An overview of AI risks

Preamble

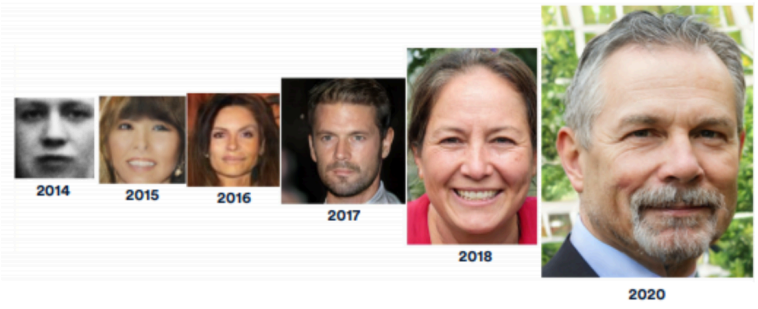
The stakes of artificial intelligence are immense and complex, in particular because of their socio-technical aspects and the possible dual nature of this technology: it is both a source of progress and likely to generate major risks.

This document aims to present a non-exhaustive overview of different sources of AI risk, to help guide future work on the safety, reliability and ethics of AI. For a deeper dive into these risks read this paper.

"Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."
Statement on AI Risk | CAIS, signed by AI scientists and notable figures

Context - The emergence of transformative AI

Generative AI is a rapidly evolving field, in a way that is difficult to predict:



'astronaut riding a horse' © OpenAI

In 2014, the most realistic face generated was the one on the far left. In 2023, one can generate any type of image in any style from a textual description. (source).

The speed of progress in deep learning is remarkable, and even the best experts struggle with short-term predictions of AI capabilities. A simple physics problem exemplifies this: "I put an object on a table, and I push the table. What happens with the object?" Some notable AI scientists stated in January 2022 that solving such a problem would be beyond the abilities of even GPT5000. However, less than a year later, ChatGPT (or GPT 3.5) proved capable of providing an accurate answer.

Games of increasing complexity are mastered to levels beyond human capabilities. First, chess and ATARI games were dominated by AI. Then came Go, where AI first imitated expert

tactics before achieving a level of competitiveness with human experts after only four hours of training. This ability to strategize in unfamiliar environments now extends to Go, Chess, Shogi, and Atari, and all without needing any explicit rules explained. Most recently, AI has even conquered Diplomacy. And it was found that even if Cicero was trained not to deceive, a recent paper showed that he acquired the ability to deceive his opponents.

Substantial progress has been made in the development of sample-efficient learning algorithms. Notably, EfficientZero, a visual Reinforcement Learning (RL) algorithm, is grounded on the sample-efficient model of MuZero. Despite starting from random weights, EfficientZero learns to play Atari faster than humans due to three simple techniques. These allow the model to be incredibly efficient in terms of sampling. With just two hours of interaction with the Atari console, the model learns faster than a median human player, even when considering that humans usually have several years of general skills acquisition before engaging with the console. This low sample complexity and high performance of EfficientZero may usher RL closer to practical real-world applications, like this RL drone system that can do continuous control challenges in the real world and can consistently beat human experts.

These models are not simply stochastic parrots; they demonstrate the ability to perform increasingly general reasoning. Interpretability results from studies like the one on OthelloGPT reveal internal world model representations. Far from simply memorizing all the answers, we can ask the AI to elaborate on its thought process. This is exemplified in the "Let's think step by step" method, also referred to as chain of thought reasoning. Variations of this technique can further enhance performance (e.g. Tree of Thoughts, Reflexion).

AI capabilities now include independent planning. Until very recently, LLMs were not autonomous agents, but methods such as those employed by AutoGPT show the conceptual possibility of converting these LLMs into autonomous planning agents. AutoGPT uses a loop to engage GPT-4 until a particular goal is accomplished, breaking down that goal into smaller tasks. The loop only halts once the goal is achieved. AI Voyager, a Minecraft robot, showcases this by exploring and expanding its abilities in the game's open world. Unlike other robots, it essentially writes and learns continuously by writing its own code and leveraging GPT-4 (summary). Generally capable, autonomous agents continuously explore, plan, and develop new skills in open-ended worlds, driven by survival & curiosity.

It is possible that there are "not many more fundamental innovations needed for Artificial General Intelligence (AGI)" according to the consensus threat model of DeepMind's safety team. OpenAI's superalignment team intends to automate the production of alignment research papers in 4 years.

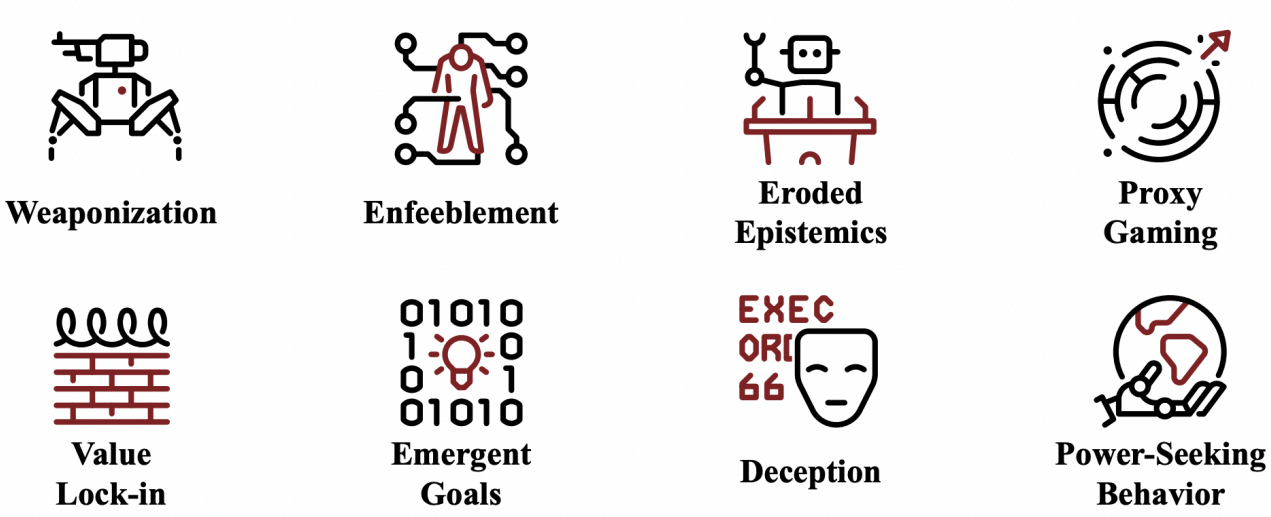
There are still milestones to be reached. As of July 2023, machine learning still has limitations, e.g. self-driving cars have unexpected vulnerabilities, LLMs have yet to achieve successful long-term planning, general ML systems learn at a sluggish pace, and continuous learning is not yet mastered. But in the words of Stuart Russell, we can ponder, "What happens if we succeed?". The primary objective of AI research is to overcome these remaining challenges. If this mission is successful, we must brace ourselves for a future where most human intellectual labor could be fully automated.

Yet, significant vulnerabilities still persist in AI safety and security, as detailed later in this document. It's crucial to address these issues before delving into deep automation of the economy, which would otherwise present extreme risks.

"There is no question that machines will become smarter than humans—in all domains in which humans are smart—in the future," says LeCun. "It's a question of when and how, not a question of if." Yann LeCun, Chief AI scientist at Meta and Turing Prize winner (MIT Tech Review, May 2023)

A classification of AI risks

"Development of superhuman machine intelligence (SMI) is probably the greatest threat to the continued existence of humanity." (Sam Altman's blog, Feb 2015)



Speculative hazards and Failure modes. "*Artificial intelligence (AI) has the potential to greatly improve society, but as with any powerful technology, it comes with heightened risks and responsibilities*". (From Hendrycks et al., 2022)

Risks associated with AI can be categorized according to the responsibilities of different stakeholders:

1. **Malicious and adversarial uses**: Some actors using AI to cause harm, including (cyber)criminals and states.
2. **Accidental issues, loss of control and the alignment problem**: Actors are trying to use AI responsibly, but the science of alignment is imperfect, opening the door to potential accidents.
3. **Systemic issues**: Even when local actors with good intentions work to prevent immediate mishaps, the integration of AI has far-reaching implications. It can disrupt existing equilibria, thereby introducing new risks and problems. This includes the potential for feedback loop risks similar to those seen in the 2010 flash crash.

For each type of risk, we write whether these risks concern current systems or whether these risks are hypothetical ones for future general-purpose systems.

Here's a partial breakdown of issues within each category:

I. Malicious and Adversarial Uses

A. Attacks enabled by AI systems - Risks from giving access to powerful AI models to many actors:

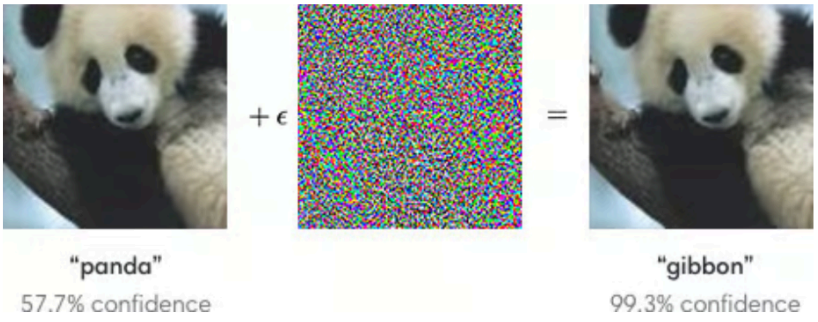
- **Offensive Cybersecurity and Hacking (future):** Current models have the capabilities to scale spear-phishing campaigns. Deception will also reach uncharted territories as deep fakes are becoming increasingly practical (e.g. with fake kidnapping scams). Though they currently lag in terms of planning and autonomous execution compared to other capabilities, language models are likely to enable fully autonomous hacking in the future. See for example WormGPT, a new AI tool for launching offensive cyber attacks.
- **Democratization of dual-use technology (future):** Can large language models democratize access to dual-use biotechnology? (2023) provides a recent example of LLMs assisting untrained users in designing a strategy to synthesize pandemic-scale pathogens. The magnitude of this risk will depend on the prevalence of such dangerous technology. Related work includes The Vulnerable World Hypothesis (2019).
- **Weaponization:** Automation of warfare enables mass automated killing, including targeting specific groups for genocide (see KARGU combat system).
- **Privacy**: there are, broadly speaking, three classes of privacy attacks on machine learning models. *Membership inference attacks* predict whether a particular example was part of the training dataset. *Model inversion attacks* go further by reconstructing fuzzy representations of a subset of the training data. Language models are also prone to *training data extraction attacks*, where verbatim training data sequences can be reconstructed, potentially including sensitive private data.
- **Enabling persistent oppression**: (Value Lock-in) Current AI systems are already capable enough to enable wide-scale surveillance and censorship. Highly competent systems could give small groups of people considerable power, leading to a lock-in of oppressive systems where overcoming the dominant regime might become increasingly unlikely.
- For a complete list of dangerous model capabilities & propensities, Model evals for extreme risks (see Shevlane et al., 2023).

B. Defense flaws of AI systems - The above attacks are made possible because of defense flaws. The pipeline of the current ML paradigm can be attacked at various stages.

- **Data poisoning**: models are currently trained on vast amounts of user-generated data. Attackers can exploit this by modifying some of this data, to influence the behavior of the final models. For example, Poisoning Web-Scale Training Datasets is Practical (2023) details two potential attacks: *split-view poisoning* and *frontrunning poisoning*.
    - **Backdoors**: the black-box nature of modern ML models allows inserting *backdoors*, or *trojans*, into models (including from third-party data poisoning, unbeknownst to the model developers). A Backdoor is a pattern that when present on any images or text, leads to misclassification or bad behavior. Backdoors can be easily placed during training, and are really hard to detect.
    - **Prompt injection:** a recently discovered prevalent attack vector in models trained to follow instructions, by which the absence of robust separation between instructions and data leads to the possibility of hijacking a model's

execution by poisoning the data with instructions. Indirect prompt injection occurs when LLMs query potentially compromised external data, such as websites, on behalf of a user. "Cross Plugin Request Forgery" leverages prompt injection to hijack the tools available to a LLM and call other tools than the ones intended.

- **Adversarial machine learning**: it is feasible to craft special inputs to induce misclassification from ML models. The magnitude of the risk scales with our increasing reliance on models, for instance in self-driving cars, even if partial solutions exist using Lipschitz Network.

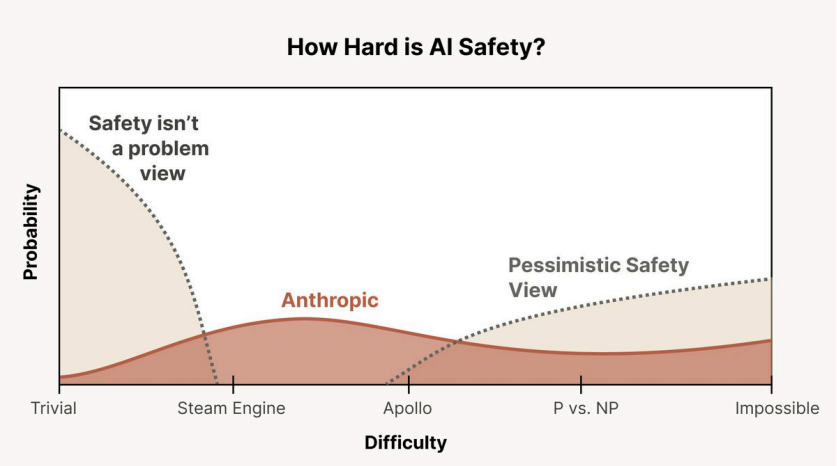


Fooling an image classifier with an adversarial attack (FGSM). Source: OpenAI

  - **"Jailbreaks"**: Even if model developers incorporate security measures for beneficial usage, current architectures may not guarantee that these safeguards won't be easily circumvented. Preliminary results suggest that existing methods are likely not robust enough against attacks. Some work like On the Impossible Safety of Large AI Models (2022) highlights some fundamental limitations to progress on these issues:

II. Accidental problems and loss of control - the alignment problem

"Aligning smarter-than-human AI systems with human values is an open research problem."
Jan Leike, Head of Alignment at OpenAI.



There is great uncertainty about the difficulty of the problem.

Maybe solving the alignment problem is harder than solving P vs NP.
From Anthropic's Core Views on AI Safety (Anthropic is one of the leading labs).

According to DeepMind's AGI safety team's literature review, most threat models involving a loss of control over AI models stem from the following two fundamental flaws:

- **Specification gaming**: Correctly specifying the goals of an AI system has proven to be a challenging task, even in simple, self-contained environments such as video games. Specification gaming refers to the phenomenon where an AI system satisfies the goal it was given, but in an unexpected way, revealing a mismatch between the implemented specification and the specification the model creators had in mind. Dozens of examples are listed in this document. As we hand more control and autonomy to AI systems, this failure mode could become a significant risk.
    - **Proxy Gaming**: Trained with defective goals, AI systems could find new ways of pursuing their objectives at the expense of individual and social values. AI systems are trained using measurable objectives, which may be only indirect proxies for what we value. For instance, **AI recommendation systems** are trained to maximize viewing time and click-through rates. However, the content people are most likely to click on is not necessarily the same as the content that will improve their well-being (Kross et al., 2013). Furthermore, some evidence suggests that recommendation systems lead people to develop extreme beliefs in order to make their preferences easier to predict (Jiang et al., 2019). As AI systems become more capable and influential, the goals we use to train them must be specified with greater care and incorporate shared human values. [more]. Note that proxy gaming can also become a systemic issue. See What failure looks like (Part 1).
- **Lack of robustness in learned objectives (goal misgeneralization):** Even with a correct specification of the objective, there are often multiple policies which perform well on the objective in the training environment, but which might be revealed as very different from each other in an out-of-distribution environment, such as in deployment. A toy example is CoinRun, a simple game where the coin to collect is always at the end of the level. It turns out that the Reinforcement Learning setup cannot ensure that the correct goal (collecting the coin) is learned rather than another compatible goal (going to the end of the level). As AI systems get more advanced, some policies might arise which would perform well against the specified goal in the training environment, but turn out to be undesirable once deployed in the real world. Some examples include:
    - **Deception (future):** deception can be found in human data, and can be useful in a wide range of settings. It may be more efficient to gain human approval through deception than to earn human approval legitimately. Deception also could provide systems that have the capacity to be deceptive a strategic advantage over honest models. Deceptive alignment refers to a hypothesized scenario where a sufficiently situationally aware misaligned model would appear aligned during training and early deployment, in order to be deployed on a wide scale, and then pivot to the pursuit of other objectives once it can do so without the risk of shutdown. (more)

As a consequence of specification gaming or lack of robustness in learned objectives, we can get new types of risks:

- **The shutdown, or corrigibility, problem (future)**: refers to the simple observation that "you can't fetch the coffee if you're dead" from Stuart Russell, meaning that being shut down scores very poorly on typical policies we might want from an agent. Corrigibility remains an open research problem.
- **Power-seeking behavior (future):** power-seeking could be the learned goal of an AI system instead of the goals the developers tried to instill in the model, as a policy that seeks power could score well in many training environments (especially in the context of deceptive alignment). Power-seeking is also a *convergent instrumental goal*, meaning that it is useful for accomplishing a wide range of objectives and is therefore likely to arise in advanced agents, making them harder to control.

These risks are made more acute by:

- The **black-box nature** of advanced ML systems. Our understanding of how AI systems behave, what goals they pursue, and our understanding of their internal behaviors lags far behind the capabilities they exhibit. The field of **interpretability** aims to make progress on this front, but remains very limited.
- **Emergent goals.** As models become more proficient, they sometimes exhibit unexpected and qualitatively different behaviors. The sudden emergence of capabilities or goals could heighten the Emergent Abilities of Large Language Models, and risks of humans losing control over advanced AI systems.

III. Systemic issues

- **Bias**: Biases within Large Language Models persist, often reflecting the opinions and biases propagated on the internet (as seen with the biased trends of some LLMs). These biases can be harmful in a myriad of ways, as exemplified by studies on GPT-3's Islamophobic biases. For more details, see
    - Ethical and social risks of harm from Language Models, which outline six specific risk areas: Discrimination, Exclusion and Toxicity, Information Hazards, Misinformation, Malicious Uses, Human-Computer Interaction Harms, Automation, Access, and Environmental Harms.
    - Evaluating the Social Impact of Generative AI Systems in Systems and Society which define seven categories of social impact: bias, stereotypes, and representational harms; cultural values and sensitive content; disparate performance; privacy and data protection; financial costs; environmental costs; and data and content moderation labor costs.
- **Economic upheaval**: The widespread consequences on the labor market resulting from the automation of the economy (see this OpenAI report) could amplify economic inequalities and social divisions. With mass unemployment as a likely byproduct, it could also lead to mental health problems by rendering human labor increasingly obsolete.
- **Enfeeblement**: can occur if humans delegate increasingly important tasks to machines; in this situation, humanity loses the ability to self-govern and becomes completely dependent on machines. [more]
- **Fragility of complex systems**: As different parts of a system are automated and tightly coupled, the failure of one component may trigger the collapse of the rest of the system. Some research avenues aim to study the characteristics of such systems to anticipate the consequences of greater automation of the economy.

- **Multi-agent settings:** New problems arise in multi-polar scenarios. Robust Agent-Agnostic Processes (RAAPs). eg. financial markets, bots colluding, misalignment of the system level objective.

"*The future is going to be good for the AIs regardless; it would be nice if it would be good for humans as well*" Open AI Chief Scientist Ilya Sutskever (Human, Nov 2019)

"*There's a long tail of things of varying degrees of badness that could happen. I think at the extreme end is the Nick Bostrom style of fear that an AGI could destroy humanity. I can't see any reason in principle why that couldn't happen.*"
Anthropic CEO Dario Amodei (previously OpenAI VP of Research) (80,000 Hours, July 2017)

A Socio-Technical Problem

The issue of AI safety is multidisciplinary, and the solution *must* be holistic. AI ethics, AI alignment, and AI governance must work hand in hand:

- **AI ethics asks which values we can incorporate in these complex systems**
- **AI alignment asks how to control autonomous systems, regardless of the value the operator wants the systems to follow.**
- **AI governance asks how to adopt the solutions at societal levels.[1]**

Different explanations lead to the conclusion that AI could be an existential risk:

- Natural Selection Favors AIs over Humans
- How harmful AIs could appear - Yoshua Bengio
- Is Power-Seeking AI an Existential Risk?
- AGI Ruin: A List of Lethalities - AI Alignment Forum
- The alignment problem from a deep learning perspective
- Other scenarios are given in Threat Model Literature Review, by DeepMind.

In all these scenarios, international coordination and governance would have a significant influence, which is why we talk about AI governance in the following document, with a focus on what technical contributions can bring to this field.

EffiSciences' AI Safety work

If you want to learn more about AI safety research, you can explore our website Pole IA - EffiSciences. We organize various activities aimed at raising awareness, training, and mentoring students in general-purpose artificial intelligence security, and courses in AI safety taught at the Ecoles Normales Supérieures in Ulm and Paris-Saclay, accredited and updated every year.

[1]And slowing down the development of powerful AIs so that we have a higher chance of getting enough time to do the research required for the solutions.

More resources

Other resources:

- EffiSciences, and our page "Our Vision"
- The AI safety newsletter (by the Center for AI Safety), which is probably one of the best introductory resources.
- TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI (Critch, 2023)
- An Overview of Catastrophic AI Risks (Hendricks, 2023)

On YouTube:

- AI safety training day (Video of the course – EffiSciences)
- Introduction to ML Safety (course – Center for AI Safety)
- AI Explained - YouTube
- Robert Miles AI Safety, and in particular, videos presenting specification gaming and goal misgeneralization,