
Self-Driving Cars	Neural nets w/ some symbolic	4/0	6/10
GenAI	Neural nets	0/0	6/5

The second autonomous system of interest includes self-driving cars, which also had a military start, albeit a short stretch of 4 years (Table 1) with the various DARPA Grand Challenge competitions. These competitions led to the unusual early spin-out of military technology, which led to the start of Google's self-driving research efforts in 2009 and their first limited commercial services in 2015 under their subsidiary Waymo. While the US military continues to conduct research in autonomous ground vehicles, there are no deployed military ground autonomous vehicles today.

Demonstrating a critical shift away from military-sponsored research, the rise of GenAI began with the 2014 publication of the first paper on generative adversarial networks, enabling the development of transformers and GenAI as we know it today [24]. GPT-3 from OpenAI was the first commercially-available LLM in 2020. Not only was there no military R&D for GenAI, the time to market was only 6 years, whereas it was 10 (4+6) for self-driving cars. While there are many other factors at play such as the rise of venture capital funding and compute power, these combined factors resulted in a dramatic decrease in time to market for self-driving cars and GenAI, bypassing the more conservative and slower military R&D process.

This rapid increase in time to market is impressive, especially since the complexity of the systems increases while the time to market drops. However, there are hidden costs. One of the reasons drones took so long to mature in the military is the level of testing that was required in the maturation stages to ensure safety of flight, not just for the drones but also manned aviation and supporting humans on the ground. Although time and resource intensive, this testing was critical to understanding the weaknesses and limitations of the systems, which led to iterative and improved designs. The ability of the military to absorb the risk and cost of early failures was indeed a catalyst for the commercial drone industry. Notably, despite the safety-critical nature of drone operations, the period of commercial R&D is on par with that of self-driving cars and GenAI.

Table 1 illustrates that self-driving cars only benefitted from a few years of military R&D, and GenAI had none. While possibly good from a taxpayer perspective, it also means that these industries did not benefit from spending longer periods of time in the mid-TRLs where the bulk of technology maturation occurs. This means that the systems were not exhaustively tested and we do not understand their failure modes. For example, we still do not know why, how, when or where self-driving cars or LLMs will hallucinate. While eradicating hallucinations is not possible, it could be possible to predict those conditions and circumstances that raise the likelihood of hallucinations so that they can be mitigated. For example, if hallucinations have a higher likelihood of occurring in tunnels due to shadows, then the headlights could possibly become much brighter to remove such shadows.

Until we can develop predictive models to these ends, GenAI is simply too dangerous to include in safety-critical systems, especially weapons. Further complicating matters is the dearth of research around AI testing and certification, both in academia and industry. Additionally, while we have developed stop-gap measures of human oversight and assistance in self-driving cars, for time-

pressured military missions that include weapons that leverage GenAI for control or targeting, remote oversight simply will not be an option due to compressed timelines. In self-driving pilot operations around the country, state agencies have determined that the risk of public injury or death is worth the deployment of experimental Agentic AI. This effectively means that the public bears the brunt of testing in order to mature a technology. This reasoning simply cannot hold for weapons deployment, and such acts would likely violate many international agreements and basic human rights.

7 Alternate View

Whenever there is a call for any kind of limitation on any technology, the most common opposing argument is that constraining technology development in any way hurts innovation and ultimately society is harmed through a lack of jobs and a lack of access to a potentially useful technology. Supporting this argument, pilot demonstrations with self-driving cars have shown just how critical real-world deployments of AI are in understanding failure modes. For example, the prevalence of phantom braking issues in self-driving applications was not known until data collection occurred [12]. Such exposure to real-world settings is critical, especially in computer vision system development where impoverished training sets for edge cases need augmentation. Deploying connectionist systems in the real world is a fundamental stepping stone to maturation. If researchers and companies were only allowed to use GenAI in weapons in simulation and not on the battlefield, it would take companies far longer to discover latent failures and gaps in functionality. Indeed, it is likely impossible that a connectionist GenAI system could be matured to the point of successful deployment only using simulation.

GenAI is the quintessential Scylla and Charybdis conundrum. It is too unreliable to deploy in autonomous systems today without close human supervision, but operational time scales will make this extremely difficult, if not impossible. On the other hand, GenAI and other similar connectionist systems in safety-critical settings cannot be successfully matured without real world exposure. In order to move forward, there are two steps that researchers and developers must embrace. The first is an organized and well-funded national program to tackle hallucinations (aka confabulations or predictive errors) in all of connectionist AI. While there are pockets of researchers addressing this problem, there are still no reliable methods or models that can predict when a hallucination is likely to occur or the magnitude of such an error. Without such models, safety-critical systems that embed connectionist AI will never be able to move beyond human supervision.

The second need is the development of a new science of testing and certification of connectionist AI. The US military has long been interested in this problem but progress has lagged over the last twenty years. In addition to the need for more quantitative methods to model AI risk and capture an AI system's ability to address uncertainty including hallucinations, the development of physical test ranges that allow safety-critical systems to gain exposure to the real world and possible corner cases under controlled conditions is needed. Such test ranges would be very expensive, but private-public partnerships would help defray the costs, and a key missing piece of the puzzle. Unless and until both of these steps are realized, there is no place for generative AI in the control or targeting of any weapon, or in any safety-critical system in general.

Acknowledgments and Disclosure of Funding

This paper was supported in part by the Office of Naval Research through the Science of Autonomy program. Dr. Cummings was an advisor to the Palantir Privacy and Civil Liberties committee at the time of acceptance. The anonymous reviewers' comments were also very helpful in improving the paper.

References

[1] Mohamed Aghzal, Erion Plaku, and Ziyu Yao. Look further ahead: Testing the limits of GPT-4 in path planning. In *IEEE 20th International Conference on Automation Science and Engineering*, 2024.

[2] Anduril Industries. Anduril partners with OpenAI to advance U.S. artificial intelligence leadership and protect U.S. and Allied Forces, Dec. 4 2024.

[3] Gabriel Y. Arteaga, Thomas B. Schön, and Nicolas Pielawski. Hallucination detection in LLMs: Fast and memory-efficient fine-tuned models. In *Northern Lights Deep Learning Conference*, 2025.

[4] B. Bauchwitz and M.L. Cummings. Individual differences dominate variation in ADAS takeover alert behavior. *Transportation Research Record: Journal of the Transportation Research Board*, 2676(5):489–499, 2022. doi: 10.1177/03611981211068362.

[5] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[6] Jacob Browning and Yann LeCun. Language, common sense, and the Winograd schema challenge. *Artificial Intelligence*, 325:104031, 2023. doi: 10.1016/j.artint.2023.104031.

[7] M Chelli, J Descamps, V Lavoué, IC Trojan, M Azar, M Decker, JL Raynie, G Clowez, P Boileau, and C Ruetsch-Chelli. Hallucination rates and reference accuracy of ChatGPT and bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research*, 26: e53164, 2024. doi: 10.2196/53164.

[8] Jaemin Cho, Abhay Zala, and Mohit Bansal. Probing the reasoning skills and social biases of text-to-image generation models. In *ICCV*, 2023.

[9] Patrick Cockburn. In Middle East wars it pays to be skeptical, 2018. URL `https://www.counterpunch.org/2018/04/23/in-middle-east-wars-it-pays-to-be-skeptical/`.

[10] K.M. Collins, A.Q. Jiang, S. Frieder, L. Wong, M. Zilka, U. Bhatt, T. Lukasiewicz, Y. Wu, J.B. Tenenbaum, W. Hart, T. Gowers, W. Li, A. Weller, and M. Jamnik. Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences*, 121 (24), 2024. doi: 10.1073/pnas.2318124121.

[11] M. Cummings and B. Bauchwitz. Driver alerting in ADAS-equipped cars: A field study. In *IEEE International Conference on Assured Autonomy*, 2023.

[12] M. Cummings and B. Bauchwitz. Identifying research gaps through self-driving car data analysis. *IEEE Transactions on Intelligent Vehicles*, pages 1–10, 2024. doi: 10.1109/TIV.2024.3506936.

[13] M. L. Cummings. Lethal autonomous weapons: Meaningful human control or meaningful human certification? *IEEE Technology and Society*, 38(10):20–26, 2019.

[14] M.L. Cummings. Automation bias in intelligent time critical decision support systems. In *AIAA 3rd Intelligent Systems Conference*, 2004.

[15] M.L. Cummings. Man vs. Machine or Man + Machine? *IEEE Intelligent Systems*, 29(5):62–69, 2014.

[16] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93, 2024. doi: 10.1093/jla/laae003.

[17] Ben Dickson. How LLMs are ushering in a new era of robotics, 2024. URL `https://venturebeat.com/ai/how-llms-are-ushering-in-a-new-era-of-robotics/`.

[18] S. Elbaum and J. Panter. AI weapons and the dangerous illusion of human control. *Foreign Affairs*, 2024.

[19] Quinn Emanuel. Cruise safety report, 2022. URL `https://assets.ctfassets.net/95kuvdv8zn1v/zKJHD7X22fNzpAJztpd5K/ac6cd2419f2665000e4eac3b7d16ad1c/Cruise_Safety_Report_2022_sm-optimized.pdf`.

[20] Jessica López Espejel, El Hassane Ettifouri, Mahaman Sanoussi Yahaya Alassan, El Mehdi Chouham, and Walid Dahhane. Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. *Natural Language Processing*, page 100032, 2023. doi: 10.1016/j.nlp.2023.100032.

[21] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

[22] A. Fiske, P. Henningsen, and A. Buyx. 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), 2019. doi: 10.2196/13216.

[23] B. Gertz. China's military working on AI weapons and systems for warfighting and 'overthrowing regimes', 2023. URL `https://www.washingtontimes.com/news/2023/aug/22/chinas-military-working-ai-weapons-and-systems-war/`.

[24] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Neural Information Processing Systems*, 2014.

[25] Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517, 2023. doi: 10.1162/tacl_a_00615.

[26] Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing uncertainty for large language models through input clarification ensembling. In *ICML'24: Proceedings of the 41st International Conference on Machine Learning*, 2024. doi: 10.5555/3692070.3692835.

[27] E. Hüllermeier and W. . Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110:457–506, 2021. doi: 10.1007/s10994-021-05946-3.

[28] Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. LLM internal states reveal hallucination risk faced with a query. In *The 7th BlackboxNLP Workshop*, 2024.

[29] Adam Tauman Kalai and Santosh S. Vempala. Calibrated language models must hallucinate. In *56th Annual ACM Symposium on Theory of Computing*, 2024.

[30] Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Saldyt, and Anil B Murthy. LLM's can't plan, but can help planning in LLM-Modulo Frameworks. In *41st International Conference on Machine Learning*, 2024.

[31] O. S. Kayhan, B. Vredebregt, and J. C. van Gemert. Hallucination in object detection — A study in visual part verification, 2021.

[32] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, page abs/1703.04977.

[33] Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. Careless whisper: Speech-to-text hallucination harms. In *FAccT '24: ACM Conference on Fairness, Accountability, and Transparency*, pages 1672 – 1681, 2024. doi: 10.1145/3630106.3658996.

[34] Ernests Lavrinovics, Russa Biswas, Johannes Bjerva, and Katja Hose. Knowledge graphs, large language models, and hallucinations: An NLP perspective. *Journal of Web Semantics*, 85: 100844, 2025. doi: 10.1016/j.websem.2024.100844.

[35] S.A. Lehr, A. Caliskan, S. Liyanage, and M.R. Banaji. ChatGPT as research scientist: Probing GPT's capabilities as a research librarian, research ethicist, data generator, and data predictor. *Proceedings of the National Academy of Sciences*, 121(35):e2404328121, 2024.

[36] Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. Are multilingual LLMs culturally-diverse reasoners? An investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 2016–2039. Association for Computational Linguistics, 2024.

[37] B. Marr. Generative AI sucks: Meta's chief ai scientist calls for a shift to objective-driven ai, 2024. URL `https://lnkd.in/eymYCQg6`.

[38] Timothy R. Mcintosh, Tong Liu, Teo Susnjak, Paul Watters, and Malka N. Halgamuge. A reasoning and value alignment test to assess advanced GPT reasoning. *ACM Transactions on Interactive Intelligent Systems*, 14(3):1–17, 2024. doi: 10.1145/3670691.

[39] G. Melotti, C. Premebida, J. J. Bird, D. R. Faria, and N. Gonçalves. Reducing overconfidence predictions in autonomous driving perception. *IEEE Access*, 10:54805–54821, 2022. doi: 10.1109/ACCESS.2022.3175195.

[40] D. Milad, F. Antaki, Jason Milad, A. Farah, Thomas Khairy, D. Mikhail, Charles-Édouard Giguère, S. Touma, Allison Bernstein, A-A Szigiato, T. Nayman, G.A Mullie, and R. Duval. Assessing the medical reasoning skills of GPT-4 in complex ophthalmology cases. *British Journal of Opthamology*, 2024. doi: 10.1136/bjo-2023-325053.

[41] Philipp Mondorf and Barbara Plank. Beyond accuracy: Evaluating the reasoning behavior of large language models - A survey. *First Conference on Language Modeling*, 2024.

[42] L.R. Newcome. Unmanned aviation: A brief history of unmanned aerial vehicles. *AIAA*, 2012. doi: 10.2514/4.868894.

[43] NHTSA. Second amended Standing General Order 2021-01, 2023. URL `https://www.nhtsa.gov/sites/nhtsa.gov/files/2023-04/Second-Amended-SGO-2021-01_2023-04-05_2.pdf`.

[44] Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *62nd Annual Meeting of the Association for Computational Linguistics*, pages 10862–10878. doi: 10.18653/v1/2024.acl-long.585.

[45] ODI. Nhtsa Part 573 Safety Recall Report 23e-029, 2023.

[46] Office of Systems Engineering and Architecture. Technology readiness assessment guidebook, Department of Defense. 2025.

[47] OpenAI. Learning to reason with LLMs, 2024. URL https://openai.com/index/learning-to-reason-with-llms/.

[48] OpenAI. OpenAI o3 and o4-mini system card, April 16 2025. URL https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf.

[49] K. Orland. Viral ChatGPT-powered sentry gun gets shut down by OpenAI, 2025. URL https://arstechnica.com/ai/2025/01/viral-chatgpt-powered-sentry-gun-gets-shut-down-by-openai/.

[50] OSD. DOD Directive 3000.09: Autonomy in weapon systems, 2023.

[51] K. Payne. An AI chatbot pushed a teen to kill himself, a lawsuit against its creator alleges, 2024. URL https://apnews.com/article/chatbot-ai-lawsuit-suicide-teen-artificial-intelligence-9d48adc572100822fdbc3c90d1456bd0.

[52] Vipula Rawte, Swagata Chakrobarty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit P. Sheth, and Amitava Das. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *2023 Conference on Empirical Methods in Natural Language Processing*.

[53] K. Ritter, A. Madhani, and D. Hazell. Pentagon video shows Russian jet dumping fuel on us drone, 2023. URL https://apnews.com/article/russia-ukraine-war-drone-us-intercept-b28e38a7cd046f8685b89a3038973e49.

[54] J. Sizemore and R. Posey. Unmanned aircraft systems (UAS) certification status, FAA report. 2006.

[55] Ashwini Kanakapura Sriranga, Ganesh Bhagwat, Thrilochan Sharma Pendyala, and Prithvi Pagala. Trigger-based pothole detection using smartphone and OBD-II. In *IEEE International Conference on Electronics, Computing and Communication Technologies*, pages 1–6, 2020. doi: 10.1109/CONECCT50063.2020.9198602.

[56] J. Tumlin, T. Chang, and L. Bohn. Protest of Cruise LLC Tier 2 advice letter (0002), 2023. URL https://www.sfmta.com/sites/default/files/reports-and-documents/2023/01/2023.01.25_ccsf_23.0125_cpuc_cruise_tier_2_advice_letter_protest_002.pdf.

[57] M. Valdenegro-Toro and D. S. Mori. A deeper look into aleatoric and epistemic uncertainty disentanglement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1508–1516, 2022. doi: 10.1109/CVPRW56347.2022.00157.

[58] Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. On the planning abilities of large language models: A critical investigation. In *37th International Conference on Neural Information Processing Systems*, 2023. doi: 10.5555/3666122.3669442.

[59] Kyle Wiggers. Microsoft claims its new tool can correct ai hallucinations, but experts advise caution, 2024. URL https://techcrunch.com/2024/09/24/microsoft-claims-its-new-tool-can-correct-ai-hallucinations-but-experts-caution-it-has-shortcomings/.

[60] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind's eye of LLMs: Visualization-of-thought elicits spatial reasoning in large language models. In *NeurIPS*, 2024.

[61] Yijun Xiao and William Yang Wang. On hallucination and predictive uncertainty in conditional language generation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *16th Conference of the European Chapter of the Association for Computational Linguistics*, page 2734–2744. Association for Computational Linguistics, 2021.