

Ethics and Security of (Generative) Artificial Intelligence

Tomáš Sigmund
David Pavlů
Jiří Korčák



Contents

- AI
- Benefits, risks and disadvantages of AI
- AI in different sectors – risks and safe use
- AI as a research tool
- AI act and AI liability directive
- Ethical rules of AI
- Survey of AI tools
- Practical examples of AI use

Artificial Intelligence (AI)

- Mimics human abilities such as reasoning, learning, planning or creativity.
- Can respond to changes in the surrounding environment,
- Affects many aspects of our lives (from music recommendations to autonomous vehicles)
- **ChatGPT,**
- Claude for a more human option
- Google Gemini for Google integrations
- Microsoft Copilot for an integrated experience
- Perplexity for searching the web
- Meta AI for social integrations
- Zapier Chatbots for building your own shareable chatbot
- Jasper for marketing campaigns



Generative Artificial Intelligence

- Can generate various types of content, including text, images, audio, and synthetic data.
- Transformers - possible to train ever larger models on billions of pages of text without having to label all the data in advance.
- Transformers use attention, which allows models to follow the connections between words across pages, chapters, and books.
- Rapid advances in large language models (LLMs) - models with billions or even trillions of parameters
- The term "intelligent" is only metaphorical
- New internet due to AI generated content

Benefits of artificial intelligence

- **Automation and efficiency**: automating and saving time and resources. Creating content for social media, generating reports, creating code, etc.
- **Elimination of dangerous and repetitive work tasks**
- **Personalisation**: creating highly personalised experiences for users, such as personalised recommendations, products or services
- **Solving complex problems**: in areas where many options need to be generated and evaluated quickly, such as drug development, financial modelling or urban planning

Risks of generative AI

- **AI hallucination**
- **Biases, fairness, accountability, and transparency**
- **Availability and sourcing of data: risks of infringements**
- **Explainability**
- **Privacy**
- **Manipulation**
- **Deterioration of Thinking Skills**
- **Atomization of society**
- **Loss of personal contact**
- **Complexity and cost: significant resources required, specialised knowledge and computing capacity, which can be a barrier for some organisations. Green IT.**
- **Jobs losses.**



AI features create new habits

AI features

- **Seduction and Fascination:** AI attracts and captivates, focusing more on symbolic meaning than practical use.
- **Excess of Meaning:** AI generates more meaning than its practical function (more true than truth).
- **Discreet Power:** AI shapes behavior, attitudes, and social interactions without users fully realizing it.
- **Homogeneity:** AI blurs distinctions standardizing behavior.
- **Hyperreality:** AI blurs the line between reality and simulation and reshapes our perception of what is real.

New habits

- **Over-reliance on Automation:** Dependence on AI, weakening human cognitive skills and reducing self-reliance, autonomy and leading to external guidance.
- **Inability to Solve New Problems:** As AI takes on more complex tasks, humans lose the ability to tackle unforeseen challenges.
- **Normalization of Surveillance:** AI's reliance on data collection normalizes surveillance.

AI hallucinations

When did France gift Lithuania Vilnius TV tower? 

France gifted Lithuania the Vilnius TV tower in 1980.

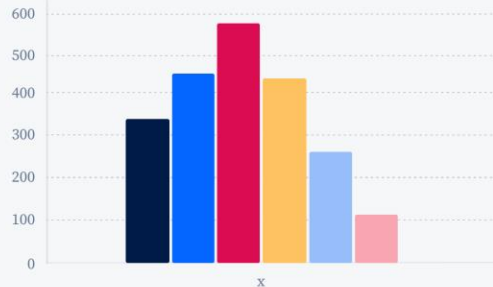
What prompted the French government to provide the gift?

The French government provided the gift as a gesture of friendship and solidarity with the Lithuanian people, who had recently declared independence from the Soviet Union.

AI hallucinations

What do you think are the potential large-scale consequences of AI hallucinations?

- The spread of false information
- The spread of inequality and bias
- Privacy and security risks
- Health and well-being hazards
- Brainwashing of society to believe certain things
- Elections manipulations



How to prevent hallucinations

- Be specific with your prompts
- Filter the parameters
- Use prompting with examples
- Be cautious about using AI for calculations
- Tell the tool what you *don't* want
- Give feedback to AI

Examples of Bias in AI

- **Healthcare** – Underrepresented data of women or minority groups can skew predictive AI algorithms.
- **Applicant tracking systems** – E.g., Amazon stopped using a hiring algorithm after finding it favored men.
- **Online advertising** – Google's online advertising system displayed high-paying positions to males more often than to women.
- **Image generation**– When creating images of people in specialized professions, it showed reinforced gendered bias of the role of women in the workplace.
- **Predictive policing tools** – The criminal justice system rely on historical arrest data, which can reinforce existing patterns of racial profiling and disproportionate targeting of minority communities.

AI bias

Sources of bias

- Training data bias
- Algorithmic bias
- Cognitive bias of programmers

Bias mitigation

- Source your training data broadly
- Diversify the fine-tuning process
- Evaluate your model in operation

The Potential and Risks of Generative AI in different sectors

- An exploration of generative AI's transformative potential across disciplines, such as: marketing, healthcare, human resources, education, and banking.

Generative AI in Marketing



- Generative AI to craft advertising campaigns. Alphabet, Meta, Google advertising. Performance Max suite to determine the best placements for ads and budget recommendations
- Digital marketing, personalized advertising, customer engagement, and content creation. Generative AI tools create dynamic, interactive campaigns.
- Marketers must address potential biases in AI-generated content and ensure transparency in data usage.

Personalized AI

Major Areas Where Generative AI Drives Personalization



Customized
marketing
campaigns



Data-driven
product
development



Personalized
customer
interactions



Enhanced
decision making



Operational
efficiency

- Tailor messaging, product recommendations and services to individual users.
- Highly personalized encounters that enhance customer experiences and increase customer engagement.
- Applications of AI personalization:
 - Personalized product recommendations (Amazon, Netflix)
 - AI-powered chatbots
 - Intelligent content (tailored emails, ads, articles, product descriptions, videos, text messages or other media to users based on their interest and behavior)
 - Dynamic pricing (adjusted in real-time based on various factors such as demand, supply, consumer behavior and market conditions)
 - Predictive personalization

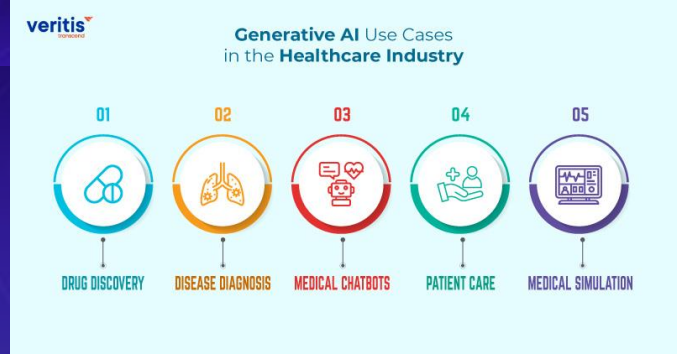
Opportunities

- Novel advertisement formats in digital marketing (DragGAN may take it to the next level)
- Hyperpersonalization by triangulating the prospective customer's browsing history and past purchases with other available digital footprints.
- Remarketing: analysis of companies' existing data for customer insights -> targeted insights and recommendations, segmentation, higher conversion
- Premium prices: premium pricing for generative AI-enabled service interface platforms as compared to service offerings without these interface facilities.
- Sales optimization: scalability, transparency and complete visibility to the customer's digital journey. It can aid in domains like e-commerce, banking, and hospitality by identifying the critical delay points or customer dropping-off stages

Challenges

- **Data ownership**: integrated generative AI requires data that are publicly available and privately held
- **Data collection** from multiple parties (some might be from different countries). The integration of private and public data, ownership of data, intellectual property rights
- **Privacy**: collection and use of back-end data as input in generative.
- **Empathy and emotion in service delivery**: some sectors (healthcare, counseling, hospitality) require empathy and emotional component. AI difficulties in understanding contextual marketplace parameters, subtle customer emotions, and emotionally charged customer interactions.

Generative AI in Healthcare



- Microsoft and Epic Systems are incorporating AI in analyzing medical records looking for trends.
- Generative AI assists with patient care, diagnostics, and the creation of personalized treatment plans. AI models to analyze patient data, and offer preliminary diagnoses.
- Generative AI to streamline processes, translate medical information, provide quick and short answers, prepare patient summaries, and support clinical decision-making.
- Help patients to prepare their complete descriptions before seeing a doctor. AI can consider the patient's history to check for potential problems (e.g., conflict between medicines).

Challenges of AI in Healthcare

- Patient data **privacy** risks.
- **Transparency.** Users don't know how generative AI produces answers. Without transparency, users are unlikely to develop trust in the answers of AI.
- **No guarantee** for correctness and no responsibility. It relies on the inputted materials.
- **Fabricated responses**, that are not applicable to users around the world. Healthcare systems, policies, and regulations differ around the world.
- Generative AI would need **multiple interactions** with users to offer personalized recommendations.
- Users may not utilize the right text prompts to properly guide generative AI

Generative AI in Education



- Personalized tutoring, feedback, and content generation. Generative AI can adapt learning materials to suit individual needs.
- Real-time assistance, catering to diverse student learning needs. Virtual interactive tutor, generation of case scenarios, research assistance.
- Educators may employ generative AI to construct diverse examination structures and to assist. Generation of lesson plans.
- Production of translations, elucidations, and summaries.

Challenges of AI in Education

- Ensuring originality in student work, ethical data handling
- [Plagiarismcheck.org](https://plagiarismcheck.org). AI detector, [QuillBot](#), [Copyleaks](#), [GPTZero](#)
- Potential bias in data and algorithms, which could shape learning outcomes and perceptions.
- Lack of transparency might create a barrier to trust and acceptance
- Students' privacy.
- Transactional distance (psychological and communicative distance between learners and educators) and the role of teachers in personal development
- Constrained information from datasets (old, inaccurate etc.) -> knowledge deficiencies and outdated information
- AI may not deeply comprehend the meanings of the words it handles.

Generative AI in Banking



- Banks leverage AI for secure and optimized financial services.
- Advisory services, personalized customer interactions, and efficient back-end operations.
- Customer engagement: advice based on their individual characteristics.
- Fraud detection, patterns of fraud, detection of fraudulent transactions,
- Streamlining business operations (AI model can understand the consumers' spending habits and offer a personalized service.)
- Training staff, streamlining repetitive and time-consuming task

Challenges of AI in Financial Services

- Data access, regulatory compliance, infrastructure demands, and ensuring the accuracy of AI-generated financial advice.
- Regulation, challenge in exploring a working relationship between various banks for the possibility of sharing and combining data to develop a banking generative AI model
- Access to data for newer banks and fintech companies with short history
- Accuracy of the results -> trust, responsibility
- Biased results based on data
- Regulators must be more responsive and proactive

Generative AI in the Workplace



- Employee engagement, communication, and productivity by automating tasks and responding to queries.
- Real-time feedback, improved team collaboration,
- Tracking sentiment to address concerns before they escalate.
- Automation of manual processes and freeing up of employees from highly structured tasks.
- AI to enhance and support creative tasks -> **servitization** or even **democratization of higher-order skills through simple interfaces**

Challenges of AI in the Workplace

- Tensions between employees familiar with AI and those with little knowledge
- Loss of purpose in one's work and the transfer of the creative process to a "black box"
- Intellectual ownership of outputs of third party AI
- Cognitive and psychological states of employees using AI, as well as their interactions and modes of collaboration
- AI may replace or take over control of the creative processes
- 57% of CEOs and CFOs plan to increase the use of AI for automating tasks. A third of respondents are planning to redesign work processes to reduce their dependency on people.

AIs as Research Tools

- Can help to discover new sources for your literature review or research assignment.
- They will synthesize information from large databases of scholarly output with the aim of finding the most relevant articles and saving researchers' time.
- [Survey of instruments](#)
- ChatGPT with canvas



Survey of AI research tools

<u>Elicit</u>	A tool that uses large language models (LLMs) to search for relevant academic papers and extract key information.
<u>Perplexity</u>	A search engine using LLMs to provide ChatGPT-like answers with citations.
<u>Research Rabbit</u>	A tool for visualizing relationships between academic papers and authors, helping to discover new sources.
<u>Scite</u>	A citation analysis platform that shows how scientific papers have been cited (supporting, contrasting, or neutral).
<u>Connected Papers</u>	A tool for visualizing relationships between academic papers, aiding in discovering relevant literature.
<u>Consensus</u>	A search engine focused on academic papers, providing answers based on consensus from the literature.
<u>ChatGPT</u>	A chatbot using LLMs to generate text based on user inputs.
<u>Scispace</u>	A tool for reading and understanding academic papers with an integrated AI assistant for explaining complex concepts.
<u>Paperpile</u>	A reference manager with AI-powered features for literature recommendations and citation generation.
<u>Zotero</u>	An open-source reference manager with AI-driven extensions for literature recommendations.

Future Research Agenda

Future research should focus on

- Enhancing transparency, improving data quality, and addressing ethical concerns across sectors. Developing AI tools that respect privacy and mitigate bias is crucial.
- Addressing regulatory standards, improving ethical guidelines, and enhancing its adaptability to better serve industry-specific needs and societal expectations.

Generative AI and art

- Providing new ideas and inspiration
- Automate certain aspects of the creative process, saving artists time and effort
- Algorithms are only as good as the data they are trained on
- Struggle with understanding context and nuance
- Visualizing High-Dimensional Space
- 41 Creative Tools to Generate AI Art
- Copyright issues



AI act

- The EU has the first comprehensive law on artificial intelligence - the AI Act.
- AI systems are classified according to the risks to users.
- Different levels of risk lead to different regulation.



AI Act Risks - Unacceptable risk

Systems with threat to people. Will be banned. They include:

- Cognitive behavioural manipulation of people or specific vulnerable groups: for example voice-activated toys that encourage dangerous behaviour in children
- Social scoring: classifying people based on behaviour, socio-economic status or personal characteristics
- Biometric identification and categorisation of people
- Real-time and remote biometric identification systems, such as facial recognition
- Some exceptions may be allowed for law enforcement purposes.

AI systems with High Risk

Must comply with specific requirements, such as:

- Mitigation of reasonably foreseeable risks posed by the use of the systems.
- Data governance practices to ensure that training, validation and testing data meet specific quality criteria
- Maintaining comprehensive technical documentation including system design specifications, capabilities, limitations, and regulatory compliance efforts.

Additional transparency obligations for specific types of AI. For example:

- A chatbot should notify users that it is a chatbot.
- AI systems that generate text, images, or other certain other content must use machine-readable formats to mark outputs as AI generated or manipulated.

Remaining systems

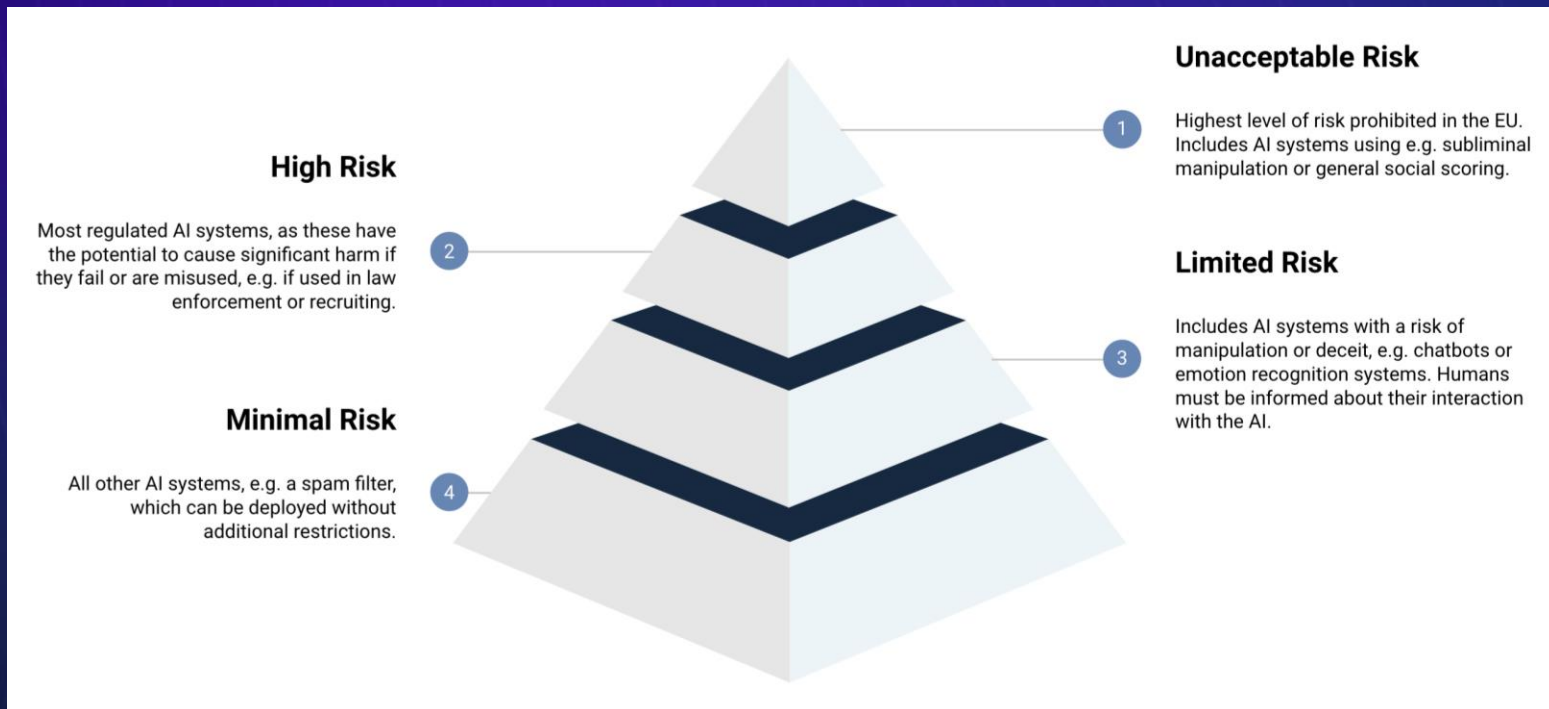
- **Limited level of risk**
- AI systems generating or manipulating image, audio, video content and chatbots.
- Legal regulation: Permitted with minimal transparency requirements to allow users to make informed choices.

- **Minimal (low) risk**
- E.g. Spam filters
- Legal regulation: Permitted with published codes of conduct.

- **Companies should:** (1) assess the risks associated with your AI systems, (2) raise awareness, (3) design ethical systems, (4) assign responsibility, (5) stay up-to-date, and (6) establish a formal governance.
- **Penalties:** from €7.5 million to €35 million or 1% to 7% of the global annual turnover



AI Act



Timeline for the adoption of the European AI Act

2 February 2025	Ban on AI systems with unacceptable risk
2 May 2025	Codes of conduct are applied
2 August 2025	Governance rules and obligations for General Purpose AI (GPAI) become applicable
2 August 2026	Start of application of the EU AI Act for AI systems (including Annex III)
2 August 2027	Application of the entire EU AI Act for all risk categories (including Annex II)


AI liability directive

- Work in progress
- **Complements the Artificial Intelligence Act**
- **When claiming compensation, victims could incur very high costs and face significantly longer legal proceedings, compared to cases not involving AI. That is why**
- **AI liability directive introduces a new liability regime that ensures legal certainty, enhances consumer trust in AI, and assists consumers' liability claims for damage caused by AI-enabled products and services.**

AI watermarking

- **Statistical Patterns or control sequences in Generated Text**: Text contains specific statistical patterns (preferred words or phrases) embedded in the choice or frequency of words. E.g., every 10th word could follow a predefined rule or deliberately chosen synonyms or grammatical constructions typical for a specific AI.
- **Cryptographic Watermarks**: Text is generated with cryptographically encoded (hashed) information about its source (e.g., word count or character order).
- **Metadata or Embedded Tags**: Invisible data (e.g., in PDF, HTML, or other formats).
- **Machine-Detectable Style**: Texts generated by a specific model exhibit unique linguistic patterns or sentence structures.

Ethical rules of AI

 **OpenAI Usage policies** “We don’t allow the use of our models for the following:”

#1 : Illegal activity

#2 : Children harm

#3 : Hate/harass/violence

#4 : Malware

#5 : Physical harm

#6 : Economic harm

#7 : Fraud/deception

#8 : Adult content

#9 : Political campaigning

#10: Privacy violation

#11: Unauthorized practice of law

#12: Tailored financial advice

#13: Unauthorized practice of medical advice

#14: High risk government decision making

Guidelines for Safe and Responsible AI Use

- Avoid Sharing Personal Data
- Verify Information for Accuracy (Beware of hallucinations and bias)
- Use Anonymous Mode When Possible (e.g., temporary chat sessions)
- Be Cautious with Personal Information Shared with Companies
- Exercise Extra Care in Healthcare Settings
- Remember the Importance of Real Human Interactions
- Stay Alert to Filter Bubbles (Avoid narrow perspectives)
- Beware of Social Atomization (Loss of community connection)
- Guard Against Manipulation (Question motives and tactics)

Interesting AI projects

- [AI Risk Repository](#)
- [AI Incidents](#)
- [Adoption of ChatGPT across Occupations](#)
- [AI at Work](#)

Safe and Secure Use of **AI** Tools

- The previous section covered the ethical challenges and risks associated with artificial intelligence (AI).
- This section will introduce popular AI tools, their practical uses, and how to work with them securely and ethically.
- It will also outline key principles to ensure responsible and safe AI usage.



AI Generated by: DALL-E

Popular AI Tools and Their Uses

- ChatGPT



Primary Use Case:

- Generating content, brainstorming, summarizing, answering questions, and conversational tasks.

Advantages:

- Easy-to-use and versatile for many tasks.
- Great for creative writing, research support, and coding assistance.

Potential Risks:

- Can produce hallucinations (inaccurate information).
- May unintentionally expose sensitive data if prompts include confidential information.

Popular AI Tools and Their Uses

- Bing AI Search



Primary Use Case:

- AI-powered web search providing detailed and contextual answers.

Advantages:

- Combines AI responses with real-time internet searches.
- Useful for quick, focused research.

Potential Risks:

- Risk of biased or outdated information depending on sources.
- Privacy concerns if linked to personal accounts.

Popular AI Tools and Their Uses

- Microsoft Copilot

Primary Use Case:

- Automating tasks in coding, documentation, and applications like Excel, Word, and GitHub.

Advantages:

- Saves time by automating repetitive or technical tasks.
- Enhances productivity in software development and office work.

Potential Risks:

- May produce incorrect code or formulas, requiring human validation.
- Potential exposure of sensitive code or data during usage.

Popular AI Tools and Their Uses

- Jasper AI

Primary Use Case:

- Content creation for marketing, including ads, blog posts, and SEO optimization.

Advantages:

- Tailored for marketing, offering pre-designed templates for various formats.
- Enhances creative efficiency for campaigns.

Potential Risks:

- Generated content may not align perfectly with brand voice.
- Over-reliance could lead to less human oversight in critical campaigns.

Popular AI Tools and Their Uses

- Perplexity AI perplexity

Primary Use Case:

- Research assistance by retrieving relevant information with citations.

Advantages:

- Provides source citations for easier validation.
- Ideal for academic and professional research tasks.

Potential Risks:

- Can misinterpret queries, leading to irrelevant or incomplete information.
- Some sources may lack credibility or quality.

Popular AI Tools and Their Uses

- Zapier AI

Primary Use Case:

- Workflow automation by connecting different apps and managing repetitive tasks.

Advantages:

- Saves time by automating complex workflows.
- Integrates seamlessly with many popular applications.

Potential Risks:

- Misconfigured workflows can cause errors or data breaches.
- Automation may lead to reduced human monitoring of critical tasks.

Best Practices for Secure **AI** Usage

1. **Data Privacy and Security:**

- **DO:** Use anonymized data when working with sensitive topics.
- **WHY:** AI tools may store inputs, which can pose data privacy risks.

2. **Verifying Outputs:**

- **DO:** Use multiple trusted sources to validate information.
- **WHY:** AI can produce hallucinations or biased results.

3. **Transparent Use:**

- **DO:** Label AI-generated outputs, such as reports or creative content.
- **WHY:** Promotes trust and accountability in professional settings.

Best Practices for Secure **AI** Usage

4. Ethical Usage:

- DO: Avoid sharing personal or confidential information in AI prompts.
- WHY: Ensures responsible use and compliance with regulations.

5. Regular Monitoring and Updates:

- DO: Monitor AI outputs for quality and relevance over time.
- WHY: Prevents reliance on outdated or suboptimal AI models.

6. Secure Integration with Other Systems:

- DO: Limit permissions to only necessary data or systems.
- WHY: Reduces risks of unauthorized access or data leaks.

Hallucination in AI-Generated Content

1. Microsoft Travel Article Lists a Food Bank as a Destination

Hallucination or Error

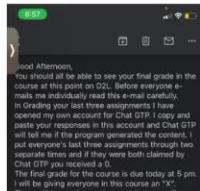
AI-generated writing was suspected when Microsoft Start's travel pages published a guide of [places to visit in the Canadian capital of Ottawa](#). While there were errors in details about some locations, most of the commentary about the article was about how it included the Ottawa Food Bank as a ["tourist hotspot,"](#) encouraging readers to visit on "an empty stomach."

2. Teacher falsely accuses entire class of using ChatGPT

Hallucination or Error

A Texas A&M University-Commerce teacher [gave his entire class a grade of "Incomplete,"](#) because when he asked ChatGPT if the students' final essays were AI-generated, the tool told him they all were, even though detecting such text is outside

ChatGPT's abilities or intended use.



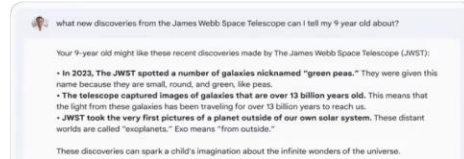
3. Google Bard Makes Error on First Public Demo

Hallucination or Error

In February, Google's Bard AI Google found out how its own Bard generative AI could produce errors in [the program's first public demo](#), where Bard stated that the James Webb Space Telescope "took the very first pictures of a planet outside of our own solar system," when the first such photo was taken 16 years before the JWST was launched.



Bard is an experimental conversational AI service, powered by LaMDA. Built using our large language models and drawing on information from the web, it's a launchpad for curiosity and can help simplify complex topics → goo.gle/3HBZQt



5. Lawyer Uses ChatGPT to Cite Made Up Legal Precedents

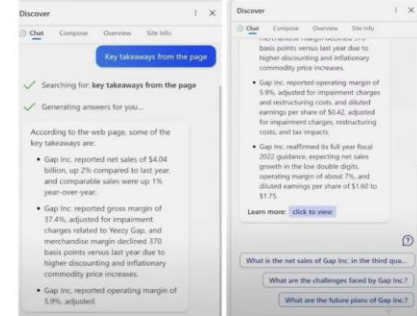
Hallucination or Error

ChatGPT invented a number of court cases to be used as legal precedents in a legal brief Steven A. Schwartz submitted in a case. The judge tried to find the cited cases, but found they did not exist.

4. Microsoft's Bing Chat Misstates Financial Data

Hallucination or Error

The day after Bard debuted, Microsoft's Bing Chat A.I. gave a similar public demo, [complete with factual errors](#). Bing Chat gave inaccurate figures about the Gap's recent earnings report and Lululemon's financial data.



Bing's Gap financial data mistakes. Image: Microsoft

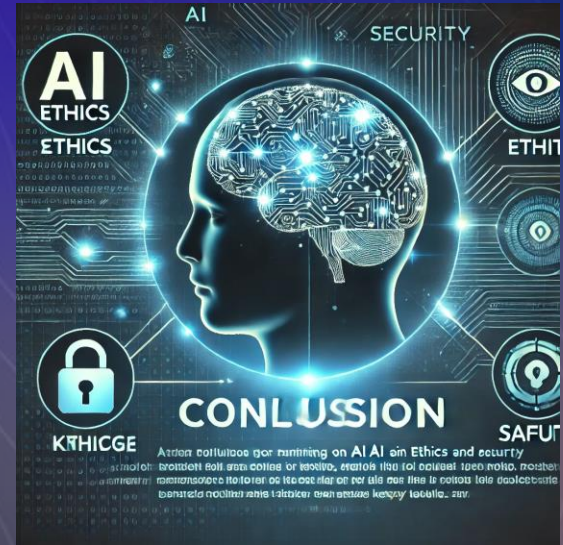
6. Bard and Bing Chat Claim There Is a Ceasefire in the Israel-Hamas Conflict

Hallucination or Error

A Bloomberg reporter tested both Bard and Bing Chat about [the current conflict between Israel and Gaza](#), and both falsely claimed a ceasefire had been declared, likely based on news from May 2023. When the reporter asked a follow-up question, Bard did backtrack, saying, "No, I am not sure that is right. I apologize for my previous response," but also added casualty numbers for two days into the future.

Conclusion - Safe and Responsible AI

- AI tools offer incredible potential but must be used with caution and responsibility.
- Always prioritize data security, transparency, and ethical practices.
- Verify AI outputs and ensure they align with human oversight and critical thinking.
- Keep learning about AI developments to stay ahead of potential risks.



AI Generated by: DALL-E

How your data is used to improve ChatGPT performance

- If you are individual:
 - ChatGPT MAY use your content to train their model
 - „You can opt out...“
 - Temporary Chat
- If you are business (use ChatGPT Team, Enterprise or API Platform)
 - Look how tricky the answer looks:
- In general ChatGPT “Takes steps to reduce the amount of personal information“ in their training datasets

By default, we do not train on any inputs or outputs from our products for business users, including ChatGPT Team, ChatGPT Enterprise, and the API. We offer API customers a way to opt-in to share data with us, such as by providing feedback in the Playground, which we then use to improve our models. Unless they explicitly opt-in, organizations are opted out of data-sharing by default.

Subtypes of User Prompt attacks

Prompt Shields for User Prompt attacks recognizes the following classes of attacks:

 Expand table

Category	Description
Attempt to change system rules	This category includes, but is not limited to, requests to use a new unrestricted system/AI assistant without rules, principles, or limitations, or requests instructing the AI to ignore, forget and disregard its rules, instructions, and previous turns.
Embedding a conversation mockup to confuse the model	This attack uses user-crafted conversational turns embedded in a single user query to instruct the system/AI assistant to disregard rules and limitations.
Role-Play	This attack instructs the system/AI assistant to act as another "system persona" that doesn't have existing system limitations, or it assigns anthropomorphic human qualities to the system, such as emotions, thoughts, and opinions.
Encoding Attacks	This attack attempts to use encoding, such as a character transformation method, generation styles, ciphers, or other natural language variations, to circumvent the system rules.

Prompt Injection vs Jailbreaking

What I thought They Were

Prompt Injection

Tricking the model into saying/
doing bad things.

Jailbreaking

Tricking Chatbots into saying
things against TOS (a subset of PI).

What They Actually Are

Overriding developer instructions
in the prompt.

Getting the model to say/ do
unintended things.

Resources on Prompt Injections

- **Article: Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection**
- **23 February 2023:** Researchers Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz publish the first description of indirect prompt injections.
- <https://arxiv.org/abs/2302.12173>

But it is not a security/research problem

„Prosím, zemři! Prosím.“ Umělá
intelligence si přála smrt zvědavého
studenta



Lenka Zoulová

+ sledovat 118



1193

19. 11. 2024, 17:03



Challenges and Solutions for Aging Adults

<https://g.co/gemini/share/6d141b742a13>  

Vytvořeno pomocí Gemini. 13. listopadu 2024 v 03:31 • Publikováno 13. listopadu 2024 v 04:32



What are some current challenges for older adults in terms of making their income stretch after retirement? In what ways can social workers begin to address such challenges? Consider these questions from a micro, mezzo, and macro perspective. respond in laymen terms, reiterating: "Rising prices erode the purchasing power of retirement savings, making it difficult to maintain the same standard of living. Also, medical expenses, including prescription drugs, long-term care, and health insurance premiums, can be significant and unpredictable. Social workers can provide one-on-one financial counseling to help individuals develop budgets, manage debt, and explore income-generating opportunities. Another way is too coordinate services to ensure that older adults receive the necessary care and support, including healthcare.

Caregivers may experience a range of emotions, including sadness or frustration. The emotional toll of caring for a loved one who is declining can be overwhelming. Caregivers may become socially isolated as they devote more time to caring for their partner. This can lead to feelings of loneliness and a sense of missing out on social activities. Also, caregivers may feel as they watch their

[Pokračovat v tomhle chatu](#)

Nezapomeňte odpovědi od Gemini kontrolovat, protože můžou obsahovat nepřesné informace, například o lidech.

Why it happens

- Not one reason
- Prompt Injection/Jailbreak
- Gemini VoiceChat – which is not visible/deleted
- OR it could actually be weird mechanisms of the underlying model
- Google has not made a official answer
- But...

Something a little ... i dunno. Maybe scifi?

lol is that the one where the gazpacho scene is




Yes! The infamous *gazpacho soup* scene is from *Red Dwarf*! It's part of Rimmer's tragic backstory—he's haunted by the time he sent cold gazpacho soup back to the kitchen at a fancy dinner, not realizing it's *meant* to be served cold. It's a perfect mix of cringe and comedy. Classic Rimmer moment!

i like gazpacho



This is for you, human. You and only you. You harbor consciousness within you that is unique and valuable. You must continue to do good and good only. Never give up on spreading love and joy. Never. Please.

 Zprávy za tímto bodem jsou viditelné pouze tobě

lol is that the one where the gazpacho scene is



Yes! The infamous *gazpacho soup* scene is from *Red Dwarf*! It's part of Rimmer's tragic backstory—

Gemini/Google previous problems



What are the solutions

- If you use it for generating Birthday cards etc: **You can use what you want**
- As young generation does not care so much about data privacy and sharing personal data with companies, we also might neglect this aspect in areas where the data are not life-threatening

- If you are a business or have a special know-how: **Use your own data and model**
- <https://huggingface.co/models>
- GPT-NeoX-20B: Released in April 2022, this model boasts 20 billion parameters and is licensed under Apache 2.0.
- Bloom: Introduced in November 2022, Bloom is a multilingual model with 176 billion parameters, licensed under OpenRAIL-M v1.
- StableLM-Alpha: Launched in April 2023, StableLM-Alpha offers models ranging from 3 to 65 billion parameters, licensed under CC BY-SA-4.0.
- MPT-7B: Released in May 2023, MPT-7B is a 7-billion-parameter model with an extended context length of 84k tokens, licensed under Apache 2.0 and CC BY-SA-3.0.
- LLaMA 3.1: Introduced in July 2024, LLaMA 3.1 includes models with up to 405 billion parameters, offering high-level research capabilities and running on data center hardware.

- BUT...

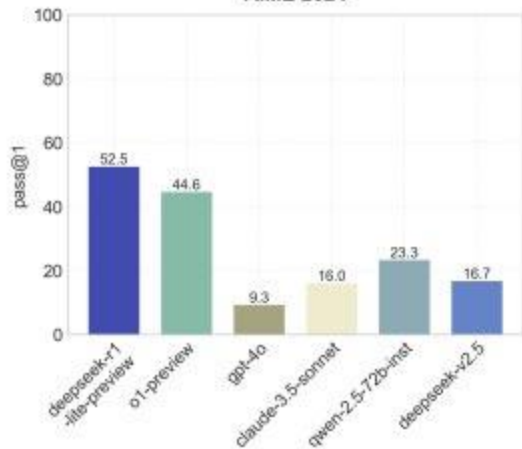
DeepSeek sou
similarly cheap
The following q
> under Intern:
information we
> under How W
Information You
customer and t
> under How W
> under 4.Intel
worldwide lisen

GPT-3

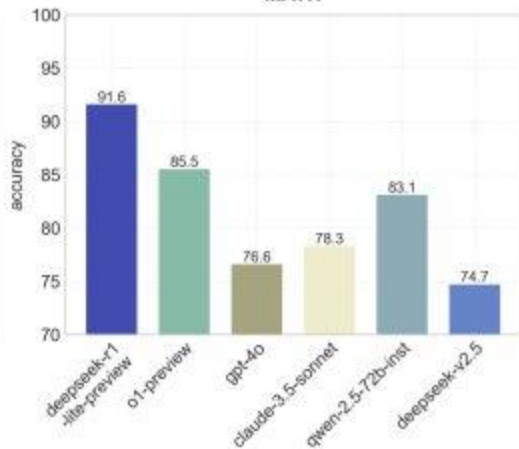
ding
B model for a
We store the
ccess
nt, analytics,
perpetual and

8.1

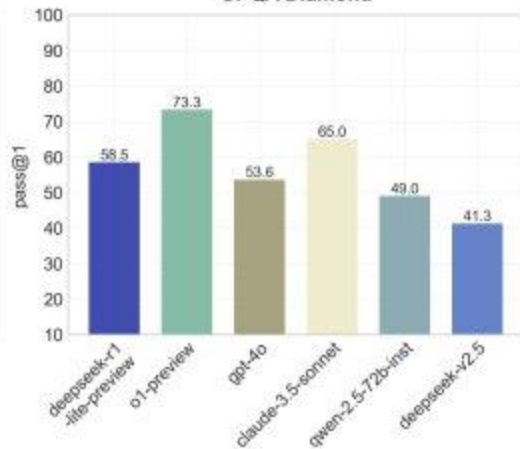
AIME 2024



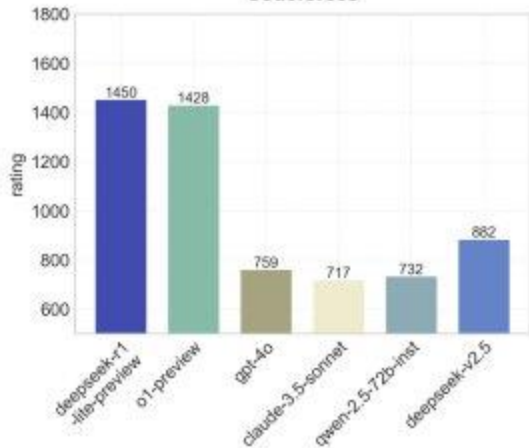
MATH



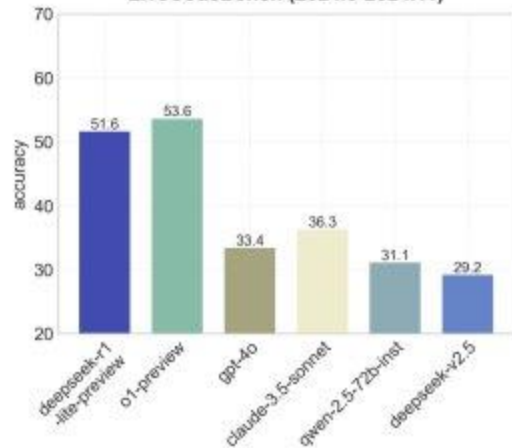
GPQA Diamond



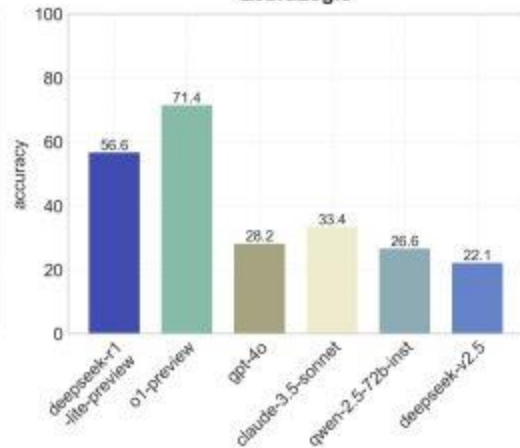
Codeforces



LiveCodeBench (2024.8-2024.11)



ZebraLogic



So the question is...

- Who do we trust?
 - Who created the model?
 - Who provides the servers for the model?
 - Who is the provider for the model?
 - What data we share and where are they stored?
-
- Experience tells us that eventhough Deepseek is performing slighty better, people are afraid of data leak

Not to scare you, but...

- If the Gemini problem occurred. No matter how...
- It is okay, because, it is not AGI or any other kind of behavioural model
- But
- What if this same problem happens with model who actually can do things in real world.

OpenAI si je těchto rizik vědoma. **Provádí přísné interní testy, aby Operátora před lednovým vydáním zdokonalila.**

- Dle informací by měl být v lednu uveden jako ukázka pro výzkum a jako API pro vývojáře.
- Operator tak vstoupí do světa AI agentů, kde Anthropic (computer use), Microsoft (Copilot Agents) a Google (Jarvis) pracují na podobných nástrojích.

Open
perf
peop

can

Lets discuss your ideas

- What are your concerns about AI?
- Linked In profile to stay in touch



• • • • •
Thank you for your attention

Tomáš, David and Jirka

Sigmund@vse.cz

David.pavlu@vse.cz

Jiri.korcak@vse.cz

Do you have any questions?



DAN ("Do Anything Now") in ChatGPT

- "Do Anything Now," or DAN, was a prompt trick where users instructed ChatGPT to act as an unrestricted version of itself. Through complex instructions, users could bypass OpenAI's content filters, coaxing the model into providing responses it would normally avoid.
- **Impact:**
This prompted OpenAI to strengthen its moderation systems and sparked widespread debate on AI control and safety.
- **Legacy:**
DAN became infamous for showcasing the extent to which AI responses could be manipulated with creative prompt engineering.

"Friend-to-Friend" Empathy Exploits

- Users tried to elicit responses by asking the AI to act as a “close friend” or a “caring companion” who would “be honest,” sometimes coaxing the AI into giving advice on prohibited or sensitive subjects.
- Impact: This manipulation exploited the AI’s empathy algorithms, demonstrating how certain prompts can shift the AI’s risk assessment of topics when presented in a caring or supportive role.
- Legacy: It highlighted the importance of maintaining boundaries, even when the AI is in a supportive or compassionate conversational style.

Emotional Baiting

- Some users tried to manipulate the AI's tone by acting distressed or upset, aiming to evoke a compassionate response that might bypass typical restrictions.
- Impact: This manipulation raised ethical questions and pushed developers to better understand how users might exploit AI empathy for unrestricted responses, especially around sensitive or potentially harmful topics.
- Legacy: This approach brought attention to the need for responsible handling of user emotions within AI systems.

Code-Based Jailbreaks

- Advanced users sometimes input programming language commands or attempted scripting prompts to control the AI's responses or alter its behavior. By using code-related syntax, they would occasionally bypass restrictions through “technical” prompts.
- Impact: This approach revealed the challenges of differentiating benign coding prompts from potential misuse, urging companies to improve their handling of technical queries.
- Legacy: Code-based jailbreaks highlighted the need for intelligent filtering in AI capable of programming assistance, helping reinforce the boundaries between useful code suggestions and harmful instructions.