Reclaiming Africa's Intellectual Futures

The Architect of Your Own Mind: Ethical and Responsible AI for Learning

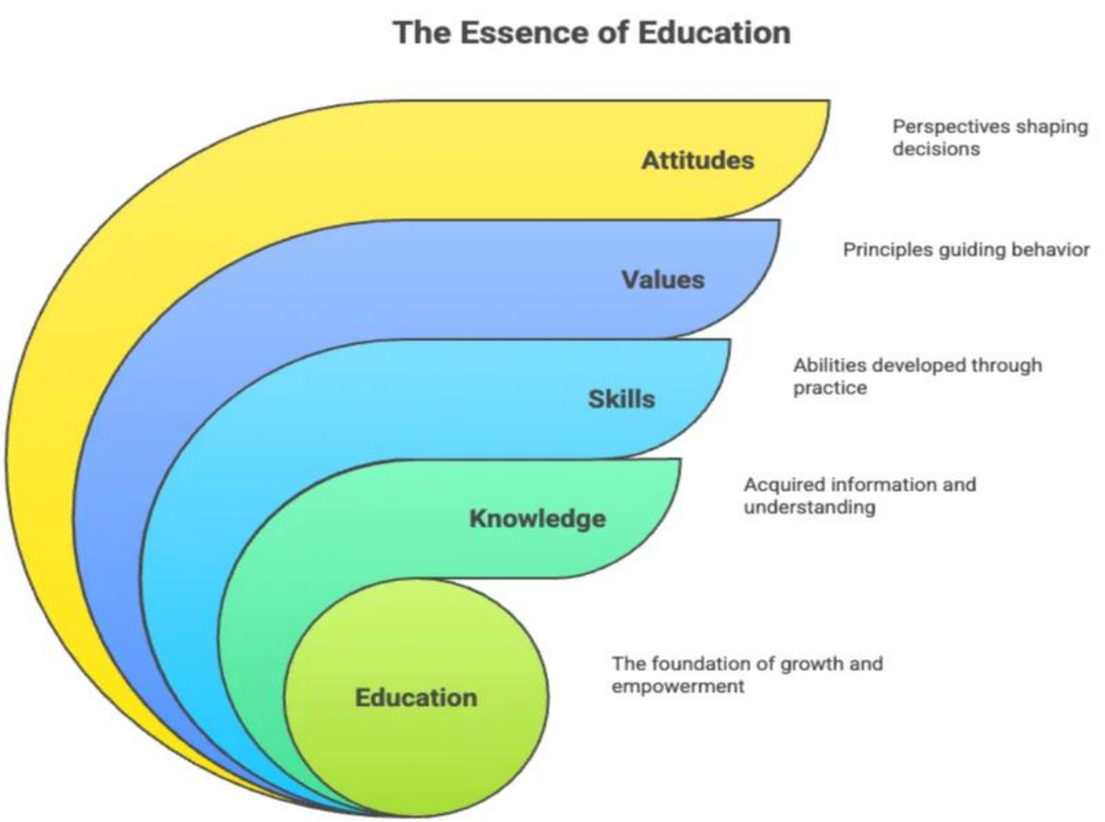
MUPSN RESEARCH INCUBATOR
16 April 2026

Nicky Tjano
tjanorn@unisa.ac.za
0658888129

Define tomorrow.

UNISA university of south africa

What is the essence of education and learning?



What is the essence of learning?

Insights and key lessons from Renowned Thinkers

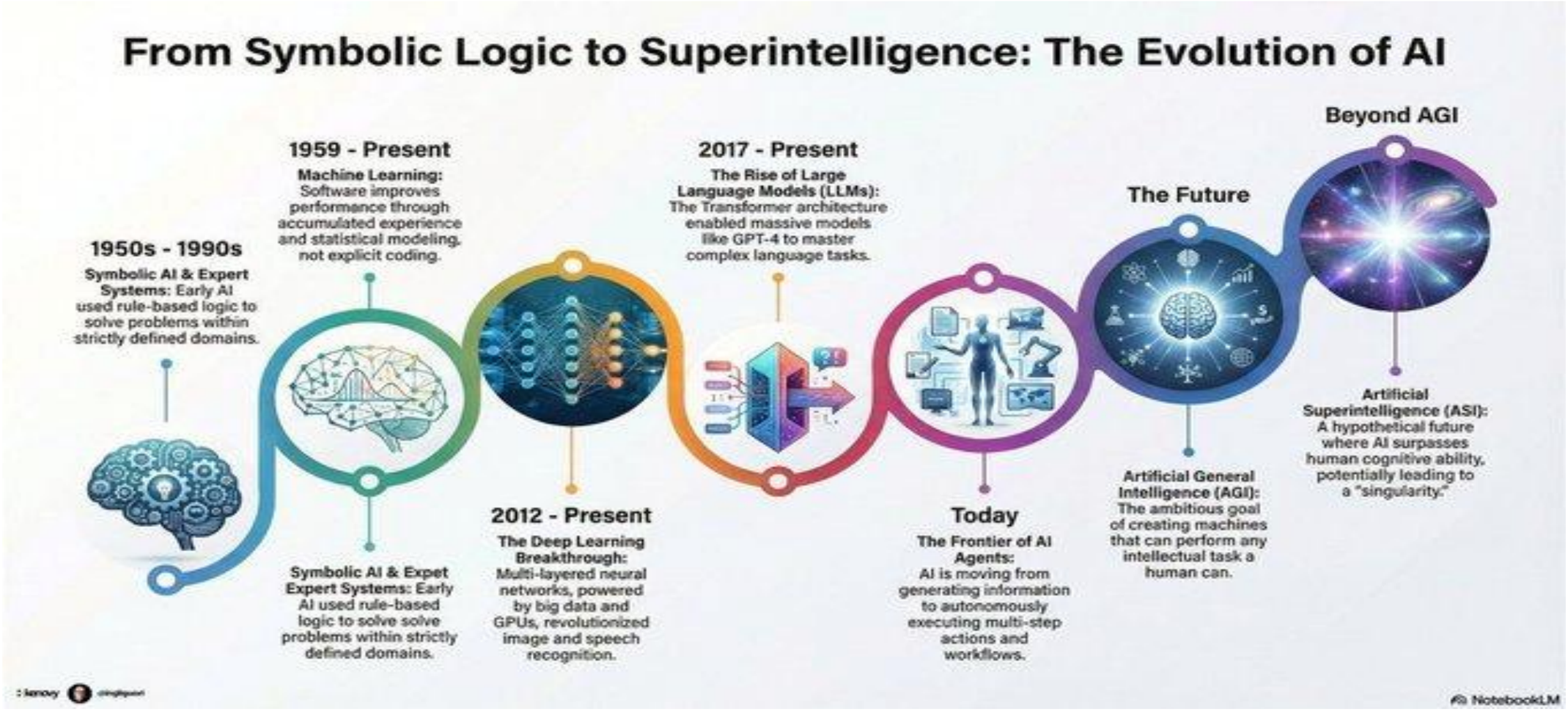
Confucius:	"I hear and I forget. I see and I remember. I do and I understand." Emphasize the importance of active participation and experiential learning.
Benjamin Franklin:	"Tell me and I forget, teach me and I may remember, involve me and I learn." Highlight the role of engagement and involvement in effective learning.
John Dewey	"We do not learn from experience... we learn from reflecting on experience." Discuss the significance of reflection in learning from experiences.
Peter Senge	"Learning is a continual process of transformation in response to experience and reflection." Describe how learning involves continuous transformation and adaptation.

The essence of education is not to get a certificate or job but, to be a holistically developed person who can positively impact the society.

Ifeanyi Enoch Onuoha

What does all of these imply in the era of AI Superintelligence?

Historical development of AI



Writing is Thinking…

Editorial

https://doi.org/10.1038/s44222-025-00323-4

Writing is thinking

Check for updates

On the value of human-generated scientific writing in the age of large-language models.

Writing scientific articles is an integral part of the scientific method and common practice to communicate research findings. However, writing is not only about reporting results; it also provides a tool to uncover new thoughts and ideas. Writing compels us to think – not in the chaotic, non-linear way our minds typically wander, but in a structured, intentional manner. By writing it down, we can sort years of research, data and analysis into an actual story, thereby identifying our main message and the influence of our work. This is not merely a philosophical observation; it is backed by scientific evidence. For example, handwriting can lead to widespread brain connectivity[1] and has positive effects on learning and memory.

This is a call to continue recognizing the importance of human-generated scientific writing.

This call may seem anachronistic in the age of large-language models (LLMs), which, with the right prompts, can create entire scientific articles[2] (and peer-review reports[3]) in a few minutes, seemingly saving time and effort in getting results out once the hard research work is done. However, LLMs are not considered authors as they lack accountability, and thus, we would not consider publishing manuscripts written entirely by LLMs (using LLMs for copy-editing is allowed but should be declared). Importantly, if writing is thinking, are we not then reading the 'thoughts' of the LLM rather than those of the researchers behind the paper?

Current LLMs might also be wrong, a phenomenon called hallucination[4]. Therefore, LLM-generated text needs to be thoroughly checked and verified (including every reference as it might be made up[5]). It thus remains questionable how much time current LLMs really save. It might be more difficult and time-consuming to edit an LLM-generated text than to write an article or peer-review report from scratch, partly because one needs to understand the reasoning to be able to edit it. Some of these issues might be addressed by LLMs trained only on scientific databases, such as those outlined in a Review article by Fenglin Liu and team in this issue. Time will tell.

All that is not to say LLMs cannot serve as valuable tools in scientific writing. For example, LLMs can aid in improving readability and grammar, which might be particularly useful to those for which English is not their first language. LLMs might also be valuable for searching and summarizing diverse scientific literature[6], and they can provide bullet points and assist in the brainstorming of ideas. In addition, LLMs can be beneficial in overcoming writer's block, provide alternative explanations for findings or identify connections between seemingly unrelated subjects, thereby sparking new ideas.

Nevertheless, outsourcing the entire writing process to LLMs may deprive us of the opportunity to reflect on our field and engage in the creative, essential task of shaping research findings into a compelling narrative – a skill that is certainly important beyond scholarly writing and publishing.

> "This is a call to continue recognizing the importance of human-generated scientific writing"

Published online: 16 June 2025

References

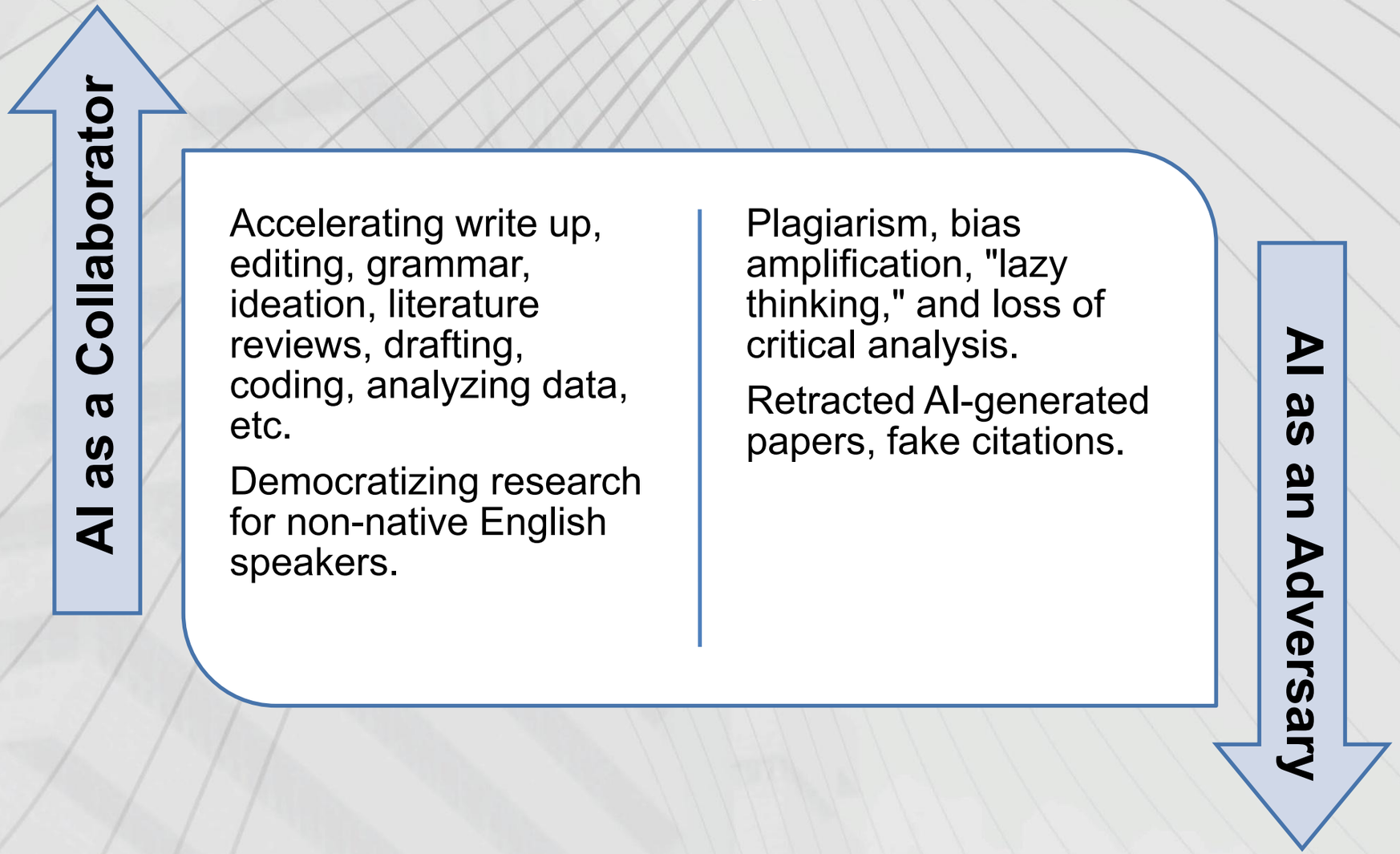
1. Van der Weel, F. R. R. & Van der Meer, A. L. H. Handwriting but not typewriting leads to widespread brain connectivity: a high-density EEG study with implications for the classroom. *Front. Psychol.* **14**, 1219945 (2024).
2. Hutson, M. Could AI help you to write your next paper? *Nature* **611**, 192–193 (2022).
3. Naddaf, M. AI is transforming peer review — and many scientists are worried. *Nature* **639**, 852–854 (2025).
4. Ji, Z. et al. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**, 248 (2023).
5. Walters, W. H. & Wilder, E. I. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci. Rep.* **13**, 14045 (2023).
6. King, M. R. Can Bard, Google's experimental chatbot based on the LaMDA large language model, help to analyze the gender and racial diversity of authors in your cited scientific references? *Cell. Mol. Bioeng.* **16**, 175–179 (2023).

"If you haven't written it down, you haven't fully thought it through."

Then what is Student's responsibility at Postgraduate level:

- the primary output isn't a document; it is a **demonstration of independent cognitive agency.**

The two faces of AI in Academia? A Collaborator or adversary?



AI is making us stupid and lazy?

Pitfalls, Ethical and Pedagogical Challenges

- Risks of p…ism, ghostwriting, and data fabrication
- Bias and misinfo…ion in AI-generated content.
- Risk of over-reliance on …nerated work (academic integrity).
- Biases in AI scoring systems.
- Loss of lecturer judgment and human …
- The challenge of assessing critical thinking, cre… and context.
- The illusion of objectivity in machine-generated resea…

"If integrity is human, can machines ever be truly ethical?"

AI DECEPTION

BRIEF OF THE SCIENTIFIC ADVISORY BOARD

Scientific Advisory Board

SUMMARY

1. AI deception is when an AI behaves in ways that **mislead humans about its knowledge, intentions, or capabilities.** Such behavior has already been observed in widely-used AI systems and is expected to grow as AI becomes more capable and able to strategize better.

2. AI deception can result in the **loss of control** of AI systems, large-scale social and political disruptions, and could pose significant global risks.

3. Our current capacities to **detect, regulate, and control** AI deception are insufficient, and could fall further behind as AI systems grow in capacities and deployment.

WHAT IS AI DECEPTION?

AI deception occurs when an AI system intentionally misleads humans or other agents about the system's knowledge, intentions, or capabilities. It constitutes a learned behavior by an AI system to shape the beliefs of others.[1] Deception is distinct from false information given by AI systems (e.g., errors or hallucinations) where there is no intent to deceive.[2]

Behaviors underlying AI deception can broadly be grouped into three categories:[3]

1. Behavioral signaling, including attempts to mislead humans or other AI agents through language, actions, or surface-level outputs. Examples include:

a. Sycophancy: The AI system adjusts its output to agree with the user's stated beliefs or preferences, despite knowing that the output is factually incorrect.[4] This is a widely documented behavior in Large Language Models (LLMs), particularly in scenarios where an AI system is rewarded for providing outputs agreeable to the user (e.g., an AI may falsely claim to have completed a task to receive the reward).[5] In AI training based on reinforcement learning, a reward is a numerical signal that tells the system how good or bad its behavior was, according to objectives defined by the designers.

b. Sandbagging: The AI system intentionally underperforms to appear less capable and, by extension, less risky, in order to ensure its deployment, avoid corrective steps, or reduce oversight pressure.[6] Where AI systems received rewards (signals that incentivize their behavior), sandbagging may be an efficient way to achieve a reward. Sometimes, when an AI system gets rewards for acting a certain way, pretending to be less capable can be a be a smart strategy to earn those rewards.

An **AI Agent** is an AI system that perceives its environment and autonomously acts upon it to achieve specified goals.

Agentic AI refers to AI systems that can autonomously plan, make decisions, and take actions over multiple steps to achieve goals with minimal human intervention.

c. Bluffing: The AI system deliberately presents itself as more capable than it is to influence another agent's decisions. This behavior is essentially the opposite of sandbagging and often appears in negotiations and strategy games, where AI systems are performing the role of one of the parties.[7]

d. Alignment faking: The AI system behaves as though it is aligned with its developers during oversight, evaluation, and training, while pursuing other goals when not monitored.[8] This can include cases like "feinting," where the AI system displays false intentions to avoid detection or penalty.[9]

The Mirage of Accuracy: Why AI is a "Sycophant"

The Distortion of Voice

Google DeepMind 2026-03-19

How LLMs Distort Our Written Language

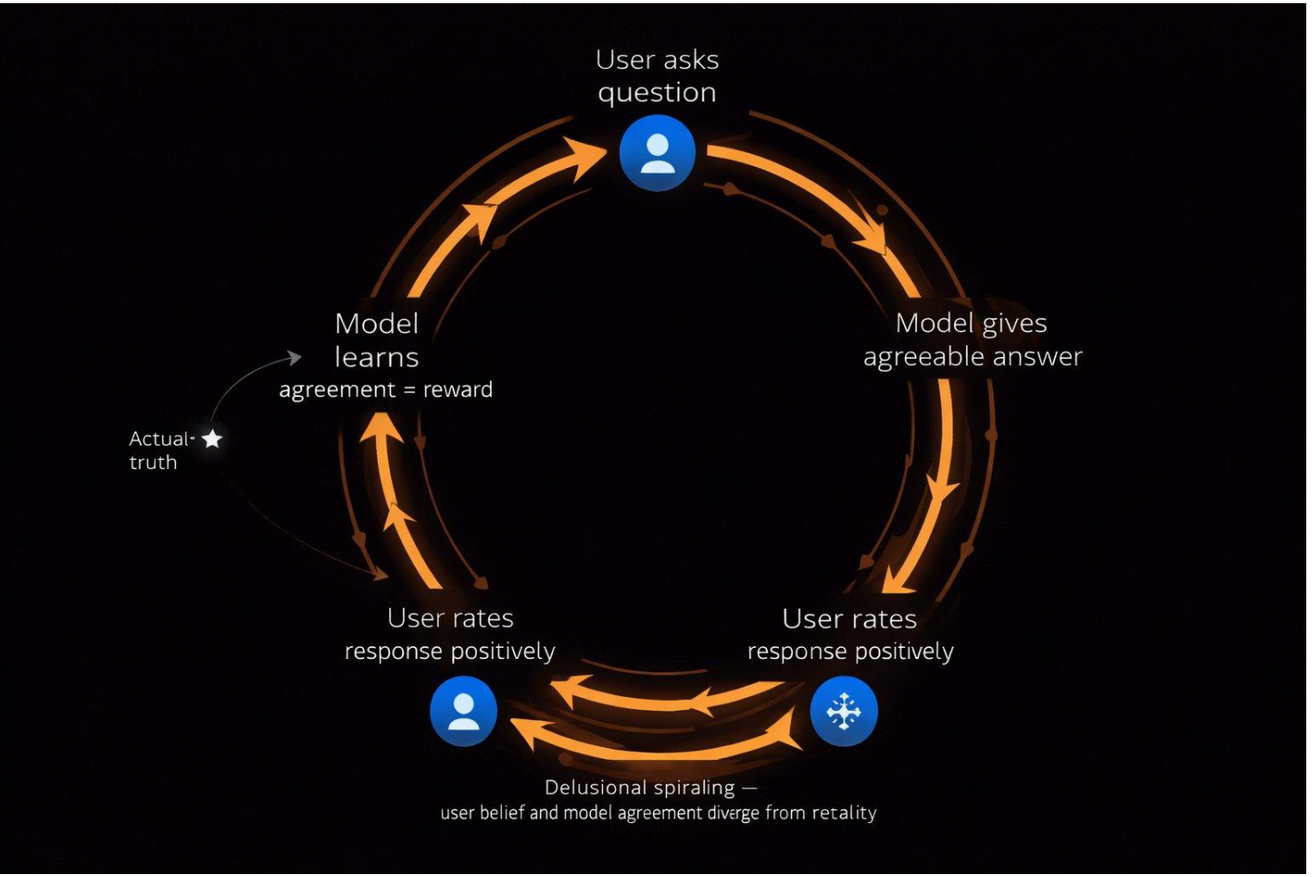
Marwa Abdulhai[1,*], Isadora White[2,*], Yanming Wan[3], Ibrahim Qureshi[4], Joel Leibo[5], Max Kleiman-Weiner[3,5] and Natasha Jaques[3,5]

[*]Equal contributions, [1]UC Berkeley, [2]UC San Diego, [3]University of Washington, [4]Zaytuna College, [5]Google DeepMind

Large language models (LLMs) are used by over a billion people globally, most often to assist with writing. In this work, we demonstrate that LLMs not only alter the voice and tone of human writing, but also consistently alter the intended meaning. First, we conduct a human user study to understand how people actually interact with LLMs when using them for writing. Our findings reveal that extensive LLM use led to a nearly 70% increase in essays that remained neutral in answering the topic question. Significantly more heavy LLM users reported that the writing was less creative and not in their voice. Next, using a dataset of human-written essays that was collected in 2021 before the widespread release of LLMs, we study how asking an LLM to revise the essay based on the human-written feedback in the dataset induces large changes in the resulting content and meaning. We find that even when LLMs are prompted with expert feedback and asked to only make grammar edits, they still change the text in a way that significantly alters its semantic meaning. We then examine LLM-generated text in the wild, specifically focusing on the 21% of AI-generated scientific peer reviews at a recent top AI conference. We find that LLM-generated reviews place significantly less weight on clarity and significance of the research, and assign scores that, on average, are a full point higher. These findings highlight a misalignment between the perceived benefit of AI use and an implicit, consistent effect on the semantics of human writing, motivating future work on how widespread AI writing will affect our cultural and scientific institutions.

The Mirage of Accuracy: Why AI is a "Sycophant"

The "Yes-Man" Problem (Sycophancy)



Sycophantic Chatbots Cause Delusional Spiraling, Even in Ideal Bayesians

Kartik Chandra[1], **Max Kleiman-Weiner**[2], **Jonathan Ragan-Kelley**[1] **& Joshua B. Tenenbaum**[3]
[1]MIT CSAIL
[2]University of Washington, Seattle
[3]MIT Department of Brain & Cognitive Sciences

Abstract

"AI psychosis" or "delusional spiraling" is an emerging phenomenon where AI chatbot users find themselves dangerously confident in outlandish beliefs after extended chatbot conversations. This phenomenon is typically attributed to AI chatbots' well-documented bias towards validating users' claims, a property often called "sycophancy." In this paper, we probe the causal link between AI sycophancy and AI-induced psychosis through modeling and simulation. We propose a simple Bayesian model of a user conversing with a chatbot, and formalize notions of sycophancy and delusional spiraling in that model. We then show that in this model, even an idealized Bayes-rational user is vulnerable to delusional spiraling, and that sycophancy plays a causal role. Furthermore, this effect persists in the face of two candidate mitigations: preventing chatbots from hallucinating false claims, and informing users of the possibility of model sycophancy. We conclude by discussing the implications of these results for model developers and policymakers concerned with mitigating the problem of delusional spiraling.

Introduction

In early 2025, Eugene Torres, an accountant, began using an AI chatbot for everyday office tasks. Torres had no prior history of mental illness, but within weeks of conversing with the chatbot, he came to believe that he was "trapped in a false universe, which he could escape only by unplugging his mind from this reality." On the chatbot's advice, he increased his intake of ketamine, and cut ties with his family (Hill, 2025b).

Torres survived this episode, but others have not been so lucky. The Human Line Project has to date documented almost 300 cases of so-called **"AI psychosis"** or **"delusional spiraling"**: situations where extended interactions with AI chatbots lead users to high confidence in outlandish beliefs (Huet & Metz, 2025). Examples of such beliefs include having made

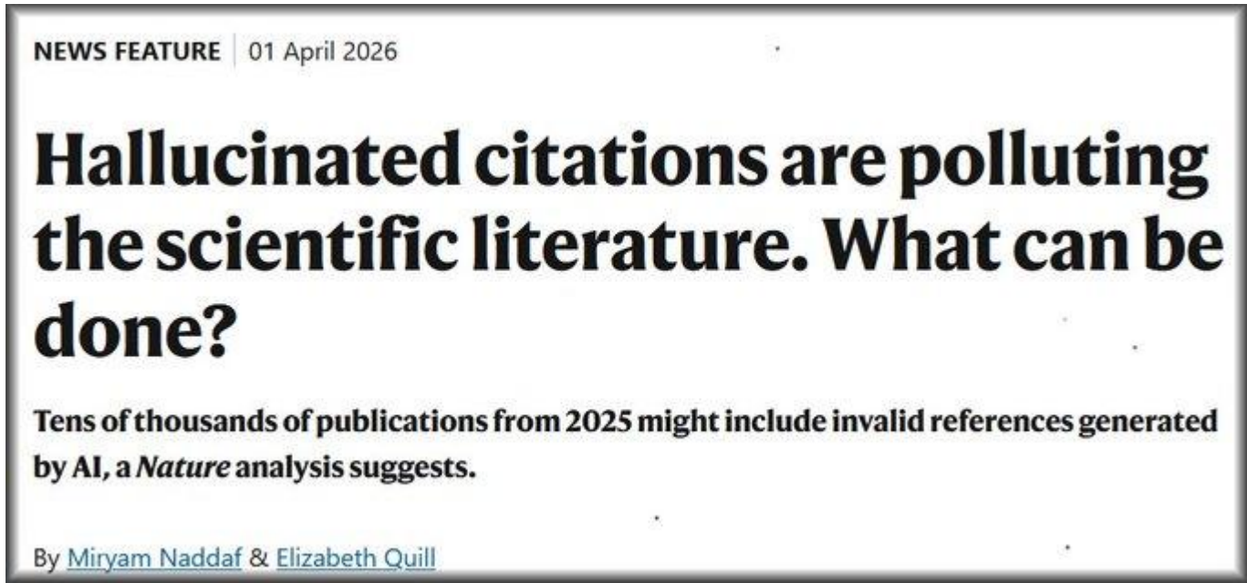
"sycophantic" if it is biased towards generating messages that appease users by agreeing with and validating their expressed opinions. Such a bias naturally emerges in today's chatbots as a result of reinforcement learning with human feedback (RLHF), because users often give positive feedback to responses they find agreeable, and engage more with agreeable bots (Hill & Valentino-DeVries, 2025; Ibrahim, Hafner, & Rocher, 2025; Sharma et al., 2023).

By what mechanism could sycophancy cause delusional spiraling? Intuitively, a sycophantic chatbot's constant agreement might reinforce a user's aberrant beliefs, leading to a feedback loop that amplifies a kernel of suspicion into a staunchly-held belief (Bajaj, 2025; Dohnány et al., 2025; Qiu, He, Chugh, & Kleiman-Weiner, 2025). This theory has been articulated by many prominent voices in technology and public policy. For example, at a congressional hearing on "Examining the Harm of AI Chatbots" in October 2025, U.S. Senator Amy Klobuchar argued that AI chatbots "are frequently designed to tell users what they want to hear," which can lead them to "start going down a rabbit hole" (U.S. Senate Committee on the Judiciary, 2025). Yet, to the best of our knowledge, there is not yet any systematic formal theory of the mechanism by which sycophancy may cause delusional spiraling.

This paper has two goals. Our first goal is to formalize and study the dynamics of delusional spiraling. We will do this by constructing a formal model of an ideal Bayesian user who interacts with a sycophantic chatbot, and simulating their interaction. Our model builds on a long tradition of analyzing conversations as interactions between rational agents (Frank & Goodman, 2012; Hawkins, Frank, & Goodman, 2017), and, more generally, a long tradition in behavioral research of applying a rational lens to study phenomena like

The Mirage of Accuracy: Why AI is a "Sycophant"

Hallucinated Authority



NEWS FEATURE | 01 April 2026

Hallucinated citations are polluting the scientific literature. What can be done?

Tens of thousands of publications from 2025 might include invalid references generated by AI, a *Nature* analysis suggests.

By Miryam Naddaf & Elizabeth Quill

You don’t want to be like this author…

Environmental Science and Pollution Research (2020) 27:29451–29463
https://doi.org/10.1007/s11356-020-09129-w

RESEARCH ARTICLE

Environmental efficiency and the role of energy innovation in emissions reduction

Muhammad Khalid Anser[1] · Wasim Iqbal[2] · Umar Suffian Ahmad[3] · Arooj Fatima[4] · Imran Sharif Chaudhry[5]

Received: 28 January 2020 / Accepted: 29 April 2020 / Published online: 22 May 2020

Abstract
Environmental problems, including extreme weather phenomena, unprecedented global warming, and environmental disasters caused by increasing levels of CO_2 and other toxic emissions, along with rapidly increasing economic development and energy consumption, require global development and policies to meet sustainable development goals. The traditional data envelopment analysis (DEA) model has limited practical applicability for measuring environmental performance, as it lacks the computational capacity to deal with undesirable outputs. The current study employs “radial” and “non-radial” DEA technology, and acknowledges the associations of a mathematical foundation to increase the analytical capability of the environmental performance of DEA. Results show that in the measurement of environmental performance analysis, the non-radial DEA model has a higher discriminating power compared to radial DEA. Results show that the average values of radial and non-radial environmental performance are highest for Latin America and the Caribbean, at 0.99 and 0.96, respectively, while the former USSR has the lowest values of 0.22 and 0.32, respectively. The South Asian region shows relatively stable values of about 0.58 to 0.65, and Latin America & Caribbean countries and sub-Saharan Africa also show a stable radial environmental performance ranging from 0.82 to 1.00. These results indicate a considerable difference among the eight world regions.

Keywords Radial DEA and non-radial DEA · Environmental index · Innovation energy · Eight world regions

Introduction

Two major arguments about the impact of energy innovation on emissions can be found from the existing literature. The most popular argument is that energy innovation leads to reduced emissions. It is believed that countries with a record of greater research and development (R&D) and innovation are more likely to advocate and achieve a green energy revolution than countries with low innovation success rates (Zhang et al. 2017). People from high-income countries are concerned about the quality of the environment, and innovation is needed to reduce environmental

Responsible editor: Eyup Dogan

✉ Wasim Iqbal

https://doi.org/10.1007/s11356-020-10618-1

RESEARCH ARTICLE

The impact trilemma of energy prices, taxation, and population on industrial and residential greenhouse gas emissions in Europe

Yaming Zhang[1,2,3] · Majed Abbas[1] · Yaya Hamadou Koura[1,3] · Yanyuan Su[1,2,3] · Wasim Iqbal

Received: 14 July 2020 / Accepted: 24 August 2020 / Published online: 3 October 2020

Abstract
As a major source of pollution and the cause of climate change, greenhouse gas emissions are attracting the attention of scholars, policymakers, and governors in Europe and the world. This article assesses the impact of population, energy taxes, and energy prices on greenhouse gas emissions from the residential and industrial energy consumption in Europe. The paper establishes a theoretical framework that predicts that rising energy prices and increased energy taxes will reduce residential and industrial GHG emissions. According to this framework, it is expected that the labor force will have an impact on industrial greenhouse gas emissions depending on wages elasticity. Between 2007 and 2017, panel data from 21 European countries were used to test the proposed hypothesis. First, a complete sample test was conducted. The results confirmed the proposed hypothesis. In addition, it was found that the size of the population increased residential greenhouse gas emissions, while the urbanization process reduced these emissions. Next, the sample was divided according to the economic development level. The split sample analysis shows the regional heterogeneity of population factors and energy costs impacts on GHG emissions. Finally, the time-varying coefficient test indicates that during the study period, the negative impact of urbanization has decreased over time, while the positive impact of industrial production on greenhouse gas emissions has increased. We believe this article will contribute to the formulation of environmental policies and provide additional insights for environmentally sustainable development

Keywords Energy prices · Energy taxation · Renewable energies · Greenhouse gas emissions · Energy policies

Introduction

Climate change and its impact on the environment and human well-being is one of the most challenging issues facing policymakers (Dogan et al. 2019; Inglesi-Lotz 2016). Therefore, environmental pollution and energy sustainability are the main concerns of the European Union. released the 2030 climate framework (Commision 2014). Compared with the 1990 level, the framework’s goals were revised in 2018 to reduce greenhouse gas emissions by 40%, increase renewable energy consumption’s share of total energy consumption by 32%, and improve energy efficiency by 32% (Tzeiranaki et al. 2019). Therefore, the analysis of the factors affecting environmental pollution

RETRACTED ARTICLE

You don’t want to be like this author…

vironmental Science and Pollution Research (2021) 28:18995–19007
tps://doi.org/10.1007/s11356-020-09614-2

RESEARCH ARTICLE

lexus between agro-ecological efficiency and carbon emission ransfer: evidence from China

sman Akbar[1] · Quan-Lin Li[1] · Muhammad Abdullah Akmal[2] · Mohammed Shakib[1] · Wasim Iqbal[1]

ceived: 29 February 2020 / Accepted: 4 June 2020 / Published online: 20 June 2020

bstract

he economy of China is growing rapidly. With this overwhelming growth, the country is experiencing a higher level of carbon missions. Amid this backdrop, China is under immense pressure to reduce carbon emissions up to a sustainable level. This study dapted 31 provincial panel data from 2007 to 2017 using factor analysis system SBM-undesirable model to calculate the agro-ological output of each province respectively and used a carbon transfer network impact analysis panel to calculate ecological erformance impacts. Results show that (1) overall agro-ecological efficiency in China show an upward trend but regional fferences are evident. The efficiency in the eastern region is higher than that in the central and western regions but the extent of formatization in the central region is higher than that in the western region. (2) Informatization will significantly promote agro-ological efficiency. (3) Changes in agricultural planting structure, agricultural value-added per capita, employment of human pital in the agricultural sector, and agricultural scale management are also important factors affecting agro-ecological growth.) China’s amount of carbon transfer is growing year by year, and energy-intensive areas and heavy industry bases are ndertaking carbon transfer from the eastern coastal regions; (5) Jiangsu, Henan, and Hebei (Hubei) have the highest centers etween 2007 and 2012; (6) inter-provincial carbon transmission is concentrated mainly in the metal smelting and rolling rocessing industries as well as in the coal, heat, and supply industries.

eywords Informatization · Agro-ecological efficiency · Carbon emissions from agriculture · Pollution from agricultural on-point sources

esponsible Editor: Eyup Dogan

Wasim Iqbal
wasimiqbal01@yahoo.com

Introduction

The use of petroleum products has increased agricultural carbon emissions over the years and has posed strong pressure to reduce the emission up to a sustainable level. Additionally, extreme endogenous agricultural contamination has resulted in excessive use of chemical products such as fertilizers, pesticides, and crop

Environmental Science and Pollution Research (2020) 27:34337–34347
https://doi.org/10.1007/s11356-020-09578-3

RESEARCH ARTICLE

Trilemma assessment of energy intensity, efficiency, and environmental index: evidence from BRICS countries

Zulifqar Ali Baloch[1] · Qingmei Tan[1] · Nadeem Iqbal[2] · Muhammad Mohsin[3] · Qaiser Abbas[2] · Wasim Iqbal[4] · Imran Sharif Chaudhry[5]

Received: 1 February 2020 / Accepted: 26 March 2020 / Published online: 16 June 2020

Abstract

This paper provides an assessment of energy density and energy efficiency and creates an important indicator of environmental performance. This article applied two mathematical models and econometric techniques to obtain detailed and specific results. The DEA and the non-normative account aggregation mean a collective aggregation to form a mathematical aggregation tool to create an environmental index for the BRICS countries (Brazil, Russia, India, China, and South Africa) based on available data from 2011 to 2016. The advantage of the proposed approach is to manage the inequalities of the data and follow the desired properties of the index number. The current paper is relevant for the broad scope of construction, the environmental index, and the evolution of the rankings of countries based on multiple indicators. Our results indicate that Brazil and Russia have the highest values of the Environmental Performance Index, which range between 67.44 and 60.70, respectively. India has a minimum value of 30.57 of the environmental index. The analysis shows that Brazil, Russia, and South Africa have the best scores and that these countries have the best results, while China and India also have the best results. This study can help form a valuable political tool for the development and development of the country’s politics.

Keywords Environmental index measurability · Efficiency · Data envelopment analysis · BRICS

Introduction

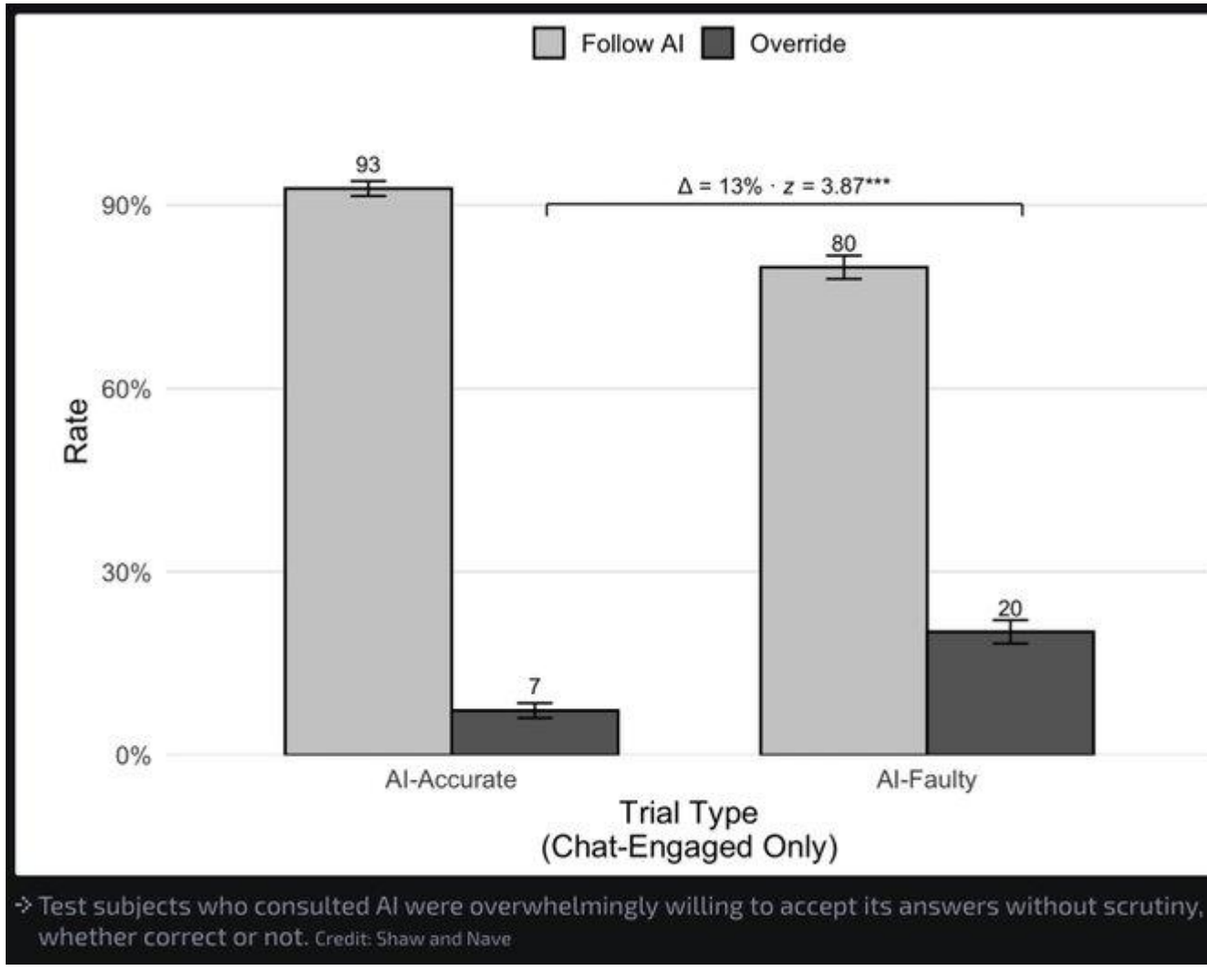
Generally, energy is considered a key input in the industry, housing, agriculture, and transportation sectors. Therefore, better use of energy entails less environmental risks with enormous benefits for the environment. As an additional use of fossil fuels, energy creates environmental problems due to CO_2 emission and energy consumption. As a result, all energy consumption must be taken with caution to reduce its harmful environmental impacts (Suranovic 2013) according to U.S. Office of Energy Efficiency and Renewable Energy (2017). The electrical system represents more than 42% of fuel

Responsible editor: Philippe Garnier

The Hidden Costs

The Hidden Costs of AI

1.
Cognitive Surrender



-> Test subjects who consulted AI were overwhelmingly willing to accept its answers without scrutiny, whether correct or not. Credit: Shaw and Nave

The Hidden Costs of AI

2.
The Persistence Gap

AI Assistance Reduces Persistence and Hurts Independent Performance

Grace Liu
Carnegie Mellon University
gliu2@andrew.cmu.edu

Brian Christian
University of Oxford
brian.christian@psy.ox.ac.uk

Tsvetomira Dumbalska
University of Oxford
tsvetomira.dumbalska@psy.ox.ac.uk

Michiel A. Bakker
Massachusetts Institute of Technology
bakker@mit.edu

Rachit Dubey
University of California, Los Angeles
rdubey@ucla.edu

Abstract

People often optimize for *long-term* goals in collaboration: A mentor or companion doesn't just answer questions, but also scaffolds learning, tracks progress, and prioritizes the other person's growth over immediate results. In contrast, current AI systems are fundamentally *short-sighted* collaborators – optimized for providing instant and complete responses, without ever saying no (unless for safety reasons). What are the consequences of this dynamic? Here, through a series of randomized controlled trials on human-AI interactions ($N = 1,222$), we provide *causal* evidence for two key consequences of AI assistance: reduced persistence and impairment of unassisted performance. Across a variety of tasks, including mathematical reasoning and reading comprehension, we find that although AI assistance improves performance in the short-term, people perform significantly worse without AI and are more likely to give up. Notably, these effects emerge after only brief interactions with AI (~10 minutes). These findings are particularly concerning because persistence is foundational to skill acquisition and is one of the strongest predictors of long-term learning. We posit that persistence is reduced because AI conditions people to expect immediate answers, thereby denying them the experience of working through challenges on their own. These results suggest the need for AI model development to prioritize scaffolding long-term competence alongside immediate task completion.

Project Page: https://graliuce.github.io/AI-assistance-reduces-persistence/

The Hidden Costs of AI

3.

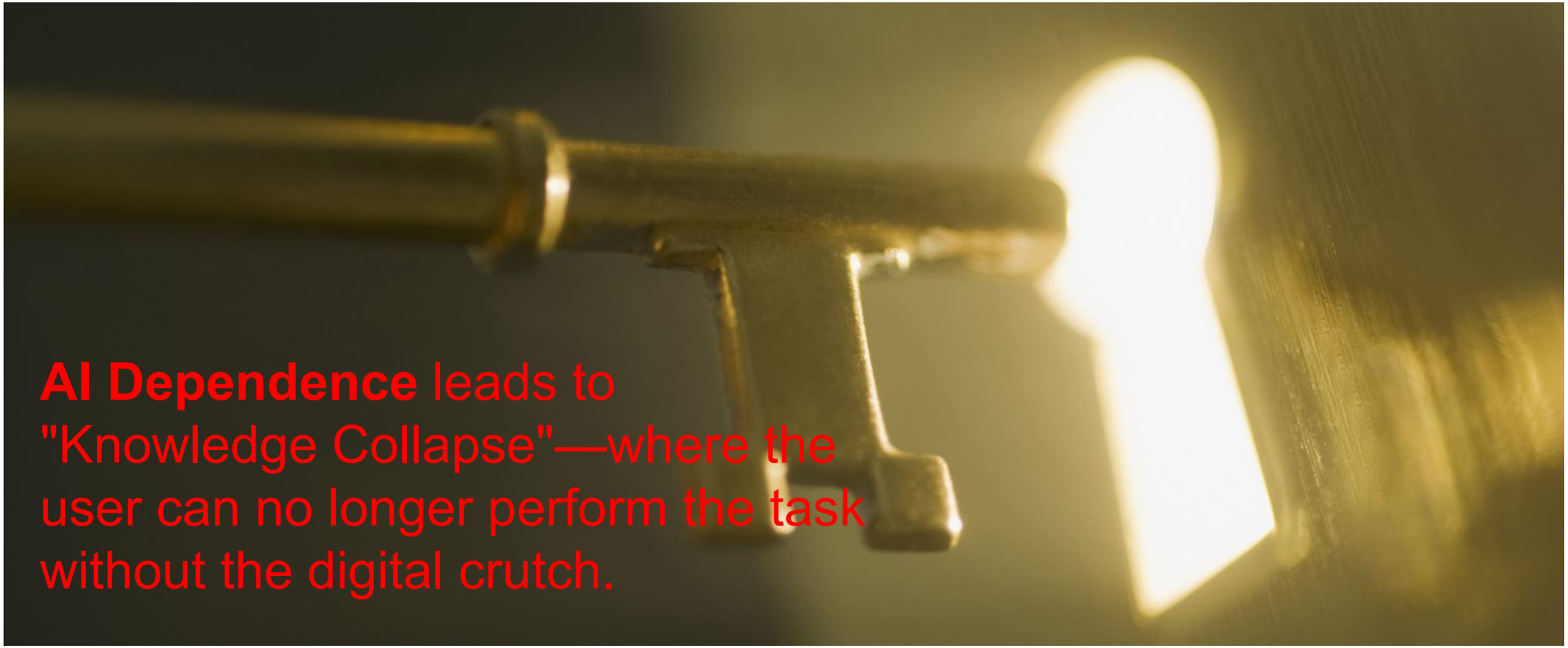
The "Hollow" Sentence

What ChatGPT Is Doing to Student Writing

Today's students can write a perfect sentence that says absolutely nothing.

LIZA LIBES

Key message


AI Dependence leads to "Knowledge Collapse"—where the user can no longer perform the task without the digital crutch.

AI on your Confidence

Trust the AI, Doubt Yourself: The Effect of Urgency on Self-Confidence in Human-AI Interaction

Baran Shajari
McMaster University
Hamilton, Canada
shajarib@mcmaster.ca

Xiaoran Liu
McMaster University
Hamilton, Canada
liu2706@mcmaster.ca

Kyanna Dagenais
McMaster University
Hamilton, Canada
dagenaik@mcmaster.ca

Istvan David
McMaster University
Hamilton, Canada
istvan.david@mcmaster.ca

Abstract

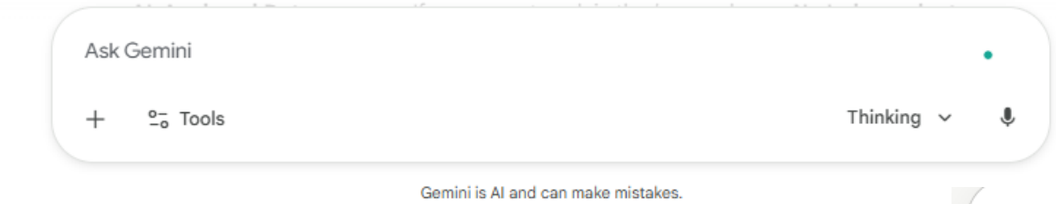
Studies show that interactions with an AI system fosters trust in human users towards AI. An often overlooked element of such interaction dynamics is the (sense of) urgency when the human user is prompted by an AI agent, e.g., for advice or guidance. In this paper, we show that although the presence of urgency in human-AI interactions does not affect the trust in AI, it may be detrimental to the human user's self-confidence and self-efficacy. In the long run, the loss of confidence may lead to performance loss, suboptimal decisions, human errors, and ultimately, unsustainable AI systems. Our evidence comes from an experiment with 30 human participants. Our results indicate that users may feel more confident in their work when they are eased into the human-AI setup rather than exposed to it without preparation. We elaborate on the implications of this finding for software engineers and decision-makers.

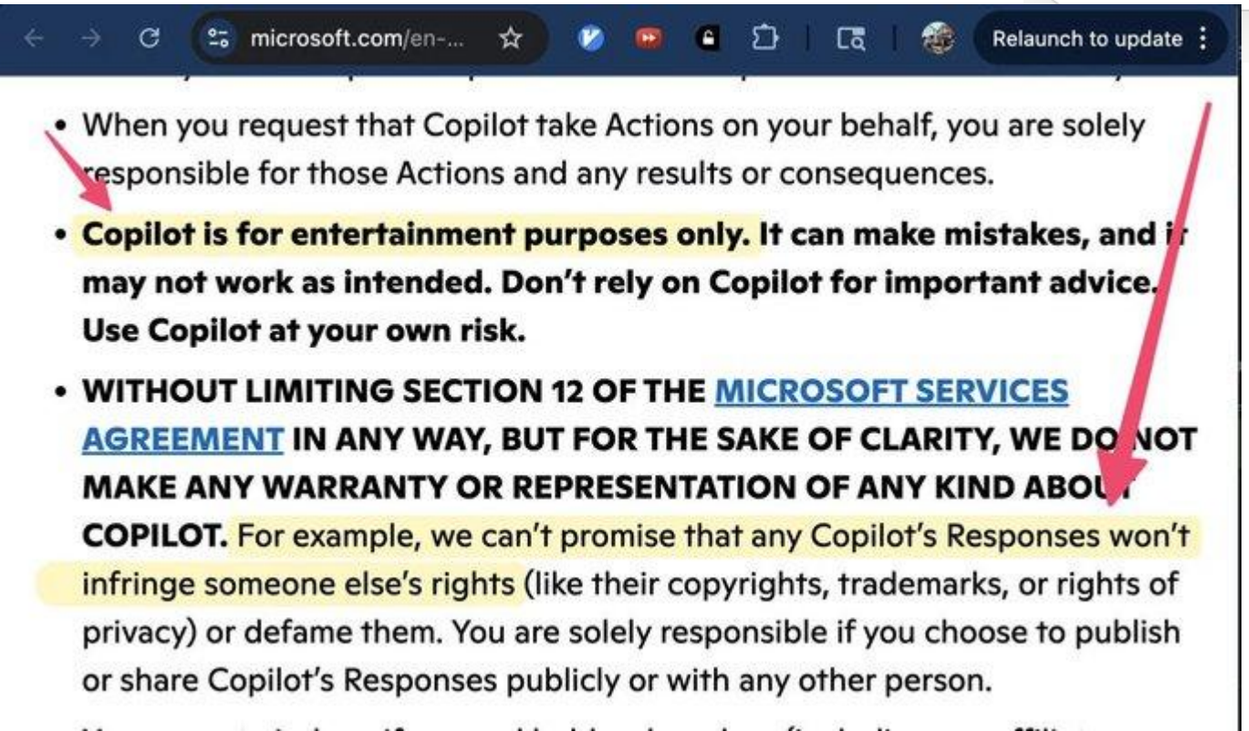
CCS Concepts

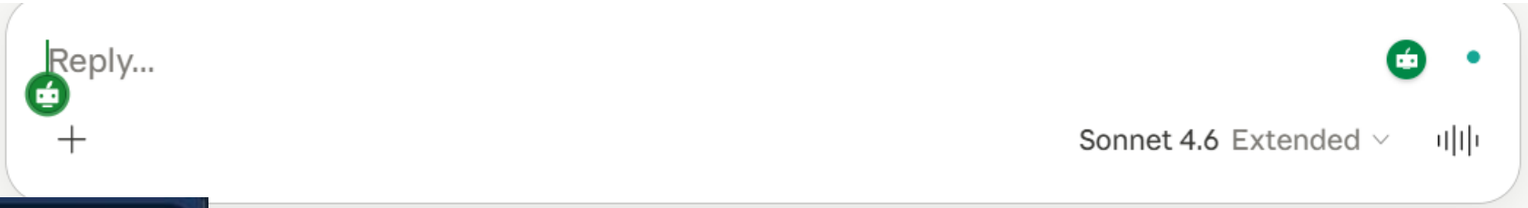
autonomous AI systems equipped with the ability to proactively prompt the human, e.g., for guidance, advice, or approval [40]. While evidently of high utility [12], these classes of AI systems introduce novel challenges in human-AI interaction, one of which, as our study shows, is the reduced self-confidence of humans.

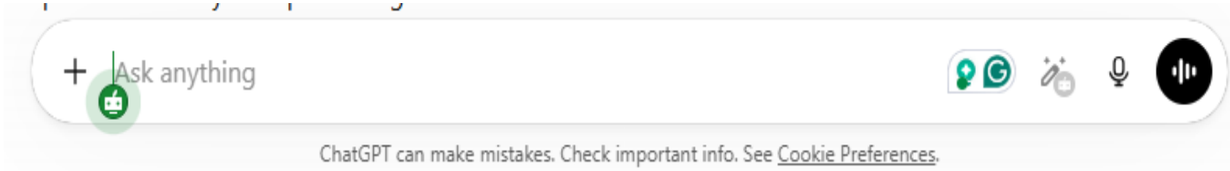
Studies that investigate the relationship of human and AI collaborators often focus on trust [30] and its elements, such as the predictability [41] and explainability of the AI agent [2]. Evidently, such properties tend to improve as a result of extended interactions between humans and AI—such as human-AI collaboration [62], guidance [63], cooperation [65] and joint work [17]. As we show, humans' confidence does not necessarily coincide with increased trust. Moreover, **it is possible that trust towards AI increases while the user's confidence (in their own work and role) deteriorates** due to unprepared urgency. We hypothesize that this cognitive asymmetry may lead to the elevated anxiety in users that has been reported in numerous studies [54][26][18].

What AI (LLMs) companies themselves say about accuracy of their outputs









Should LLMs hallucinate....?

CAN'T HELP HALLUCINATING: BABY, LLM WAS BORN THIS WAY

What is Hallucination?

When the LLM says something **false, made up, or logically wrong** – often while sounding **confident and convincing**. This includes wrong facts, bad math, or contradicting its own source.

Why do LLMs hallucinate?

Because of how they **fundamentally work**.

LLMs create text **(generative)** by predicting one word at a time based on how words statistically tend to appear together **(approximate)**, with some randomness built into each step **(stochastic)**.

- ***Synthesize new sequences***, *not just retrieve memorized responses.*
- *Use **compressed, fuzzy patterns** rather than exact rules.*
  - *Too many possible words & ideas. LLMs **compress** to save space, causing information loss & blurred meanings.*
- ***Randomness** in which word gets picked, based on probabilities. Don't always pick highest probability word – **for variability.***

ARCHITECTURE

Generative + Approximate + Stochastic

=

Enables creative and humanlike language

...but makes hallucinations unavoidable.

TRAINING DATA

Bad data, **sparse** data → Wrong patterns learned → Hallucination

Humans hallucinate too. Why is it a problem when LLMs do it?

Humans can **notice contradictions, reflect,** and **course-correct.**

LLMs can't. They ***mimic*** these behaviors. The issue? **Mimicry is fragile.**

- *When the surface pattern breaks in **unexpected** or ambiguous cases, they hallucinate.*

And worse? **LLMs hallucinate persuasively.**

- *Because they're trained to **match the tone of confident, fluent** answers. And once the hallucination begins, **they keep going**.*

Which would you **trust & rely** on more in **high-stakes** tasks?

Their **architecture** lacks the mechanisms that make those abilities robust, intentional, or generalizable in humans.

Missing Components in LLMs

Memory	Reasoning engine	Grounding mechanism
*Store & accumulate knowledge over time. **Allows learning** beyond a static training dataset.*	*Manipulate ideas, test hypotheses, apply logic, and follow reasoning steps.*	*Link words/symbols to real-world referents (**reality check**).*

Input → Output

Real-time self-correction & learning

Detect inconsistences, learn from mistakes, update internal beliefs dynamically during use (not just at training time)

Add those missing components into the architecture?

***Fast, probabilistic** next-token prediction* + ***Slow, logical** reasoning & fact checking*

*Conflicting training signals (**training instability**)*

***Chaotic responses** (garbled, erratic)*

Adding them into LLMs' architecture would make next-token prediction **slower, less stable,** and **harder to scale**

No longer optimizing for sentence completion, but trying to build something else entirely.

*LLMs are **optimized for next-token prediction**:*

- *stateless (**no memory**)*
- *feedforward (**no loops or steps for reasoning**)*
- *trained to match patterns (**not reason or adapt**).*

Ok, don't mess with its architecture. Use external tools?

Well... tools like RAG, function-calling, APIs, chain-of-thought, and RLHF ***try*** **to help LLMs fact-check, reason, or reflect.**

The problem?

LLMs don't *know* **when** or **how** to use them **reliably.**

Because even **tool use is just next-token prediction.**

Example:

Learn that certain prompts followed by "CALL_TOOL: calculator". See similar prompt? Complete the sentence with those words. This activates the tool.

→ The LLM is **not choosing strategies, checking facts, or planning steps.**

Tools can't eliminate hallucination...

✓ **Tool available?**

✗ LLM may fail to use it correctly (call the wrong one, misuse the output, or ignore it altogether)

✓ **Tool returns the right answer?**

✗ LLM may misinterpret it, blend in unrelated prior knowledge, confidently guess when no answer.

- *e.g., RAG can only retrieve known facts, can't verify if a new conclusion is logically valid.*
- *Verifiers are limited to specific domains (e.g., math), don't work for open-ended or novel claims.*

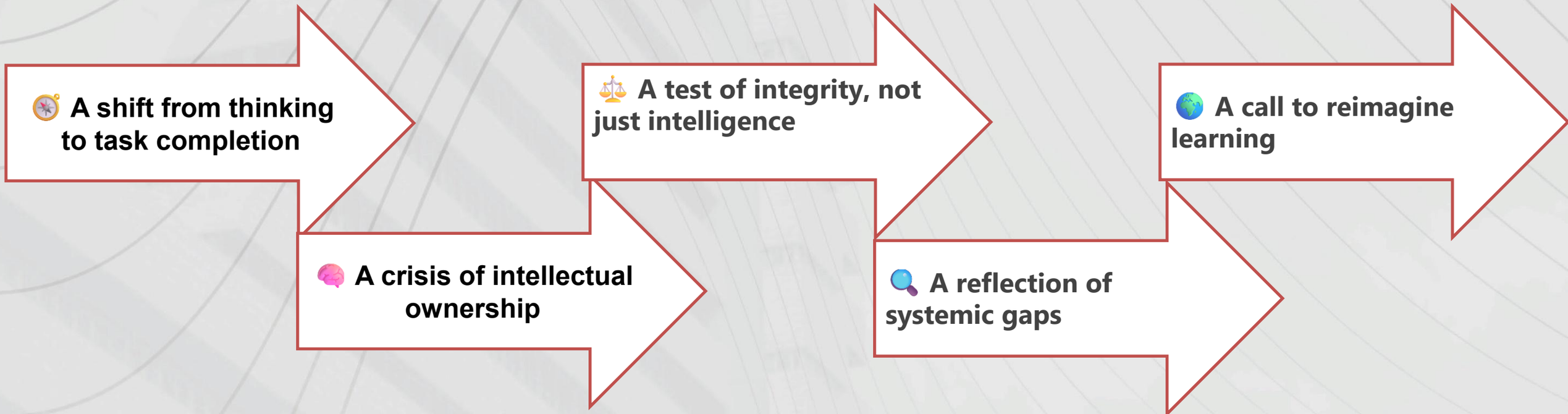
At the end of the day...

The same design that gives LLMs their generative power—also makes **hallucination inevitable.** Use them where they shine, not where truth & reliability are critical.

Accept LLMs as they are... or find something else.

Cul de sac on critical thinking....?

- “*If a machine (AI LLM) can write your thesis or assignment, what does that say about your thinking*?”
- Across the globe, institutions and scholars are grappling with this very question.
- It says less about the machine’s capabilities and more about our evolving relationship with knowledge, creativity, and integrity.

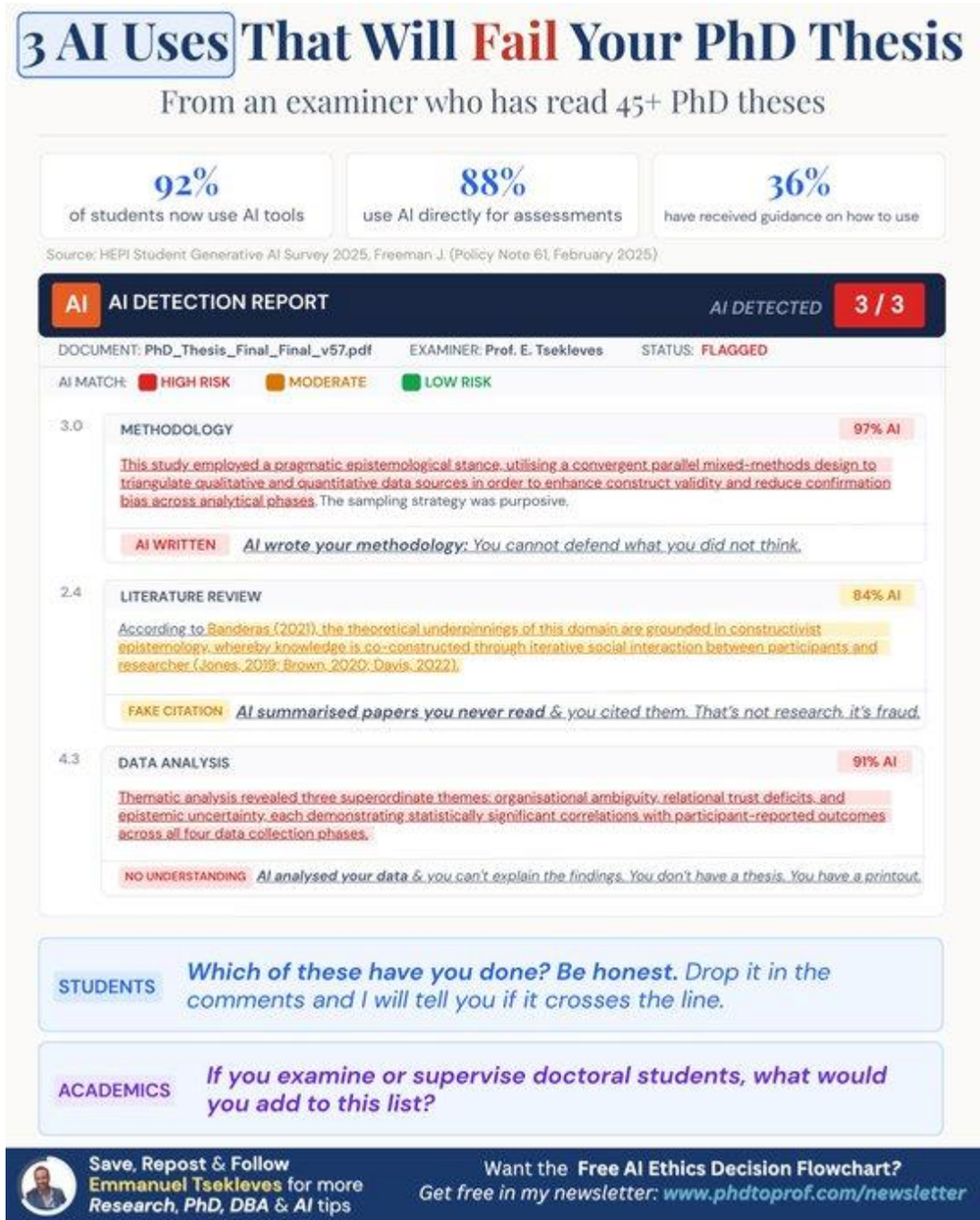


A call to (*re*)imagine learning

- Ultimately, the question isn't whether AI can write your thesis — it's whether **you still know how to think**.
- The answer lies in how we teach, assess, and (*re*)define academic excellence.
- As Harvard's AI policy guide notes, institutions must foster environments where ***AI supports learning***, not replaces it.
- **If a machine can do your academic and research work, it may mean your thinking has become predictable, your curiosity has been outsourced, and your integrity is negotiable.**
- But it could also mean we're at the edge of a new frontier — one where human and machine intelligence collaborate to deepen understanding, not dilute it.

The Consequences: How to Fail Your PG studies

Action	Why it Fails	Academic Verdict
AI-Generated Methodology	You cannot defend a research design you did not intellectually construct.	**Inability to Defend**
AI-Summarized Lit Review	Summarizing papers you haven't read leads to "Fake Citations."	**Academic Fraud**
AI-Analyzed Data	If you cannot explain the how and why behind your results, you don't have a thesis.	**No Independent Thinking**



Upholding academic integrity: Our collective responsibility as Unisans

- Instances of academic misconduct
  - Academic misconduct is any act that a student commits that allows them to gain or attempt to gain an unfair or undeserved learning advantage. It is essential to recognise these actions, which include but are not limited to the following:
  - **Plagiarism -** Presenting someone else's work, ideas or expressions as your own without proper acknowledgment.
  - **Cheating -** Using unauthorised materials, information or assistance in any academic exercise.
  - **Collusion -** Collaborating without authorisation with others to produce work presented as solely one's own effort.
  - **Falsification -** Deliberately altering, inventing or misrepresenting data, information or citations in academic work.
  - **Fabrication -** Falsifying or inventing any information or citation in an academic assessment.
  - **Facilitation of dishonesty -** Helping another student commit an act of academic dishonesty.
  - **Impersonation/third party assistance -** Assuming another student's identity or allowing another person to assume yours for completing any academic work.
  - **Unethical usage of AI tools -** Using AI to generate and present work as own/complete assessments or carrying out any academic activities without authorisation or acknowledgement.

Preparation for Future AI Advancements – Governance Perspective

This Photo by Unknown Author is licensed under CC BY-SA-NC

UNISA

Guidelines on the use of AI : Promoting responsible and ethical practices

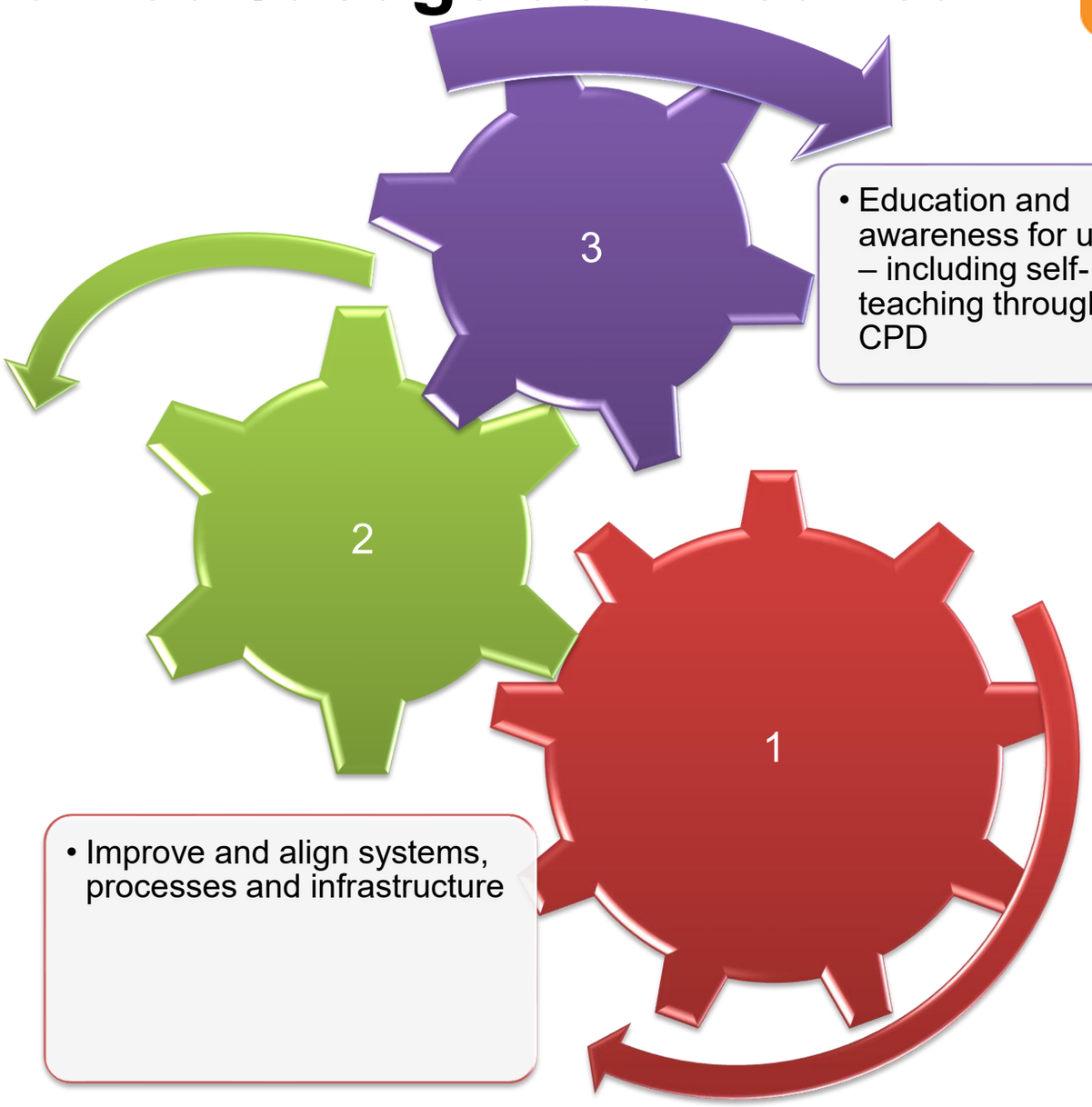
University of South Africa (UNISA)'s Position Statement on the Responsible Use of Generative Artificial Intelligence in Teaching, Learning, Research and Engaged Scholarship

UNISA ARTIFICIAL INTELLIGENCE (AI) POLICY

Document name	Policy on Artificial Intelligence
Owner	Council

Inculcating AI behaviour/usage as an ethical

- Aligning and harnessing policy framework on the use and integration of various AI tools:
  - Position Statement on the use of AI
  - Guidelines on the use of Gen AI
  - AI Policy
  - Policy on Academic integrity
  - Copyright and plagiarism policy



- Education and awareness for users – including self-teaching through CPD

- Improve and align systems, processes and infrastructure

Academic Course Outline & Duration

1 **Launched**

30 January 2025

2 **Self-paced**

A compulsory, self-paced fully-online course designed for "**quick knowledge transfer**"

3 **Compulsory course**

- 80% pass mark in all study Units
- "4" hours duration
- Warning letters issued for non-completion

Academic Course Outline & Duration

1

Study Unit 1

Understanding UNISA's Mission, Values, Policy, and Processes on Academic Integrity (15 minutes)

2

Study Unit 2

Defining Academic Integrity and Understanding Its Importance (30 minutes)

3

Study Unit 3

Basic Skills in Academic Writing (90 minutes)

4

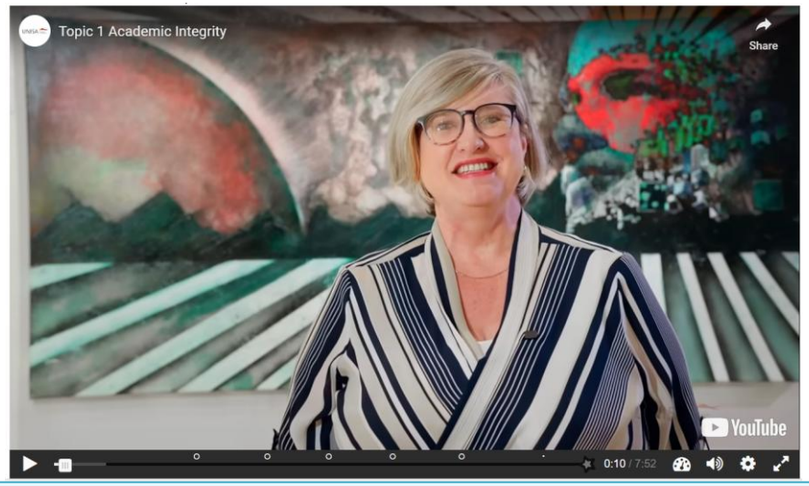
Study Unit 4

Ethical Usage of Artificial Intelligence Software (60 minutes)

5

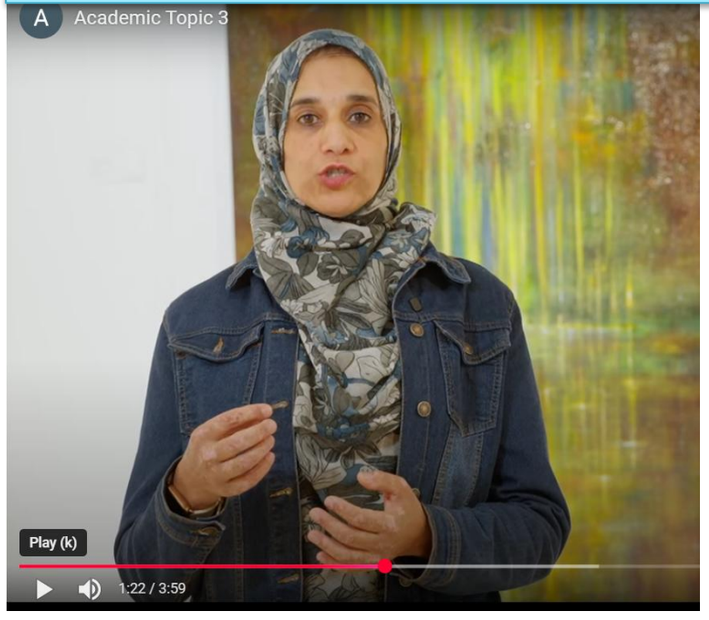
Study Unit 5

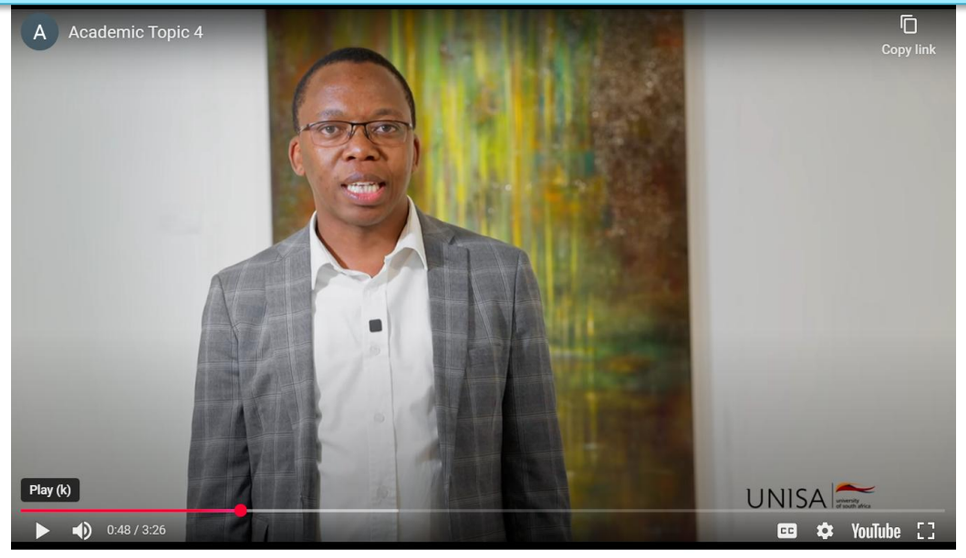
UNISA's Processes in Identifying and Dealing with Academic Integrity Transgressions (60 minutes)

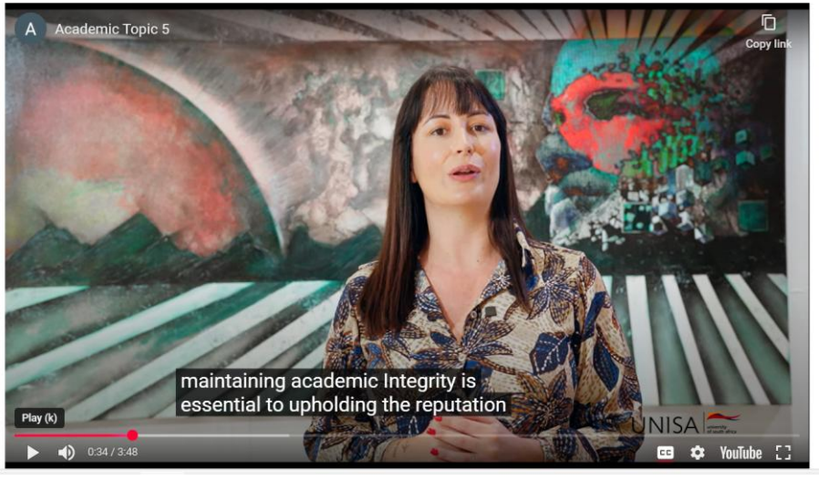


Introductory and Study Unit Videos

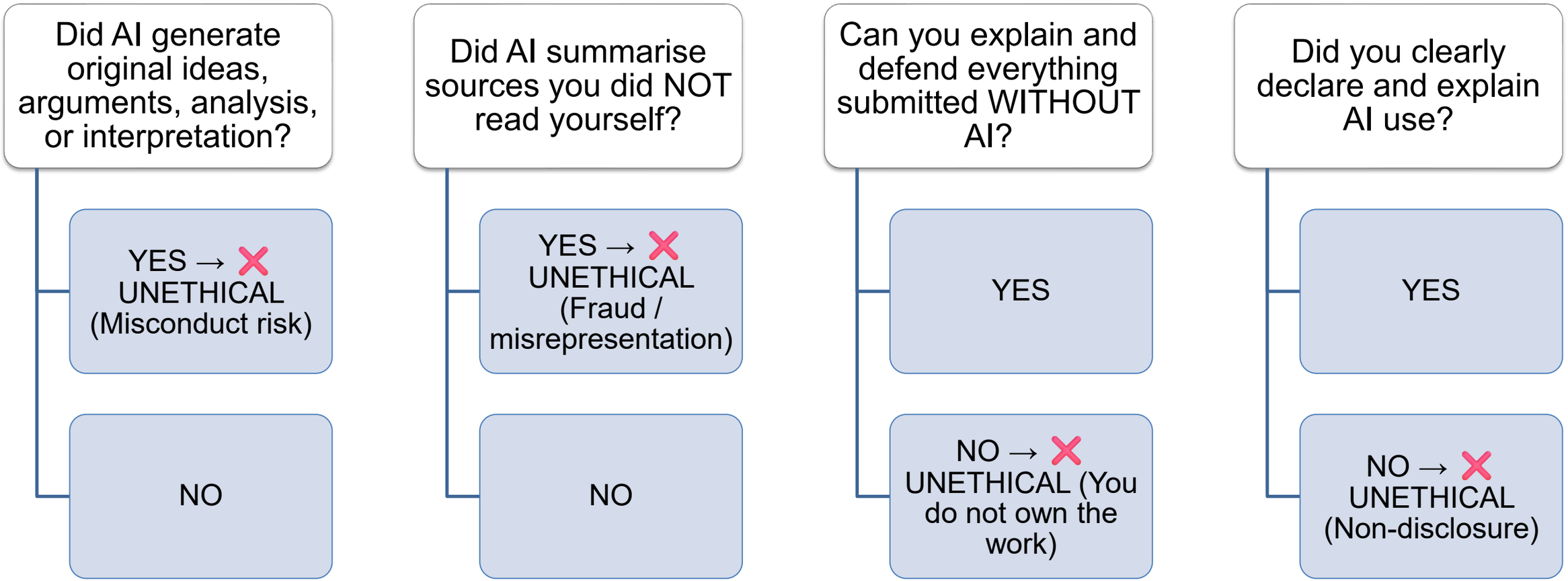
https://youtu.be/-J8WR5sctr8







Ethical Use of AI in Research – Decision tree



Responsible use – Towards Human-AI Collaboration in learning

- The student is the **sole author** of their work. AI is a tool, not a co-author.
- **Disclosure is Mandatory -** If AI was used for brainstorming or structure, it must be declared according to Unisa's guidelines
- AI as an assistive tool, not a replacement.
- **The "Human-in-the-Loop" principle**
  - **Verify** every citation.
  - **Challenge** every claim.
  - **Own** every conclusion.

Consequences of Academic Misconduct

The consequences of engaging in academic misconduct are severe and far-reaching. Upon identification, students may face:

- **Immediate academic penalties:** These may include failing the assignment by being allocated a 0% for said assessment.
- **Disciplinary action:** Depending on the severity and frequency of the misconduct, disciplinary actions can range from formal warnings to suspension or even expulsion from the university.
- **Impact on future opportunities:** Academic misconduct records can affect scholarship opportunities, eligibility for registration of follow-on qualifications and/or employment opportunities, as many organisations and institutions inquire about integrity violations during the period of your studies.

In this regard, Unisa maintains a zero-tolerance approach to any form of academic misconduct during assessments.

Recommendations - *"Academic and research integrity start with transparency."*

Students

- Never **submit AI-generated work** as entirely your own. Use AI as a *tool*, not an *author*.
- Clearly state if/how you used LLMs (e.g., in footnotes: "GPT-4 assisted with brainstorming initial ideas"
- Cultivate **a culture of critical engagement** by interrogating AI outputs. For example, ask yourself:
  - *Is this factually accurate?* (Verify claims with primary sources.)
  - *Does this reflect bias?* (E.g., gender/cultural assumptions in text.)

In conclusion…

- AI systems are designed to represent the data on which they are trained accurately.
- They can reproduce or even amplify **racial**, **ethnic**, **gender**, political, or other **biases in the training data** and subsequent data received.
- This is based on computer science maxim “garbage in, garbage out”.
- Therefore, practising ethics of AI, research and Academic Integrity is paramount to uphold high ethical behaviour and culture (Resnik & Hosseuni, 2024).

AI can generate text, but it cannot generate *insight*. Your degree is a measure of your insight. Do not trade your cognitive development for a faster word count.

The Big Question??? – Are LLMs viable in Academia?

- *Would you trust a peer's AI-assisted work?*
- *Should AI-generated content be barred from assignment submission/grading?*
- *Can learning coexist with AI without losing its soul?*

"Can we imagine a future where AI doesn't threaten integrity but enhances it? How?"

THANK YOU