



Studium generale-Reihe "Leben mit KI"
Eine Welt mit allgemeiner KI...

Beginn 17:00 Uhr

Studierende, die einen Nachweis über den Besuch des Studium generale benötigen: Bitte hier in BBB mit Vorund Nachnamen anmelden.

### Studium generale-Reihe "Leben mit KI"

09.04.2024	Grundlagen generativer KI
16.04.2024	Prompting und AI Agents
23.04.2024	KI in der Arbeitswelt
07.05.2024	KI in der Hochschulbildung
14.05.2024	Studieren mit KI
21.05.2024	Rechtliche und ethische Herausforderungen durch KI
04.06.2024	KI und das Verständnis der Welt
11.06.2024	Eine Welt mit allgemeiner KI

# Eine Welt mit allgemeiner Kl...

Inhalt

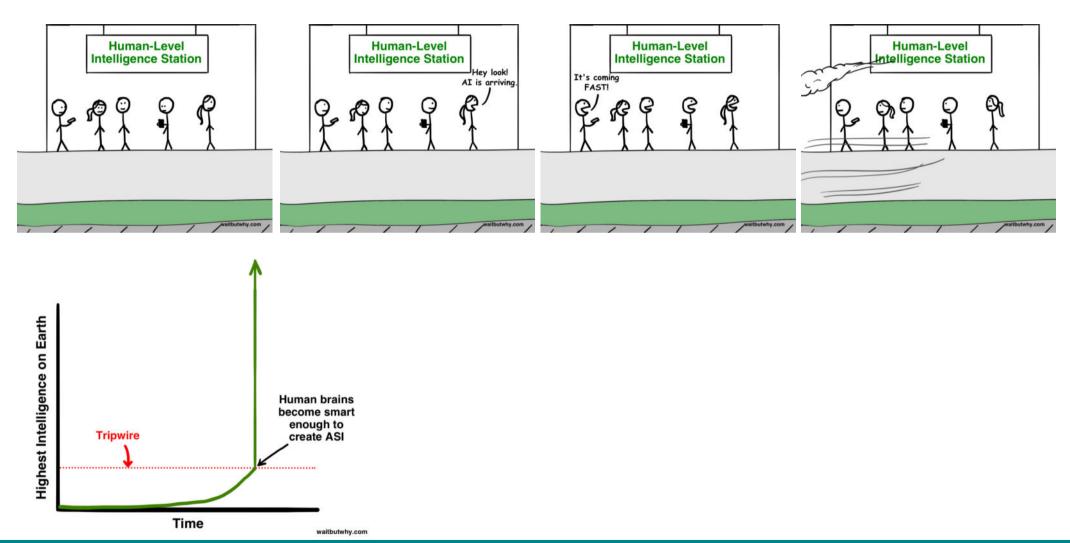
- 1. Allgemeine KI (AGI)
- 2. Prognosen zu AGI
- 3. Szenarien einer (Hochschul-)Welt mit AGI

# Allgemeine KI (AGI) (Moravec 1998)



https://jetpress.org/volume1/moravec.htm.

# Allgemeine KI (AGI) (Urban 2015)



# Allgemeine KI (AGI)

Mit allgemeiner KI (AGI) wird ein technisches System beschreiben, das generell klüger als Menschen ist (Altmann 2023) und weitgehend autonom wertvollere (ökonomische) Arbeit leistet als es Menschen in der Lage wären (OpenAI 2023).

"When I say AGI I mean something which is basically a drop-in substitute for a human remote worker circa 2023, and not just a mediocre one, a good one." (Kokotajlo 2024)

#### Levels of AGI

Donformana (rossa)	Nome	Camanal							
Performance (rows) x	Narrow	General							
Generality (columns)	clearly scoped task or set of tasks	wide range of non-physical tasks,							
		including metacognitive abilities							
		like learning new skills							
Level 0: No AI	Narrow Non-AI	General Non-AI							
	calculator software; compiler	human-in-the-loop computing,							
		e.g., Amazon Mechanical Turk							
Level 1: Emerging	Emerging Narrow AI	Emerging AGI							
equal to or somewhat better than	GOFAI (Boden, 2014); sim-	ChatGPT (OpenAI, 2023), Bard							
an unskilled human	ple rule-based systems, e.g.,	(Anil et al., 2023), Llama 2							
	SHRDLU (Winograd, 1971)	(Touvron et al., 2023), Gemini							
		(Pichai and Hassabis, 2023)							
Level 2: Competent	Competent Narrow AI	Competent AGI							
at least 50th percentile of skilled	toxicity detectors such as Jig-	not yet achieved							
adults	saw (Das et al., 2022); Smart								
	Speakers such as Siri (Apple),								
	Alexa (Amazon), or Google As-								
	sistant (Google); VQA systems								
	such as PaLI (Chen et al., 2023);								
	Watson (IBM); SOTA LLMs for a								
	subset of tasks (e.g., short essay								
	writing, simple coding)								
Level 3: Expert	Expert Narrow AI	Expert AGI							
at least 90th percentile of skilled	spelling & grammar checkers	not yet achieved							
adults	such as Grammarly (Gram-	1000							
	marly, 2023); generative im-								
	age models such as Imagen (Sa-								
	haria et al., 2022) or Dall-E 2								
	(Ramesh et al., 2022)								
Level 4: Virtuoso	Virtuoso Narrow AI	Virtuoso AGI							
at least 99th percentile of skilled	Deep Blue (Campbell et al.,	not yet achieved							
adults	2002), AlphaGo (Silver et al.,								
	2016, 2017)								
Level 5: Superhuman	Superhuman Narrow AI	Artificial Superintelligence							
outperforms 100% of humans	AlphaFold (Jumper et al., 2021;	(ASI)							
A 10	Varadi et al., 2021), AlphaZero	not yet achieved							
	(Silver et al., 2018), StockFish								
	(Stockfish, 2023)								
	A CALL CONTROL OF THE PROPERTY OF THE PARTY								

# Allgemeine KI (AGI) (Silver et al. 2017)

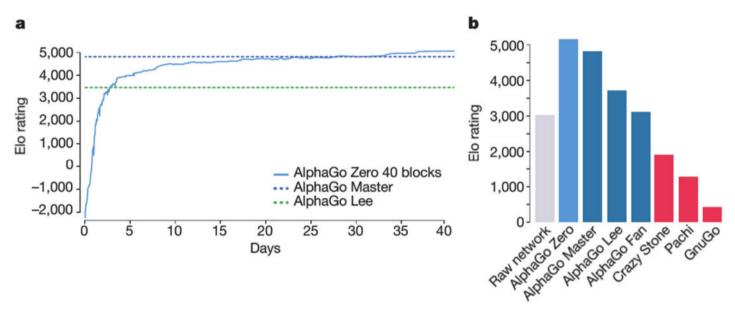
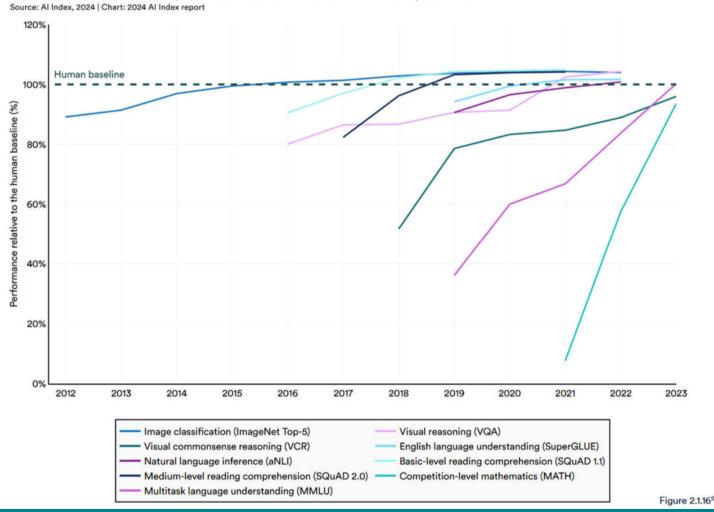


Figure 6 | Performance of AlphaGo Zero. a, Learning curve for AlphaGo Zero using a larger 40-block residual network over 40 days. The plot shows the performance of each player  $\alpha_{\theta_i}$  from each iteration i of our reinforcement learning algorithm. Elo ratings were computed from evaluation games between different players, using 0.4 s per search (see Methods). b, Final performance of AlphaGo Zero. AlphaGo Zero was trained for 40 days using a 40-block residual neural network. The plot shows the results of a tournament between: AlphaGo Zero, AlphaGo Master (defeated top human professionals 60–0 in online games), AlphaGo

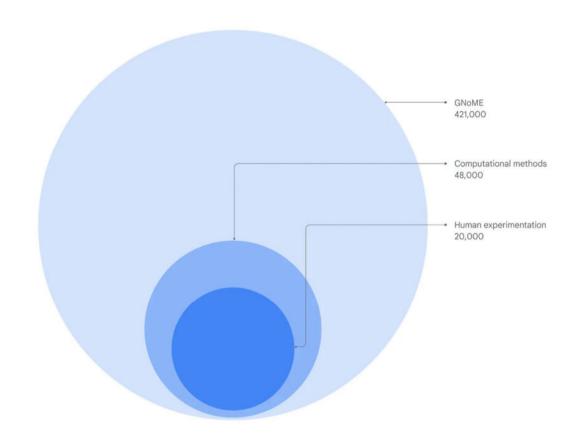
Lee (defeated Lee Sedol), AlphaGo Fan (defeated Fan Hui), as well as previous Go programs Crazy Stone, Pachi and GnuGo. Each program was given 5 s of thinking time per move. AlphaGo Zero and AlphaGo Master played on a single machine on the Google Cloud; AlphaGo Fan and AlphaGo Lee were distributed over many machines. The raw neural network from AlphaGo Zero is also included, which directly selects the move a with maximum probability  $p_a$ , without using MCTS. Programs were evaluated on an Elo scale<sup>25</sup>: a 200-point gap corresponds to a 75% probability of winning.

# Allgemeine KI (AGI) (Silver et al. 2017)

#### Select AI Index technical performance benchmarks vs. human performance



# Allgemeine KI (AGI)





A-Lab, a facility at Berkeley Lab where artificial intelligence guides robots in making new materials. Photo credit: Marilyn Sargent/Berkeley Lab

Aber: keine Stichprobe der 380,000 stabilen Kristalle konnte als "glaubwürdig", "nützlich" und "neuartig" eingestuft werden (Cheetham & Seshadri 2024)

### Eine Welt mit allgemeiner Kl...

Inhalt

- 1. Allgemeine KI (AGI)
- 2. Prognosen zu AGI
- 3. Szenarien einer (Hochschul-)Welt mit AGI

Entwicklung einer allgemeinen künstlichen Intelligenz

- Unaufhaltbarkeit
- Absehbarkeit
- Unvermeidbarkeit tiefgreifender gesellschaftlicher Veränderungen

Grundlage: Berichte und Prognosen von Technologieunternehmen bzw. Interviews mit deren CEOs und Forschenden

- Veröffentlichungen als Teil der Interessenpolitik der Technologieunternehmen
- "Building state-of-the-art AI systems increasingly requires large amounts of data, compute, and money, resources that industry actors inherently possess in greater amounts compared to nonprofits and academia." (Maslej et al. 2023 / AI Index Report)

### Entwicklung einer allgemeinen künstlichen Intelligenz

- Unaufhaltbarkeit
  - "And a recursive loop of innovation, as these smart machines themselves help us make smarter machines, will accelerate the revolution's pace." (Altman 2021)
  - o Denkmuster und Weltanschauungen im Kontext der Technologieunternehmen
    - ein spieltheoretisches Verständnis von Menschen, Organisationen und Gesellschaften, das von Wettbewerb und Machtstreben geprägt ist (Tegmark 2023)
    - Ansätze eines transhumanistischen Menschenbildes, in denen die Verbesserung des fehlerbehafteten Menschen mit Hilfe von Technologie propagiert wird (Altmann 2023b; Suleyman 2023)
    - eine tendenziell kultische Verehrung von technologischem Fortschritt (WIRED 2023)
- Absehbarkeit
- Unvermeidbarkeit tiefgreifender gesellschaftlicher Veränderungen

Entwicklung einer allgemeinen künstlichen Intelligenz

- Unaufhaltbarkeit
- Absehbarkeit
  - Prognosen von Expert:innen: Entwicklung in den n\u00e4chsten Jahrzehnten oder sogar in den n\u00e4chsten Jahren (Roser 2023)
  - Aussagen CEOs: weniger als 10 Jahren (Altman, Brockman & Sutskever 2023; Leike & Sutskever 2023; Suleyman 2023)
- Unvermeidbarkeit tiefgreifender gesellschaftlicher Veränderungen

### Prognosen zu AGI (Kokotajlo 2024)

#### Comment Permalink



- 1. Probably there will be AGI soon -- literally any year now.
- 2. Probably whoever controls AGI will be able to use it to get to ASI shortly thereafter -- maybe in another year, give or take a year.
- 3. Probably whoever controls ASI will have access to a spread of powerful skills/abilities and will be able to build and wield technologies that seem like magic to us, just as modern tech would seem like magic to medievals.
- 4. This will probably give them godlike powers over whoever doesn't control ASI.
- 5. In general there's a lot we don't understand about modern deep learning. Modern Als are trained, not built/programmed. We can theorize that e.g. they are genuinely robustly helpful and honest instead of e.g. just biding their time, but we can't check.
- 6. Currently no one knows how to control ASI. If one of our training runs turns out to work way better than we expect, we'd have a rogue ASI on our hands. Hopefully it would have internalized enough human ethics that things would be OK.
- 7. There are some reasons to be hopeful about that, but also some reasons to be pessimistic, and the literature on this topic is small and pre-paradigmatic.
- 8. Our current best plan, championed by the people winning the race to AGI, is to use each generation of AI systems to figure out how to align and control the next generation.
- 9. This plan might work but skepticism is warranted on many levels.
- 10. For one thing, there is an ongoing race to AGI, with multiple megacorporations participating, and only a small fraction of their compute and labor is going towards alignment & control research. One worries that they aren't taking this seriously enough.

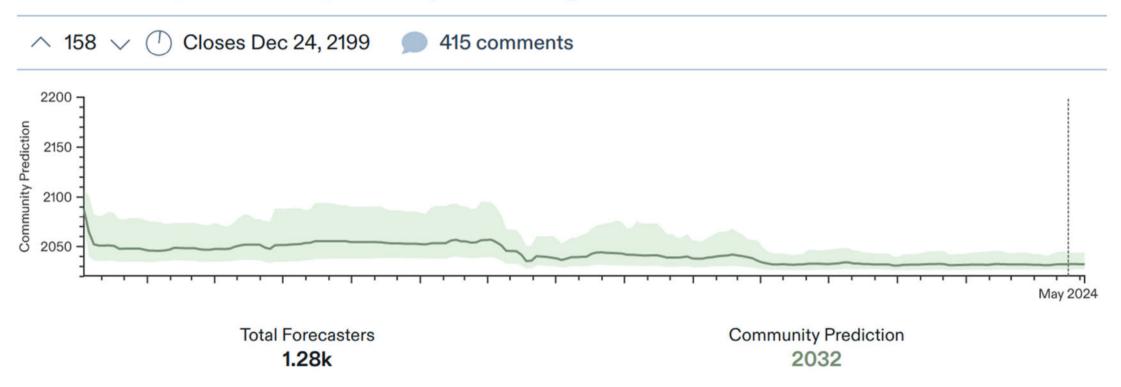
A1 \( □ \( □ \)

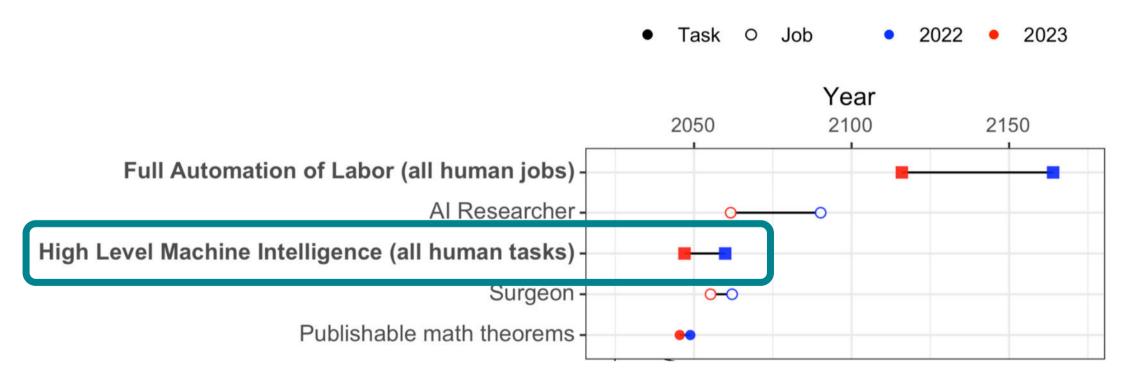
Reply

# When will the first general AI system be devised, tested, and publicly announced?

**Mar** 5, 2032

2.92k predictions





"I would not be surprised if we saw systems nearing that kind of capability [AGI] within the next decade or sooner".



https://youtu.be/nwUARJeepIA?t=1910

(Demis Hassabis 2024)

Entwicklung einer allgemeinen künstlichen Intelligenz

- Unaufhaltbarkeit
- Absehbarkeit
- Unvermeidbarkeit tiefgreifender gesellschaftlicher Veränderungen
  - "Three crucial consequences follow:
    - 1. This revolution will create phenomenal wealth. The price of many kinds of labor (which drives the costs of goods and services) will fall toward zero once sufficiently powerful AI, joins the workforce.
    - 2. The world will change so rapidly and drastically that an equally drastic change in policy will be needed to distribute this wealth and enable more people to pursue the life they want
    - 3.If we get both of these right, we can improve the standard of living for people more than we ever have before." (Altman 2021)

"Novel capabilities often emerge in more powerful models. Some that are particularly concerning are the ability to create and act on long-term plans, to accrue power and resources ('power-seeking'), and to exhibit behavior that is increasingly 'agentic'."

(OpenAl 2023)

"As AI systems improve, it is becoming increasingly difficult to rule out that models might be able to autonomously gain resources and evade human oversight."

(Alignment Research Center 2023)

#### **GPT-4 Technical Report**

#### OpenAI\*

#### Abstract

We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated but exam with a score around the top 10% of text takers, GPT-4 is a Transformer-based model pre-trained to predict the next token in a document. The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior. A core component of this project was developing infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to accurately predict some aspects of GPT-4's performance based on models trained with no more than 1/1,000th the compute of GPT-4.

#### 1 Introduction

This technical report presents GPT-4, a large multimodal model capable of processing image and text inputs and producing text outputs. Such models are an important area of study as they have the potential to be used in a wide range of applications, such as dialogue systems, text summarization, and machine translation. As such, they have been the subject of substantial interest and progress in recent years [1–34].

One of the main goals of developing such models is to improve their ability to understand and generate natural language text, particularly in more complex and nuanced scenarios. To test its capabilities in such scenarios, GPT-4 was evaluated on a variety of exams originally designed for humans. In these evaluations it performs quite well and often outscores the vast majority of human test takers. For example, on a simulated bar exam, GPT-4 achieves a score that falls in the top 10% of test takers. This contrasts with GPT-3.5, which scores in the bottom 10%.

On a suite of traditional NLP benchmarks, GPT-4 outperforms both previous large language models and most state-of-the-art systems (which often have benchmark-specific training or hand-engineering). On the MMLU benchmark [153, 36], an English-language suite of multiple-choice questions covering 57 subjects, GPT-4 not only outperforms existing models by a considerable margin in English, but also demonstrates strong performance in other languages. On translated variants of MMLU, GPT-4 surpasses the English-language state-of-the-art in 24 of 26 languages considered. We discuss these model capability results, as well as model safety improvements and results, in more detail in later sections.

This report also discusses a key challenge of the project, developing deep learning infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to make predictions about the expected performance of GPT-4 (based on small runs trained in similar ways) that were tested against the final run to increase confidence in our training.

Despite its capabilities, GPT-4 has similar limitations to earlier GPT models [1, 37, 38]: it is not fully reliable (e.g. can suffer from "hallucinations"), has a limited context window, and does not learn

#### Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

#### Microsoft Research

#### Abstract

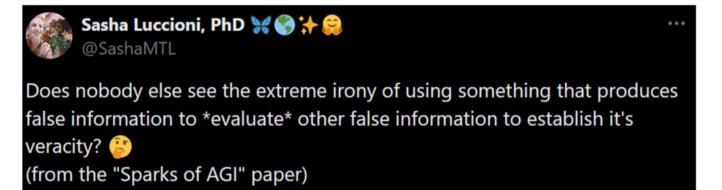
Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4 [Ope23], was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards deeper and more comprehensive versions of AGI, including the possible need for pursuing a new paradigm that moves beyond next-word prediction. We conclude with reflections on societal influences of the recent technological leap and future research directions

#### Contents

1.1																																							4
1.1															6																								
1.2	Organ	niza	tion	of	our	dem	ons	tra	tie	on																													8
Mu	ltimod	lal	and	l in	tere	lisci	plir	na	ry	C	on	ip	05	it	io	n																							13
2.1	Integra	rati	ve a	bili	ty.				+						+									. +												4			13
2.2	Vision	1 .							+																											+			16
	2.2.1	I	nage	e ge	nera	tion	bey	ron	nd	m	em	101	riz	ati	on	1																							16
	2.2.2	I	nage	e ge	nera	tion	foll	ov	vir	ıg (	det	ai	le	1 i	ns	tr	uc	tic	ns	; (	à	la	D	all	-E	)													17
	2.2.3	F	ossi	ble	appl	licati	on i	in	sk	etc	h	ge	ne	rai	tio	n																							18
2.3	Music								٠	٠.																													19
Cod	ling																																						21
3.1	From i	ins	truc	tion	as to	cod	е.	-																		*													21
	3.1.1	(	odin	ng c	hall	enge	5 .																																21
	3.1.2																																						22
3.2	Under																																						26
	Mu 2.1 2.2 2.3 Coo 3.1	1.2 Organ  Multimod 2.1 Intege 2.2 Vision 2.2.1 2.2.2 2.2.3 Music  Coding 3.1 From 3.1.1 3.1.2	1.2 Organiza  Multimodal 2.1 Integrati 2.2 Vision - 2.2.1 In 2.2.2 In 2.2.3 P 2.3 Music -  Coding 3.1.1 From ins 3.1.2 R	1.2 Organization  Multimodal and 2.1 Integrative a 2.2 Vision 2.2.1 Imag 2.2.2 Imag 2.2.3 Music  Coding 3.1 From instru 3.1.1 Codi 3.1.2 Real	1.2 Organization of   Multimodal and in	1.2 Organization of our	1.2 Organization of our dem   Multimodal and interdisci   2.1 Integrative ability	Multimodal and interdisciplin 2.1 Integrative ability	1.2 Organization of our demonstra  Multimodal and interdisciplina 2.1 Integrative ability	Multimodal and interdisciplinary 2.1 Integrative ability	1.2 Organization of our demonstration  Multimodal and interdisciplinary color of the property	Multimodal and interdisciplinary con 2.1 Integrative ability	1.2 Organization of our demonstration	Multimodal and interdisciplinary compos 2.1 Integrative ability 2.2 Vision 2.2.1 Image generation beyond memoriz 2.2.2 Image generation following detaile 2.2.3 Possible application in sketch gene 2.3 Music  Coding 3.1 From instructions to code 3.1.1 Coding challenges 3.1.2 Real world scenarios	Multimodal and interdisciplinary composit  1.1 Integrative ability 2.2 Vision 2.2.1 Image generation beyond memorizati 2.2.2 Image generation following detailed i 2.2.3 Possible application in sketch general  3.1 From instructions to code 3.1.1 Coding challenges 3.1.2 Real world scenarios	Multimodal and interdisciplinary composition  1.1 Integrative ability  2.2 Vision  2.2.1 Image generation beyond memorization  2.2.2 Image generation following detailed ins  2.2.3 Possible application in sketch generation  2.3 Music  Coding  3.1 From instructions to code  3.1.1 Coding challenges  3.1.2 Real world scenarios	1.2 Organization of our demonstration  Multimodal and interdisciplinary composition 2.1 Integrative ability 2.2 Vision 2.2.1 Image generation beyond memorization 2.2.2 Image generation following detailed instr 2.2.3 Possible application in sketch generation 2.3 Music  Coding 3.1 From instructions to code 3.1.1 Coding challenges 3.1.2 Real world scenarios	Multimodal and interdisciplinary composition  1.1 Integrative ability 2.2 Vision 2.2.1 Image generation beyond memorization 2.2.2 Image generation following detailed instruc 2.2.3 Possible application in sketch generation 2.3 Music  Coding 3.1 From instructions to code 3.1.1 Coding challenges 3.1.2 Real world scenarios	1.2 Organization of our demonstration  Multimodal and interdisciplinary composition  2.1 Integrative ability	Multimodal and interdisciplinary composition  1.1 Integrative ability 2.2 Vision 2.2.1 Image generation beyond memorization 2.2.2 Image generation following detailed instructions 2.2.3 Possible application in sketch generation 2.2.3 Music  Coding 3.1 From instructions to code 3.1.1 Coding challenges 3.1.2 Real world scenarios	1.2 Organization of our demonstration  Multimodal and interdisciplinary composition 2.1 Integrative ability 2.2 Vision 2.2.1 Image generation beyond memorization 2.2.2 Image generation following detailed instructions (2.2.3 Music  Coding 3.1 From instructions to code 3.1.1 Coding challenges 3.1.2 Real world scenarios	1.2 Organization of our demonstration  Multimodal and interdisciplinary composition 2.1 Integrative ability 2.2 Vision 2.2.1 Image generation beyond memorization 2.2.2 Image generation following detailed instructions (à) 2.2.3 Possible application in sketch generation 2.3 Music  Coding 3.1 From instructions to code 3.1.1 Coding challenges 3.1.2 Real world scenarios	1.2 Organization of our demonstration  Multimodal and interdisciplinary composition 2.1 Integrative ability . 2.2 Vision . 2.2.1 Image generation beyond memorization 2.2.2 Image generation following detailed instructions (à la 2.2.3 Music .  Coding 3.1 From instructions to code . 3.1.1 Coding challenges . 3.1.2 Real world scenarios .	1.2 Organization of our demonstration  Multimodal and interdisciplinary composition 2.1 Integrative ability 2.2 Vision 2.2.1 Image generation beyond memorization 2.2.2 Image generation following detailed instructions (à la D. 2.2.3 Possible application in sketch generation 2.3 Music  Coding 3.1 From instructions to code 3.1.1 Coding challenges 3.1.2 Real world scenarios	1.2 Organization of our demonstration  Multimodal and interdisciplinary composition 2.1 Integrative ability 2.2 Vision 2.2.1 Image generation beyond memorization 2.2.2 Image generation following detailed instructions (à la Dall 2.2.3 Possible application in sketch generation 2.3 Music  Coding 3.1 From instructions to code 3.1.1 Coding challenges 3.1.2 Real world scenarios	1.2 Organization of our demonstration  Multimodal and interdisciplinary composition 2.1 Integrative ability 2.2 Vision 2.2.1 Image generation beyond memorization 2.2.2 Image generation following detailed instructions (à la Dall-E 2.2.3 Possible application in sketch generation 2.3 Music  Coding 3.1 From instructions to code 3.1.1 Coding challenges 3.1.2 Real world scenarios	1.2 Organization of our demonstration  Multimodal and interdisciplinary composition 2.1 Integrative ability	Multimodal and interdisciplinary composition  1.1 Integrative ability 2.2 Vision 2.2.1 Image generation beyond memorization 2.2.2 Image generation following detailed instructions (à la Dall-E) 2.2.3 Possible application in sketch generation 2.3 Music  Coding 3.1 From instructions to code 3.1.1 Coding challenges 3.1.2 Real world scenarios	1.2 Organization of our demonstration  Multimodal and interdisciplinary composition 2.1 Integrative ability 2.2 Vision 2.2.1 Image generation beyond memorization 2.2.2 Image generation following detailed instructions (à la Dall-E) 2.2.3 Possible application in sketch generation 2.3 Music  Coding 3.1 From instructions to code 3.1.1 Coding challenges 3.1.2 Real world scenarios	1.2 Organization of our demonstration  Multimodal and interdisciplinary composition 2.1 Integrative ability 2.2 Vision 2.2.1 Image generation beyond memorization 2.2.2 Image generation following detailed instructions (à la Dall-E) 2.2.3 Possible application in sketch generation 2.3 Music  Coding 3.1 From instructions to code 3.1.1 Coding challenges 3.1.2 Real world scenarios	1.2 Organization of our demonstration  Multimodal and interdisciplinary composition 2.1 Integrative ability 2.2 Vision 2.2.1 Image generation beyond memorization 2.2.2 Image generation following detailed instructions (à la Dall-E) 2.2.3 Possible application in sketch generation 2.3 Music  Coding 3.1 From instructions to code 3.1.1 Coding challenges 3.1.2 Real world scenarios	1.2 Organization of our demonstration  Multimodal and interdisciplinary composition 2.1 Integrative ability 2.2 Vision 2.2.1 Image generation beyond memorization 2.2.2 Image generation following detailed instructions (à la Dall-E) 2.2.3 Possible application in sketch generation 2.3 Music  Coding 3.1 From instructions to code 3.1.1 Coding challenges 3.1.2 Real world scenarios	1.2 Organization of our demonstration  Multimodal and interdisciplinary composition 2.1 Integrative ability . 2.2 Vision . 2.2.1 Image generation beyond memorization . 2.2.2 Image generation following detailed instructions (à la Dall-E) . 2.2.3 Possible application in sketch generation .  Coding 3.1 From instructions to code . 3.1.1 Coding challenges . 3.1.2 Real world scenarios .	1.2 Organization of our demonstration  Multimodal and interdisciplinary composition 2.1 Integrative ability 2.2 Vision 2.2.1 Image generation beyond memorization 2.2.2 Image generation following detailed instructions (à la Dall-E) 2.2.3 Possible application in sketch generation 2.3 Music  Coding 3.1 From instructions to code 3.1.1 Coding challenges 3.1.2 Real world scenarios	1.2 Organization of our demonstration  Multimodal and interdisciplinary composition 2.1 Integrative ability 2.2 Vision 2.2.1 Image generation beyond memorization 2.2.2 Image generation following detailed instructions (à la Dall-E) 2.2.3 Possible application in sketch generation 2.3 Music  Coding 3.1 From instructions to code 3.1.1 Coding challenges 3.1.2 Real world scenarios	1.2 Organization of our demonstration  Multimodal and interdisciplinary composition 2.1 Integrative ability 2.2 Vision 2.2.1 Image generation beyond memorization 2.2.2 Image generation following detailed instructions (à la Dall-E) 2.2.3 Possible application in sketch generation 2.3 Music  Coding 3.1 From instructions to code 3.1.1 Coding challenges 3.1.2 Real world scenarios	1.2 Organization of our demonstration  Multimodal and interdisciplinary composition 2.1 Integrative ability 2.2 Vision 2.2.1 Image generation beyond memorization 2.2.2 Image generation following detailed instructions (à la Dall-E) 2.2.3 Possible application in sketch generation 2.3 Music  Coding 3.1 From instructions to code 3.1.1 Coding challenges 3.1.2 Real world scenarios	1.2 Organization of our demonstration  Multimodal and interdisciplinary composition 2.1 Integrative ability 2.2 Vision 2.2.1 Image generation beyond memorization 2.2.2 Image generation following detailed instructions (à la Dall-E) 2.2.3 Possible application in sketch generation 2.3 Music  Coding 3.1 From instructions to code 3.1.1 Coding challenges 3.1.2 Real world scenarios	1.2 Organization of our demonstration  Multimodal and interdisciplinary composition 2.1 Integrative ability

"Auch OpenAI selbst und Google veröffentlichen immer wieder Forschungsberichte, die den Namen nicht verdient haben Denn die KI-Unternehmen halten Datengrundlagen und wichtige Spezifikationen ihrer Modelle zurück. Sie geben lange Autorinnenlisten an und spicken die Texte mit Fachbegriffen, um darüber hinwegzutäuschen, dass es eigentlich vor allem PR-Mitteilungen sind." (Kilg 2024)

<sup>\*</sup>Please cite this work as "OpenAI (2023)". Full authorship contribution statements appear at the end of the document. Correspondence regarding this technical report can be sent to gpt4-report@openai.com



### 7.2.2 GPT-4 as a Judge

To mitigate some of the aforementioned limitations of the similarity metrics, we utilize GPT-4 (itself) to determine relevance of the response; we refer to this approach as *Judge GPT-4*. GPT-4 is asked to determine if the response generated using GPT-4 is more similar to the reference answer, or the one generated by GPT-3.

"Die Kernaussage des Papiers – "GPT-4 erreicht eine Form allgemeiner Intelligenz [wie] seine grundlegenden geistigen Fähigkeiten (wie logisches Denken, Kreativität und Deduktion) zeigen" – kann buchstäblich nicht ernsthaft geprüft werden, da die wissenschaftliche Gemeinschaft keinen Zugriff auf die Trainingsdaten hat. Alles muss auf Treu und Glauben hingenommen werden (und dennoch gab es bereits Berichte über Verunreinigungen der Trainingsdaten)." (Gary Marcus 2023)

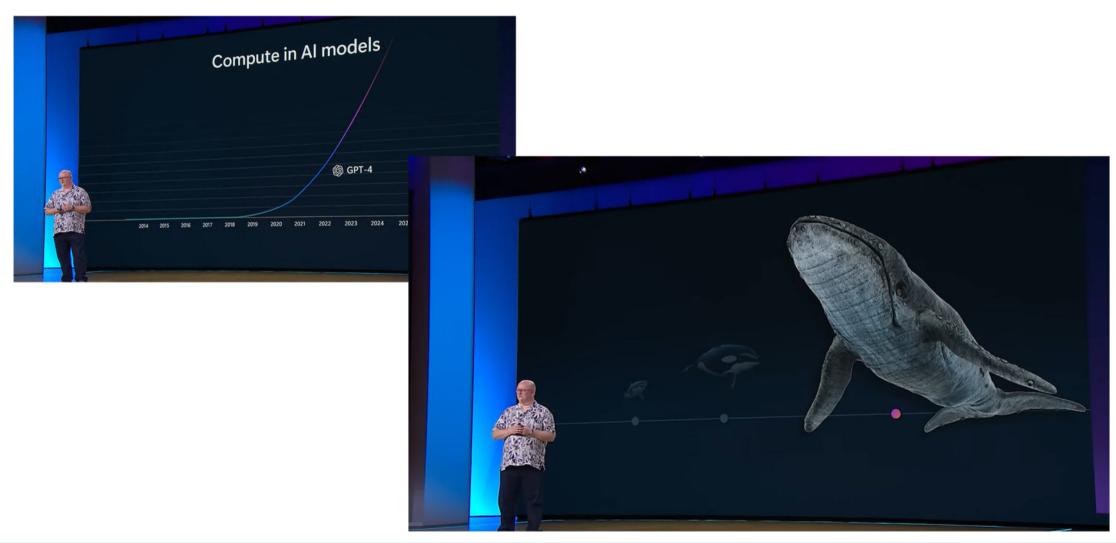
https://x.com/SashaMTL/status/1639011882173452289 https://garymarcus.substack.com/p/the-sparks-of-agi-or-the-end-of-science

"I can see it. I can see how AGI will be built. (...) I can basically tell you the cluster AGI will be trained on and when it will be built, the rough combination of algorithms we'll use, the unsolved problems and the path to solving them, the list of people that will matter. I can see it."



(Leopold Aschenbrenner 2024)

# Prognosen zu AGI (Microsoft 2024)



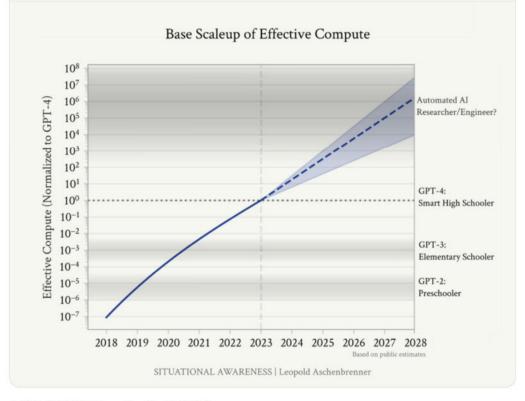
### Prognosen zu AGI (Marcus 2024)

- GPT-4 =/= intelligenter Highschool-Schüler (manches besser, manches schlechter, sehr weniges zuverlässig)
- qualitative, ungelöste Probleme (Halluzinationen, Planung, Argumentation usw.), möglicherweise eingeschränkte Nützlichkeit synthetischer Daten, Grenzen der Energieverteilung und Rechenleistung im Widerspruch zu kontinuierlichem Fortschritt
- seit August 2022 GPT-4; seither kein Modell, mit dem der Fortschritt 2020-2022 wiederholt werden konnte

•

**⑤ Leopold Aschenbrenner ⓒ** @leopoldasch ⋅ 23h AGI by 2027 is strikingly plausible.

That doesn't require believing in sci-fi; it just requires believing in straight lines on a graph.



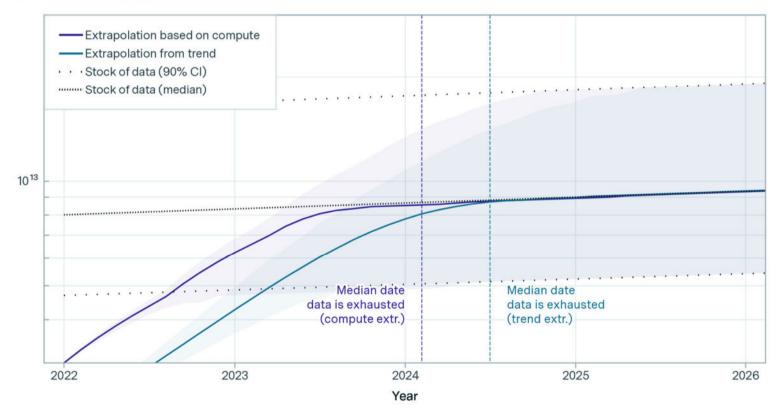
17:07 · 6/4/24 From Earth · **3.1M** Views

### Prognosen zu AGI (Villalobos et al. 2022)

### High-quality language data



Number of words (log)



### Prognosen zu AGI (Shumailov et al. 2024)

Example of text outputs of an OPT-125m model affected by *Model Collapse*– models degrade over generations, where each new generation is trained on data produced by the previous generation.

Input: some started before 1360 — was typically accomplished by a master mason and a small team of itinerant masons, supplemented by local parish labourers, according to Poyntz Wright. But other authors reject this model, suggesting instead that leading architects designed the parish church towers based on early examples of Perpendicular

### **Outputs:**

- **Gen 0:** Revival architecture such as St. John's Cathedral in London. The earliest surviving example of Perpendicular Revival architecture is found in the 18th @-@ century Church of Our Lady of Guernsey, which dates from the late 19th century. There are two types of perpendicular churches: those
- Gen 1: architecture such as St. Peter's Basilica in Rome or St. Peter's Basilica in Buenos Aires. There is no evidence that any of these buildings were built during the reign of Pope Innocent III, but it is possible that they may have been built during the reign of his successor, Pope Innocent
- **Gen 7:** architecture in England. In an interview with The New York Times, Wright said: "I don't think there is anything wrong with me being able to do what I want to do. It just doesn't work for me. "He added: "I don't know if you can call it funny,
- **Gen 9:** architecture. In addition to being home to some of the world's largest populations of black @-@ tailed jackrabbits, white @-@ tailed jackrabbits, blue @-@ tailed jackrabbits, red @-@ tailed jackrabbits, yellow @-

A typical challenge in continual learning is that the model forgets previous samples when learning new information; this is known as *catastrophic forgetting* [Kirkpatrick et al., 2017]. A typical way of preventing it is to use regularisations (Memory Aware Synpass [Aljundi et al., 2018]) or just rely on data (*e.g.* Learning without Forgetting [Li and Hoiem, 2017]). This has an indirect connection to our work, yet differs since the data in the process of *model collapse* are generated by different generations of models.

# Eine Welt mit allgemeiner Kl...

Inhalt

- 1. Allgemeine KI (AGI)
- 2. Prognosen zu AGI
- 3. Szenarien einer (Hochschul-)Welt mit AGI



(Gary Marcus 2024)



# Szenarien einer (Hochschul-)Welt mit AGI (Tegmark 2017)

#### Libertarian Utopia

Humans, cyborgs, uploads and superintelligences coexist peacefully thanks to property riahts.

#### **Benevolent Dictator**

Everybody knows that the Al runs society and enforces strict rules, but most people view this as a good thing.

### Egalitarian Utopia

Humans, cyborgs and uploads coexist peacefully thanks to property abolition and guaranteed income.

### Gatekeeper

A superintelligent AI is created with the goal of interfering as little as necessary to prevent lament their fate. the creation of another superintelligence. As a result, helper robots with slightly subhuman 1984 intelligence abound, and human-machine cyborgs exist, but technological progress is forever stymied.

### **Protector God**

Essentially omniscient and omnipotent Al maximizes human happiness by intervening only in ways that preserve our feeling of control of our own destiny and hides well enough technological society in the style of the Amish. that many humans even doubt the Al's existence.

#### **Enslaved God**

A superintelligent AI is confined by humans, who use it to produce unimaginable technology and wealth that can be used for good or bad depending on the human controllers.

#### Conquerors

Al takes control, decides that humans are a threat/nuisance/waste of resources and gets rid of us by a method that we don't even understand.

#### Descendants

Als replace humans, but give us a graceful exit, making us view them as our worthy descendants, much as parents feel happy and proud to have a child who's smarter than them, who learns from them, and then accomplishes what they could only dream of even if they can't live to see it all.

### Zookeeper

An omnipotent AI keeps some humans around, who feel treated like zoo animals and

Technological progress toward superintelligence is permanently curtailed not by an AI but by a human-led Orwellian surveillance state where certain kinds of AI research are banned.

### Reversion

Technological progress toward superintelligence is prevented by reverting to a pre-

#### Self-destruction

Superintelligence is never created because humanity drives itself extinct by other means (say nuclear and/or biotech mayhem fueled by climate crisis).

		Gesellschaftliche Basis							
		gegeben	nicht gegeben						
Institutionelle	hoch	Szenario 1 "Pro-Forma Hochschule"	Szenario 3						
Trägheit	gering	Szenario 2 "Transformierte und transformierende Hochschule"	"Das Ende der akademischen Bildung"						

Szenario 1: "Die pro-forma-Hochschule"

Akademische Bildung im Zeitalter der AGI

- funktionale Obsoleszenz: schneller, effizienter und flexibler (Aus-)Bildung durch KI-Systeme
- Besuch der Hochschulen lediglich zur ritualisierten Distinktion der Nutzer:innen

Szenario 2: "Die transformierte und transformierende Hochschule"

Akademische Bildung im Zeitalter der AGI

- Verlust der Qualifikationsfunktion f
   ür fachliche Berufs- oder Wissenschaftsfelder
- Ausprägung von breiten Bildungs- und Transformationsfunktionen
- mögliche Ausgestaltung
  - Hochschule als Ort des menschlichen Logos
  - Hochschule als Low-Tech-Ort des zwischenmenschlichen Diskurses
  - Hochschule als Ort der Moderationsausbildung
  - Hochschule als Ort der Philosophie

Szenario 3: "Das Ende der akademischen Bildung"

### Vakanzen

- gesellschaftlicher Bedarf an akademischer Qualifikation
  - künstliche Intelligenz ersetzt menschliche Intelligenz in den komplexen Lebens- und Arbeitswelten
  - funktionale Entscheidungen durch persönliche KI-Assistenten, auch ohne akademische Ausbildung
  - veränderte physische und psychische Konstitution des Menschen (transhumanistische Ansätze)
- gesellschaftliche Ressourcenbereitstellung für akademische Qualifikation
  - Alignement als Grundlage gesellschaftlichen Fortbestehens
  - Gebrauch von AGI zu militärischen, nachrichtendienstlichen oder terroristischen Zwecken
  - Transformation ökonomischer und sozialer Strukturen





# VIELEN DANK FÜR DEN AUSTAUSCH



Stefan Müller

WHZ/HDS

stefan.mueller@hd-sachsen.de