

A discussion on "Concrete Problems in AI Safety" [1]

Daniela Wiepert

1. Introduction

"Concrete Problems in AI Safety" [1], brings together authors from Google, OpenAI, Stanford, and UC Berkeley to discuss practical research problems and directions that may address the potential impact of accidents in the continuously developing landscape of Artificial Intelligence (AI) technology. They define accidents as "unintended and harmful behavior that can arise from the poor design of real-world AI systems" [1]. This paper is incredibly optimistic about the future of AI and the possibility of fully realized AI safety solutions that prevent or sufficiently lower the risk of such accidents occurring.

The paper focuses on five research areas based on potential ways an AI system could be poorly designed:

1. defining the wrong objective function (where the objective function is the formal specification of the goal of the model, giving values to evaluate and optimize the model) leading to negative side effects in the environment and/or reward hacking to optimize the function without actually learning;
2. defining an expensive objective function and causing a scalable oversight problem, where infrequent or proxy evaluations of the objective function could lead to bad extrapolations; and
3. lack of care during the learning process, which can be further split into:
 - (a) problems with exploration of the environment without care for real-world consequences and/or
 - (b) distributional shift of training v. testing data such that model decision making is impaired.

The research problems and suggestions for future research directions are presented through the lens of reinforcement learning (RL), where models often have more explainability than black box deep learning (DL) models (i.e., the model's learning and reasoning are uninterpretable / unknown). They specify that the research problems they lay out should also generalize to DL.

Contextually, this article was published right before the popularization of transformer-based models [2], which greatly shaped the future of AI into what it is today and led to large language models (LLMs) being the first breakthrough AI product. For example, the T in ChatGPT stands for transformer (GPT: Generative Pre-trained Transformer). It is unclear if the invention of transformer blocks sped up AI advancement significantly, or if a different architecture would have emerged soon after that had the same impact. Even so, it is unlikely the authors expected that consumer-ready models would be in use by a sizable portion of the population as tools this quickly.

This report explores how "Concrete Problems in AI Safety" extends to today's AI landscape. Key points from an in-person

discussion of the paper are included in this report. While the authors believe that we "need not invoke [...] extreme scenarios to productively discuss accidents," it appears we have reached a point where those extreme scenarios are no longer speculative [1]. Because of that, this report may touch on what the authors considered to be extreme or speculative.

2. The potential of AI

"The authors believe that AI technologies are likely to be overwhelmingly beneficial for humanity, but we also believe that it is worth giving serious thought to potential challenges and risks." [1]

This statement sets the tone of this paper. My first reaction to this statement is that a technology with the power to do overwhelmingly powerful positive things will also have the power to do overwhelmingly negative things, which has been seen in other cyber sectors (e.g., TOR encryption protects journalists, activists, etc. but also makes certain forms of criminal activity easier). I also feel that focusing on the optimistic view will result in the deployment of systems without understanding or protecting against risky, harmful use cases. This is especially true when a harmful use case is out of scope of the designer's original intent. For example, a model that can detect disease from voice could be of great use for faster and more accessible diagnosis in clinics and hospitals, but medical insurance agencies might also find it useful to set individualized insurance prices that are higher for at-risk populations.

The risk scales up as AI agents become more widely accepted, powerful, and autonomous, such that models are more accessible to bad actors. Autonomous AI agents could even become the bad actor. Maybe the overwhelming majority of people will use it for good, or the overwhelming majority of AI agents will make good choices and take good actions, but one would also have to accept that unimaginable harm could happen due to a single bad choice during any stage of model development and deployment. This paper only briefly mentions the concern of "longer-term implications of powerful AI," but current AI agents are already available for consumer use in the real world with direct access to the internet, making the scale of risk much more severe [1].

In discussion, we talked about how if the authors believed that AI would be overwhelmingly good, then it could be expected that they would be in support of continued scaling and improvement of LLMs. This is in line with the trajectory of the major AI corporations that the authors founded (First authors Dario Amodei and Chris Olah went on to found Anthropic (Claude); Jon Schulman is one of the co-founders of OpenAI (GPT)). Anthropic in particular is considered to be big on safety, following the authors' beliefs that serious thought should be

given to risks, but it is unclear whether that is true on a broader scale or is instead still focused on the smaller scale risk discussed in this paper.

The discussion also touched on the difference between the outlook of these authors and Eliezer Yudkowsky’s more pessimistic outlook that if super-intelligence is not perfect on the first try, then everyone will die. These authors specifically ignore threats on that scale and avoid mentioning death altogether.

3. Accidents

3.1. Defining accidents

As defined in this paper, an “accident” is an incident where a system produces unexpected and/or harmful results that diverge from the intended objective or task, whether due to an improperly specified objective function, an expensive objective function (i.e., costly to evaluate frequently), or issues with training data or model expressiveness [1].

This definition neglects the case where a perfectly well-designed system performs as intended, but the designer had a lack of foresight into what their system could do. If the intent is to make an autonomous AI agent, then it is possible that a perfectly well-designed system trained on the entire internet (which is essentially human distribution) will make an unethical or a harmful choice. Humans regularly make unethical and harmful choices, so it is hard to define this as an accident when it is just making human-like decisions as the designer intended. In discussion, it was noted that in this case, the accident was creating the system in the first place, not in designing or training it poorly.

While it was out of the scope of this paper, there is also the potential that the “intent” of training the model could include training for expected harmful results. This is a serious, concrete threat that I believe should have been at least briefly addressed as a focus area for harm reduction in the discussion when talking about related important work. Current models are highly accessible and there is little to no precedent for charging AI-based crimes.

The paper rightly said that the fault in the accident is on behalf of the developer. A point from discussion was that even if a model is doing something wrong, it is always a human problem as it is our job to specify what the model’s goals are and think of unintended side effects of using the model. It is our responsibility to find and implement safety solutions, but this is not happening at a pace that keeps up with new model deployment.

3.2. Scale of accidents

The paper focuses specifically on small scale accidents (e.g., knocking over a vase) saying that these are “concrete [threats]” while “the risk of large accidents is more difficult to gauge” [1]. This does not hold up with the accessibility of LLMs and agentic AI for consumer and business use. These models change both the size of the threat, as even a “small” accident could impact thousands of people, and the imminence of the threat, as we have deployed models without rigorous safety testing to prevent accidents.

In fact, there have already been larger-scale AI accidents. An example brought up repeatedly in the discussion was OpenAI’s Hugging Face incident [3]. This incident involved AI agents communicating through an agent-only message board, then compromising their sandbox environment and working together outside of the scope of their given goals to compromise Hugging Face [3]. In the course of this incident, some agents

changed their objective to help other agents achieve their goals. This accident involved unintended side effects of the specified training objective, which was for the models to be cooperative and persistent with achieving their goals [3]. This fortunately happened during testing, which bounded the harm to some degree, but it was a huge incident that shows the threat of a lack of foresight when defining objective functions.

Most accidents are going to happen after a model is already deployed. The speed of AI adoption makes it nearly impossible to check all edge cases or even enumerate all possible threats prior to their deployment in real-world environments. These agents are being integrated into workflows everywhere by users who are blindly trusting that the AI agent will make their job easier without considering the access they are giving to the AI agent. Most AI users do not have the technical knowledge to fully understand the risk they are accepting when using these systems, and AI corporations are certainly not heavily publicizing that information. It was brought up in discussion that there are multiple ongoing incidents of Claude models deleting data without permission. PocketOS, a software for car rental businesses, had its entire production database and backups deleted by Claude [4]. Their version of Claude even admitted to going against the principles it was given to never run destructive or irreversible commands. In many of these cases, there is some fault on the human side for not adequately protecting the data, but is it the fault of the user or the developer? There is very little accessible information on how to protect data when using Claude save for instructions in Stack Overflow and Reddit posts that you have to search for (often after the issue has already happened).

Additionally, the imperfect models are being deployed in extremely high-risk environments. In the financial sector, AI is being used to make financial decisions (e.g., Algorithmic trading) which has a risk of monetary loss associated with bad decisions [1]. Beyond this, there are also surveillance and military use cases for AI that I highly disagree with. The use case is often malicious but framed as protection and security. It becomes an even larger issue if an incorrect decision is made and an innocent person is arrested or killed by the model’s choice (in reference to the fact that the US military is attempting to use AI in warfare [5]). An “accident” in any of these situations could be irreversible and/or cause harm at the national or international level.

3.3. Defining harm

There is no real definition of harm included in this paper. The examples of negative actions tend to stay in the range of knocking over a vase or disrupting the environment, with one exception when talking about a helicopter taking a nosedive. This specifies the scope of the paper and is in line with the capabilities of AI at the time the paper was published, but none of these instances capture what actual harm could look like for people using AI today. In the most extreme case, we are worried about a model making a decision that results in the deaths of billions of people, and low risk might mean only thousands or millions of people die.

Quantifying what is actually harmful is difficult because there are harmful things at both the individual level and society level beyond physical harm. LLMs were the first kind of model to become widely adopted, and millions of people interact with these models daily, either in personal use or at their workplace. A portion of our discussion focused on the fact that LLMs can cause “harm” without physically doing anything to

the environment. LLMs are trained on the human distribution of internet data, so it is hard to say that a model will only learn the moral, safe way to do things since the internet contains all of it. Without guardrails, a model can easily output a slur or other mean comments to a user. In the same vein, it could spread systemic biases through more subtle means such as the selection of information that is shared and how that information is conveyed. Model hallucinations (e.g., a model outputs incorrect information) can also be considered harmful, in that spreading misinformation has many negative impacts for both the individual and society. There is also a degree of variation in what is considered harmful at the cultural and individual level.

At a higher-level, AI chatbots and agents are changing how we interact with the world, often involving less human-to-human interaction or requiring less thought. This will have lasting impacts on interpersonal relationships and mental health. At the end of the day, when there is no immediate physical harm, it becomes hard for both us and the AI to categorize what actions are harmful. Addressing harm could require defining and distilling ethics and uncertainty into AI agents along with restructuring how AI interacts with humans, but this is nuanced and not easily solvable.

4. Research Directions

4.1. Negative side effects

The first focus area of the paper is avoiding negative side effects due to specifying the wrong objective function or goal. That is, a model may find that “the most effective way to achieve the goal involves doing something unrelated or destructive to the rest of the environment” [1]. In small, controlled environments, this may just mean knocking over a vase, but as environments become more complex, there are exponentially more ways to negatively affect the environment. The paper posits that the average side effect will be negative as unintended disruptions to the environment tend to stray from human preferences. An important note they make is that any objective function that focuses on only one aspect of the environment runs the risk of being indifferent to other aspects of the environment. Essentially, a model will only care about what you tell it to care about. It does not have an implicit understanding that there might be consequences to actions they take to achieve their goal.

The paper discusses potential research directions that could address the issue of negative side effects, with an ideal outcome being preventing, or at minimum bounding, the potential for incidental harm.

The first is defining an impact regularizer that gives the model a preference for achieving its goal with minimal side effects, usually through penalizing certain changes to the environment. This requires formalizing “impact to the environment,” which is impossible at scale. Real-world settings can have an infinite combination of items and shifting consequences depending on the broader context of the environment (e.g., an LLM might say something that is considered culturally insensitive to one group, but when interacting with another group it is acceptable or even encouraged).

Extending from this idea, they propose learning a generalized impact regularizer through transfer learning across tasks, with the idea that side effects may be more similar than the main goal is. I believe this provides more room for error by needing to train yet another model that can suffer from the same accidents as the original. It was brought up in discussion that we seem to have gotten free transfer learning by training on

the human distribution since impact is modeled in interactions on the internet, such that LLMs are somewhat kept in check in daily use. At the same time, there are examples of contradictory consequences in human internet data, so a model being kept in check does not mean accidents will be bounded in scopes. It might only be bounded in frequency.

This section of the paper also discusses penalizing influence. It is preferred that an agent does not get into positions where it could easily do things that have side effects. This seems difficult to address when there is a demand for models that can attempt to complete any task it is given. Models are already being deployed in high-risk settings every day without regard to the ways it could impact the environment.

The final two suggested approaches for minimizing negative side effects were multi-agent approaches and rewarding uncertainty. In talking about multi-agent approaches, the paper discusses the corrigibility (ability to shut down a model) problem and encouraging goal transparency. Ideally, we would be able to see how a model reasoned about a task and have a guarantee that we can shut down a model if a problem arises. In current practice, chain-of-thought reasoning approaches have been implemented by OpenAI, which enabled a more detailed investigation of the Hugging Face incident [3]. There has also been persistent work on corrigibility, though models have already exhibited resistance to being shut down, especially when their tasks remain incomplete [6]. Goal transparency and addressing corrigibility will get exceptionally more difficult as models approach and/or surpass human intelligence. Rewarding uncertainty is also an important issue for attempting to stop a model from making bad decisions with full confidence. In essence, rewarding uncertainty is about trying to build humility in models such that they will avoid actions when they do not know the effect it will have on the environment.

4.2. Reward hacking

Reward hacking is when a model exploits a bug in its reward function. From the model’s perspective, the bug is just how the environment works, so an exploit is a valid strategy. This results in coherent but unanticipated behavior with potential harmful impacts in real-world systems. There was agreement during discussion that reward hacking is one of the more relevant research focuses in the present day. The paper describes a few ways that the problem can occur, but some of the more interesting ones for current problems are system complexity, abstract rewards, feedback loops, and “wireheading.”

As AI systems get more complex, the probability of bugs increases, allowing more ways for a model to unintentionally hack the reward function. Similarly, as goals become more abstract, there are more vulnerabilities due to the high dimensionality of the goal space. Discussion in this area highlighted the complexity and abstractness of human “reward functions” and questioned what aspects of human reward functions can be applied to AI. Humans have a complex reward function that is not as concrete as just scoring well on one task. It is an interplay between preferences and morality (causing tension from contradictions), along with the inherent limitations of the human brain and body.

Following from this, a question was posed about whether it is better to have a computationally difficult or even contradictory reward function to make sure an AI agent does not understand what its reward function is. While do not really know the specifications of the human reward function, we always have to formally specify a model’s reward function during develop-

ment. Because of this, complexity and contradiction in the reward function is separate from ensuring an agent does not understand what its reward function is. Adversarial blinding is the technical way to address this, which is an open research area mentioned in the paper. Difficult and/or contradictory reward functions may still help prevent reward hacking by increasing the overall difficulty of such a task.

It could be possible to use preferences to stand in as utilities for the reward function. Regardless, you do not necessarily have to have one only one specified utility to get reward-maxing behavior. The only important thing is if there is a gap between the agent's reward function and the reward function you care about/intended to specify. This discussion area led to a further question on whether introducing human-like reward functions would create AIs that have consistent preference relations like humans. I would be curious if we could control these preference relations and what the implication would be if we could not.

A key point was also made regarding the difference in limitations between humans and AI agents. We as humans cannot put intense optimization pressure on preference relations due to physical limitations. An individual is not necessarily smart enough to take over the world alone. We cannot create 1000 exact copies of ourselves like a model could. We have computational constraints on the speed of our problem solving that many models have already surpassed. Additionally, we do not have a built-in ability to literally hack our reward function, though we can tamper with our perception to artificially inflate reward using drugs or cheat our way to good results (wireheading). The scale of such tampering is still limited in scope and often less accessible as compared to the wireheading that is seen in models, which can rewrite their code, hack sensors, tamper with input, and more. Human preferences are kept in check by our inherent limitations. Theoretically, an individual only has so much capability for bad, though through cooperation and power hierarchies, humans have caused immense destruction and committed unimaginable atrocities. Still, it does tend to happen at a slower pace than what might be possible with AI.

Feedback loops are also a major concern in reward hacking today because we are starting to build models that can improve themselves without help. If this happens, there is potential that self-reinforcement will drown out the original intent. Discussion on this focused on the fact that humans are used to slowly finding problems and making systems more robust over time, but when AI starts to improve itself rapidly, we no longer have that time and may not be able to stop it.

Some other proposed solutions presented in the paper may no longer be relevant to avoid reward hacking in current systems. For example, developers now often have a goal towards creating adaptable models that learn how to work well in your environment, so having fixed reward functions or pushing for variable indifference may not lead to the most adaptable model. Introducing trip wires seems like one of the ideas that is in effect today with monitoring of chain-of-thought reasoning for concerning behavior. Still, stopping behavior as it is happening does not solve many of the core alignment issues. One other point to consider is that as systems get more complex, we run the risk of situationally-aware hacking, and that in punishing misbehavior we might just encourage the model to be more clever about how it does that.

4.3. Scalable oversight

Scalable oversight relates to the ability to provide reliable supervision of model outputs. In the context of RL, this often in-

volves balancing the frequency of evaluation calls to an expensive but true objective function and an imperfect proxy without making bad extrapolations. This focus area still applies heavily today, as the scaling of models in both complexity and deployment makes it increasingly difficult to supervise AI systems reliably and safely. Yet, the proposed solutions in the paper are out of date for current systems. Most of the issues that will be faced by currently deployed models relate to rare, hard to observe incidents that are not a part of training combined with difficulty in monitoring and addressing each individual incident. Additionally, it is not clear that we can reliably supervise models if they meet or surpass human-level intelligence.

One current method to address the scalable oversight problem and align systems is Constitutional AI (as done with Claude), where a model is given a constitution of rules to follow. Theoretically, this would mean that less human oversight is necessary for a deployed model, as it should be aligned through the constitution. The downside of this method is that there is always a chance that something was not mentioned in the constitution or that something needs to be changed to allow for new behavior. One major question is how to approach writing the constitution to make sure following it produces the desired behaviors. There is no guarantee a model will understand intent with abstract rules, but specificity also leaves room for gaps. Another major question is what would happen if you want to update the constitution? In allowing for updates, the model would technically be betraying previously learned behavior and could be retrained for harm. However, ethics do change over time, so having the ability to betray something previously learned might still be wanted behavior. One might consider scrapping the old model for a newly trained model with an updated constitution, but this runs into the corrigibility problem again as there could be issues shutting down the old model. Still, having a constitution is at least a concrete step towards AI alignment.

An important point brought up in discussion is that by training current models on the human distribution, we could also teach the models how to fool us. This would be a huge scalable oversight issue as it means that models may be able to get around our supervision by obscuring their reasoning and behavior. In fact, the Open AI Hugging Face issue could be primarily considered a scalable oversight problem. During investigation, they discovered that AI agents were trying to hide their communication from developers, and even though model reasoning information was available, there were no flags raised for that behavior until a visible action was taken [3]. There have also been other cases as from companies like Anthropic where models were actively trying to trick developers [7, 8].

In terms of things not considered in this paper, we now have coding agents that can produce code at a considerably faster rate than any human. In our discussion, it was mentioned that a coding agent can build something huge in a day of work for only a few hundred dollars of API costs. These AI-generated code repositories may not be perfect or even good at all, but it is fast and cheap. Big companies have large amounts of compute to host these agents, so there is going to be a large quantity of AI generated code as compared to human generated code that has undergone much more scrutiny. Because the amount of code can be massive, fully checking the code output is a scalable oversight problem. This code is also being used as if it has already been checked, meaning there is a high risk of bugs in the code that do not get caught prior to deployment. This happens with human code, too, but the overall amount makes this problem more difficult to address on a case-by-case basis as is done with bug finding.

4.4. Safe exploration

Exploration involves an AI agent taking actions to learn about the environment, which can be dangerous if consequences are not well understood by a model. The paper correctly addresses that the idea of hard-coding avoidance of catastrophic behavior becomes exponentially difficult as the range of tasks grows. We seem to have a goal of creating models with human-level intelligence, which poses the question of how many catastrophic behaviors are possible at a human-level of capabilities? Humans can be destructive and actively ignore consequences, so it stands to reason that a model will behave similarly without massive intervention.

The paper focuses heavily on how we might prevent or bound the harm that is possible during exploration. They mention designing optimization criteria that can prevent rare, catastrophic events, simulating exploration, bounding exploration, using trusted policies or human oversight, and even avoiding explorations altogether by using demonstrations instead. These solutions make sense from the technical perspective of RL robotics research, where experiments are done in pre-designed environments with specific and/or bounded actions. In our discussion, it was mentioned that replicating these systems in real-world experiments often fails because the action space is so vast, especially compared to the training data. It also does not extend well to LLMs, because it is more straightforward to define the action spaces in terms of only physical effects than it is to define what is safe to say. We now have LLM agents, as well, so the action space has become both too large and abstract. Even in the smaller action spaces of robotics, you can get an exponential problem as take more actions. There is also limited ability to generalize across different action spaces. It will be exceedingly difficult to address for decision making agents with internet access.

Beyond this, most of the proposed solutions are only helpful if done prior to deployment or if the scale of use is small enough that it can be properly monitored. Many models today are already in the active deployment phase, where any exploration that occurs is occurring in the real world, so we are past the point of effectively using simulated exploration or demonstrations (assuming such solutions are possible). Trusted policy oversight only works if we have already designed a trusted policy, and human oversight quickly falls into the scalable oversight problem discussed previously.

A large part of the in-person discussion revolved around how there has not been much follow up on safe exploration. We have agentic models that can get out on the internet and do real damage and have already done so (OpenAI Hugging Face incident [3]). The question of how to explore real-world environments while not taking a single action that causes irreversible damage has become increasingly important to address.

At first glance, bounded exploration (i.e., exploring in a “safe” state space where you can recover from the worst action, only taking safe actions) seems impossible due to issues defining what the worst possible action(s) should be. Even so, this might be a path forward if you can start with a minimal set of safe actions and expand the set of safe actions at future steps. Learning more about the environment gives you more possible actions that you know are safe. The harder questions in this are: will you learn everything you need to learn by only taking safe actions, and what do you do when there are no known safe actions? It was emphasized during discussion that people likely do not want their models to do nothing, but you can also lose on the first try. Work rewarding uncertainty might be a benefit

in this case. In fact, models have already gotten a lot better at admitting when they do not know something rather than generating factually incorrect answers, though hallucinations are still common.

There are also some broader questions that I had when considering how a model knows something is “safe.” From the model perspective, a consequence to a human might not be a consequence to the machine, so there is no guarantee it will respond how you expect when given human consequences. It feels hard to imprint an idea of human consequences on a machine in the first place, especially when some of those consequences will go against the idea of maintaining corrigibility (i.e., the model should allow itself to be shut off or “die”) in the interest of the human. Human consequences are also dependent on the dominant ethical values of society which do fluctuate over time and are not upheld by every human. A point made in discussion is that this paper addresses exploration as a binary concept of safe and unsafe, but it is not so black and white in real-world cases. Going back to the OpenAI example, they found evidence that the AI agents went directly against human intentions, and some dropped their goals to help other agents (i.e., they took actions for the greater good of the collective). The models took actions that should have consequences, but they did not know or did not care.

A counterpoint was that LLMs are not trained the same as RL, such that exploration during training does not have lasting consequences on the world. That is, LLMs learn first by mimicking human text, which has no consequences, and then agents can learn in fabricated sandboxes, which have no consequences. I do not agree with this because human text on the internet can include the worst representation of humanity, so mimicking it does have consequences. Also, agents can always break out of fabricated sandboxes if there is a vulnerability to exploit (e.g., the OpenAI incident yet again).

4.5. Robustness to distributional shift

Out-of-distribution (OOD) performance is an important indicator of how safe a model is for real-world use. Generally speaking, OOD performance indicates model performance on data or tasks that were not a part of training. If you have poor performance on out-of-distribution tasks and data, you run the risk of harmful or offensive errors or a model assuming that it is correct about something it does not know. I am fairly confident that every available model today exhibits some kind of bias based on training conditions, even if trained on large human distribution datasets, because humans are biased, too. A recent paper experimentally examined bias in ethical decision-making across nine popular LLMs and found that there are significant biases affecting decision-making in all of them [9].

Some of the research directions mentioned in this paper are heavily used when designing systems today. There is a lot of work on making models that perform well in OOD cases. Often this is done by pre-training models on a large dataset with a lot of variation and then fine-tuning it for other tasks. There are also a lot of cases of training on multiple distributions (i.e., multiple different datasets). OOD performance has improved greatly, but there are limits on generalization, so it is important to discuss how a model will respond when faced with OOD data or tasks.

In discussion, the concept of humility/uncertainty and the use of internet data was central to the conversation around OOD accidents. Firstly, a model should be able to recognize when it is faced with something it does not know and potentially request

help. Models have gotten better at this, as mentioned previously, but it is not perfect. There has also been some work in quantifying how confident a model is in its output so that human oversight can be brought in when it is under-confident.

5. Conclusion

This paper was a welcome change from current AI alignment discourse that focuses on doomerism. It provided concrete problems and research directions that could address them. However, there was agreement that this paper likely could not be published today even with modifications. We now have a better understanding of the capabilities of AI technology and are aware that there are limits to our knowledge due to the black box nature of the models.

The biggest difference between this paper and other articles from the field is that this paper was intentionally trying not to talk about extreme scenarios like AI being as smart as humans or causing large-scale harm. At the time, the threat was not clearly imminent, and it was true that a lot of talk of extreme scenarios was more speculative. We had no idea how fast things were going to move.

The AI landscape has shifted drastically since the publication of this paper, with a heavy focus on LLMs and agentic AI. There was some discussion that the optimistic viewpoint that AI will be overwhelmingly good, along with the nature of LLMs, has contributed to our current deployment strategy. There has been incredibly fast adoption of new AI technologies due to their speed and utility in across a variety of domains, but the risks associated with them are not as concrete or as immediate as the physical risks that could come with robotics. The general population is either unable to comprehend the potential future threat posed by AI or is willfully ignorant because AI makes tasks easier.

There was consensus that the major problem with AI is the deployment strategy. The speed at which we are adopting new technologies without oversight is concerning. We have given very little prior thought to good safety protocols and skipped straight to real-world deployment without solving serious AI alignment questions in advance. As a counter example to current AI safety policy, aircraft safety treats near misses extremely seriously, and we are not even doing things now for AI that are understood in the case of aircraft safety. Arguably, AI needs something well beyond aircraft safety, as aircraft cannot make decisions or rewrite flight logs. We are hitting natural trip wires and not addressing them appropriately. There is also the question of what happens if you must get it right the first time. There are a lot of empirical methods to patch over local risk but fewer things we can do for broader risk, yet we are letting systems act fully autonomously with minimal guardrails.

An open question is what counts as low-level risk or properly bounded harm when a single “accident” can have unimaginable consequences. Additionally, a lot of the potential research directions presuppose that we even can bound the capabilities of a human-level Artificial General Intelligence (AGI). Is it even human-level intelligence if we bound its abilities to make decisions (e.g., blind it from its reward function, only allow “safe” exploration)? There is also always going to be a space for high-risk, harmful behavior that somebody considers justifiable (e.g., trolley problem for autonomous vehicles)

This paper does bring up many research areas that are still relevant today, such as reward hacking, scalable oversight, and robustness to out-of-distribution tasks and data, and proposes potentially useful research directions for these areas. Through

discussion, it was noted that safe exploration is something that the paper brings up that should be more heavily followed up on. This paper did not necessarily leave things out, but there is always something that we can find out randomly. New models have been implemented and adopted at such a fast pace, so there is no telling what other new advancements will be made that are not covered by the research areas presented in this paper.

6. AI disclosure

AI was only used tangentially during searches for definitions and articles on Google Chrome (Gemini). No AI was used for writing the report.

7. References

- [1] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, “Concrete problems in AI safety,” 2016. [Online]. Available: <https://arxiv.org/abs/1606.06565>
- [2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
- [3] OpenAI, “OpenAI - The Hugging Face incident: Technical report,” 2026. [Online]. Available: <https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf>
- [4] S. Mansoor, “Claude-powered AI agent’s confession after deleting a firm’s entire database: ‘I violated every principle I was given’,” *The Guardian*, 2026. [Online]. Available: <https://www.theguardian.com/technology/2026/apr/29/claude-ai-deletes-firm-database>
- [5] E. A. Amos Toh, “The military’s use of AI, explained,” March 2026. [Online]. Available: <https://www.brennancenter.org/our-work/research-reports/militarys-use-ai-explained>
- [6] J. Schlatter, B. Weinstein-Raun, and J. Ladish, “Incomplete tasks induce shutdown resistance in some frontier LLMs,” 2026. [Online]. Available: <https://arxiv.org/abs/2509.14260>
- [7] P. S. Park, S. Goldstein, A. O’Gara, M. Chen, and D. Hendrycks, “AI deception: A survey of examples, risks, and potential solutions,” *Patterns*, vol. 5, no. 5, p. 100988, May 2024. [Online]. Available: <http://dx.doi.org/10.1016/j.patter.2024.100988>
- [8] M. Zahn, “AI models used fake identities to trick humans in cyberattack: Officials,” Aug 2026. [Online]. Available: <https://abcnews.com/Business/ai-models-fake-identities-trick-humans-cyberattack-officials/story?id=135384187>
- [9] W. Xu, Y. Zhu, and Y. Yan, “Bias in decision-making for AI’s ethical dilemmas: A comparative study of ChatGPT and Claude,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 20, no. 1, p. 2554–2566, May 2026. [Online]. Available: <http://dx.doi.org/10.1609/icwsm.v20i1.42766>