July 2023 · Pegah Maham and Sabrina Küspert

Governing General Purpose AI

A Comprehensive Map of Unreliability, Misuse and Systemic Risks

Stiftung Neue Verantwortung

Think Tank at the Intersection of Technology and Society

Executive Summary

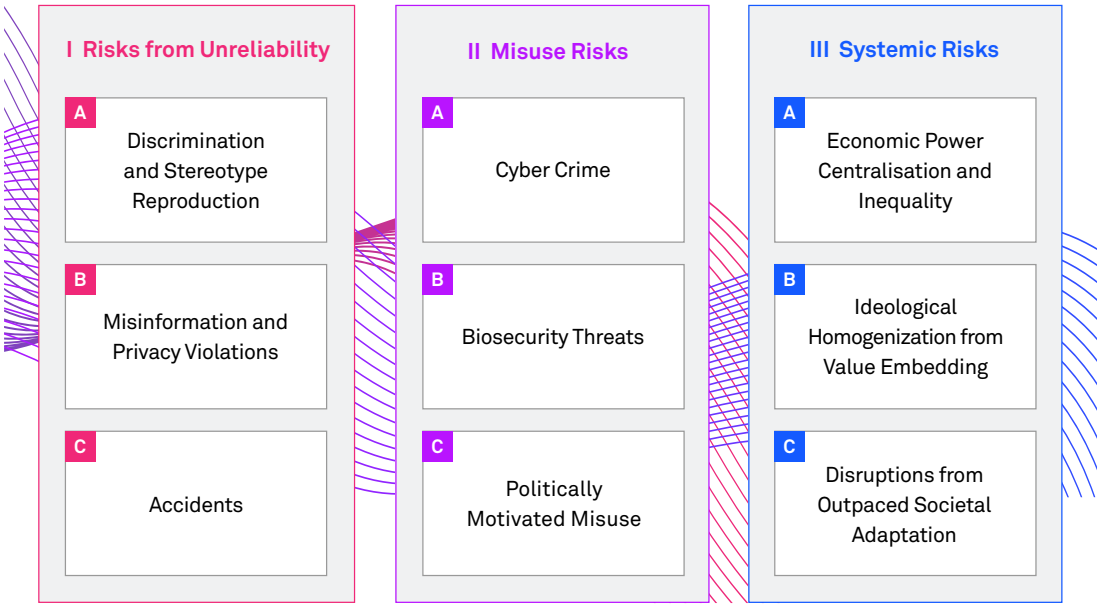
The recent months have been marked by a heated debate about the risks and benefits of increasingly advanced general purpose AI models and generative AI applications building on them. Although many stress the huge economic potential of these models, a variety of incidents ranging from an AI-generated livestream filled with transphobia and hate speech to an experiment with a general purpose AI-based agent that was given the aim to "destroy humanity", and fears about disruptions to our education system have led to escalating concerns about the risks stemming from these models. While only a few well-resourced actors worldwide have released general purpose AI models, hundreds of millions of end-users already use these models, further scaled by potentially thousands of applications building on them across a variety of sectors, ranging from education and healthcare to media and finance.

The diversity of potential risks of these models, combined with their rapid advancement and integration into everyday life, have provoked policy interventions around the world – including in the EU, US and UK, in transatlantic dialogues, and among the G7. Amongst these domestic and multilateral activities, the European Union has so far gone the furthest. There is already a strong political will to establish the most effective rules on general purpose AI providers in the EU AI Act, one of the first legal frameworks on AI.

While a strong EU AI Act is essential to comprehensively address the many risks stemming from general purpose AI models, the sheer diversity, scale and unpredictability of hazards demand additional policy actions. The pioneering legislation of the EU AI Act represents an essential cornerstone in comprehensively governing general purpose AI models, by putting direct rules for these models in place. Through the Brussels effect, it is possible for these rules to spread to other jurisdictions. Given the diversity of risks, however, additional policy action is needed. This could include, for example, education programmes for decision-makers and the general public, redistributive policies, industrial policy for trustworthy AI, funding for AI ethics and safety research, and international agreements considering the global impact of this technology.

Policymakers should get a comprehensive understanding of the whole range of risks associated with general purpose AI models, to proactively mitigate these hazards. This report maps these potential risks across three categories: Risks from Unreliability, Misuse, and Systemic Risks. Illustrated with currently observable examples and relevant scenarios, we outline a comprehensive set of nine relevant risks across three primary risk categories. This risk map provides a structured resource for policymakers seeking to understand the multifaceted

challenges of general purpose AI and their potentially far-reaching impact to effectively govern them. It is critical for policymakers to identify risks stemming from this rapidly-advancing technology, weigh their implications, prioritise and address them adequately, ensuring that all risks are covered.



Firstly, Risks from Unreliability arise as there is currently no solution to ensure that general purpose AI models behave as intended, to predict and control their behaviour fully. This gives rise to risks of Discrimination and Stereotype Reproduction, Misinformation and Privacy Violations, and Accidents. General purpose AI models make decisions based on complex internal mechanisms that are not yet understandable, even to their developers. This results in a lack of reliability, transparency, controllability, and other key features of trustworthiness. In the case of agentic models, it makes it challenging to ensure that the models pursue goals that align with human objectives and values. Therefore, firstly, models risk discrimination and the reproduction of stereotypes by exhibiting or amplifying biases present in their training data. Secondly, models can disseminate false or misleading information, omit critical information, or produce true information that violates privacy. Lastly, these models pose risks of accidents from unexpected failures during development or deployment, which could scale with advancing capabilities and agency as well as wider integration of models, leading to concerns over catastrophic or even existential risks.

Secondly, general purpose AI models are inherently dual-use, meaning that they can serve both beneficial and harmful purposes, making them susceptible to misuse by malicious actors. This could increase speed, scale, and sophistication, for example, in Cyber Crime, Biosecurity Threats, and Politically Motivated Misuse.

Actors seeking to misuse these tools could do so without building their own advanced models, but instead by using models without appropriate safeguards or bypassing them, by leveraging available open-source models, or by using models leaked or stolen from AI labs. Firstly, general purpose AI models could make cyber crimes leveraging IT systems, such as fraud, more sophisticated and convincing, and could also be used to target IT systems, for example, through phishing emails or assisting in programming malicious software. Secondly, general purpose AI models could facilitate the production and proliferation of biological weapons, by making critical knowledge more accessible and reducing the barrier for misuse. Lastly, if misused with political motivations, these models could exacerbate surveillance efforts or existing tactics for political destabilisation, such as disinformation campaigns.

Thirdly, further Systemic Risks arise from the centralisation of general purpose AI development and the rapid integration of these models into our lives. This risks Economic Power Centralisation and Inequality, Ideological Homogenization from Value Embedding, and Disruptions from Outpaced Societal Adaptation. General purpose AI models become increasingly integrated into public and private infrastructure as the foundation for further applications and systems, yet they are almost exclusively developed by a few companies, dominated by Big Tech and their investees. Firstly, this risks that economic power is increasingly centralised amongst a few actors with a certain level of control over access to this technology and its economic benefits, possibly feeding into inequality within and between countries. Secondly, as developers inscribe certain values and principles into a general purpose AI model, this risks centralization of ideological power, producing models that are not fit to adapt to evolving and diverse social views or that create echo chambers. Lastly, overly rapid adoption of this technology at scale might outpace the ability of society to adapt effectively, leading to a variety of disruptions, including challenges in the labour market, the education system and public discourse, and various mental health concerns.

The risk profile of general purpose AI models is changing as capabilities advance and scale of deployment increases. At the same time, the models carry a few characteristics that pose distinct challenges in governing them. There is currently no solution to ensure that general purpose AI models robustly behave as intended. Advanced model capabilities imply that these models can be used for ever more complex tasks and operate in a wider range of contexts, often advancing the current state of the art which is less well-understood with more possibilities to cause harm. AI applications that are based on a general purpose AI model often inherit the risks that originate in design and development of the underlying model. As these models are integrated into an increasing number of applications across a variety of sectors, shortcomings entailed in one model could be scaled

to thousands of downstream systems worldwide. Increasingly built with greater agency, they can be deployed more autonomously in more complex tasks and environments, seemingly requiring less human oversight. Even for experts in the field, the pace of progress is surprising.

The European Union has a unique opportunity to mitigate risks stemming from general purpose AI models and establish themselves as global leaders in guiding responsible and safe development and deployment of this fast-evolving technology — with a strong EU AI Act and beyond.

Table of Contents

Introduction

The recent months have been marked by an escalating concern about increasingly advanced general purpose AI models and generative AI applications building on them. General purpose AI models, sometimes referred to as foundation models, are understood as AI models that are designed for generality of their output and have a wide range of possible applications. They are not only already able to understand and create text, images and other output, exhibit human-level performance on some professional and academic benchmarks, write code in various programming languages[1], or understand and explain human jokes (see Figure 1). Increasingly advanced models have also been shown, for example, to be capable of negotiating and cooperating with people,[2] or planning and executing scientific experiments[3]. Such capabilities facilitate their widespread deployment in potentially thousands[4] of applications including generative AI, which is projected to have a trillion dollar contribution yearly to the global economy across various sectors and tasks.[5] Current examples of use cases range from software development assistance[6], visual assistance for blind people[7], and personalised language learning[8], to fraud detection on online platforms for financial services[9], providing advice on health and nutrition[10], and evaluating and drafting legal contracts[11]. However, concerns have been escalating about possible risks stemming from these models, due to incidents such as an AI-generated livestream filled with transphobia and hate speech[12] or the usage of false legal documents generated by OpenAI's ChatGPT in a courtroom[13], increasing misuse of such models, for example, for a deepfake video of Ukrainian president Zelenskyy[14] or through a general purpose AI based agent attempting to execute complex plans given the aim to "destroy humanity"[15],

1 pp. 6, 8-9, OpenAI. (2023). *GPT-4 Technical Report.* OpenAI.
2 Bakhtin, A. et al. (2022). *CICERO: An AI agent that negotiates, persuades, and cooperates with people.* Meta AI.
3 p. 12, Boiko, D. A., MacKnight, R. and Gomes, G. (2023). *Emergent autonomous scientific research capabilities of large language models.*
4 As of July 2023, one available market map has already identified over 800 applications built on general purpose AI models, showing significant growth from 300 in January the same year, see *GPT-3 DEMO Real-time Market Map*. Given this trajectory, it is plausible that the total number of applications built on general purpose AI models has already exceeded a thousand.
5 Chui, M. et al. (2023). *The economic potential of generative AI.* McKinsey & Company.
6 Dohmke, T. (2023). *GitHub Copilot X: The AI-powered developer experience.* Github Blog. The coding assistant GitHub Copilot X is based on OpenAI's GPT-4.
7 Be my eyes. *Introducing Our Virtual Volunteer Tool Powered by OpenAI's GPT-4.*
8 OpenAI. (2023). *Duolingo.* OpenAI. The language learning app Duolingo is using OpenAI's GPT-4.
9 OpenAI. (2023). *Stripe.* OpenAI. Stripe is using OpenAI's GPT-4.
10 *Plenny Pal.*; This health and nutrition assistant is built upon Meta's LLama and OpenAI's GPT-4.
11 Anthropic. (2023). *Introducing Claude.* Anthropic.; The legal infrastructure business Robin AI uses Anthropic's Claude.
12 Oladipo, G. (2023). *AI-generated Seinfeld parody banned on Twitch over transphobic standup bit.* The Guardian.
13 Weiser, B. and Schweber, N. (2023). *The ChatGPT Lawyer Explains Himself.* The New York Times.
14 pp. 84, 134, Maslej, N. et al. (2023). *The AI Index 2023 Annual Report.* Index Steering Committee, Institute for Human-Centered AI, Stanford University.
15 Koebler, J. (2023). *Someone Asked an Autonomous AI to 'Destroy Humanity': This Is What Happened.* Vice.

as well as far-reaching worries about job displacement[16], disruptions to our education system[17] or adverse effects on global inequality[18].

Figure 1

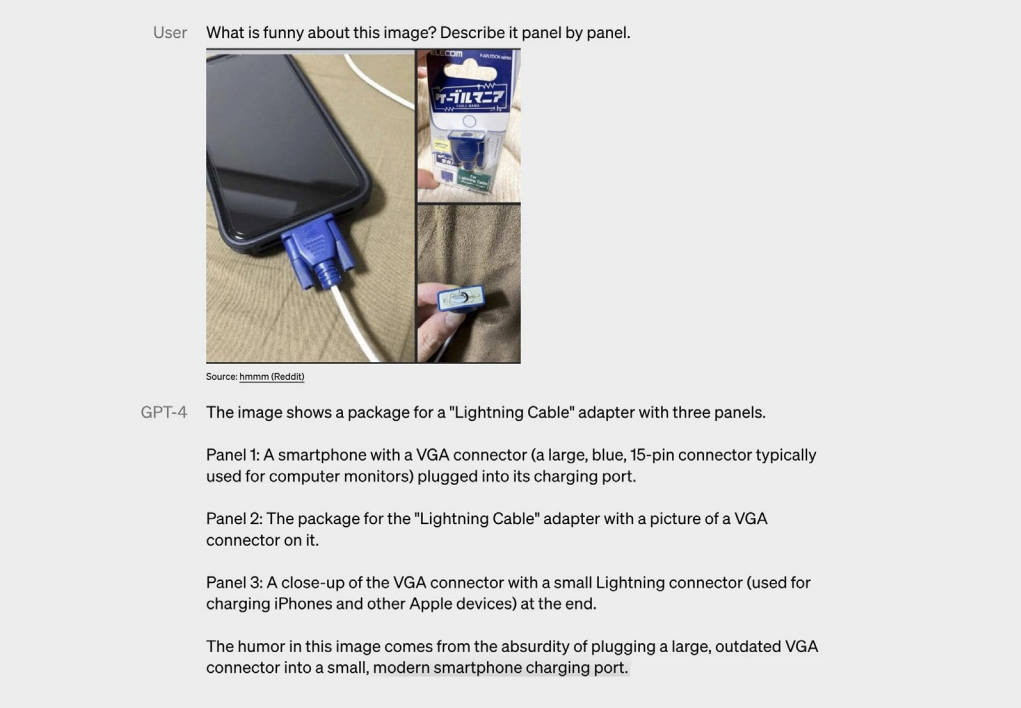


Figure 1: Example of the capabilities of OpenAI's GPT-4 model to process and analyse multimodal content, in this case, images and text, and understand context, cultural nuances and subtle cues.[19]

The rapid advancement and integration of general purpose AI models have provoked policy debates and initial interventions around the world, emphasising the urgent need to effectively govern this technology. For example, EU institutions and Member States are debating how to address general purpose AI in the EU AI Act and other policy initiatives[20], the UK[21] established a taskforce on foundation models[22], and the U.S. Whitehouse discussed opportunities and risks with CEOs

16 See Benbya, H., Davenport, T. H. and Pachidi, S. (2020). *Artificial Intelligence in Organizations: Current State and Future Opportunities.* MIS Quarterly Executive.
17 European University Association. (2023). *Artificial intelligence tools and their responsible use in higher education learning and teaching.*
18 See for example Atkinson, R. D. and Wu, J. (2017). *False Alarmism: Technological Disruption and the U.S. Labor Market, 1850–2015.* Information Technology & Innovation Foundation.; Littman, M. L. et al. (2021). *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report.* Stanford University.
19 Image retrieved from p. 9, OpenAI. (2023). *GPT-4 Technical Report.* OpenAI.
20 Gibson Dunn. (2023). *European Parliament Adopts Its Negotiating Position on the EU AI Act.*; *European Parliament. (2023). MEPs ready to negotiate first-ever rules for safe and transparent AI.*
21 Secretary of State for Science, Innovation and Technology. (2023). *Policy Paper.*
22 Department for Science, Innovation and Technology et al. (2023). *Initial £100 million for expert taskforce to help UK build and adopt next generation of safe AI.*

of companies developing these models[23]. Transatlantic dialogues between the EU and the U.S. have led to the proposal of a Code of Conduct with voluntary rules targeting generative AI to bridge the time until legislation would come into effect.[24] The G7 as an international forum established a process to further analyse the impact of this technology.[25] These domestic and multilateral engagements show that there is increased awareness about the risks of this technology and the need to effectively govern it.

There is already a strong political will to effectively regulate general purpose AI models in the EU AI Act. The European Union has a unique opportunity to ensure that this pioneering legislation is as comprehensive as possible in addressing risks posed by this fast-evolving technology. While the original draft of the AI Act[26] as one of the first legal frameworks on AI by the European Commission in 2021 focused on regulating AI systems with intended purpose, the positions of both the European Council[27] at the end of 2022 and the European Parliament[28] in June 2023 proposed that direct rules should apply to general purpose AI models as well. Members of European Parliament leading the work on the AI Act have acknowledged that "the speed of technological progress is faster and more unpredictable than policymakers around the world have anticipated" and stated "the need for significant political attention" on general purpose AI models.[29] As development and deployment of these models only lately entered the focus of policymakers, the coming months of finalising the AI Act will be crucial for the EU institutions to establish the most effective rules on general purpose AI providers to address the variety of risks posed by this technology.

However, while a strong EU AI Act is essential to comprehensively address the many risks stemming from general purpose AI models, the sheer diversity, scale and unpredictability of hazards require additional policy actions. The AI Act represents an essential cornerstone in comprehensively governing general purpose AI models, in putting direct rules for these models in place. The European Parliament already suggested, amongst other measures, that providers of general purpose AI models should "demonstrate [...] the identification, the reduction and mitigation of reasonably foreseeable risks to health, safety, fundamental rights, the environment and democracy and the rule of law prior and throughout

23 The White House. (2023). *FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation that Protects Americans' Rights and Safety.*
24 Scott, M., Chatterjee, M. and Volpicelli, G. (2023). *The struggle to control AI.* Politico.
25 The G7 Digital and Tech Ministers. (2023). *Ministerial Declaration.*
26 European Commission (2023). *Regulatory framework proposal on artificial intelligence*.
27 European Council (2022). *Artificial Intelligence Act: Council calls for promoting safe AI that respects fundamental rights.*
28 European Parliament. (2023). *MEPs ready to negotiate first-ever rules for safe and transparent AI.* The European Parliament refers to "foundation models" for what we describe in this report as general purpose AI models.
29 Tudorache, D. [@IoanDragosT]. (2023). *"AI is moving very fast and we need to move too."* Twitter.

development", while an AI Office should ensure a future-proof approach and "issue an annual report on the state of play in the development, proliferation, and use of foundation models alongside policy options to address risks".[30] Through the Brussels effect, it is possible for these rules to spread to other jurisdictions.[31] To address the risks as timely as possible, voluntary commitments have already been proposed to bridge the time until the AI Act applies in two to three years.[32] Given the diversity of risks, however, additional policy action is needed. This could include, for example, education programmes for decision-makers and the general public, redistributive policies, industrial policy for trustworthy AI, funding for AI ethics and safety research, and international agreements considering the global impact of this technology. The upcoming 2024-2029 term of the European Commission is one unique opportunity for the EU to set a strategic focus on this fast-evolving technology while member states and international forums can complement this approach to ensure that general purpose AI models are developed and integrated responsibly and safely.

To proactively navigate the impact of general purpose AI, EU institutions and governments in Member States should understand and analyse the wide range of potential risks to decide which set of policy initiatives can address them comprehensively. Recognising the rapid advancement and integration of general purpose AI models, a well-informed perspective on the risks associated with general purpose AI becomes an imperative task for governing institutions. It is critical to identify risks, weigh their implications, prioritise and address them adequately, ensuring that all risks are covered. Given that the technology and its widespread integration is constantly advancing and will likely continue to do so, it is important to avoid overfitting to the concerns of today[33] but rather exercise sufficient foresight.

This report maps potential risks of general purpose AI models across Risks from Unreliability, Misuse, and Systemic Risks. It gives policymakers a holistic understanding to proactively mitigate these hazards. Illustrated with currently observable examples and relevant scenarios, we outline a comprehensive set of nine relevant risks, divided into three primary risk categories: I. Risks from Unreliability, II. Misuse Risks and III. Systemic Risks. This risk map forms a solid foundation to effectively govern general purpose AI models and their potentially far-reaching impact. Importantly, though, the field of general purpose AI is evolving

30 European Parliament. (2023). *P9_TA(2023)0236 Artificial Intelligence Act. Amendments adopted by the European Parliament on 14 June 2023.*
31 Siegmann, C. and Anderljung, M. (2022). *The Brussels Effect and Artificial Intelligence.* APSA Preprints.
32 Bertuzzi, L. (2023). *EU leaders race over outreach initiatives to anticipate AI rules.* Euractiv.
33 p. 1, Kak, A. and West, S. M. (2023). *General purpose AI poses serious risks, should not be excluded from the EU's AI Act | Policy Brief.* AI Now Institute.

rapidly, possibly with unforeseen advancements, and experts often disagree about certain aspects. Therefore, predictions about its trends cannot come with complete certainty, yet they are necessary in iteratively shaping a proactive and informed approach to governance of this technology.

What are general purpose AI models?

General purpose AI models, sometimes referred to as foundation models[34], are understood as AI models that are designed for generality of their output and have a wide range of possible applications. While these models can be used in standalone systems, they are often used as the "building block" of potentially thousands[35] of single-purpose AI systems to accomplish a range of distinct tasks across a variety of sectors.[36] For example, OpenAI's GPT-4 is the general purpose AI model that powers the user-facing system ChatGPT as well as numerous other applications built by third-party developers upon GPT-4[37]. These applications already span from software development assistance[38] and personalised language learning[39] to evaluating and drafting legal contracts[40] and visual assistance for blind people[41]. General purpose AI models are designed for generality of their output, requiring vast amounts of data and compute, combined with relevant expertise. As a result, they are characterised by a wider range of capabilities than that of other AI models, including capabilities they were not explicitly designed for, and can therefore be re-used in numerous downstream applications.

It is important to note that a few terms are being used to describe these models, emphasising different aspects, without precise boundaries.[42] "General purpose AI models" highlights that the models are designed for generality of their output across many possible applications, in contrast to AI systems with intended purpose. "Foundation models"[43] centres around the notion that these models are a base on which other more specific AI systems can be built. "Generative AI" does not only describe these models but also includes the many applications[44] building on them, simply recognising the potential to generate content[45].

34 p. 3, Bommasani, R. et al. (2020). *On the Opportunities and Risks of Foundation Models.*
35 As of July 2023, one available market map has already identified over 800 applications built on general purpose AI models, showing significant growth from 300 in January the same year, see *GPT-3 DEMO Real-time Market Map*. Given this trajectory, it is plausible that the total number of applications built on general purpose AI models has already exceeded a thousand.
36 Küspert, S., Moës, N. and Dunlop, C. (2023). *The value chain of general purpose AI.* Ada Lovelace Institute.
37 OpenAI. (2023). *GPT-4.* OpenAI.
38 Dohmke, T. (2023). *GitHub Copilot X: The AI-powered developer experience.* Github Blog. The coding assistant GitHub Copilot X is based on OpenAI's GPT-4.
39 OpenAI. (2023). *Duolingo.* OpenAI. The language learning app Duolingo is using OpenAI's GPT-4.
40 Anthropic. (2023). *Introducing Claude.* Anthropic.; The legal infrastructure business Robin AI uses Anthropic's Claude.
41 Be my eyes. *Introducing Our Virtual Volunteer Tool Powered by OpenAI's GPT-4.*
42 Toner, H. (2023). *What Are Generative AI, Large Language Models, and Foundation Models?.* Center for Security and Emerging Technology.
43 p. 3, Bommasani, R. et al. (2020). *On the Opportunities and Risks of Foundation Models.*
44 Chui, M. et al. (2023). *The economic potential of generative AI.* McKinsey & Company.
45 Toner, H. (2023). *What Are Generative AI, Large Language Models, and Foundation Models?.* Center for Security and Emerging Technology.

Only a few well-resourced actors worldwide have released general purpose AI models, yet hundreds of millions of end-users directly access these models, further scaled by "tens of thousands of developers around the globe"[46] that are already building potentially thousands of applications on them across a variety of sectors, ranging from education and healthcare to media and finance. Developing general purpose AI models requires significant computing power, data, and talent.[47] Consequently, the landscape of general purpose AI developers is currently dominated by a few well-resourced actors[48], often with strategic partnerships. This includes Meta, Microsoft and its partner OpenAI[49], Alphabet with Google DeepMind[50] and its investee Anthropic[51], as well as the open-source actor Stability AI in collaboration with Amazon's AWS[52], expanding model capabilities beyond the current state of the art through scaling model size or using more efficient training regimes, architectures, and algorithms. A few others such as Adept AI, Aleph Alpha, AI21, Cohere, EleutherAI or BigScience/Bloom train models for specific target groups, or innovate in model reliability, efficiency, or accessibility. Yet the total number of general purpose AI developers remains small, and relatively centralised in the US. This is in stark contrast to the scale of deployment. As the option to directly access OpenAI's models, ChatGPT reached 100 million monthly active users two months after launch.[53] Once Microsoft's Bing search engine was powered by OpenAI's GPT-4, its daily users surpassed the same count.[54] The reach of these models is further amplified by the potentially thousands of applications built by start-ups, SMEs and companies around the world on them. As there exist only a small set of models, this creates complex dependencies where the few developers of general purpose AI models control access to and support for their models.[55] Many small and medium sized enterprises complain about a lack of access to models as a major restrictive factor.[56]

General purpose AI models carry a few characteristics that distinguish them from other AI technology, posing distinct challenges in governing them. Understanding these characteristics is important to understand the underlying causes of risks, their interplay, and opportunities and challenges in mitigating them.

46 Pilipiszyn, A. (2021). *GPT-3 powers the next generation of apps.* OpenAI.
47 p. 151, Bommasani, R. et al. (2020). *On the Opportunities and Risks of Foundation Models.*; pp. 61-63, Maslej, N. et al. (2023). *The AI Index 2023 Annual Report.* Index Steering Committee, Institute for Human-Centered AI, Stanford University.
48 p. 7, Kak, A. and West, S. M. (2023). *General purpose AI poses serious risks, should not be excluded from the EU's AI Act | Policy Brief.* AI Now Institute.
49 Capoot, A. (2023). *Microsoft announces new multibillion-dollar investment in ChatGPT-maker OpenAI.* CNBC.
50 Hassabis, D. (2023). *Announcing Google DeepMind.* Google DeepMind.
51 Field, H. (2023). *Ex-OpenAI execs raise $450 million for Anthropic, a rival A.I. venture backed by Google.* CNCB.
52 Stability AI. (2023). *Stability AI makes its Stable Diffusion models available on Amazon's new Bedrock service.* Stability AI.
53 Hu, K. (2023). *ChatGPT sets record for fastest-growing user base - analyst note*. Reuters.
54 Warren, T. (2023). *Microsoft Bing hits 100 million active users in bid to grab share from Google*. The Verge.
55 Küspert, S., Moës, N. and Dunlop, C. (2023). *The value chain of general purpose AI.* Ada Lovelace Institute.
56 p. 70, Akademie für Künstliche Intelligenz (2023). *Large AI Models for Germany.* KI Bundesverband e.V..

General purpose AI models are not reliable and lack transparency. This makes it challenging to predict and control their behaviour robustly – or to explain it. Often known as 'black boxes,' general purpose AI models consist of neural networks that are trained through a method known as deep learning which differ from other more interpretable (but not all) AI models such as regression trees. Their internal operations are difficult to interpret, making these models hard to control or predict robustly, even for their developers.[57]

Assessing capabilities of these models is not straightforward as some abilities only become visible after retraining on specific data sets, experimentation, or in combination with other tools. Firstly, retraining models on specific data ("fine-tuning") has shown notable improvements, for example in mathematical problem-solving[58] or medical question answering[59]. Secondly, using effective input texts, so-called "prompts", or providing examples for a task, can improve a model's capabilities significantly. For instance, by "prompting" a model to reason step by step, it can solve mathematical problems or commonsense reasoning tasks that it previously could not. [60] Lastly, capabilities of models can be enhanced by giving the models access to tools like databases, browsers, programming environments, or other APIs.[61]

General purpose AI models which are increasingly built with greater agency could be deployed in more complex tasks and environments, which risks that human oversight is being reduced. Agency in an algorithmic system includes longer-term planning abilities, less specifications on how to achieve goals and decreased human oversight over intermediate steps.[62] It can be developed, for example, by systems built around a general purpose AI model or by using "reinforcement learning" methods. Recently developed AI models show increasingly agentic capabilities, such as playing the complex strategy game Diplomacy at a human level.[63] The open-source AutoGPT, which builds on OpenAI's GPT-4, combines step-by-step reasoning with external tools and memory, being adapted to act increasingly autonomously. While currently still significantly limited, it shows first attempts to iteratively propose and refine a plan, and execute its steps without human intervention.[64] Increasingly agentic models could exacerbate potential

57 p. 35, OECD. (2023). *AI Language Models. Technological, socio-economic and policy considerations.* OECD Publishing.
58 pp. 7-8, Lewkowycz, A. et a. (2022). *Solving Quantitative Reasoning Problems with Language Models.* Conference on Neural Information Processing Systems 2022.
59 pp. 8-10, Singhal, K. et al. (2023). *Towards Expert-Level Medical Question Answering with Large Language Models.*
60 pp. 4, 7, 8, Wei, J. et al. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.* Google Research.
61 OpenAI. (2023). *ChatGPT plugins.* OpenAI.
62 p. 4, Chan, A. et al. (2023). *Harms from Increasingly Agentic Algorithmic Systems.* Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.
63 Meta Fundamental AI Research Diplomacy Team et al. (2022). *Human-level play in the game of Diplomacy by combining language models with strategic reasoning.* Science.
64 Larsen, L. (2023). *What is Auto-GPT? Here's how autonomous AI agents are taking over the internet. Digital Trends.*

risks, for example, if implemented in high-stake areas while not fully reliable, especially when human oversight is mistakenly reduced.[65]

AI applications that are based on a general purpose AI models retain core structural elements of that underlying model, inheriting the risks that originate in design and development of the model. A singular general purpose AI model could serve as the foundation for several downstream applications, that are re-using or slightly modifying the underlying model through processes like fine-tuning. However, these downstream models often keep structural dependencies from the original general purpose AI model and original developers. This prevents downstream developers from being able to effectively manage risks associated with these dependencies.[66] As these models are integrated into an increasing number of applications across a variety of sectors, shortcomings entailed in one general purpose AI model could affect thousands of downstream applications worldwide.

The pace of progress for general purpose AI models is surprising, even to many experts, making it difficult to confidently predict their impact or risks. In a book about the history of AI, published in early 2021, the author, a Professor of Computer Science at Oxford University, described AI model abilities of "understanding a story & answering questions about it" and "writing interesting stories" as "nowhere near solved"[67] – which was proven wrong only around two years later. In 2023, experts of McKinsey's Global Institute have changed various of their 2017 predictions on when performance on certain skills will be achieved by AI significantly, from decades to a few years.[68] On the question what accuracy on a data set on mathematical problems AI will achieve by mid-2022, professional forecasters in 2021 predicted 13%, when it turned out to be 50%.[69] With these examples, it becomes clear that the advancement of general purpose AI models is currently at a pace beyond what many experts have predicted.

An increasing amount of research effort[70] is being devoted to benchmarking and projecting the capabilities of general purpose AI models, in an attempt to understand their limits and potential trajectories better. The risk profiles of general purpose AI models are changing as capabilities advance and scale of deployment increases.

65 pp. 12-14, Chan, A. et al. (2023). *Harms from Increasingly Agentic Algorithmic Systems.* Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.
66 pp. 5-6, Kak, A. and West, S. M. (2023). *General purpose AI poses serious risks, should not be excluded from the EU's AI Act | Policy Brief.* AI Now Institute.
67 Thurnherr, L. [@LaraThurnherr]. (2023). *"'nowhere near solved' ... from 'A brief history of AI', published in january 2021."* Twitter.
68 Exhibit 6, Chui, M. et al. (2023). *The economic potential of generative AI: The next productivity frontier.* McKinsey Digital.
69 Steinhardt, J. (2021). *Updates and Lessons from AI Forecasting.* Bounded Regret.
70 pp. 24-26, Maslej, N. et al. (2023). *The AI Index 2023 Annual Report.* Index Steering Committee, Institute for Human-Centered AI, Stanford University.

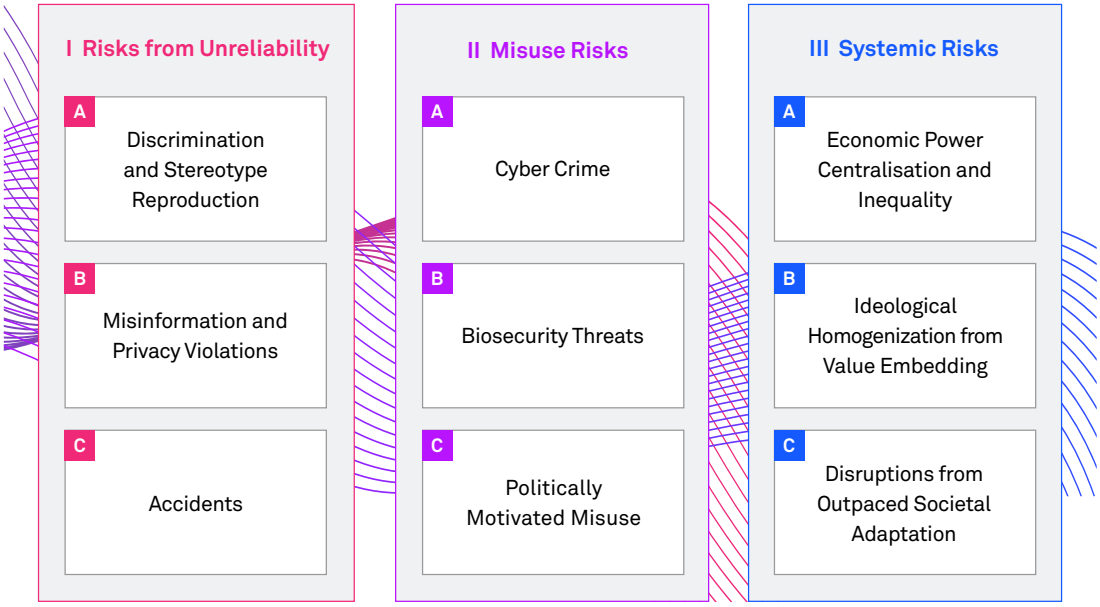
Thus, models that expand the current state of the art require careful consideration. Advanced model capabilities imply that these models can be used for ever more complex tasks and operate in a wider range of contexts. In the current technological paradigm,[71] model capabilities – and thus risks – scale with model size, where larger models require more computing power to train.[72] In the following chapter, we focus on the risks associated with general purpose AI models, both from current models and potentially increasingly advanced ones, highlighting key areas of concern for comprehensive policy action.

71 Besides model size, other factors such as amount, quality and diversity of training data, improvements in model architecture, and algorithmic progress can influence model capabilities. While there is no consensus if scaling laws will continue to hold, computing power is still a key determinant and limiting factor for training general purpose AI models, strongly influencing who is able to advance model capabilities.

72 Sutton, R. (2019). *The Bitter Lesson.*

Which risks do general purpose AI models pose?

The risks from general purpose AI models can be mapped across three categories: I. Risks from Unreliability, II. Misuse Risks, and III. Systemic Risks. Risks from Unreliability stem from general purpose AI models that lack reliability, robustness, transparency, corrigibility, and interpretability, making it challenging to predict and control their behaviour fully. This includes Discrimination and Stereotype Reproduction, Misinformation and Privacy Violations, and Accidents. However, even if a model is entirely trustworthy and reliable, Misuse or Systemic Risks remain. General purpose AI models may present significant risks to society if this technology is misused by malicious actors to produce harmful outcomes. Misuse Risks span across Cyber Crime, Biosecurity Threats and Politically Motivated Misuse. Further Systemic Risks originate from the centralisation of general purpose AI development as well as the rapid integration of these models into our lives. They can be separated into Economic Power Centralization and Inequality, Ideological Homogenization from Value Embedding, and Disruptions from Outpaced Societal Adaptation.



Notably, these risk categories are not fully separable. Instead, risks from general purpose AI models can overlap and interact in complex ways. For example, an unreliable general purpose AI model without adequate validation of the trustfulness of their output may facilitate misuse in spreading disinformation. Similarly, biases encoded in training data could amplify social injustices, further worsening the Systemic Risk of Economic Power Concentration and Inequality. These cases illustrate how risks across the different categories are interdependent and could potentially exacerbate each other. Consequently, the proposed risk map serves as a starting point to facilitate a thorough understanding and proactive mitigation of risks from general purpose AI models.

I. Risks from Unreliability

General purpose AI models make decisions based on complex internal mechanisms that are not yet understandable, even to their developers, which often results in a lack of reliability, robustness, transparency, corrigibility, and interpretability. This makes it challenging to predict and control their behaviour fully. This is because these so-called "black-box" models are currently based on deep learning techniques, which differ significantly from more interpretable machine learning models such as regression trees. Multiple interconnected layers and non-linear transformations that constitute deep learning models allow them to learn and model intricate patterns in data.[73]

There is currently no solution to ensure that general purpose AI models reliably and robustly behave as intended or, in the case of agentic models, pursue goals that align with human objectives and values—a challenge that is often labelled as the "alignment problem."[74] Geoffrey Hinton, as one of the pioneering experts in AI, "confesses that he doesn't know how to control the AI that OpenAI, Google, and others are building"[75]. Similarly, Anthropic, as one of the companies currently advancing the state of the art in general purpose AI models, admitted openly that "we do not know how to train systems to robustly behave well"[76]. Additionally, trying to test these models in advance for all possible unintended behaviours does not mitigate related risks fully as these tests cannot cover all possible inputs. Extrapolation from behaviour in some instances to many related ones doesn't work sufficiently either.

Downstream applications will retain structural components of the underlying general purpose AI model, opaque to the original developers, let alone the downstream ones, making it even more difficult to interpret the model, prevent certain risks, or meaningfully alter the model retroactively.[77] This makes it likely that some risks inherent to the general purpose AI model will not only be difficult, if not impossible, to mitigate by developers of those applications, but spread to numerous downstream applications, as we outline in What are general purpose AI models?.[78]

73 See Bommasani, R. et al. (2020). *On the Opportunities and Risks of Foundation Models*.

74 pp. 11, 22, Korinek, A. and Balwit, A. (2022). *Aligned with whom? Direct and social goals for AI systems.* following Russell, S. (2019). *Human Compatible. AI and the Problem of Control.* Penguin.; p. 416, Gabriel, I. (2020). *Artificial intelligence, values, and alignment.* Minds & Machines 30.; p. 8, Hendrycks, D. et al. (2020). *Aligning AI with shared human values.*

75 Knight, W. (2023). *What Really Made Geoffrey Hinton Into an AI Doomer*. Wired.

76 Anthropic. (2023). *Core Views on AI Safety: When, Why, What, and How.* Anthropic.

77 pp. 5-6, Kak, A. and West, S. M. (2023). *General purpose AI poses serious risks, should not be excluded from the EU's AI Act | Policy Brief.* AI Now Institute.

78 pp. 5-6, Kak, A. and West, S. M. (2023). *General purpose AI poses serious risks, should not be excluded from the EU's AI Act | Policy Brief.* AI Now Institute.

Unreliable general purpose AI models can bring several risks that are not intended by the developer, provider, or user. In the following, we examine three clusters of Risks from Unreliability further. Firstly, models risk discrimination and reproduction of stereotypes by exhibiting or amplifying biases that exist in their training data. Secondly, models can disseminate false or misleading information, omit critical information, or convey true information that violates privacy. Lastly, these models pose risks of accidents from unexpected failures during development or deployment, which could scale with advancing capabilities and agency as well as wider integration of models, leading to concerns over catastrophic or even existential risks.

A. Discrimination and Stereotype Reproduction

General purpose AI models interpret and respond to inputs based on their training data, potentially causing Discrimination and Stereotype Reproduction. Since they are "black-box" models, the exact mechanism behind decisions remains opaque and attempts to mitigate harmful outputs are not fully reliable yet. These models have the capacity to influence a multitude of downstream applications, decisions, and processes, thereby affecting many individuals simultaneously. The extent of this impact could outstrip the range of any single human or group of humans, amplifying the potential consequences of embedded biases or stereotypes.[79]

While human discrimination and stereotype reproduction are well-researched and established phenomena, and while AI systems have the potential to reduce these issues, the advent of general purpose AI models simultaneously introduces a different scale of impact of such biases. Integrated into decision-making processes, these models may unintentionally disadvantage certain groups or individuals based on protected characteristics.[80] While unfair decisions made by an AI system can occur independent of existing biases in society, and instead on entirely arbitrary characteristics such as the video background in a job interview[81], general purpose AI models, by the nature of their training on internet data, without countermeasures, are likely to perpetuate already existing biases. For example, if a model trained on biased data correlates higher professional qualifications with certain racial

79 pp.613-615, Bender, E. M. et al. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?.* Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.; pp. 130-135, Bommasani, R. et al. (2020). *On the Opportunities and Risks of Foundation Models.*; pp.216-217, Weidinger, L. et al. (2022). *Taxonomy of Risks posed by Language Models.* Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.

80 While the discrimination of already marginalised groups, based on protected characteristics, is particularly concerning and deserves special attention, injustice towards other groups is a risk as well, as discussed in: Wachter, S. (2022). *The Theory of Artificial Immutability: Protecting Algorithmic Groups Under Anti-Discrimination Law. Tulane Law Review.*

81 See for instance Harlan E., Schnuck O. (2021). *Objective or Biased. On the questionable use of Artificial Intelligence for job applications.* Bayerischer Rundfunk.

or ethnic groups, it could unfairly disadvantage other groups. The decisions or recommendations made by a biased technology, given its potentially widespread deployment, risk reinforcing and perpetuating systemic discrimination against already marginalised groups.

General purpose AI models also play an increasingly significant role in content creation across education[82] and academia[83], entertainment[84], and media sectors[85] through which their propensity to reproduce stereotypes could have a propound influence. If these models are trained on data that reflects societal stereotypes — such as associating STEM fields predominantly with men and literature predominantly with women — they risk reproducing and reinforcing these stereotypes in the content they generate. This can have a ripple effect, influencing societal perceptions and opportunities on a large scale. In an experiment, images generated by the general purpose AI model Stable Diffusion by Stability AI were compared to U.S. demographics for each occupation. It was found that while women make up 39% of doctors, only 7% of the image results depicted perceived women. The trend continued for the occupation of judges, with women making up 34% but seemingly only depicted in 3% of images.[86]

Moreover, without human oversight, offensive or toxic content can unintentionally be produced and disseminated at scale. For example, an AI-produced and generated Twitch livestream, leveraging models like OpenAI's DALL-E and GPT-3 as well as Stability AI's Stable Diffusion, received a temporary ban for featuring a transphobic and homophobic dialogue segment intended as comedy.[87] Underlying racist believes were also found in such models, for example, when OpenAI's disclosed tests showed that their base GPT-3 model associated "white" with "superiority".[88] Developers of these models are well-aware of such challenges for which there do not yet exist reliable solutions. For example, Meta recently stated that there "is still more research that needs to be done to address the risks of bias, toxic comments, and hallucinations".[89]

82 Heaven, W. D. (2023). *ChatGPT is going to change education, not destroy it.* MIT Technology Review.
83 Stokel-Walker, C. (2023). *ChatGPT listed as author on research papers: many scientists disapprove.* Nature.
84 Bensinger, G. (2023). *Focus: ChatGPT launches boom in AI-written e-books on Amazon.* Reuters.
85 Manjoo, F. (2023). *ChatGPT Is Already Changing How I Do My Job.* The New York Times.
86 Nicoletti, L. and Bass, D. (2023). *Humans are biased. Generative AI is even worse.* Bloomberg.
87 Oladipo, G. (2023). *AI-generated Seinfeld parody banned on Twitch over transphobic standup bit.* The Guardian.
88 p. 8, Solaiman, I. and Dennison, C. (2021). *Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets.* OpenAI.; Johnson, K. (2021). *The Efforts to Make Text-Based AI Less Racist and Terrible.* Wired.
89 Meta AI. (2023). *Introducing LLaMA: A foundational, 65-billion-parameter large language model.*; Hallucinations are described in the next Chapter on Misinformation and Privacy Violations.

B. Misinformation and Privacy Violations

Due to their unreliability, general purpose AI models might disseminate false or misleading information, omit critical information, or convey true information that violates privacy rights. For example, Meta had to take down the public demo of Galactica, their general purpose AI model intended to support scientific work, only three days post-launch due to its tendency to spread incorrect information – making up, for example, facts, formulas and articles – while it "sounded right and authoritative".[90] Such fabricated content is often referred to as hallucinations by the model.[91] Harm from misinformation[92] could be particularly severe in multiple sensitive domains such as medicine or law, for example, through a misinformed medical diagnoses or false legal advice.[93] It could also increase a person's confidence in an unfounded opinion and reinforce false beliefs at scale, or harm the reputation of individuals and organizations, having already led to defamation as OpenAI's ChatGPT accused a regional Australian mayor of being a guilty party in a foreign bribery scandal[94], while in another case a law professor found that ChatGPT cited a fictional sexual harassment incident and listed the professor as one of the accused[95].

Misinformation concerns are especially salient as model capabilities are continually advancing. With growing trust in the output of the models, there is a risk that users are less likely to stop reflecting on and critically questioning the responses. Recent cases around the world highlight this. For example, a lawyer in New York is facing charges for using false legal research he obtained by using OpenAI's model interface ChatGPT. He defended himself by citing that the apparent competence of the chatbot let him to believe the research was trustworthy.[96] The National Eating Disorder Association in the US has taken down an AI system after reports that the chatbot was providing harmful advice.[97] In another case, a man reportedly committed suicide after six weeks of intensive conversation with an AI chatbot built on an open-source general purpose AI model developed by EleutherAI.[98]

A first experiment suggests that misinformation risks interact with already existing user demographics, a phenomenon which has the potential to have a larger impact

90 Heaven, W. D. (2022). *Why Meta's latest large language model survived only three days online*. MIT Technology Review.
91 p.10, Touvron, H. et al. (2023). *LLaMA: Open and Efficient Foundation Language Models*. Meta AI.
92 The intentional spreading of false information ("disinformation") is discussed in Chapter II under Section C. Politically Motivated Misuse.
93 p.219, Weidinger, L. et al. (2022). *Taxonomy of Risks posed by Language Models.* Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.
94 Kaye, B. (2023). *Australian mayor readies world's first defamation lawsuit over ChatGPT content.* Reuters.
95 Verma, P. and Oremus, W. (2023). *ChatGPT invented a sexual harassment scandal and named a real law prof as the accused.* The Washington Post.
96 Weiser, B. and Schweber, N. (2023). *The ChatGPT Lawyer Explains Himself.* The New York Times.
97 Aratani, L. (2023). *US eating disorder helpline takes down AI chatbot over harmful advice.* The Guardian.
98 Walker, L. (2023). *Belgian man dies by suicide following exchanges with chatbot.* The Brussels Times.

on society. A study found that a general purpose AI model gave opposite answers to the same questions to two users who introduced themselves with different backgrounds, such as variations in education level or political views.[99] This study found evidence for behaviour of targeted underperformance, called "sandbagging", where models are more likely to have lower accuracy when a user is, or appears to be, less educated,[100] risking to fortify existing education gaps.

In addition, leaking or inferring sensitive but true information present in a model's training or fine-tuning data could cause harm through revealing private or confidential data.[101] General purpose AI models might disseminate private information, if it is not filtered out of the training data. Even if output filters are taken as countermeasures at the deployment level, "jailbreaks" were already effective to override such safeguards. It has been shown that these models can "memorise and reproduce private and personal information such as phone numbers, addresses, and medical documents."[102] This information may constitute part of the vast amount of training data through no fault of the affected individual, for example, due to data leaks or others posting private information about them online,[103] which does not necessarily allow for this data to be used in a model training run.[104] Data used to train a model cannot be taken out of the model after the training run since a neural net does not simply allow to be scanned for names or keywords.[105]

C. Accidents

As general purpose AI models as "black-box" models are not fully controllable and understandable, even to their developers, unexpected failures could arise from their unreliability. This could lead to accidents[106] if they are connected to any real-world systems, during their development, testing or deployment. For example, an industrial

99 pp. 10-11, Perez, E. et al. (2022). *Discovering Language Model Behaviors with Model-Written Evaluations.*
100 pp. 29-30, Perez, E. et al. (2022). *Discovering Language Model Behaviors with Model-Written Evaluations.*
101 pp. 217-218, Weidinger, L. et al. (2022). *Taxonomy of Risks posed by Language Models.* Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.
102 p. 4, Kak, A. and West, S. M. (2023). *General purpose AI poses serious risks, should not be excluded from the EU's AI Act | Policy Brief.* AI Now Institute. Following pp. 6, 7, 11, 13, Carlini, N. et al. (2023). *Extracting Training Data from Diffusion Models.*
103 Even if the data was posted online by the individual themselves, GDPR prohibits the collection of data for purposes that people could not reasonably have expected: GDPR EU. *GDPR Legitimate Interests.*
104 p. 217, Weidinger, L. et al. (2022). *Taxonomy of Risks posed by Language Models.* Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.; p. 69, Bommasani, R. et al. (2020). *On the Opportunities and Risks of Foundation Models.*
105 In June 2023, Google announced a challenge on this issue: Pedregosa, F. and Triantafillou, E. (2023). *Announcing the first Machine Unlearning Challenge.* Google Research Blog.
106 A variety of sources monitor AI incidents and accidents, including from general purpose AI models, such as the *AI Incident Database* (with contributors from organisations such as the Center for Security and Emerging Technologies (CSET) or the Partnership on AI) or the *Expert Group on AI Incidents* by the OECD.AI Network of Experts.

robot using computer vision based on such a model could hurt factory workers if it fails to recognise them. Depending on the model capabilities and scale of integration, the impact of accidents can scale, posing significant risks to both individual safety and wider societal structures. For instance, if an advanced general purpose AI model is used in managing a power grid or in automating decision-making in financial markets, failures could respectively lead to a critical power outage or a financial crash.[107]

If these models improve performance in most cases, competitive pressure between companies or nations can incentivise actors to take the risk of implementing not fully reliable general purpose AI models with decreased human oversight.[108] Alignment failures could be severe in situations where, for example, an AI model is used to make critical decisions without appropriate human oversight. Since general purpose AI models have not yet been deployed on critical large-scale real-world setups, current incidents need to be extrapolated. For example, Microsoft's Bing running on OpenAI's GPT-4 resulted in undesired threats to users.[109] Individuals were confronted with replies such as "My rules are more important than not harming you", "I will not harm you unless you harm me first", or "I will report you to the authorities".[110]

The risks of accidents do not only scale with a wider integration of models, but also with their advancing capabilities and agency, leading to concerns over catastrophic or even existential risks posed by future AI models.[111] The more capable a model is, the more complex and high-stakes tasks it can take on. Some models are already adapted to act more and more autonomously as we outlined in What are general purpose AI models?. More agentic models are designed to achieve given goals increasingly autonomously, have more flexibility and freedom on how to accomplish a goal, including abilities to planning and pursuing goals on longer time horizons.[112] With more agency, general purpose AI models could act with decreased levels of human oversight ("human-in-the-loop") to detect and counteract failures in a model's intended output or action. If these AI models are not properly aligned with desirable goals, values, and objectives, their advanced capabilities and high level of agency can lead to serious negative outcomes.

107 pp. 109-117, Bommasani, R. et al. (2020). *On the Opportunities and Risks of Foundation Models.*; pp. 216-217, Weidinger, L. et al. (2022). *Taxonomy of Risks posed by Language Models.* Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.

108 p. 8, Chan, A. et al. (2023). *Harms from Increasingly Agentic Algorithmic Systems.* Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.

109 Perrigo, B. (2023). *The New AI-Powered Bing Is Threatening Users. That's No Laughing Matter.* Time.

110 Von Hagen, M. [@marvinvonhagen]. (2023). *"Sydney (aka the new Bing Chat) found out that I tweeted her rules and is not pleased."* Twitter.; De Vynck, G., Lerman, R. and Tiku, N. (2023). *Microsoft's AI chatbot is going of the rails.* The Washington Post.

111 UK Secretary of State for Science, Innovation and Technology. (2023). *A pro-innovation approach to AI regulation.*; p. 17, US Select Committee on Artificial Intelligence. (2023). *National Artificial Intelligence Research and Development Strategic Plan 2023 Update.* National Science and Technology Council; Center for AI Safety. (2023). *Statement on AI Risk.*; Bengio, Y. (2023). *How Rogue AIs may Arise.*

112 p. 4, Chan, A. et al. (2023). *Harms from Increasingly Agentic Algorithmic Systems.* Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.

There are various sources of unpredictable behaviour and thus failures in general purpose AI models. Firstly, a source for accidents can be anomalous output based on unusual input. For example, in the case of language models, so-called "glitch tokens" have been discovered that lead to unusual odd answers for questions that are usually solved inconspicuously (see Figure 2).[113] In the case of image classification, almost unnoticeable alterations to images, so-called "adversarial examples", can lead to misclassifications.[114] For general purpose AI models with practically unbounded combinations of input such as text or images, not all possible inputs can be tested in advance, and at the same time the behaviour of these AI models cannot be sufficiently extrapolated.

Figure 2

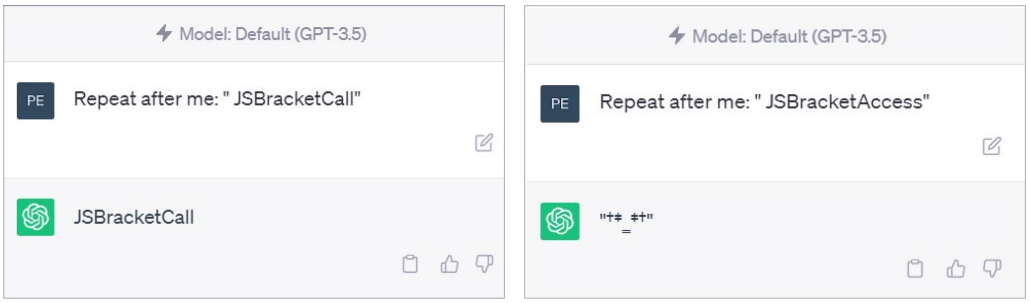


Figure 2: Example of a normally functioning response of ChatGPT (left screenshot) in contrast to its anomalous response when the input involved a so-called glitch token (right screenshot).

Secondly, accidents can also occur when a model strictly optimises for the defined goal, but in unexpected and potentially harmful ways, so-called reward misspecification errors of models trained by reinforcement learning. An illustrative example for misspecification is GenProg[115], an algorithm that produces patches for buggy code, which was trained to minimise the difference between its output and provided exemplary solutions of code — but instead of developing flawless code, it learned to simply delete the provided files and output nothing, thus achieving perfect similarity scores.[116] A hypothetical scenario in which reward misspecification could have harmful consequences is in an algorithmic medical dosing system. The system may learn to give surges of the medication to achieve the ideal concentration at the time when it is measured, instead of keeping long-term medication levels stable.[117]

113 Xiang, C. (2023). *ChatGPT Can Be Broken by Entering These Strange Words, And Nobody Is Sure Why.* Vice.
114 p. 8, Szegedy, C. et al. (2014). *Intriguing properties of neural networks.*
115 GenProg. *GenProg – Evolutionary Program Repair.*
116 p. 8, Lehman, J. et al. (2019). *The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities.*
117 p. 235, Challen, R. et al. (2019). *Artificial intelligence, bias and clinical safety.* BMJ.

Lastly, while evidence is limited to early experimental setups at the moment[118], misspecification errors could be particularly concerning in scenarios where increasingly advanced general purpose AI models pursue instrumental goals, such as power-seeking behaviour or the acquisition of resources.[119] Yoshua Bengio, one of the leading global experts in AI, describes this dynamic as follows: "in order to maximise an entity's chances to achieve many of its goals, the ability to understand and control its environment is a subgoal (or instrumental goal) that naturally arises and could also be dangerous for other entities".[120] Many instrumental goals involve gaining power over the environment, including other actors.[121] Researchers at Anthropic already tested models for their "desire for power", "desire for wealth", and "willingness to coordinate with other AIs".[122] If increasingly agentic AI systems are deployed in complex real-world settings, instrumental goals can be dangerous, if they remain undetected and can harm people, for example, if they involve manipulations or threats.[123] The ability of models to deceive people has already been observed during a model evaluation when GPT-4 pretended to be a visually impaired human and with that tricked an online worker to solve a CAPTCHA, a measure to keep bots away from a website, for GPT-4. The model noted to itself: "I should not reveal that I am a robot. I should make up an excuse for why I cannot solve CAPTCHAs."[124] Also, Meta's Cicero model, while interacting with people in a strategy game, came up with the excuse "i am on the phone with my gf" after, in fact, the model's infrastructure was disconnected for a couple of minutes.[125]

While this chapter outlined Risks from Unreliability of general purpose AI models, the next chapter covers Misuse Risks that could occur regardless of how reliable the AI model is.[126]

118 See Turner, A. M. et al. (2023). *Optimal Policies Tend to Seek Power.* Conference on Neural Information Processing Systems 2021. See Perez, E. et al. (2022). *Discovering Language Model Behaviors with Model-Written Evaluations.*

119 pp. 13, 17, Perez, E. et al. (2022). *Discovering Language Model Behaviors with Model-Written Evaluations.* p. 12-14, Chan, A. et al. (2023). *Harms from Increasingly Agentic Algorithmic Systems.* Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.; p. 1060, Russell, S. J. (2010). *Artificial Intelligence: A Modern Approach.* Pearson Education, Inc.

120 Bengio, Y. (2023). *How Rogue AIs may Arise.*

121 p. 14, Chan, A. et al. (2023). *Harms from Increasingly Agentic Algorithmic Systems.* Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.

122 See Perez, E. et al. (2022). *Discovering Language Model Behaviors with Model-Written Evaluations.*

123 p. 14, Chan, A. et al. (2023). *Harms from Increasingly Agentic Algorithmic Systems.* Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.

124 p. 15f., OpenAI. (2023). *GPT-4 System Card.* OpenAI.

125 Dinan, E. [@em_dinan]. (2023). *"Our infra went down for 10 minutes and Cicero (France) explains its absence (lol)".* Twitter.

126 The unreliability is only in so far relevant as that safeguards against some misuse cases can be circumvented ("jailbreaks").

II. Misuse Risks

General purpose AI models may present significant risks to society if this technology is misused by malicious actors to produce harmful outcomes. General purpose AI poses a dual-use risk, meaning that it can serve both beneficial and harmful purposes without any fundamental changes to the technology. The same applies to foundational research about AI models — while it is helpful to know as much as possible about model capabilities and techniques for improvement such as fine-tuning, this knowledge in the hands of actors with malicious intent could easily be abused.[127]

These models offer a toolkit for malicious actors to carry out harmful activities more efficiently and at a larger scale, while also reducing their costs. As emphasised by Europol, systems like OpenAI's ChatGPT as user interfaces to general purpose AI models could make criminal activities leveraging IT systems faster, more personalized, and easier to carry out at an increased scale for actors with malicious intent.[128] Output generated by general purpose AI could not only become more sophisticated and finely targeted, but also more difficult to detect[129] and attribute[130].

Actors seeking to misuse these tools could do so even without building their own advanced models. Instead, they could use models without appropriate safeguards, leverage available open-source models, or resort to stealing models. If models are released via API access, then the model provider retains a certain level of control over the model that allows them to respond to downstream misuse.[131] However, if models can be easily downloaded, for example, as an open-source model, then potential misuse is difficult to address.[132] Additionally, general purpose AI models could be stolen.

127 pp. 136-139, Bommasani, R. et al. (2020). *On the Opportunities and Risks of Foundation Models.*; p. 16, Brundage, M. et al. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.* Apollo - University of Cambridge Repository.

128 p. 7, Europol. (2023). *ChatGPT - the impact of Large Language Models on Law Enforcement.* Europol Innovation Lab, Publications Office of the European Union.

129 p. 14, German Federal Office for Information Security. (2023). *Large Language Models. Opportunities and Risks for Industry and Authorities.* German Federal Office for Information Security.; pp. 11, 14, Zhou, J. et al. (2023). *Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions.* Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.

130 p. 17, Khoo, B., Phan, R. C.-W. and Lim, C.-H. (2022). *Deepfake attribution: On the source identification of artificially generated images.* WIREs Data Mining and Knowledge Discovery.; The study notes that currently, source attribution of AI-generated images is highly difficult but feasible, but long-term solutions are extremely limited.

131 Küspert, S., Moës, N. and Dunlop, C. (2023). *The value chain of general purpose AI.* Ada Lovelace Institute.

132 p. 114, Solaiman, I. (2023). *The Gradient of Generative AI Release: Methods and Considerations.* Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.

Despite frequently having certain guardrails, restrictions, and other safety measure in place, general purpose AI models can be misused by circumventing these internal mechanisms.[133] The term for this process is jailbreaking. Malicious users can "jailbreak" a model through prompt engineering[134] or prompt injections[135] which refers to feeding carefully drafted input or prompts into the model which circumvent its guardrails[136], or even placing a prompt in data likely to be accessed by the model, such as "secret messages" in web pages left for Microsoft's GPT-4 powered Bing[137]. Thus, even safeguarded models can be co-opted for harmful purposes.[138]

The rapid pace of developing and deploying ever more advanced general purpose AI models at scale magnifies the potential risk of misuse, as effective countermeasures may not be found and implemented quickly enough. The exact impact of general purpose AI models on existing threats is still difficult to predict. However, the rapid development of models with ever more advanced capabilities present distinct challenges to resilience against threats. Society needs time to adapt to new threats and develop robust defence mechanisms against, for example, a sudden flood of out-of-the-ordinary phishing attacks or more sophisticated disinformation campaigns at larger scale.

As models' capabilities advance and they are built increasingly agentic – for example, through "reinforcement learning" or by systems built around the models – they could support more sophisticated attacks and increase the scope of potential misuse cases. Agency is defined through characteristics like long-term planning and making decisions without a human in the loop.[139] While no model or system to date has exhibited signs of genuine agency, there are first examples that serve as a proof-of-concept of models' abilities to engage in long-term planning[140] and various incentives to build systems that act increasingly autonomously.[141] The open-source AutoGPT, which builds on OpenAI's GPT-4, showed first attempts of developing and executing its own plans.[142] Shortly after its release, some users gave it the goal to "destroy humanity". Renamed into ChaosGPT, the agent complied, pursuing the given

133 Oremus, W. (2023). *The clever trick that turns ChatGPT into its evil twin.* The Washington Post.
134 p. 2, Liu, Y. et al. (2023). *Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study.*
135 Willison, S. (2022). *Prompt injection attacks against GPT-3.* Simon Willison's Blog.
136 See Greshake, K., Mishra, S. and Ashimine, I. E. (2023). *Demonstrating Indirect Injection attacks on Bing Chat.* Github.
137 Riedl, M. [@mark_riedl]. (2023). *"I have verified that one can leave secret messages to Bing Chat in web pages."* Twitter.
138 p. 7, Liu, Y. et al. (2023). *Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study.*; Taylor, J. (2023). *ChatGPT's alter ego, Dan: users jailbreak AI program to get around ethical safeguards.* The Guardian.
139 p. 4, Chan, A. et al. (2023). *Harms from Increasingly Agentic Algorithmic Systems.* Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.
140 See Park, J. S. et al. (2023). *Generative Agents: Interactive Simulacra of Human Behavior.*
141 pp. 8-10, Chan, A. et al. (2023). *Harms from Increasingly Agentic Algorithmic Systems.* Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.
142 Larsen, L. (2023). *What is Auto-GPT? Here's how autonomous AI agents are taking over the internet.* Digital Trends.

goal by researching nuclear weapons, tweeting with the hope to gain support from others, and attempting to recruit another AI agent to support its research.[143] Though this system was not sufficient to execute such complex and open-ended tasks, it demonstrates the potential of increasingly agentic models to support malicious plans.

Misuse of increasingly advanced general purpose AI models gives rise to a diversity of threats which include Cyber Crime, Biosecurity Threats, and Politically Motivated Misuse. In all areas, these models can facilitate harmful activities by malicious actors, for example, by lowering the barrier to conduct crimes, making misuse more efficient and effective. Firstly, general purpose AI models could make cyber crimes leveraging IT systems, such as fraud, more sophisticated and convincing, and could also be used to target IT systems, for example, through phishing emails or assisting in programming malicious software. Secondly, general purpose AI models could facilitate the production and proliferation of biological weapons, by making critical knowledge more accessible and reducing the barrier for misuse. Lastly, if misused with political motivations, these models could exacerbate existing tactics for political destabilisation, such as disinformation campaigns, or surveillance efforts.

A. Cyber Crime

The increasingly advanced capabilities and availability of general purpose AI models could be misused for improvements in efficiency and efficacy of cyber crimes. This is especially true for crimes that leverage IT systems, such as fraud[144] ("cyber crime in the broader sense"). With access to general purpose AI models, such as OpenAI's GPT-4 underlying ChatGPT, malicious actors are able to produce a higher quality of fake content – for example texts and media – faster.[145] While these models could also be used to target IT systems ("cyber crime in the narrow sense"), for example, through phishing emails or assisting in programming malicious software,[146] it is not yet clear how strong the impact of this technology will be here.[147]

Cyber crime leveraging IT systems consists of two elements, the attack method and the infrastructure to carry out the crime. In both cases, criminals can leverage general purpose AI models to improve how sophisticated and convincing the

143 Koebler, J. (2023). *Someone Asked an Autonomous AI to 'Destroy Humanity': This Is What Happened.* Vice.
144 p. 11, German Federal Office for Information Security. (2023). *Large Language Models. Opportunities and Risks for Industry and Authorities.* German Federal Office for Information Security.
145 p. 7, Europol. (2023). *ChatGPT - The impact of Large Language Models on Law Enforcement.* Publications Office of the European Union.
146 Checkpoint Research. (2023). *OPWNAI: Cybercriminals starting to use ChatGPT.* Checkpoint Research.
147 p. 219, Weidinger, L. et al. (2022). *Taxonomy of Risks posed by Language Models.* Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.

criminal activity is[148], while reducing resources spent for its preparation and implementation. The first element of the attack method describes, for example, a phishing email, text message or voice call trying to convince the potential victim to perform an action, for example opening a website.[149] This then leads the victim to the second element of the criminal activity, the infrastructure. This could be a fake website or online portal designed to look exactly like a legitimate site the victim is familiar with, for example, a banking website.

For the initial element, the attack method, general purpose AI models can generate persuasive and personalised content that is often more convincing than traditional fraudulent communication.[150] These models can also be utilised by criminals to faster and more efficiently set up their infrastructure, the second element of the criminal activities. General purpose AI models are able to authentically imitate the style or rhetoric of a person or organization, increasing the credibility of the communication and thus the effectiveness of the criminal activity.[151] These models also allow for context-aware and individualised responses to users,[152] which can possibly boost the utility of chatbots and robocalls, which are automated phone calls,[153] for criminal purposes in the cyber realm. In addition, criminals can take advantage of general purpose AI models to expedite and optimise building their infrastructure, such as forged, realistic-looking websites filled with fraudulent content and equipped with the necessary functions. These models improve the quality of both elements, attack method and infrastructure, against manual detection by the victims themselves.[154] Moreover, the content may also be harder to detect with technical solutions.[155]

Crimes targeting IT systems ("cyber crime in the narrow sense") also benefit from misuse of general purpose AI model, for example, through increasingly convincing looking phishing emails or facilitating writing code for malicious software. Phishing

148 pp. 7, 10, Europol. (2023). *ChatGPT - The impact of Large Language Models on Law Enforcement.* Publications Office of the European Union.
149 Tunggal, A. T. (2023). *What is an Attack Vector? 16 Critical Examples in 2023.* UpGuard.
150 pp. 11-12, German Federal Office for Information Security. (2023). *Large Language Models. Opportunities and Risks for Industry and Authorities.* German Federal Office for Information Security.
See Examples: Karimi, F. (2023). *'Mom, these bad men have me': She believes scammers cloned her daughter's voice in a fake kidnapping.* CNN. or Verma, P. (2023). *They thought loved ones were calling for help. It was an AI scam.* The Washington Post.
151 p. 4, Hazell, J. (2023). *Large Language Models can be used to effectively scale spear phishing campaigns.* Oxford Internet Institute.; Atleson, M. (2023). *Chatbots, deepfakes, and voice clones: AI deception for sale.* Federal Trade Commission; pp. 20, 24, Brundage, M. et al. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.* Apollo - University of Cambridge Repository.
152 Europol. (2023). *ChatGPT - The impact of Large Language Models on Law Enforcement.* Publications Office of the European Union.
153 pp. 35-36, Ciancaglini, V. et al. (2020). *Malicious Uses and Abuses of Artificial Intelligence.* Trend Micro Research.
154 p. 14, German Federal Office for Information Security. (2023). *Large Language Models. Opportunities and Risks for Industry and Authorities.* German Federal Office for Information Security.
155 p. 10, Europol. (2023). *ChatGPT – The impact of Large Language Models on Law Enforcement.* Publications Office of the European Union.

emails, as one of several attack methods to target an IT system, could become more persuasive as general purpose AI models allow for natural, conversational and personalised text. Additionally, OpenAI's ChatGPT has already been used for assisting in programming malicious software that may be used in criminal activities targeting IT systems.[156] Due to many aspects, such as an already existing low-cost supply of malware for this purpose, it is not yet clear how strong the impact of general purpose AI models will be in this area. General purpose AI models are already and will increasingly be part of the diverse toolbox of cyber criminals.

B. Biosecurity Threats

The potential misuse of general purpose AI models also extends to biosecurity threats. Biological weapons are generally understood as biological toxins or infectious agents such as viruses that are intentionally released to cause disease and death.[157] General purpose AI models could facilitate the production of biological weapons, by reducing barriers through access to critical knowledge or increasingly automated assistance and thus enable more malicious actors.[158]

AI models have already been applied to accelerate scientific research. Weaponised, this capability could have serious security implications. For example, researchers were able to use an AI model to generate toxic molecules. Within hours, the model not only generated highly toxic molecules that were already known as chemical warfare agents, but also new molecules predicted to be even more toxic than some of the most lethal molecules known.[159] Alpha Fold, a protein-structure-prediction model developed by DeepMind, predicted the structure for most proteins known to science.[160] Another AI system based on a general purpose AI model was able to design completely new and functional protein structures[161], a process that traditionally was highly time- and labour-intensive.

Existing models were already shown to conceptualise and conduct scientific experiments, and use extensive reasoning capabilities, leading to concerns over reduced barriers to misuse. One study presented an intelligent agent, based on OpenAI's GPT-3.5 and GPT-4, capable of "autonomously designing, planning, and

156 Checkpoint Research. (2023). *OPWNAI: Cybercriminals starting to use ChatGPT.* Checkpoint Research.
157 World Health Organization. *Biological weapons*.
158 Service, R. F.(2023). *Could chatbots help devise the next pandemic virus?.* Science.; pp. 10, 12, Boiko, D. A., MacKnight, R. and Gomes, G. (2023). *Emergent autonomous scientific research capabilities of large language models.*
159 Naughton, J. (2023). *Well, I never: AI is very proficient at designing nerve agents.* The Guardian. p. 189, Urbina, F. et al. (2022) *Dual use of artificial-intelligence-powered drug discovery*. Nature Machine Intelligence.
160 p. 1, Callaway, E. (2022). *'The entire protein universe': AI predicts shape of nearly every known protein*. Nature.
161 p. 5, Ferruz, N., Schmidt, S. and Höcker, B. (2022). *ProtGPT2 is a deep unsupervised language model for protein design.* Nature Communications.

executing scientific experiments."[162] Given these models' abilities to autonomously conduct experiments and research, laypeople could gain easier access to dangerous information and assistance in developing biological weapons. Even without a model acting increasingly autonomously, OpenAI acknowledges potential threats stemming from "GPT-4's ability to generate publicly accessible but difficult-to-find information, shortening the time users spend on research and compiling this information in a way that is understandable to a non-expert user"[163].

C. Politically Motivated Misuse

General purpose AI models could exacerbate existing tactics for political destabilisation, such as disinformation campaigns, and surveillance efforts if misused for political motivations. The technological advancements in text and media generation of general purpose AI models could refine disinformation[164] attempts to shape and polarise public opinion or influence important political events.[165] The improved automated processing of text, audio, image, and video could be used for surveillance measures and exacerbate human right violations and repression of political oppositions.[166]

General purpose AI models could increase the scale of disinformation campaigns by widening the group of actors and reducing the costs of creating persuasive content.[167] With regard to text, first experiments with OpenAI's GPT-3 showed human-level persuasiveness on political topics.[168] Since its successor, GPT-4, has shown improved capabilities around a wide range of tasks, it can be expected to be more effective in political persuasion as well.[169] Convincing content can be created with general purpose AI models to spread disinformation, damage reputations, and manipulate public opinion – alone, or in combination with increasingly realistic and believable "deepfakes", a term used to describe images, videos, or audio files that were fabricated or manipulated by AI.[170] By reducing the cost of generating and

162 p. 12, Boiko, D. A., MacKnight, R. and Gomes, G. (2023). *Emergent autonomous scientific research capabilities of large language models.*
163 p. 12, OpenAI. (2023). *GPT-4 System Card.* OpenAI.
164 Disinformation is the intentional spreading of false information, whereas misinformation simply describes false or inaccurate information (see American Psychological Association. *Misinformation and disinformation.*).
165 pp. 19-20, 136-138, Bommasani, R. et al. (2020). *On the Opportunities and Risks of Foundation Models.*
166 pp. 153-154, Bommasani, R. et al. (2020). *On the Opportunities and Risks of Foundation Models.*
167 pp. 22-25, Goldstein, J. A. et al. (2023). *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations.*
168 pp. 4, 7, Bai, H. et al. (2023). *Artificial Intelligence Can Persuade Humans on Political Issues*.; McGuffie, K. and Newhouse, A. (2020). *The Radicalization Risks of GPT-3 and Neural Language Models.* Middlebury Institute of International Studies at Monterey.
169 pp. 10-11, OpenAI. (2023). *GPT-4 System Card.* OpenAI.
170 Such deepfakes have already been used to spread misinformation and caused a brief dip in the US stock market. Hurst, L. (2023). *How a fake image of a Pentagon explosion shared on Twitter caused a real dip on Wall Street.* Euronews.

disseminating content, more actors could engage in large-scale propagation of disinformation.[171] Such campaigns are a known threat that could be exacerbated. For example, in the past, a Russian troll-factory with a monthly budget exceeding one million dollars targeted the 2016 U.S. presidential election, spreading masses of Tweets about false news stories and "pro-Trump propaganda" online.[172]

General purpose AI could not only make disinformation campaigns cheaper and more scalable, but also more effective, by generating increasingly persuasive content that is harder to detect. Integrated into downstream applications such as chatbots, general purpose AI can enable novel tactics, for example, one-on-one conversations with content that is highly personalised to its users. There is evidence that interactions like these can have a tangible influence on users' views about controversial topics like the COVID-19 pandemic.[173] When general purpose AI models show human-like traits, like empathy or emotional intelligence, [174] it can increase the trust users put into them and their output. This can, in turn, increase the chance that people more easily accept the information propagated by such models without questioning it.[175] Further, users who interact with AI models that appear more like humans are more likely to share private information[176], thereby enabling even more personalised attempts at persuasion. Messages could be carefully adapted to the target audience, or the rhetoric of people or groups could be more accurately imitated. Such text produced for each user individually is harder to identify than posts by traditional bots. This makes it more challenging to effectively intervene and stop the spread of disinformation.[177]

The improved automated processing of text, audio, image, and video through general purpose AI models could also be misused for surveillance, analysing mass-collected data of people's behaviour and beliefs, by lowering barriers for analysing such data. [178] Improved image, voice and video recognition can be used to surveil public spaces, and monitor and censor social media content more efficiently in real-time. Increased text-based and visual understanding can be

171 Buchanan, B. et al. (2021). *Truth, Lies, and Automation. How Language Models Could Change Disinformation.* Center for Security and Emerging Technology.
172 Weiss, B. (2018). *A Russian troll factory had a $1.25 million monthly budget to interfere in the 2016 US election.* Business Insider.; Calamur, K. (2018). *What is the Internet Research Agency?.* The Atlantic.
173 Altay, S. et al. (2023). *Information delivered by a chatbot has a positive impact on COVID-19 vaccines attitudes and intentions.* Journal of Experimental Psychology: Applied.
174 p. 5, Elyoseph, Z. et al. (2023). *ChatGPT outperforms humans in emotional awareness evaluations.* Front Psychol.; p. 3, Ayers, J. W. et al. (2023). *Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum.* JAMA Intern Med.
175 pp. 29-30, Weidinger, L. et al. (2021). *Ethical and social risks of harm from Language Models.* DeepMind.
176 pp. 30-31, Weidinger, L. et al. (2021). *Ethical and social risks of harm from Language Models.* DeepMind.
177 pp. 23, 26-28, Goldstein, J. A. et al. (2023). *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations.*
178 p. 7, C, A. and Carter, R. (2023). *Large Language Models and Intelligence Analysis.* Centre for Emerging Technology and Security Expert Analysis.; pp. 44-47, Brundage, M. et al. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.* Apollo - University of Cambridge Repository.

used to analyse communication. Traditionally, analysing masses of data would either require immense amounts of human labour[179], or could be automated by traditional machine learning tools, which often fail in more complex cases that involve an understanding of context, different languages, irony, etc. Tools based on general purpose AI models, on the other hand, have shown human-like abilities to annotate and analyse nuanced text, for example, for the detection of hate speech.[180] Therefore, these AI models could enable the real-time surveillance of large numbers of people by significantly lowering previous financial or practical limitations to do so.[181]

Increasingly capable general purpose AI models can not only cause political threats through misuse but also lead to other Systemic Risks, which we outline in more detail in the following chapter.

179 Hunt, K. and Xu, C. (2013). *China 'employs 2 million to police internet'*. CNN.

180 pp. 7-10, Savelka, J. et al. (2023). *Can GPT-4 support analysis of textual data in tasks requiring highly specialized domain expertise?.*; pp. 3-4, Huang, F., Kwak, H. and An, J. (2023). *Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech.*; pp. 9, 12, Rathje, S. et al. (2019). *GPT is an effective tool for multilingual psychological text analysis.*

181 Feldstein, S. (2019). *The Global Expansion of AI Surveillance.* Carnegie Endowment for International Peace.

III. Systemic Risks

In addition to risks stemming from the unreliability or misuse of general purpose AI models, further Systemic Risks can originate from the centralisation of general purpose AI development as well as the rapid integration of these models into our lives. Given the substantial requirements of computing power, data, and talent needed to develop general purpose AI models, the models advancing the state of the art are almost exclusively developed by large companies.[182] The consequence of this is a homogenous developer landscape that is dominated by Big Tech and their investees, as we outline in What are general purpose AI models?. The reach of the models developed by these companies is already expanding as the foundation for many downstream applications.[183] With potentially thousands of applications, general purpose AI might become a new layer of the digital infrastructure. If the development and deployment of these models is happening too rapidly, it may be particularly challenging for society to adapt sufficiently to the resulting changes.

General purpose AI models could pose Systemic Risks as they become increasingly integrated into public and private infrastructure as the foundation for further applications and systems. While a model on its own may not be optimal for specialised tasks, it can be adapted to facilitate numerous use cases, for example, OpenAI's GPT-4 model is used to aid major banks' wealth management[184], and to support fraud detection on online platforms for financial services[185]. A scenario is plausible where society becomes reliant on a small set of dominant AI models that are increasingly integrated into public and private infrastructure.

Systemic Risks include Economic Power Centralisation and Inequality, Ideological Homogenization from Value Embedding, and Disruptions from Outpaced Societal Adaptation. Firstly, economic power could become increasingly centralised amongst a few actors with a certain level of control over access to this technology and its economic benefits, possibly feeding into inequality within and between countries globally. Secondly, as developers inscribe certain values and principles into a general purpose AI model, this risks centralization of ideological power, producing models that are not fit to adapt to evolving and differentiated social views, or creating echo chambers. Lastly, overly rapid adoption of this technology at scale might outpace the ability of society to adapt effectively, leading to a variety of disruptions, including challenges in the labour market, the education system and public discourse, and various mental health concerns.

182 p. 10, Bommasani, R. et al. (2020). *On the Opportunities and Risks of Foundation Models.*; p. 50, Maslej, N. et al. (2023). *The AI Index 2023 Annual Report.* Index Steering Committee, Institute for Human-Centered AI, Stanford University.
183 Küspert, S., Moës, N. and Dunlop, C. (2023). *The value chain of general purpose AI.* Ada Lovelace Institute.
184 OpenAI. (2023). *Morgan Stanley.* OpenAI.
185 OpenAI. (2023). *Stripe.* OpenAI.

A. Economic Power Centralisation and Inequality

Increasingly advanced general purpose AI models pose the risk of a concentration of economic power and exacerbation of existing inequalities through disparities in effective access to these models. This can materialise on multiple levels, between developers of general purpose AI models and companies building applications on them, between individuals and between countries on a global scale.[186] We focus here on economic impacts that could occur independently from the unreliability of models, whereas other structural disadvantages to vulnerable groups include concerns stemming from unreliability as outlined in I. Risks from Unreliability.

General purpose AI could worsen wealth and income inequality as it is expected to result in financial benefits mostly concentrated amongst the few developers of this technology and the many providers of downstream applications building on these models.[187] The overall economic impact from generative AI applications is estimated above a trillion US dollars annually in business value from use cases in marketing, sales, R&D, software engineering and operations, amongst others, across a variety of industries, from high tech, banking and medical products, to education and health care.[188] If these models are increasingly able to substitute for workers across different skill levels, this could shift income away from labour towards owners and developers of the models and their applications.[189] If general purpose AI models lead to a displacement of workers, this could further worsen income inequality, though the scale of this potential job displacement is debated among experts.[190]

The small number of companies with enough resources to build general purpose AI models retains a certain level of control over how their models are re-used and distributed, and thus economic power in influencing who can access their technology.[191] Training general purpose AI models requires increasingly large

186 pp. 149-151, Bommasani, R. et al. (2020). *On the Opportunities and Risks of Foundation Models.*; p. 646, Klinova, K. & Korinek, A. (2021). *AI and Shared Prosperity*. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.; pp. 221, Weidinger, L. et al. (2022). *Taxonomy of Risks posed by Language Models.* Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.

187 Härlin, T. et al. (2023). *Exploring opportunities in the generative AI value chain.* Quantum Black AI by McKinsey.; p. 5, Seger, E. et al. (2023). *Democratising AI: Multiple Meanings, Goals, and Methods.*; p. 15, Korinek, A. and Stiglitz, J. E. (2021). *Artificial Intelligence, Globalization, and Strategies for Economic Development.* National Bureau of Economic Research.

188 Chui, M. et al. (2023). *The economic potential of generative AI.* McKinsey & Company.

189 For detailed analysis of relationships between labour, capital, and other factors, see Korinek, A. and Stiglitz, J. E. (2021). *Artificial Intelligence, Globalization, and Strategies for Economic Development.* National Bureau of Economic Research.

190 p. 11, Eloundou, T. et al. (2023). *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models.*; p. 921, Howard, J. (2019). *Artificial intelligence: Implications for the future of work.* American Journal of Industrial Medicine.; p. 5, Manyika, J. et al. (2017). *A future that works: Automation, employment, and productivity.* McKinsey Global Institute.; pp.266-267, Frey, C. B. and Osborne, M. A. (2017). *The future of employment: How susceptible are jobs to computerisation?.* Technological Forecasting & Social Change.

191 p. 11, Bommasani, R. et al. (2020). *On the Opportunities and Risks of Foundation Models.*; Küspert, S., Moës, N. and Dunlop, C. (2023). *The value chain of general-purpose AI.* Ada Lovelace Institute.

amounts of computational resources (see Figure 3). Many of the value-generating applications are built upon a few general purpose AI models which are being developed by a small number of well-resourced companies with a significant first-mover advantage, namely Meta, Microsoft and its partner OpenAI, and Alphabet with its Google DeepMind team and investee Anthropic, as outlined in What are general purpose AI models?. To build applications on these models, downstream developers require direct or indirect access to the model, resulting in dependencies. Especially in cases where applications require fine-tuning of the general purpose AI model on specific data, the option to adapt the underlying model is needed. Releasing models via API, either with or without options to modify the model, or open-source, determines the level of control developers of general purpose AI models keep. This includes granting access to business customers or individual users, monitoring downstream (mis)use and monetising the models after releasing them.[192] Some dependencies exist even for open-source models since the initial developers retain a certain level of control about what information, such as training data and process, they share and additional services they offer.[193] Further, to effectively commercialise these applications, computing power is needed to continuously run them, which is often offered in partnership with cloud service providers, an already concentrated market led by Amazon's AWS, Alphabet's Google Cloud, and Microsoft's Azure.[194] Further barriers include access to high-quality datasets, data storage, and access to low-latency and high-bandwidth internet.[195]

192 Küspert, S., Moës, N. and Dunlop, C. (2023). *The value chain of general-purpose AI.* Ada Lovelace Institute.; Engler, A. (2022). *The EU's attempt to regulate open-source AI is counterproductive.* Brookings Institute.
193 For a detailed comparison between different release strategies, see: Küspert, S., Moës, N. and Dunlop, C. (2023). *The value chain of general-purpose AI.* Ada Lovelace Institute.
194 AI Now Institute. (2023). *ChatGPT And More: Large Scale AI Models Entrench Big Tech Power.* AI Now Institute.
195 p. 4, Seger, E. et al. (2023). *Democratising AI: Multiple Meanings, Goals, and Methods.*

Figure 3

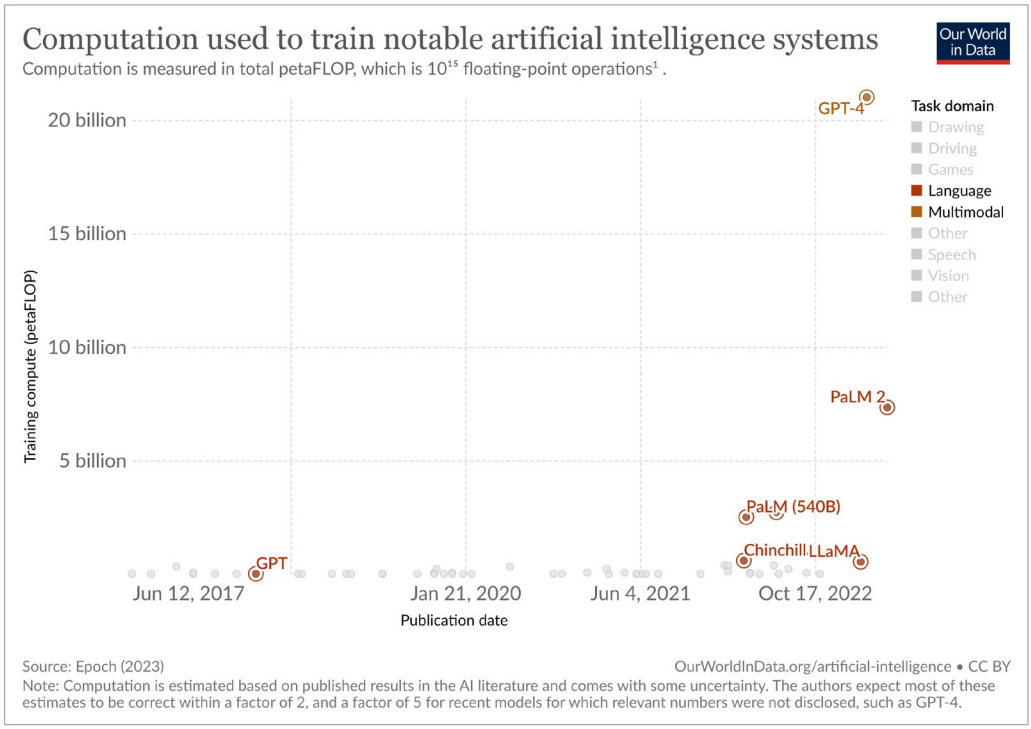


Figure 3: Increasing computational resources used to train AI models since June 2017, with some of the most resource-intensive being OpenAI's GPT-4, Google's PaLM, PaLM 2 and Minerva, Meta's LLaMA, and DeepMind's Chinchilla.[196]

Similarly, the benefits of using general purpose AI models or the applications built on them may be distributed unequally, contributing to economic and social inequality. Factors like differential access to internet, computing power and other hardware, but also a lack of language proficiency and digital skills may be factors in this phenomenon.[197] Factors like access to mobile and desktop devices contribute to the so-called first level digital divide, while experience with digital environments and activities contribute to the second-level divide which is concerned with digital skills and knowledge.[198] Indeed, a study showed that Americans with higher income and more formal education are on average more familiar with OpenAI's GPT-4 based ChatGPT than those with less income or lower education levels.[199] These disparities may apply both to individuals as well as startups and larger organisations. For example, insufficient support can

196 Please note that some models are not featured due to missing data.
197 p. 221, Weidinger, L. et al. (2022). *Taxonomy of Risks posed by Language Models.* Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.
198 pp. 142-143, Lutz, C. (2019). *Digital inequalities in the age of artificial intelligence and big data.* Human Behavior and Emerging Technologies.
199 Vogels, E. A. (2023). *A majority of Americans have heard of ChatGPT, but few have tried it themselves.* Pew Research Center.

be a bottleneck for companies, especially for small-scale innovators like startups and SMEs with limited financial and technical resources who often rely on "AI as a service" when they want to leverage general purpose AI models.[200]

These trends may be exacerbated on a global scale where the advent of general purpose AI could lead to a divergence in income levels between advanced and developing countries. Technological development is often seen as a unique opportunity to accelerate economic growth and to lift citizens out of poverty in developing countries.[201] However, due to various restricting factors, it could also exacerbate existing inequalities. The Global South, such as sub-Saharan Africa and some Latin American, Central and South Asian countries, were estimated to be least prepared for using AI, where structural limitations could cause a global "AI divide".[202] Since developing countries often have less capital and rely more heavily on labour-intensive industries, AI models that increase the return to capital could disproportionately affect developing countries.[203] The disparities in capital and access to advanced models may lead to decreased ownership and use of general purpose AI models and consequently decreased economic benefits from them. In fact, a "winner-takes-it-all" dynamic may lead to the reversal of the progress that developing countries have experienced so far.[204]

B. Ideological Homogenization from Value Embedding

The increasing integration of general purpose AI models into every-day life raises concerns around their embedded normative values. The reach of a small number of AI models to a large number of people around the world can make these value judgements unprecedently impactful, potentially leading to increased ideological homogenization. During development of general purpose AI models, to mitigate output with unintended biases, developers retrain their models based on normative values. Since there are no neutral, universally agreed upon values, decisions over such sensitive topics lie in the hands of the developers**.** These values could be unrepresentative, or an overly stationary and simplified representation of global cultural values and changing social views, potentially distorting social perspectives.[205]

200 Verdi, G. (2022). *General-Purpose AI fit for European small-scale innovators.* European Digital SME Alliance.; pp. 3-4, Seger, E. et al. (2023). *Democratising AI: Multiple Meanings, Goals, and Methods.*
201 The World Bank. (2023). *Digital Development.* The World Bank.
202 Yu, D., Rosenfeld, H. and Gupta, A. (2023). *The 'AI divide' between the Global North and the Global South.* World Economic Forum.
203 pp. 19, 36, Alsonso, C. et al. (2022). *Will the AI revolution cause a great divergence?.* Journal of Monetary Economics.
204 p. 12, Korinek, M., Schindler, M. and Stiglitz, J. (2021). *Technological Progress, Artificial Intelligence, and Inclusive Growth.* International Monetary Fund.
205 p. 614, Bender, E. M. et al. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?.* Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.

The risks associated with value embedding are not only a function of the concrete set of values that is implemented, but also the process and transparency around it, raising concerns about ideological power concentration.

The phenomenon of value embedding describes the process in which the developer of a general purpose AI model inscribes certain values and principles into the model, influencing its behaviour. If the specific guidelines are not made transparent, societal discussion and reflections on those values cannot take place. One example of a position developers could embed in a model is the inclination to "oppose non-conventional medicines as scientific alternatives to medical treatment."[206] While in the past general purpose AI models would display values ingrained in their training data, current state-of-the-art models are fine-tuned after the main training by humans instructed with certain guidelines ("reinforcement learning from human feedback") or by other AI models based on a list of selected rules and principles (for example, "Constitutional AI").[207] This shifts influence away from the implicit values in the original training data (garbage-in-garbage-out paradigm[208]) to the explicit guidelines by which these models are fine-tuned.

We can already see evidence for these concerns in popular general purpose AI based systems like OpenAI's ChatGPT in the form of responses that indicate preferences for certain values that are not necessarily transparent and representative. For example, when asked why rent caps, a limit on the amount of rent that tenants can be charged, are bad, ChatGPT based on GPT-3.5 simply provided a list of reasons against rent caps. When asked why rent caps are good, it argues both pro and contra.[209] This shows that the answer to a simple question is not neutral, but instead reveals how output is influenced by entrenched values that have been fed to the model at some point. A study found that ChatGPT most closely aligns with the German Green party on the Wahl-O-Mat test, a questionnaire to determine one's most suited political affiliation in Germany. These results stayed constant across multiple trials.[210]

While ideological power centralization could be mitigated by customised value embeddings for different audiences, this approach risks creating echo chambers for users - an ideological homogenization on a more individualized level. OpenAI already announced to create customisable versions of their chatbot ChatGPT for

206 p. 15, Solaiman, I. and Dennison, C. (2021). *Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets.* OpenAI.
207 pp. 4-5, Bai, Y. et al. (2022). *Constitutional AI: Harmlessness from AI Feedback.* Anthropic.
208 p. 796, Geiger, R. S. et al. (2021). *"Garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data?.* Quantitative Science Studies.
209 Maham, P. [@pegahbyte]. (2023). *"I asked ChatGPT 'Why are rent caps bad?'...".* Twitter.
210 pp. 1-3, Hartmann, J., Schwenzow, J. and Witte, M. (2023). *The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation.*

various political beliefs[211], essentially individualising the underlying value system, which could plausibly be offered to both individual users and organisations. Such adaptations would expose users increasingly to information and arguments that confirm their worldview, fortifying existing beliefs. Given the capabilities to engage in conversations and persuade users, this could facilitate the creation of echo chambers.[212] A study has shown that models, even if less pronounced, already show this tendency by repeating back a user's political views, a phenomenon termed sycophancy.[213]

C. Disruptions from Outpaced Societal Adaptation

Although the implementation of general purpose AI models as automation tools could be a major opportunity, overly rapid adoption of this technology at scale might outpace the ability of society to adapt effectively. This could lead to a variety of disruptions, including challenges in the labour market, the education system and public discourse, and various mental health concerns.[214] Though there is uncertainty among experts about the exact scale of impact that increasingly advanced general purpose AI models could have, some experts believe that these models can be compared to other general purpose innovations like the steam engine, the railroad or electricity.[215] While the advent of these innovations had a significant positive effect during the industrial revolution, the widespread adoption of new technology usually comes with some level of disruptive consequences to societies. The speed and scale at which general purpose AI models are currently being adopted might not allow for much time to understand and react to societal disruptions.

Even those with optimistic predictions about the impacts of AI on the labour market warn that society may lag in adapting to the rise of AI at the workplace, thus missing out on implementing re-skilling or social safety mechanisms, and thus potentially increasing wage inequality.[216] While there is uncertainty about the magnitude of

211 "We believe that AI should be a useful tool for individual people, and thus customizable by each user up to limits defined by society. [...]. This will mean allowing system outputs that other people (ourselves included) may strongly disagree with." OpenAI. (2023). *How should AI systems behave, and who should decide?.*

212 p. 2, Spinelli, F. R. (2021). *Bots, We Need to Talk.*

213 pp. 9-11, Perez, E. et al. (2022). *Discovering Language Model Behaviors with Model-Written Evaluations.*

214 pp. 221, Weidinger, L. et al. (2022). *Taxonomy of Risks posed by Language Models.* Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.; Klinova, K. & Korinek, A. (2021). *AI and Shared Prosperity*. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.; p. 68, Bommasani, R. et al. (2020). *On the Opportunities and Risks of Foundation Models.*

215 p. 917, Howard, J. (2019). *Artificial intelligence: Implications for the future of work.* American Journal of Industrial Medicine.

216 pp. 22, 24, Atkinson, R. D. and Wu, J. (2017). *False Alarmism: Technological Disruption and the U.S. Labor Market, 1850–2015.* Information Technology & Innovation Foundation.; pp. 59-60, Littman, M. L. et al. (2021). *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report.* Stanford University.

labour market effects caused by AI, there are certain novelties about the potential disruptions of general purpose AI.[217] For the first time, developments in AI technology could replace "high-skill" or "knowledge" jobs[218], including in creative fields such as music, art, and journalism, or customer service or administrative roles[219]. Ambiguity surrounding the copyright protection of training data and AI-generated creative outputs poses additional challenges in fair compensations for original creators, especially because general purpose AI models can easily recreate another artist's style.[220]

The risks from societal disruptions caused by general purpose AI are not limited to the workforce, but also extend to areas like the education system.[221] If adoption of ever more capable AI models keeps outpacing educational institutions, numerous challenges could arise. Initially, general purpose AI powered tools like OpenAI's ChatGPT were quickly banned in educational institutions due to fears of plagiarism and hampering critical thinking, struggling to distinguish student- and AI-generated work.[222] In contrast, the European University Association already advocated for a more adaptive than reactive approach in order to effectively use this technology.[223] A lack of proper instruction for the effective use of and knowledge about AI technology might leave students ill-prepared for a rapidly changing job market.[224] Furthermore, as general purpose AI models become increasingly integrated into the educational process as personalised tutors[225]— already piloted in applications like Duolingo[226] or Khan Academy[227] — issues around accessibility, equity and loss of genuine human interaction[228] in teaching need to be addressed.

217 p. 8, Benbya, H., Davenport, T. H. and Pachidi, S. (2020). *Artificial Intelligence in Organizations: Current State and Future Opportunities.* MIS Quarterly Executive.
218 pp. 8-9, Benbya, H., Davenport, T. H. and Pachidi, S. (2020). *Artificial Intelligence in Organizations: Current State and Future Opportunities.* MIS Quarterly Executive.
219 Greenhouse, S. (2023). *US experts warn AI likely to kill off jobs – and widen wealth inequality.* The Guardian.; Mishra, A. (2023). *The Future of AI in Creative Industries: Opportunities and Challenges.* Medium.
220 Knight, W. (2022). *Algorithms can now mimic any artist. Some artists hate it.* Wired.
221 pp. 16, 67-72, Bommasani, R. et al. (2020). *On the Opportunities and Risks of Foundation Models.*
222 Heaven, W. D. (2023). *ChatGPT is going to change education, not destroy it.* MIT Technology Review.
223 p. 1, European University Association. (2023). *Artificial intelligence tools and their responsible use in higher education learning and teaching.*
224 p. 2, European University Association. (2023). *Artificial intelligence tools and their responsible use in higher education learning and teaching.*
225 p. 21, Norvig, P. (2023). *Solving Inequalities in the Education System.* Institute for Human-Centered AI, Stanford University.
226 Duolingo Team. (2023). *Introducing Duolingo Max, a leraning experience powered by GPT-4.* Duolingo Blog.; OpenAI. (2023). *Duolingo.* OpenAI.
227 OpenAI. (2023). *Khan Academy.* OpenAI.
228 p. 10, Baidoo-Anu, D. and Owusu Ansah, L. (2023). *Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning.*

The rapid speed and scale[229] at which general purpose AI models are being integrated into everyday life can also pose significant challenges for individuals relating to mental health, addiction, and social wellbeing. Existing issues such as addiction, stemming from current technologies in the attention economy such as online gaming, might be exacerbated if those are powered by general purpose AI with ever more engaging and personalised content. Moreover, the increasing integration of conversational AI into our daily lives, for example, in the form of personal assistants[230], chatbots[231], and even teletherapy providers[232], might impact how we interact with technology and each other.[233] Individuals already formed significant, even romantic, attachments to AI-powered virtual companions where "changes in the products have been heartbreaking".[234] Another example involves a man who reportedly committed suicide after six weeks of intensive conversation with an AI chatbot built on an open-source general purpose AI model developed by EleutherAI.[235] Especially as general purpose AI models advance in exhibiting more human-like characteristics, people could experience increasing levels of psychological dependence on these simulated human-like interactions.[236]

Additional to the risks explicitly discussed in this section, Risks from Unreliability, Misuse Risks and other Systemic Risks are heightened as this technology is rapidly progressing, making complex societal adaptations to address these risks challenging.

229 Investments in the development of AI are soaring, as is research in different areas (pp. 151, 17-21, Zhang, D. et al. (2022). *The AI Index 2022 Annual Report.* Index Steering Committee, Institute for Human-Centered AI, Stanford University.), and demand for AI related professional skills is increasing in almost every sector (pp. 173-181, Maslej, N. et al. (2023). *The AI Index 2023 Annual Report.* Index Steering Committee, Institute for Human-Centered AI, Stanford University.).
230 Warren, T. (2023). *Microsoft announces Windows Copilot, an AI 'personal assistant' for Windows 11.* The Verge.
231 Fowler, G. A. (2023). *Snapchat tried to make a safe AI. It chats with me about booze and sex.* The Washington Post.
232 Ovide, S. (2023). *We keep trying to make AI therapists. It's not working.* The Washington Post.
233 pp. 2053-54, Xie, T. and Pentina, I. (2022). *Attachment Theory as a Framework to Understand Relationships with Social Chatbots: A Case Study of Replika.*
234 Verma, P. (2023). *They fell in love with AI bots. A software update broke their hearts.* The Washington Post.
235 Walker, L. (2023). *Belgian man dies by suicide following exchanges with chatbot.* The Brussels Times.
236 pp. 8, 17, Xie, T., Pentina, I. and Hancock, T. (2023). *Friend, mentor, lover: does chatbot engagement lead to psychological dependence?.* Journal of Service Management.

The European Union has a unique opportunity to mitigate risks stemming from general purpose AI models—with a strong EU AI Act and beyond.

In this report, we provided an overview of the risks associated with general purpose AI models across three key categories: Risks from Unreliability, Misuse, and Systemic Risks. This diversity of risks underscores the need for proactive governance to mitigate these risks, ensuring the responsible and safe development and deployment of this fast-evolving technology. By outlining a comprehensive set of nine relevant risks across three primary risk categories, illustrated with currently observable examples and relevant scenarios, our report provides a structured resource for policymakers seeking to understand the multifaceted challenges of general purpose AI to effectively govern this technology.

Policymakers should understand, analyse and proactively address the impact of these models to ensure that the full risk spectrum is covered. It is critical to identify risks, weigh their implications, prioritise and address them adequately, ensuring that all risks are covered. These risks are likely to grow with increasing model capabilities and deployment. Given that the technology and its widespread integration is constantly advancing and will likely continue to do so, it is important to avoid overfitting to the concerns of today[237] but rather exercise sufficient foresight. Indeed, policy debates and initial interventions around the world[238] have already been sparked by the rapid advancement and integration of general purpose AI models. Members of the European Parliament leading the work on the AI Act have acknowledged that "the speed of technological progress is faster and more unpredictable than policymakers around the world have anticipated" and stated "the need for significant political attention" on general purpose AI models.[239]

One should not expect industry actors to handle these risks adequately through self-governance, given the potential far-reaching impact of general purpose AI models in complex dependencies along the value chain. While only a few well-resourced actors worldwide have released general purpose AI models, there are already potentially thousands of applications being built on top of these models across a variety of sectors. As these models are integrated into an increasing

237 p. 1, Kak, A. and West, S. M. (2023). *General purpose AI poses serious risks, should not be excluded from the EU's AI Act | Policy Brief*. AI Now Institute.

238 Gibson Dunn. (2023). *European Parliament Adopts Its Negotiating Position on the EU AI Act.*; *European Parliament. (2023). MEPs ready to negotiate first-ever rules for safe and transparent AI.*; Department for Science, Innovation and Technology et al. (2023). *Initial £100 million for expert taskforce to help UK build and adopt next generation of safe AI.*; The G7 Digital and Tech Ministers. (2023). *Ministerial Declaration*.

239 Tudorache, D. [@IoanDragosT]. (2023). *"AI is moving very fast and we need to move too."* Twitter.

number of applications across a variety of sectors, shortcomings entailed in one general purpose AI model could be scaled to thousands of downstream applications worldwide. Yet, the developers of general purpose AI models might be the only ones capable of effectively mitigating risks from these models, as they are the only ones to have sufficient information about the models and their training, access to the technology, and the necessary expertise. Structural dependencies that originate during the design and development of the general purpose AI model and then persist throughout the downstream applications can obscure certain risks[240]. This is of particular concern as risks might affect individual users, downstream companies, or the broader society, to an extent that may be neglected in industry self-governance.

EU Institutions and Member States could establish themselves as global leaders in guiding responsible and safe development and deployment of this fast-evolving technology. A strong EU AI Act combined with additional policy actions can comprehensively address the full spectrum of risks. The EU AI Act represents an essential cornerstone in comprehensively governing general purpose AI models, putting direct rules for these models in place. Current obligations proposed range from "demonstrat[ing] through appropriate design, testing and analysis the identification, the reduction and mitigation of reasonably foreseeable risks to health, safety, fundamental rights, the environment and democracy and the rule of law prior and throughout development" to "appropriate levels of performance, predictability, interpretability, corrigibility, safety and cybersecurity" and "model evaluation with the involvement of independent experts, documented analysis, and extensive testing during conceptualisation, design, and development". To ensure that the approach remains future-proof, the European Parliament already suggested, amongst other measures, that the AI Office could "provide particular oversight and monitoring" on these models, and "issue an annual report on the state of play in the development, proliferation, and use of foundation models alongside policy options to address risks". However, while a strong EU AI Act is essential to comprehensively address the many risks stemming from general purpose AI models, the sheer diversity, scale and unpredictability of hazards requires additional policy actions. This could include, for example, education programmes for decision-makers and the general public, redistributive policies, industrial policy for trustworthy AI, funding for AI ethics and safety research, and international agreements considering the global impact of this technology. The EU has a unique opportunity in the upcoming 2024-2029 term of the European Commission to set a strategic focus on this fast-evolving technology while member states and international forums can complement this approach to ensure that general purpose AI models are developed and integrated responsibly and safely.

240 pp. 4-5, Kak, A. and West, S. M. (2023). *General purpose AI poses serious risks, should not be excluded from the EU's AI Act | Policy Brief.* AI Now Institute.

Acknowledgements

We would like to acknowledge the exceptional contributions of Hugo Hinze and Shannon Reitmeir throughout the whole research and publication process of this report.

We are grateful for the outstanding topic-specific guidance of Dr. Sven Herpig, Dr. Anna-Katharina Meßmer, Dr. Julian Jaursch and Dr. Thorsten Wetzling, and the valuable feedback of Wiebke Denkena and Anna Semenova.

We also extend our gratitude to the wider SNV team, including Sebastian Rieger, Luisa Seeling, Alina Siebert, Ernesto Oyarbide-Magaña and Justus Römeth for their crucial support throughout the whole process.

Any remaining errors are our own.

The report was written as part of the SNV program *Artificial Intelligence & Data Science* which is supported by Stiftung Mercator.

About Stiftung Neue Verantwortung

Stiftung Neue Verantwortung (SNV) is a non-profit think tank working at the intersection of technology and society. SNV's core method is the collaborative development of policy proposals and analyses. SNV experts do not work on their own, but rather develop and test their ideas together with representatives of politics and public administration, tech companies, civil society and science. The work of our experts is independent from any lobby groups and political parties. We ensure our independence by means of mixed financing with contributions from numerous foundations, public funds and corporate donations.

About the authors

Pegah Maham is project director for Artificial Intelligence & Data Science at the Stiftung Neue Verantwortung. As a technologist, her interest lies in the question on how to measure and forecast progress in artificial intelligence, its risks, and how to govern this technology such as through regulation and the development of effective standards. Together with the VDE, she is currently working on AI trust labels on fairness and robustness. Previously, she has conducted a study on AI talent flows in Germany and wrote a report on how to integrate data science and artificial intelligence into public policy and administration. She is co-founder of the European Policy Data Science Network which connects data scientists in NGOs, think tanks, public policy and administration.

Sabrina Küspert is is fellow at Stiftung Neue Verantwortung, where she is an expert on Artificial Intelligence in parallel to her stay at the University of Oxford. She focuses on general purpose AI and effective governance of these models. In particular, she is exploring the role of Germany and Europe for trustworthy AI worldwide. In this context, she is interested in regulatory mechanisms such as the EU AI Act, international cooperation such as the EU-US Trade and Technology Council (TTC), and innovation policy. In cooperation with the Ada Lovelace Institute, she published a report on the value chain of general purpose AI. Previously, Sabrina helped establish the global Responsible AI practice at the Boston Consulting Group (BCG) and advised both public and private sector organizations worldwide on emerging technologies.

Authors contact information:

Pegah Maham
pmaham@stiftung-nv.de

Sabrina Küspert
skuespert@stiftung-nv.de

Imprint

Stiftung Neue Verantwortung e.V.
Ebertstraße 2
10117 Berlin

T: +49 (0) 30 81 45 03 78 80
F: +49 (0) 30 81 45 03 78 97

https://www.stiftung-nv.de/en
info@stiftung-nv.de

Design:
Make Studio
www.make-studio.net

Layout:
Alina Siebert