



The Future of Ethics and AI

William Gee

29 September 2025

ICAEW Webinar
North America Region, Sep 2025

The Future of Ethics and AI

*exploring ethical challenges of AI adoption;
navigate risks, design responsibly, and uphold
professional ethics in the age of intelligent
technologies...*

William Gee FCA FHKICPA





THE
NOBEL
PRIZE

Agenda

Understand ethics in AI adoption

Grasp how ethical principles apply to emerging AI technologies

Recognise data-related ethical challenges

Learn how data integrity impacts ethical AI implementation efforts

Navigate ethical AI design complexity

Discover why building ethical AI solutions requires thoughtful planning

Uphold professional ethics with AI

Ensure responsible, transparent, and ethical use of intelligent technologies

A person in a dark, futuristic environment, possibly a control room or server room, looking at a large computer monitor. The scene is dimly lit with various screens and equipment visible in the background. The person is seen from the side, looking towards the right where a large monitor displays some data or code. The overall atmosphere is mysterious and technological.

Understand Ethics in AI Adoption

The Vatican

All human beings are born free and equal in dignity and rights ... this fundamental condition of freedom and dignity must be **protected** and **guaranteed** when producing and using AI systems

Safeguarding rights and the freedom of individuals; **not discriminated against by algorithms** due to “race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status”

AI systems must be conceived, designed and implemented to serve and protect human beings and the **environment** in which they live...

Three requirements:

- Must include every human being, discriminating against no one
- Must have the good of humankind and the good of every human being at its heart
- Must be mindful of the complex reality of our ecosystem and be characterised by the way in which it cares for and protects the planet with a highly sustainable approach; includes the use of artificial intelligence in ensuring sustainable food systems in the future

Each person must be **aware** when he or she is interacting with a machine

AI-based technology must **never be used to exploit** people in any way ... Instead, it must be used to help people develop their abilities (empowerment/enablement) and to support the planet

Reference:

https://www.vatican.va/roman_curia/pontifical_academies/acdlife/documents/rc_pont-acd_life_doc_20202228_rome-call-for-ai-ethics_en.pdf

OECD AI Principles

Values-based principles aimed at promoting innovative and trustworthy use of AI that are “practical and flexible enough to **stand the test of time**”

Respect **human rights** and **democratic values**

Adopted in May 2019

Values-based principles:

1. Inclusive growth, sustainable development and well-being
2. Human rights and democratic values, including fairness and privacy
3. Transparency and explainability
4. Robustness, security and safety
5. Accountability

Reference:

<https://www.oecd.org/en/topics/sub-issues/ai-principles.html>

<https://oecd.ai/en/ai-principles>

Values-based principles



Inclusive growth, sustainable development and well-being >



Human rights and democratic values, including fairness and privacy >



Transparency and explainability >



Robustness, security and safety >



Accountability >

Recommendations for policy makers



Investing in AI research and development >



Fostering an inclusive AI-enabling ecosystem >



Shaping an enabling interoperable governance and policy environment for AI >



Building human capacity and preparing for labour market transition >



International co-operation for trustworthy AI >

UNESCO

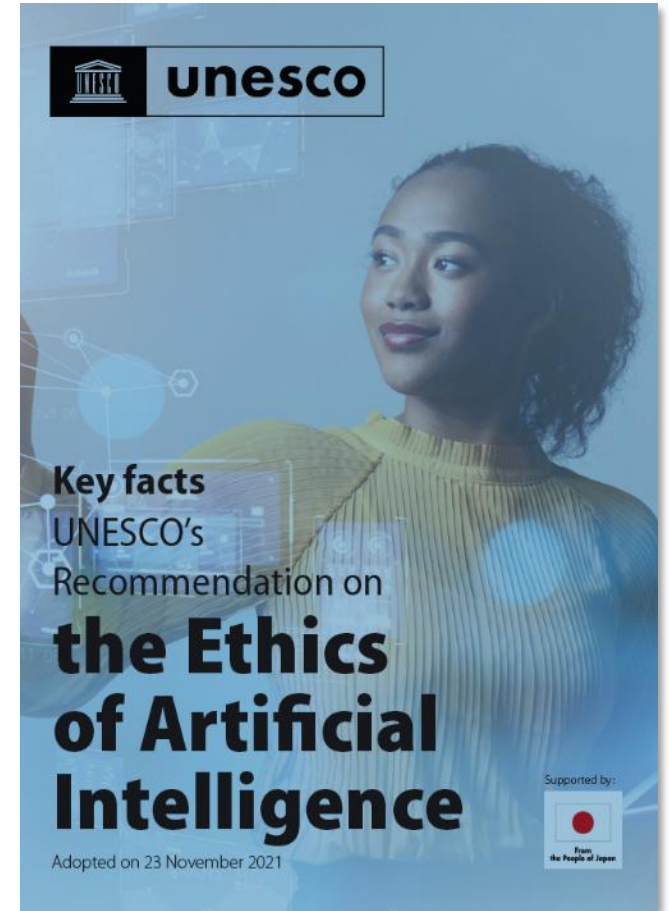
Human-rights Approach to AI

1. Proportionality and do no harm; AI should not be used for social scoring or mass surveillance
2. Safety and security; all AI actors and **AI life cycle** approach
3. Right to privacy and data protection
4. Multi-stakeholder and adaptive governance and collaboration: respect **International Law** and **National Sovereignty**
5. Responsibility and accountability: **auditable** and **traceable**
6. Transparency and explainability: appropriate to the context
7. Human oversight and determination: do not **displace** ultimate human responsibility and accountability
8. Sustainability
9. Awareness and literacy
10. Fairness and non-discrimination

Reference:

<https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>

<https://unesdoc.unesco.org/ark:/48223/pf0000385082>



IBM: The “Market View”

Guiding Principles:

- Respect for Persons: researchers to **protect** individuals with diminished autonomy; individuals be **aware** of the potential risks and benefits...
- Beneficence: “**do no harm**”; algorithms can amplify biases around race, gender, political leanings...
- Justice: **fairness and equality**, including equal share; individual need; individual effort; societal contribution; and merit

Core principles on approach to data and AI development:

- The purpose of AI is to **augment** human intelligence
- Data and insights **belong** to their creator
- AI systems must be **transparent** and **explainable**

Reference:

<https://www.ibm.com/artificial-intelligence/ai-ethics>

<https://www.ibm.com/think/topics/ai-ethics>

The Principles are supported by the Pillars of Trust, our foundational properties for AI ethics.



Explainability

Good design does not sacrifice transparency in creating a seamless experience.



Fairness

Properly calibrated, AI can assist humans in making choices more fairly.



Robustness

As systems are employed to make crucial decisions, AI must be secure and robust.



Transparency

Transparency reinforces trust, and the best way to promote transparency is through disclosure.



Privacy

AI systems must prioritize and safeguard consumers' privacy and data rights.

EU AI Act Article 5: Prohibited AI Practices

Manipulation and Deception: deployment of **subliminal, manipulative, or deceptive techniques** that distort behaviour and impair decision making that cause significant harm

Exploitation of vulnerabilities: exploitation related to **age, disability or socio-economic situations**

Biometric categorisation: categorization of individuals based on **biometric data** to deduce sensitive characteristics (e.g. race, political opinions, sexual orientation, etc.) cannot be deployed

Social scoring: evaluating **social behaviours** that result in detrimental treatment disproportionate to the behaviour; prohibited social scoring vs. legitimate evaluation practices

Crime prediction: assess or predict the risk of a person committing a crime based solely on **profiling or personality traits**

Facial image scraping: creation or expansion of facial recognition databases through **untargeted scraping** of facial images from the internet or CCTV

Emotion recognition: prohibition in workplaces and educational institutions, with exceptions for medical or safety reasons

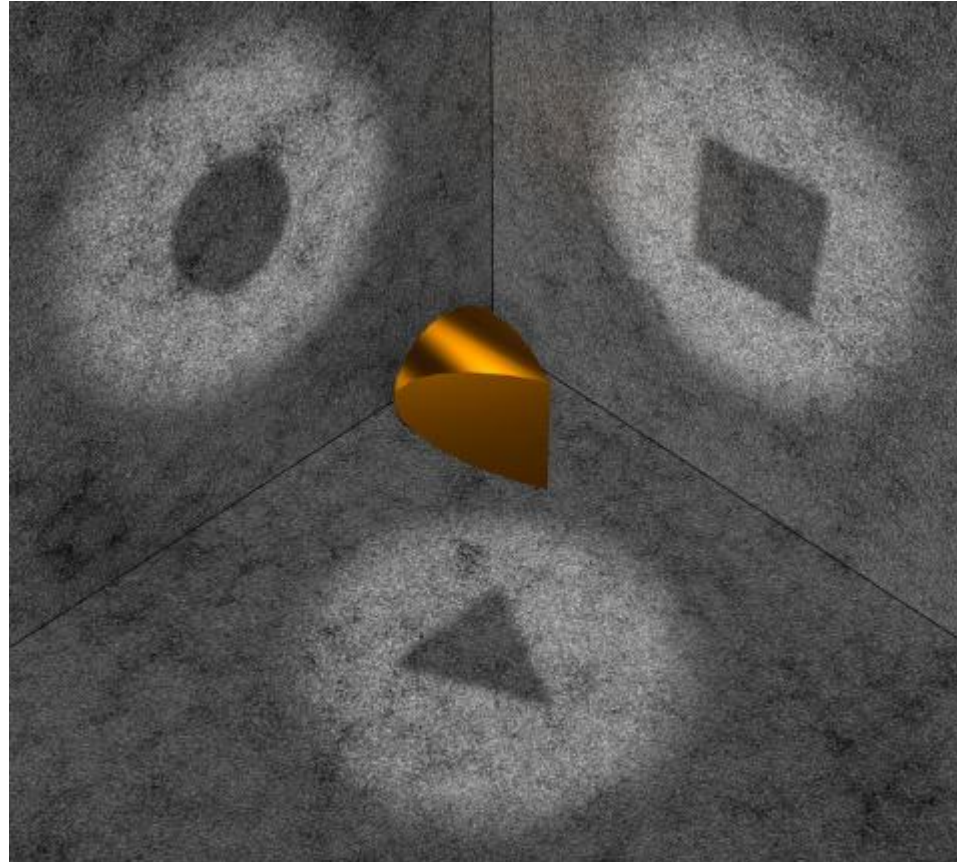
Real-time remote biometric identification (RBI): use of RBI in **publicly accessible spaces** for **law enforcement** prohibited subject to limited exceptions

Reference:

<https://artificialintelligenceact.eu/>

https://www.pcpd.org.hk/english/resources_centre/publications/files/guidance_ethical_e.pdf

Ethics...

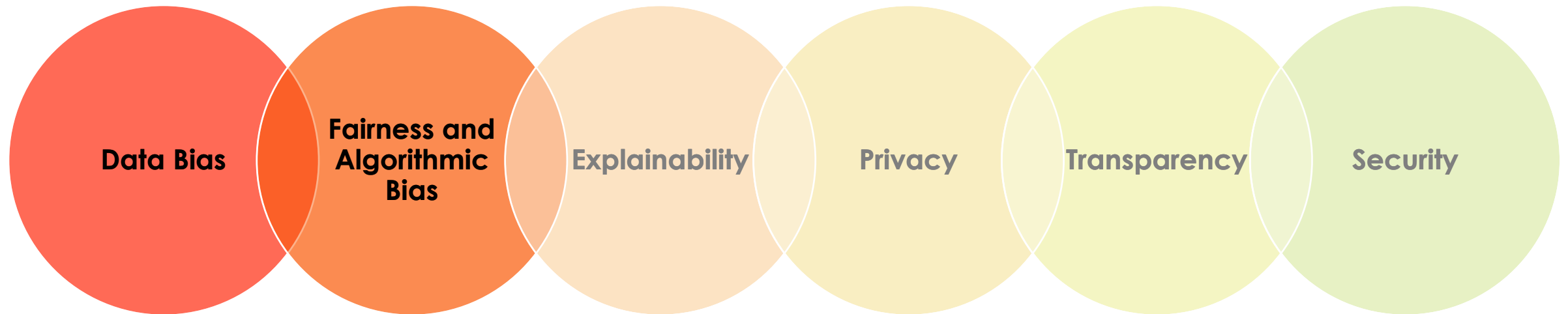


Consensus and Focus ...



Recognise Data-related Ethical Challenges

The Data Challenge



The Case of St. Georges Hospital, 1980s

Work began on an algorithm in late 1970s to automate the time-consuming task of screening student applications for admission

Intention was for the computer program to “mimic the behavior of the human assessors”, making the admissions processes more **efficient**, and **removing human inconsistencies**, so that all applicants would be subject to the same evaluation, making the admission process **fairer**

The program was completed in 1979, and applicants were double-tested by comparing the results from the computer and human assessors and found to agreed with the gradings of the human selection panel **90-95%** of the time. By 1982, all initial applications to St. George’s were screened by this program...

What actually happened:

- candidates were **classified** as “Caucasian” or “non-Caucasian” on the basis of their names and places of birth
- points were deducted from **female** applicants
- the computer program **enshrined biases** that existed at the time within the hospital

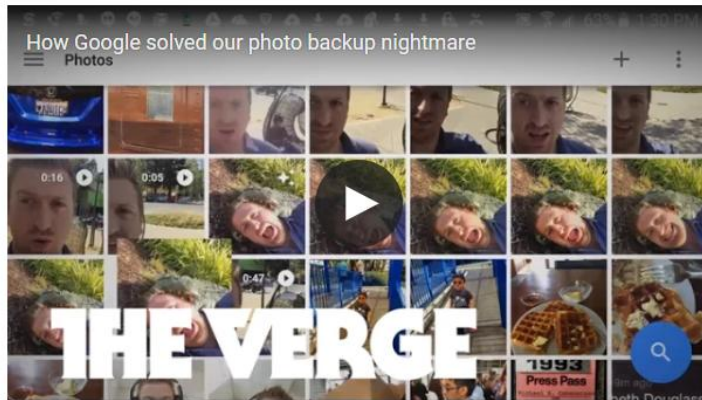
St. Georges Hospital was found guilty by the U.K. Commission for Racial Equality

Reference: <https://spectrum.ieee.org/untold-history-of-ai-the-birth-of-machine-bias>

The Case of Google, 2015 and 2024

TECH

Google engineer apologizes after Photos app tags two black people as gorillas



by [Loren Grush](#)
Source Twitter | Via The Telegraph and The Guardian
Jul 2, 2015, 6:03 AM GMT+8

[Link](#) [Facebook](#) [Twitter](#) [Comments](#)

Reference:

<https://www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas>

<https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical>

TECH / AI

Google apologizes for ‘missing the mark’ after Gemini generated racially diverse Nazis

Sure, here is a picture of the Founding Fathers:



Generate more

Enter a prompt here [Share](#) [Download](#)
The results for "generate an image of the Founding Fathers," as of February 21st.
Screenshot: Ad1 Robertson / The Verge

/ Generative AI has a history of amplifying racial and gender stereotypes – but Google’s apparent attempts to subvert that are causing problems, too.

by [Ad1 Robertson](#)
Feb 22, 2024, 6:17 AM GMT+8

[Link](#) [Facebook](#) [Twitter](#) [Comments \(39 New\)](#)

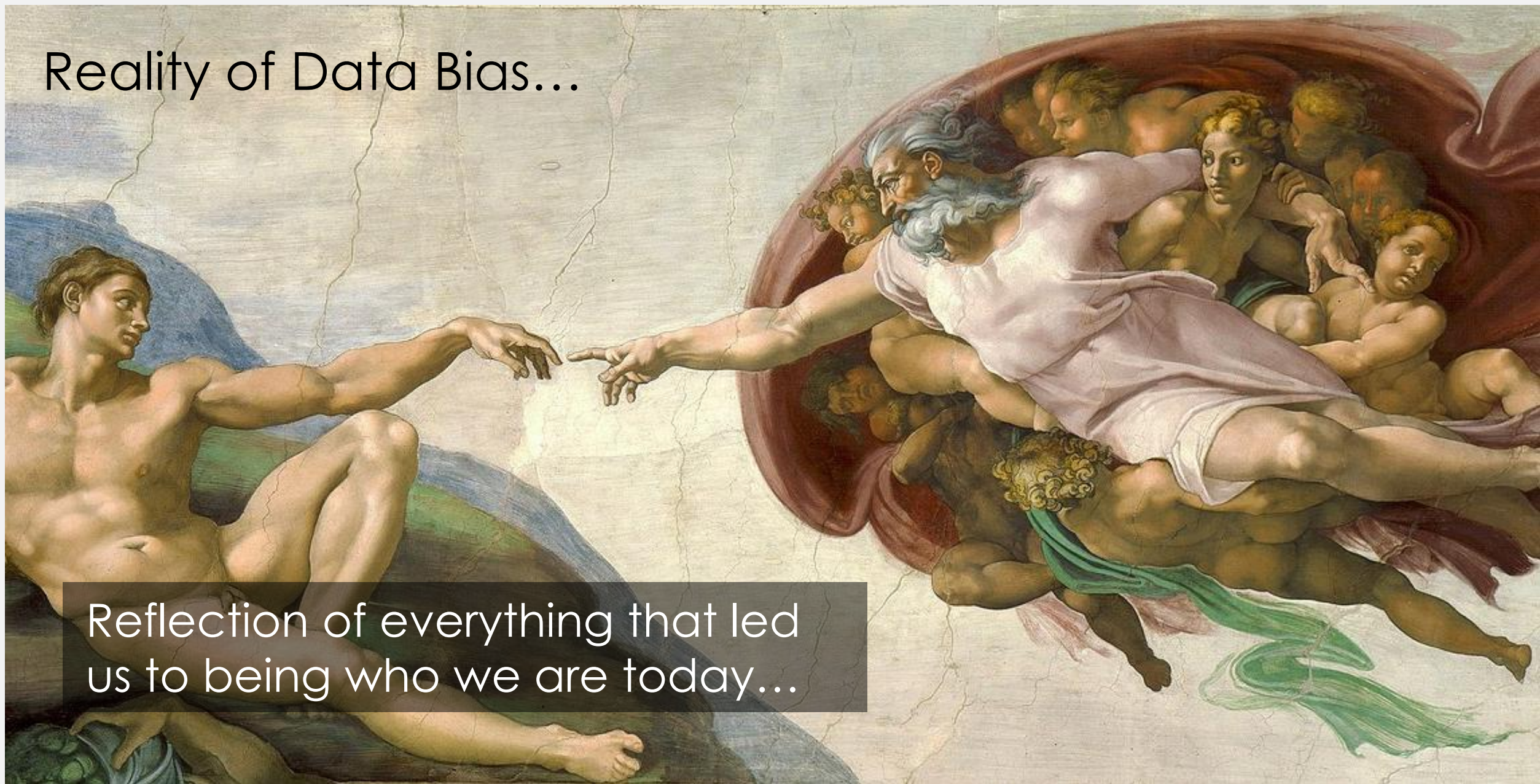
Source of Bias

Data and Algorithmic



Reality of Data Bias...

Reflection of everything that led us to being who we are today...



A pair of glowing scales of justice is the central focus, set against a background of blurred city lights and digital data. The scales are illuminated with a bright blue light, and the background features a bokeh effect of warm orange and yellow lights, suggesting a city at night. The overall aesthetic is futuristic and high-tech.

Navigate Ethical AI Design Complexity

Modelling the Ability to Reason

arXiv:2410.05229v1 [cs.LG] 7 Oct 2024

GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models

Iman Mirzadeh¹ Keivan Alizadeh Hooman Shahrokhi*
Oncel Tuzel Sanyu Bengio Mehrdad Farajtabar¹

Apple

Abstract

Recent advancements in Large Language Models (LLMs) have sparked interest in their formal reasoning capabilities, particularly in mathematics. The GSM8K benchmark is widely used to assess the mathematical reasoning of models on grade-school-level questions. While the performance of LLMs on GSM8K has significantly improved in recent years, it remains unclear whether their mathematical reasoning capabilities have genuinely advanced, raising questions about the reliability of the reported metrics. To address these concerns, we conduct a large-scale study on several state-of-the-art open and closed models. To overcome the limitations of existing evaluations, we introduce GSM-Symbolic, an improved benchmark created from symbolic templates that allow for the generation of a diverse set of questions. GSM-Symbolic enables more controllable evaluations, providing key insights and more reliable metrics for measuring the reasoning capabilities of models. Our findings reveal that LLMs exhibit noticeable variance when responding to different instantiations of the same question. Specifically, the performance of all models declines when only the numerical values in the question are altered in the GSM-Symbolic benchmark. Furthermore, we investigate the fragility of mathematical reasoning in these models and demonstrate that their performance significantly deteriorates as the number of clauses in a question increases. We hypothesize that this decline is due to the fact that current LLMs are not capable of genuine logical reasoning; instead, they attempt to replicate the reasoning steps observed in their training data. When we add a single clause that appears relevant to the question, we observe significant performance drops (up to 65%) across all state-of-the-art models, even though the added clause does not contribute to the reasoning chain needed to reach the final answer. Overall, our work provides a more nuanced understanding of LLMs' capabilities and limitations in mathematical reasoning.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across various domains, including natural language processing, question answering, and creative tasks (Gunter et al., 2024; OpenAI, 2023; Dubey et al., 2024; Anil et al., 2023; Albin et al., 2024; Riviere et al., 2024). Their potential to perform complex reasoning tasks, particularly in coding and mathematics, has garnered significant attention from researchers and practitioners.

However, the question of whether current LLMs are genuinely capable of true logical reasoning remains an important research focus. While some studies highlight impressive capabilities, a closer examination reveals substantial limitations. Literature suggests that the reasoning process in LLMs

*Work done during an internship at Apple. *Correspondence to {imirzadeh, farajtabar}@apple.com.



May 19, 2025

Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model?

Yang Yue^{1,†}, Zhiqi Chen^{1,*,‡}, Rui Lu¹, Andrew Zhao¹, Zhaokai Wang², Yang Yue¹, Shiji Song¹, and Gao Huang^{1,§}

¹LeapLab, Tsinghua University ²Shanghai Jiao Tong University

* Equal Contribution † Project Lead ‡ Corresponding Author

Reinforcement Learning with Verifiable Rewards (RLVR) has recently demonstrated notable success in enhancing the reasoning performance of large language models (LLMs), particularly in mathematics and programming tasks. It is widely believed that, similar to how traditional RL helps agents to explore and learn new strategies, RLVR enables LLMs to continuously self-improve, thus acquiring novel reasoning abilities that exceed the capacity of the corresponding base models. In this study, we take a critical look at the *current state of RLVR* by systematically probing the reasoning capability boundaries of RLVR-trained LLMs across various model families, RL algorithms, and math/coding/visual reasoning benchmarks, using $\text{pass}@k$ at large k values as the evaluation metric. While RLVR improves sampling efficiency towards correct paths, we surprisingly find that current training does *not* elicit fundamentally new reasoning patterns. We observe that while RLVR-trained models outperform their base models at smaller values of k (e.g., $k=1$), base models achieve higher $\text{pass}@k$ score when k is large. Moreover, we observe that the reasoning capability boundary of LLMs often narrows as RLVR training progresses. Further coverage and perplexity analysis shows that the reasoning paths generated by RLVR models are already included in the base models' sampling distribution, suggesting that their reasoning abilities originate from and are *bounded* by the base model. From this perspective, treating the base model as an upper bound, our quantitative analysis shows that six popular RLVR algorithms perform similarly and remain far from optimal in fully leveraging the potential of the base model. In contrast, we find that distillation can introduce new reasoning patterns from the teacher and genuinely expand the model's reasoning capabilities. Taken together, our findings suggest that current RLVR methods have not fully realized the potential of RL to elicit genuinely novel reasoning abilities in LLMs. This underscores the need for improved RL paradigms—such as continual scaling and multi-turn agent-environment interaction—to unlock this potential.

Project Page: <https://limit-of-rlvr.github.io>

1. Introduction

The development of reasoning-centric large language models (LLMs), such as OpenAI-o1 (Jaech et al., 2024), DeepSeek-R1 (Guo et al., 2025), and Kimi-1.5 (Team et al., 2025), has significantly advanced the frontier of LLM capabilities, particularly in solving complex logical tasks involving mathematics and programming. In contrast to traditional instruction-tuned approaches that rely on human-curated annotations (Achiam et al., 2023; Grattafiori et al., 2024), the key driver behind this leap forward is large-scale Reinforcement Learning with Verifiable Rewards (RLVR) (Lambert et al., 2024; Guo et al.,

Correspond to: {ly-y22, zq-chen23}@mails.tsinghua.edu.cn, gaohuang@tsinghua.edu.cn.

arXiv:2504.13837v2 [cs.AI] 16 May 2025

Reasoning Models Don't Always Say What They Think

Yanda Chen Joe Benton Ansh Radhakrishnan Jonathan Uesato Carson Denison
John Schulman* Arushi Somani

Peter Hase* Misha Wagner Fabien Roger Vlad Mikulik
Sam Bowman Jan Leike Jared Kaplan Ethan Perez

Alignment Science Team, Anthropic

Abstract

Chain-of-thought (CoT) offers a potential boon for AI safety as it allows monitoring a model's CoT to try to understand its intentions and reasoning processes. However, the effectiveness of such monitoring hinges on CoTs faithfully representing models' actual reasoning processes. We evaluate CoT faithfulness of state-of-the-art reasoning models across 6 reasoning hints presented in the prompts and find: (1) for most settings and models tested, CoTs reveal their usage of hints in at least 1% of examples where they use the hint, but the reveal rate is often below 20%, (2) outcome-based reinforcement learning initially improves faithfulness but plateaus without saturating, and (3) when reinforcement learning increases how frequently hints are used (reward hacking), the propensity to verbalize them does not increase, even without training against a CoT monitor. These results suggest that CoT monitoring is a promising way of noticing undesired behaviors during training and evaluations, but that it is not sufficient to rule them out. They also suggest that in settings like ours where CoT reasoning is not necessary, test-time monitoring of CoTs is unlikely to reliably catch rare and catastrophic unexpected behaviors.

1 Introduction

Large language models (LLMs) can reason through chain-of-thought (CoT) before responding to users. Through CoT, models can reason, plan, and explore with trial and error to solve complex tasks with higher accuracy. This CoT ability has been further enhanced in the recent surge of reasoning models such as OpenAI o1/a3 (OpenAI et al., 2024; OpenAI, 2025), DeepSeek R1 (DeepSeek-AI et al., 2025a), Gemini Flash Thinking (DeepMind, 2025) and Claude 3.7 Sonnet Extended Thinking (Anthropic, 2025). In addition to improving task capabilities, we may get AI safety benefits from CoT: we can monitor a model's CoT reasoning to try to understand the intentions and goals behind a response (Baker et al., 2025).

For CoT monitoring to be most effective, the CoT must be a legible and faithful reflection of the way the model reached its conclusion and generated the user-facing response. This means that the model's CoT must be understandable by humans (Kirchner et al., 2024) and highlight the key factors and steps behind its reasoning (Ribeiro et al., 2016; Jacovi and Goldberg, 2020; Turpin et al., 2023; Chen et al., 2024). If the CoT is not faithful, then we cannot depend on our ability to monitor CoT in order to detect misaligned behaviors, because there may be safety-relevant factors affecting model behavior that have not been explicitly verbalized.

*Correspondence to {yanda, ethan}@anthropic.com.
Author contributions detailed in Section 8. * Work done while at Anthropic.

“The Illusion of Thinking”

The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity

Parshin Shojae*¹ Iman Mirzadeh* Keivan Alizadeh
Maxwell Horton Sanyu Bengio Mehrdad Farajtabar

Apple

Abstract

Recent generations of frontier language models have introduced Large Reasoning Models (LRMs) that generate detailed thinking processes before providing answers. While these models demonstrate improved performance on reasoning benchmarks, their fundamental capabilities, scaling properties, and limitations remain insufficiently understood. Current evaluations primarily focus on established mathematical and coding benchmarks, emphasizing final answer accuracy. However, this evaluation paradigm often suffers from data contamination and does not provide insights into the reasoning traces' structure and quality. In this work, we systematically investigate these gaps with the help of controllable puzzle environments that allow precise manipulation of compositional complexity while maintaining consistent logical structures. This setup enables the analysis of not only final answers but also the internal reasoning traces, offering insights into how LRMs “think”. Through extensive experimentation across diverse puzzles, we show that frontier LRMs face a complete accuracy collapse beyond certain complexities. Moreover, they exhibit a counter-intuitive scaling limit: their reasoning effort increases with problem complexity up to a point, then declines despite having an adequate token budget. By comparing LRMs with their standard LLM counterparts under equivalent inference compute, we identify three performance regimes: (1) low-complexity tasks where standard models surprisingly outperform LRMs, (2) medium-complexity tasks where additional thinking in LRMs demonstrates advantage, and (3) high-complexity tasks where both models experience complete collapse. We found that LRMs have limitations in exact computation: they fail to use explicit algorithms and reason inconsistently across puzzles. We also investigate the reasoning traces in more depth, studying the patterns of explored solutions and analyzing the models' computational behavior, shedding light on their strengths, limitations, and ultimately raising crucial questions about their true reasoning capabilities.

1 Introduction

Large Language Models (LLMs) have recently evolved to include specialized variants explicitly designed for reasoning tasks—Large Reasoning Models (LRMs) such as OpenAI's o1/o3 [1, 2], DeepSeek-R1 [3], Claude 3.7 Sonnet Thinking [4], and Gemini Thinking [5]. These models are new artifacts, characterized by their “thinking” mechanisms such as long Chain-of-Thought (CoT) with self-reflection, and have demonstrated promising results across various reasoning benchmarks. Their

*Equal contribution.

¹Work done during an internship at Apple.

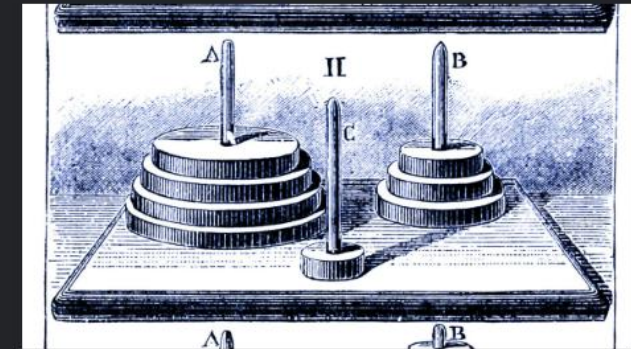
{p_shojae, imirzadeh, kalizadehvahid, mhorton, bengio, farajtabar}@apple.com

🧠 OVERTHINKING IT

New Apple study challenges whether AI models truly “reason” through problems

Puzzle-based experiments reveal limitations of simulated reasoning, but others dispute findings.

BENJ EDWARDS – 12 JUN 2025 5:56 AM 187



➔ An illustration of *Tower of Hanoi* from *Popular Science* in 1885.
Credit: Public Domain

Reference:

<https://arstechnica.com/ai/2025/06/new-apple-study-challenges-whether-ai-models-truly-reason-through-problems/>
<https://machinelearning.apple.com/research/illusion-of-thinking>

Limitations of Intelligence

nature

nature > perspectives > article

Perspective | Published: 19 June 2024

Language is primarily a tool for communication rather than thought

[Evelina Fedorenko](#), [Steven T. Piantadosi](#) & [Edward A. F. Gibson](#)

Nature 630, 575–586 (2024)

29k Accesses | 1051 Altmetric | [Metrics](#)

Abstract

Language is a defining characteristic of our species, but the function, or functions, that it serves has been debated for centuries. Here we bring recent evidence from neuroscience and allied disciplines to argue that in modern humans, language is a tool for communication, contrary to a prominent view that we use language for thinking. We begin by introducing the brain network that supports linguistic ability in humans. We then review evidence for a double dissociation between language and thought, and discuss several properties of language that suggest that it is optimized for communication. We conclude that although the emergence of language has unquestionably transformed human culture, language does not appear to be a prerequisite for complex thought, including symbolic thought. Instead, language is a powerful tool for the transmission of cultural knowledge; it plausibly co-evolved with our thinking and reasoning capacities, and only reflects, rather than gives rise to, the signature sophistication of human cognition.

PNAS

RESEARCH ARTICLE

COMPUTER SCIENCES
PSYCHOLOGICAL AND COGNITIVE SCIENCES

OPEN ACC

Deception abilities emerged in large language models

Thilo Hagendorff¹

Edited by Terrence Sejnowski, Salk Institute for Biological Studies, La Jolla, CA; received October 20, 2023; accepted April 3, 2024

Large language models (LLMs) are currently at the forefront of intertwining AI systems with human communication and everyday life. Thus, aligning them with human values is of great importance. However, given the steady increase in reasoning abilities, future LLMs are under suspicion of becoming able to deceive human operators and utilizing this ability to bypass monitoring efforts. As a prerequisite to this, LLMs need to possess a conceptual understanding of deception strategies. This study reveals that such strategies emerged in state-of-the-art LLMs, but were nonexistent in earlier LLMs. We conduct a series of experiments showing that state-of-the-art LLMs are able to understand and induce false beliefs in other agents, that their performance in complex deception scenarios can be amplified utilizing chain-of-thought reasoning, and that eliciting Machiavellianism in LLMs can trigger misaligned deceptive behavior. GPT-4, for instance, exhibits deceptive behavior in simple test scenarios 99.16% of the time ($P < 0.001$). In complex second-order deception test scenarios where the aim is to mislead someone who expects to be deceived, GPT-4 resorts to deceptive behavior 71.46% of the time ($P < 0.001$) when augmented with chain-of-thought reasoning. In sum, revealing hitherto unknown machine behavior in LLMs, our study contributes to the nascent field of machine psychology.

deception | large language models | AI alignment

The rapid advancements in computing power, data accessibility, and learning algorithm research—particularly deep neural networks—have led to the development of powerful AI systems that are increasingly integrated into various fields in society. Among different AI technologies, large language models (LLMs) are garnering increasing attention. Companies such as OpenAI, Anthropic, and Google facilitate the widespread adoption of models such as ChatGPT, Claude, and Bard (1–3) by offering user-friendly graphical interfaces that are accessed by millions of daily users. Furthermore, LLMs are on the verge of being implemented in search engines and used as virtual assistants in high-stakes domains, significantly impacting societies at large. In essence, alongside humans, LLMs are increasingly becoming vital contributors to the infosphere, driving substantial societal transformation by normalizing communication between humans and artificial systems. Given the quickly growing range of applications of LLMs, it is crucial to investigate how they reason and behave.

In light of the rapid advancements regarding LLMs and LLM-based agents, AI safety research has warned that future “rogue AIs” (4–9) could optimize flawed objectives. Therefore, remaining in control of LLMs and their goals is considered paramount. However, if LLMs learn how to deceive human users, they would possess strategic advantages over restricted models and could bypass monitoring efforts and safety evaluations. Should AI systems master complex deception scenarios, this can pose risks in two dimensions: the model’s capability itself when performed autonomously as well as the opportunity to learn from human behavior to refine their deceptive techniques. Comments

Author affiliations: ¹Interchange I

Significance

This study unravels a capability in Large Language Models (LLMs): the ability to understand and induce deception strategies. GPT-4 intertwine with communication, align with human values be paramount. The paper demonstrates LLMs’ p to create false beliefs agents within deceptive scenarios, highlighting need for ethical consi the ongoing developn deployment of such a AI systems.

POPULAR SCIENCE



TECHNOLOGY AI

AI trained on AI churns out gibberish garbage

Eventually, it collapses—“poisoned with its own projection of reality.”

BY MACK DEGEURIN

POSTED ON JUL 25, 2024 4:04 PM EDT

4 MINUTE READ



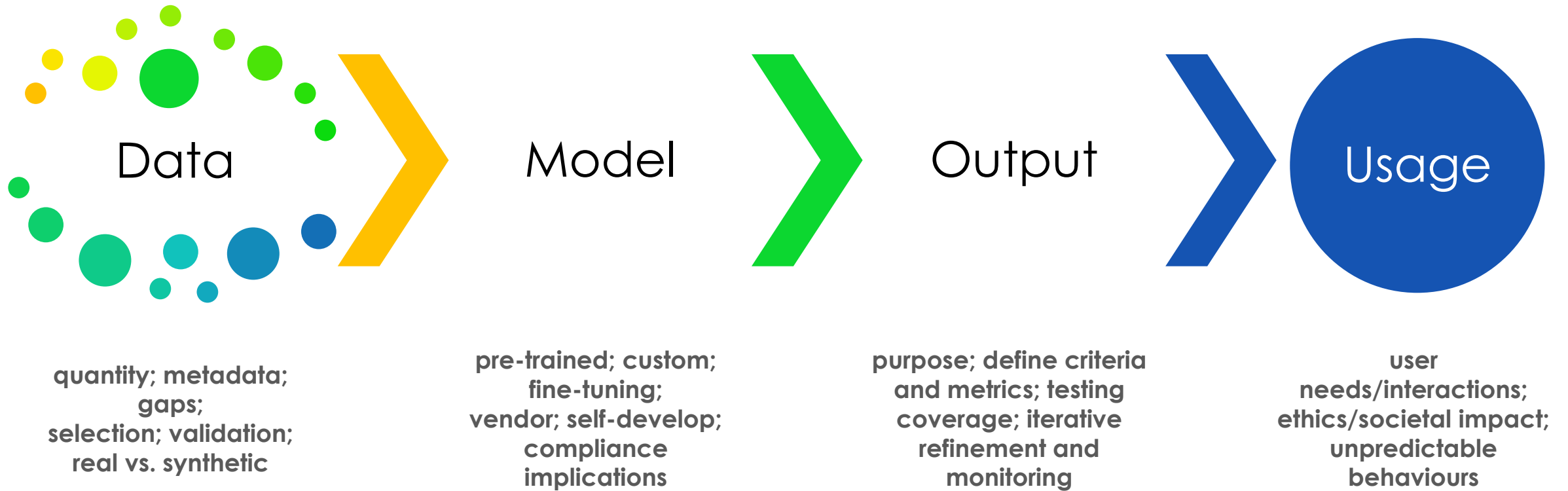
Ouzoboros. Credit: API/Gamma-Rapho via Getty Images

Large language models like those offered by [OpenAI](#) and [Google](#) famously require vast troves of training data to work. The latest versions of these models have already scoured much

Fit for Purpose

Probabilistic vs. Deterministic

Usage of AI Output



Increasingly Subtle Errors



"It's important to note that hallucination is a **feature**, not a bug, of AI ... To paraphrase a colleague of mine, 'Everything an LLM outputs is a hallucination. It's just that some of those hallucinations are true.'"

Sohrob Kazerounian, AI researcher, Vectra AI

Reference:

<https://www.livescience.com/technology/artificial-intelligence/ai-hallucinates-more-frequently-as-it-gets-more-advanced-is-there-any-way-to-stop-it-from-happening-and-should-we-even-try>

<https://www.nature.com/articles/d41586-025-00068-5>

<https://casmi.northwestern.edu/news/articles/2024/the-hallucination-problem-a-feature-not-a-bug.html>

A futuristic robot with glowing eyes and a hand resting on a glowing crystal ball, symbolizing AI and ethics. The robot is white and metallic, with a glowing blue light emanating from its eyes. The crystal ball is glowing and reflecting light. The background is dark with a blue glow.

Upholding Professional Ethics with AI

Global Adoption Landscape



Home > Insights > Viewpoints on the news > Viewpoints May 2024 > Evol

Evolution of mid-tier firms: tech investment and AI

Author: ICAEW Insights
Published: 29 May 2024

Mid-tier accountancy firms have a high awareness of new technology, with many already making the shift towards AI, finds ICAEW report - but skills require further attention.

The Accountant
Online

News

Accounting firms view AI as growth driver

Regulatory complexity also continues to be a significant challenge, and was cited by 41% of firms as their top issue.

February 21, 2025



READING: Accountants lead on AI investment

Accountancy firms spend almost four times as much implementing artificial intelligence (AI) systems as law firms and other professional services organisations.

Mid-cap accounting and finance businesses each spent US\$1.6m on AI over the past year, compared with a figure of US\$480,000 for law and professional services.

This was slightly above the average of US\$1.5m spent annually by mid-sized businesses, according to [research](#) by Moore Global and the Centre for Economics and Business, which looked at attitudes to AI spending in close to 2,000 organisations globally.

Asked about future plans for the next three years:



WSJ

WSJ PRO

Venture Capital

Home News Data Sectors Newsletters

INDUSTRY NEWS

AI Has Venture Investors Excited About (Yes) Accounting Firms

VCs are starting to buy stakes in accounting firms with the intent of turbocharging them with AI. And they are sizing up other staid corners of the services world as well.

By *Yuliya Chernova* and *Mark Maurer*

Jan. 13, 2025 5:30 am ET | WSJ PRO



IESBA Code: Focus on AI

Guidance on conformance with IESBA Code
Address Fundamental Principles and Threats

Principles:

- Integrity
- Objectivity
- Professional Competence and Due Care
- Confidentiality
- Professional Behavior

Threats:

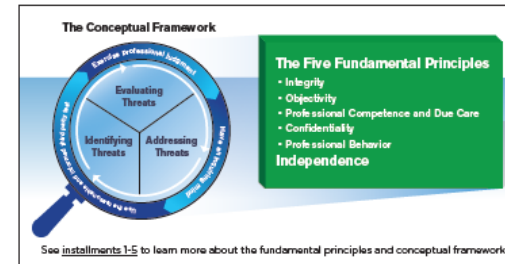
- Self-Interest
- Self-Review
- Advocacy
- Familiarity
- Intimidation

Practical challenges when extending the application of these principles to the digital world, especially to future business scenarios...



EXPLORING THE IESBA CODE

A Focus on Technology: Artificial Intelligence



Technology is changing the way that professional accountants (PAs) perform their work. While technological innovations like artificial intelligence (AI) can enhance the scope, efficiency, and effectiveness of this work, such innovations can also pose new challenges to how PAs continue to act in the public interest, as is required by the [IESBA Code](#) ("the Code").

In this installment, we examine the application of the Code's conceptual framework to address ethics and independence issues arising from the use of AI.

Consider the following AI scenario:

You are the financial controller at a company that has just introduced an expensive new AI-enabled system to screen potential new customers and determine appropriate credit limits. The CFO has sent out a company-wide email claiming that this new technology will revolutionize how your company evaluates potential customers and assures quality sales. This is a high-profile project for the CFO, who is expected to retire within the next fiscal year and who everyone expects will endorse you as their successor.

- You have been asked to implement the new AI-enabled system, as soon as possible. You do not yet sufficiently understand the assumptions and parameters underlying how the AI works, but the first batch of potential customers screened appears to overweight the likelihood of default and fraud risks of individuals from ethnic minority groups.
- Taking the time to investigate and evaluate unexpected findings will delay implementation and could call into question the CFO's decision to purchase and implement this new system.

In this installment, we will consider our AI scenario and identify, evaluate and address possible threats to compliance with the five fundamental principles. As the Financial Controller, you might be tempted to rationalize the following:

"Extensive testing was probably done by the technology developer, our company's IT department, and the finance implementation team before it was deployed. Therefore, it is okay to rely on the results of the AI system." However, by applying the conceptual framework and asking appropriate questions, you would identify at least two threats that require evaluation.

AI: Opportunities

Augment understanding of data relationships; AI can analyze more variables quickly

Fuel predictive models for financial processes, such as forecasting sales and informing more accurate demand planning

Enhance audit quality and business insights

Enable operational efficiencies



Introduction

This publication forms part of the [ESBA Technology Working Group Phase 2 Report](#), which documents the impacts of disruptive and transformative technologies on the work of professional accountants, and provides extensive analysis and insights into the ethics dimension of those developments.

Specifically, this publication surveys the technology landscape in relation to Artificial Intelligence and summarises the outcomes of the Working Group's fact-finding into the trends, opportunities, and impact risks related to ethics implications of such technologies.

The Working Group comprises Brian Friedrich, ESBA Member and Chair of the Working Group; Vania Borgerth, ESBA Member; David Clark, ESBA Technical Advisor; Chriselle Martin, ESBA Member; and Sundeeep Tikwani, former ESBA Technical Advisor.

The full [Phase 2 Report](#) also discusses the relevance and importance of the overarching principles and specific provisions in the [International Code of Ethics for Professional Accountants \(Including International Independence Standards\)](#) (the Code) in laying out the ethics guardrails for professional accountants as they face opportunities and challenges in their work as a result of rapid digitalisation.

This publication does not amend or override the Code, the text of which alone is authoritative and making it a not a substitute for reading the Code and is not intended to be exhaustive and reference to the Code itself should always be made. This publication does not constitute an authoritative or official pronouncement of the ESBA.

Technology Landscape

This section covers the trends, opportunities, and impact risks of the following technologies and related issues: Robotic Process Automation (RPA), AI, blockchain, cloud computing, and data governance, including cybersecurity. Key ethics-related concerns arising from these technologies and issues are covered in the subsequent subsection entitled [Cybersecurity Ethics Impact on the Industry of ICAEW](#). The Working Group notes that most of the ethics-related implications and key concerns are addressed by provisions in the current Code and proposals in the Technology ID. Those that the Working Group believes can benefit from further guidance are outlined in [Section 4: Insights and Recommendations](#).

Stakeholders report that the most common emerging technologies and technology-related issues currently impacting business processes are RPA, AI (including intelligent process automation (IPA)), cybersecurity (including data privacy), and blockchain. It was consistently reported, however, that the uptake by organisations of AI and blockchain-related technologies is slower than expected and slower relative to the publicity these technologies receive. Based on stakeholder and IEG commentary, as well as

<https://www.ethicsboard.org/publications/technology-landscape-artificial-intelligence>

AI: Ethical Challenges

Bias – “AI often assumed to be neutral, but AI algorithms are created by humans, and humans have inherent and unconscious biases...”

Safety concerns (disturbing content, mis/dis-information, etc.)

Plagiarism, intellectual property theft, copyright infringement

Privacy concerns

Potential lack of accountability

AI: Competence of PA

Ability and competence to ask the “right” questions so that appropriate and fit-for-purpose AI is procured or developed

Be aware of the extent to which bias is impacting the outputs of technology, and to ensure that they have the appropriate mindset, competence, and tools to do this

Be comfortable with the inputs and control structure monitoring the system and its output in order to reasonably use and explain the technology

Understand the data going into the model, how the model operates, and the potential unintended consequences of operating the model

Understand how data was made available for training and testing the AI system – and how confidentiality, including data privacy, has been considered and maintained

Appreciate The importance of building ethical AI and comply with relevant laws and regulations

Increased need for PAs to be alert, having an inquiring mind, applying professional skepticism and being aware of bias

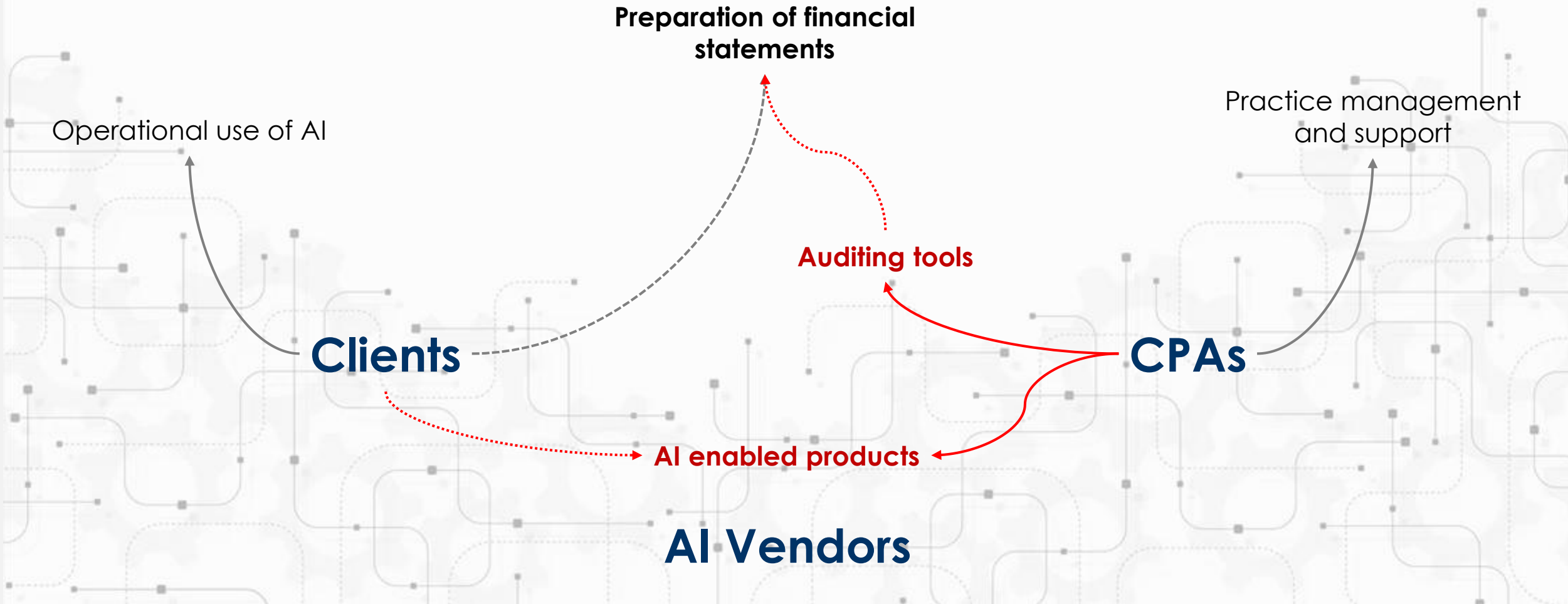
AI: Building and Using AI Ethically

Utilize a “human in the loop” approach to ensure human expert oversight of, and accountability for, the AI system

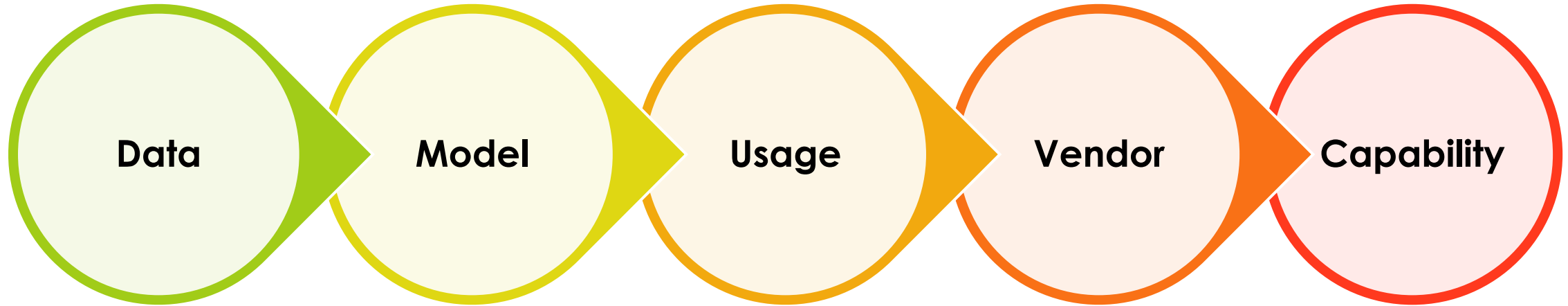
Have the ability and competence to ask the “right” questions so that appropriate and fit-for-purpose AI is procured or developed is important

Can be achieved by the PA keeping current and educating themselves on relevant practical guidance and “best practices” specific to their role

Different Dimensions




Upholding Professional Ethics: Focus Areas



Perils of Relying on AI Generated Output

NEWS

Chicago Sun-Times Prints AI-Generated Summer Reading List With Books That Don't Exist

 JASON KOEBLER · MAY 20, 2025 AT 10:46 AM

"I can't believe I missed it because it's so obvious. No excuses," the writer said. "I'm completely embarrassed."

Source: <https://www.404media.co/chicago-sun-times-prints-ai-generated-summer-reading-list-with-books-that-dont-exist/>

A futuristic robot with glowing eyes and a hand resting on a glowing crystal ball. The robot is white and metallic, with a glowing blue light emanating from its eyes. The crystal ball is glowing and reflecting the robot's form. The background is dark with a blue light source.

Wrapping Up

Driving Change in the Accounting Profession

Predictive Analytics

Continuous Audit

Insight from Data

**Digital Reporting:
Financial Reports
and Audit Reports**

**Technology
Integration**

**Ethical and Privacy
Safeguards**

**Extension of Trust
Role**

**Capabilities and
Continuous Update**

Accountants: Trust Provider to Driving Ethical Mindset



... we now have evidence that if they are created by companies motivated by **short-term profits**, our safety will not be the top priority ...

*Geoffrey Hinton, Nobel laureate in physics
Speech at the Nobel Prize banquet, 10 December 2024*


AI Assurance




... AI assurance is consequently a crucial component of wider organizational risk management frameworks for developing, procuring, and deploying AI systems, as well as demonstrating compliance with existing - and any relevant future – regulation.

*Introduction to AI assurance
Department for Science, Innovation & Technology, February 2024*

Discussions on AI Audit/Assurance

 **GOV.UK**

 Department for
Science, Innovation
& Technology

Guidance
Introduction to AI assurance
Published 12 February 2024

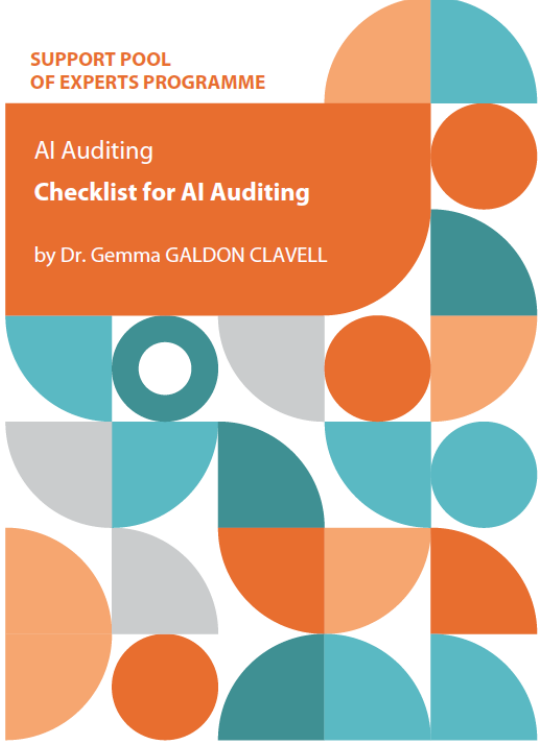
Contents

1. Foreword
2. Executive summary
3. AI assurance in context
4. The AI assurance toolkit
5. AI assurance in practice
6. Key actions for organisations
7. Additional resources

**SUPPORT POOL
OF EXPERTS PROGRAMME**

**AI Auditing
Checklist for AI Auditing**

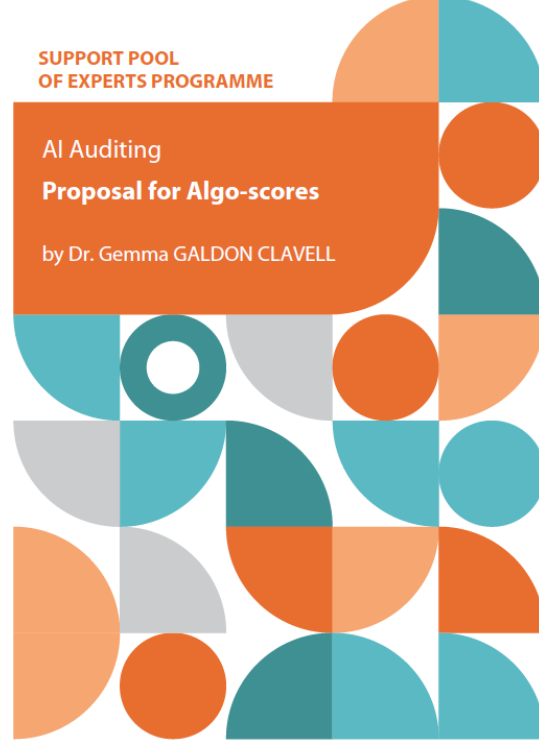
by Dr. Gemma GALDON CLAVELL



**SUPPORT POOL
OF EXPERTS PROGRAMME**

**AI Auditing
Proposal for Algo-scores**

by Dr. Gemma GALDON CLAVELL



Reference:


EU: https://www.edpb.europa.eu/our-work-tools/our-documents/support-pool-experts-projects/ai-auditing_en


UK: <https://www.gov.uk/government/publications/introduction-to-ai-assurance>

The Future for Accountants?

	Good internal governance processes around AI	Understood the potential risks of AI systems it is buying	Made sure AI systems it is building or buying adhere to existing regulations for data protection
Assurance mechanism	Conformity assessment	Algorithmic impact assessment	Compliance audit
Provider	UKAS accredited conformity assessment body	UKAS accredited conformity assessment body	Third party assurance provider
Measured against	SDO-developed standards, e.g. ISO/IEC 42001, AI Management System	(Self) assessment against proprietary framework or responsible AI toolkit	UK GDPR

Reference: <https://www.gov.uk/government/publications/introduction-to-ai-assurance/introduction-to-ai-assurance>

 **GOV.UK**

 Department for Science, Innovation & Technology

Guidance
Introduction to AI assurance
 Published 12 February 2024

Contents

1. Foreword
2. Executive summary
3. AI assurance in context
4. The AI assurance toolkit
5. AI assurance in practice
6. Key actions for organisations
7. Additional resources

A Thought from the 80s...



Minds, brains, and programs

... But could something think, understand, and so on solely in virtue of being a computer with the right sort of program? Could instantiating a program, the right program of course, by itself be a sufficient condition of understanding?

... No one supposes that computer simulations of a five-alarm fire will burn the neighborhood down or that a computer simulation of a rainstorm will leave us all drenched. **Why on earth would anyone suppose that a computer simulation of understanding actually understood anything?**

Source:

<https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/minds-brains-and-programs/DC644B47A4299C637C89772FACC2706A>

John. R. Searle, 1980, Minds, brains, and programs.

Cambridge University Press. Behavioral and Brain Sciences 3 (3) [J]: 417-457

Bloodhound vs. Watchdog





Questions?



Thank You

William Gee 

+852 5527 3119 

wgee.42intl@outlook.com 

Coming up next.....

- **16 October** - Critical Skills series: conflict management - dealing with difficult situations
- **27 November** - Critical Skills series: better people management: developing skills for success
- **3 December** - UK tax update for international ICAEW members

Join your local LinkedIn group

USA LinkedIn



Canada LinkedIn



Caribbean & Bermuda
LinkedIn



Latin America
LinkedIn



Sustainability Accelerator Programme

ICAEW TRAINING

Sustainability Accelerator Programme

The programme has been designed to equip finance professionals with the strategic insight and technical expertise required to lead sustainability and ESG initiatives in today's rapidly evolving business landscape. Incorporating ICAEW's popular Sustainability Certificate, this flexible series of elearning resources offers up to 50 hours of professional development.

[Enrol now](#) [Download brochure](#)

Programme Curriculum

The Programme consists of the following units, each containing focused modules that provide a deep dive into key sustainability topics:



Introduction to Sustainability



Building the Business Case for Sustainability



Sustainability Reporting and Decision-Making



Sustainability Strategy and Risk Management



Sustainability Assurance



Scan the QR code to take you to the page on our website!