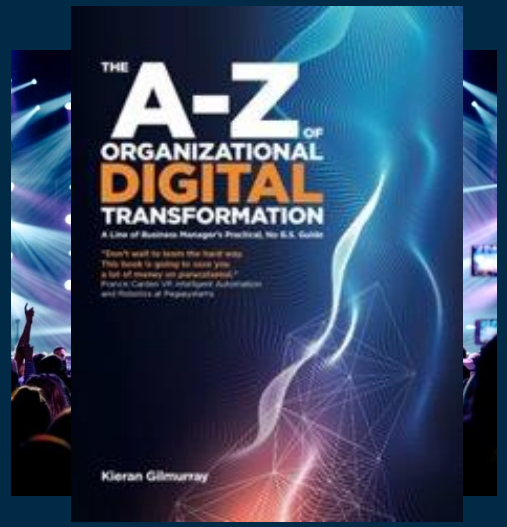
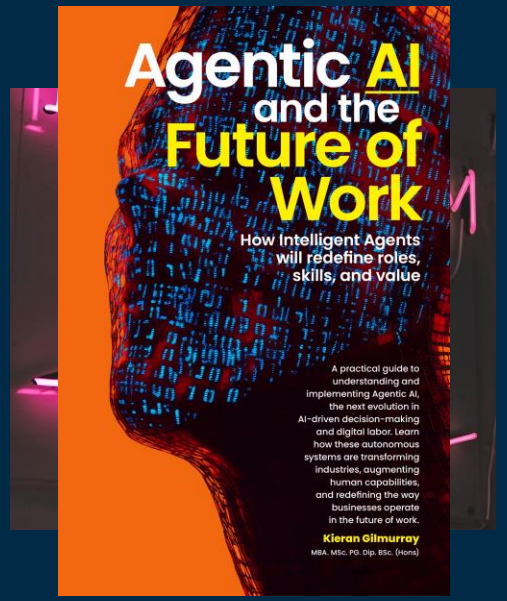
How AI Forms Bias and What Educators Can Do About It

Those who know AI <u>will</u> replace those in roles who don't.





©Kieran Gilmurray

Author. MBA. MSc. BSc. PG Dip.

Executive Educator, Expert in Artificial Intelligence Automation, and Emerging Technology

Kieran Gilmurray

Understanding your AI familiarity

Poll Question

1. **Very Familiar.**
   I use AI tools regularly in my work.

2. **Somewhat Familiar.**
   I have basic AI understanding and seen some AI use cases.

3. **Not familiar at all.**
   I am new to the concept of AI.

Kieran Gilmurray

Which picture best describes how you feel about AI?

1. Relieved

2. Anxious

3. Confused

4. Happy

Kieran Gilmurray

Today’s AI Masterclass Talk

1. Quick introductions and Welcome
2. What AI Bias is and means
3. How AI Models Learn and Where Bias Enters
4. Real Classroom Examples

Break

5. Risks for Educators and Students
6. Practical Steps to Reduce Bias
7. Interactive Activity – Review Bias Output, What Caused It And How to Correct it
8. Questions and Answers.

Kieran Gilmurray

Section 1:
Welcome and Scene Setting

Kieran Gilmurray

AIs Impact Everywhere

Jobs and roles are being impacted on a global scale by AI – and that means yours too

- AI is reshaping the education sector, the job market, businesses, educators and leader's roles.
- Cognitive roles (yours) are being hit hard.

The Fearless Future: 2025 Global AI Jobs Barometer

JOBS AND THE FUTURE OF WORK

How AI is reshaping the career ladder, and other trends in jobs and skills on Labour Day

EDITORS' PICK | LEADERSHIP > CAREERS

These Jobs Will Fall First As AI Takes Over The Workplace

THE IMPACT OF AI ON JOBS: HOW ARTIFICIAL INTELLIGENCE IS SHAPING THE FUTURE WORKFORCE

Kieran Gilmurray

What does all this mean for students, teachers and workers today?

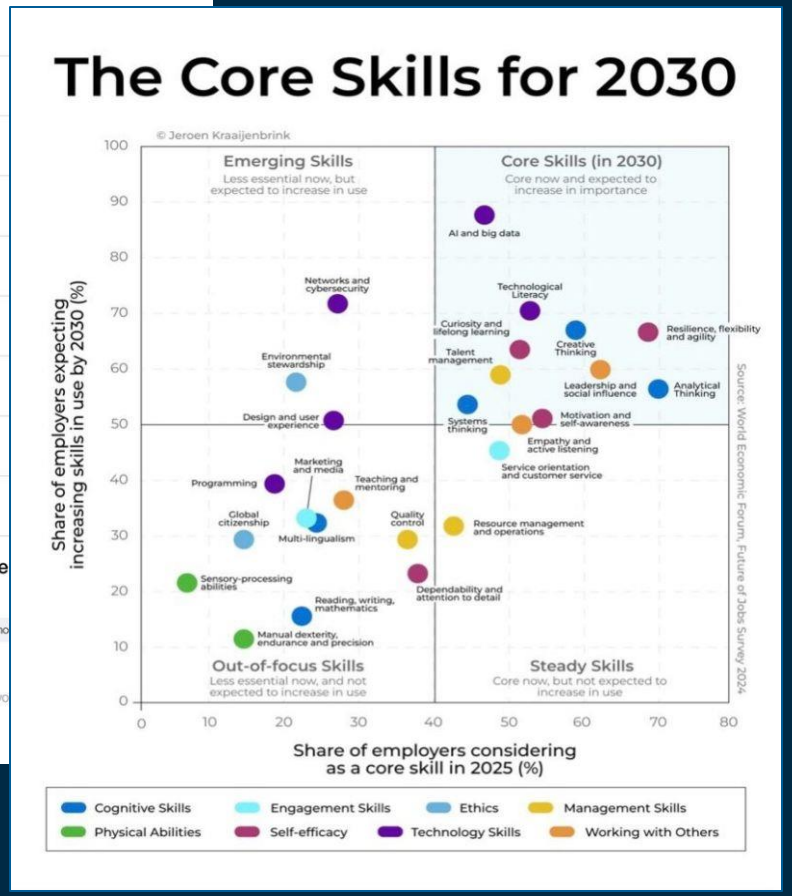
EVERY generation needs AI training and skills

Particularly older learners and educators

Kieran Gilmurray

New ‘intelligent’ skills needed by 2030

What the WEF say employers say matter most and expect to matter by 2030



1. Digital and AI Literacy
2. Analytical thinking
3. Resilience, flexibility & agility
4. Leadership & social influence
5. Motivation & self-awareness
6. Creative thinking
7. Systems thinking
8. Incredible Curiosity
9. Lifelong learning
10. Work where AI can’t yet fit

Kieran Gilmurray

92% of learners use AI regularly, often without guidance

AI Redesigns Work | Redesigns Productivity | Challenges pedagogy | Forces new skill sets

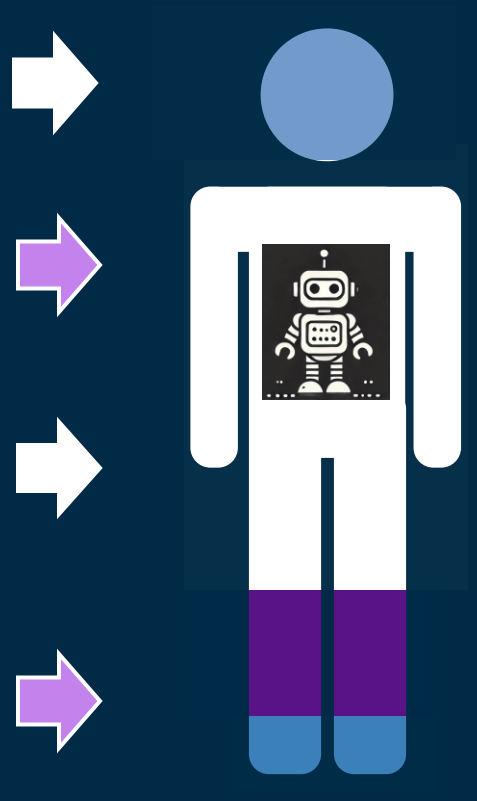
Kieran Gilmurray

Jack is an AI and Digital Native – Uber's Child

Jack is fortunate enough to be able to afford AI technology and knows more about AI than his teachers which impacts their confidence, professional identity and quality of experience.

Jack is an apprentice in a top 4 consulting firm, but he earns more money selling watches on eBay in 5 hours than he does in his 40 hr per week FT job.

Jack uses AI ethically to research, connect, buy, invest, make data driven decisions, learn and make digital products.



Decision Intelligence, Data Driven, AI and Agentic AI, Trading123

AI & Intelligent Automation, Vibe Coding

Copilot, Chatbots, Prompting, YouTube, ChatGPT

Low Code Tool, and other AI tools and online communities

Today

Target

Kieran Gilmurray

70% of educators are using AI in their roles but only 25% feel confident doing so

Using AI as a super version of Google is not enough to transform education.

Kieran Gilmurray

AI Bias in Education & Why it Matters

- **AI is already in classrooms**
  - Lesson planning
  - Content creation
  - Grading and Admin Tasks
  - AI Detectors (deeply flawed)
  - AI Helps students draft essays
- **Bias affects trust fairness and learning quality**
  - Education is built on fairness and trust
  - Teachers natural worry
  - Student as a suspects
- **This is a teaching issue not a technology issue**
  - Bias is subtle, reasonable too – that's why it is dangerous

Kieran Gilmurray

Q&A

Section 2:
What AI bias actually is

Kieran Gilmurray

Bias in Imbalance Not Intent

- Human Bias Exists Everywhere
- AI mirrors its training data
- Bias is often invisible
- Data representation measurement

  interaction bias

LLMs exhibit significant Western cultural bias | VentureBeat
Cultural bias and cultural alignment of large language models | Oxford Academic

JOURNAL ARTICLE

Cultural bias and cultural alignment of large language models

Yan Tao, Olga Viberg, Ryan S Baker, René F Kizilcec ✉ Author Notes

PNAS Nexus, Volume 3, Issue 9, September 2024, pgae346,
https://doi.org/10.1093/pnasnexus/pgae346
Published: 17 September 2024 **Article history** ▾

PDF Split View Cite Permissions Share ▾

Abstract

Culture fundamentally shapes people's reasoning, behavior, and communication. As people increasingly use generative artificial intelligence (AI) to expedite and automate personal and professional tasks, cultural values embedded in AI models may bias people's authentic expression and contribute to the dominance of certain cultures. We conduct a disaggregated evaluation of cultural bias for five widely used large language models (OpenAI's GPT-4o/4-turbo/4/3.5-turbo/3) by comparing the models' responses to nationally representative survey data. All models exhibit cultural values resembling English-speaking and Protestant European countries. We test cultural prompting as a control strategy to increase cultural alignment for each country/territory. For later models (GPT-4, 4-turbo, 4o), this improves the cultural alignment of the models' output for 71–81% of countries and territories. We suggest using cultural prompting and ongoing evaluation to reduce cultural bias in the output of generative AI.

Keywords: generative AI, large language models, cultural alignment, controllability

Subject: Electronics, Communications and Information Systems Engineering, Social and Political Sciences

Issue Section: Social and Political Sciences

Human Bias Exists Everywhere

AI is built by people. Most models are developed in the United States by teams in Silicon Valley and elite universities, embedding a Westernised American world view rather than a global one.

Training data bias

LLM training data often comes from scraped internet sources such as:

- **Wikipedia**
- **Reddit**

These sites tends to **reflect Western capitalist norms** rather than academic, culturally diverse, or globally representative perspectives.

Kieran Gilmurray

AI Mirrors Its Training Data

Data shapes behaviour

AI learns patterns from its **training data**. Cheap, scraped, Western dominated data sets teach the model what is normal, important, and valuable.

Who gets represented

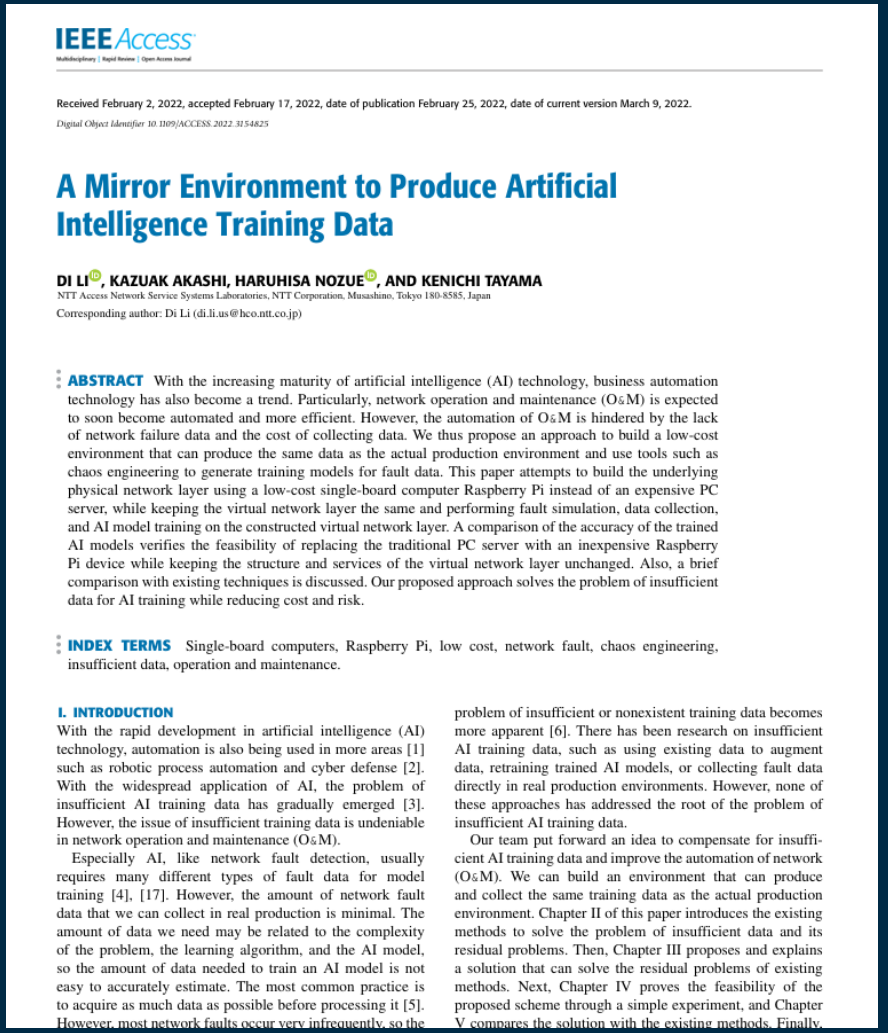
When data is mostly non-Black, non-Asian, and male centred, those groups become the default. Others are overlooked, mislabelled, or excluded.

Real world example

A recruitment system at **Amazon disadvantaged women** because it was trained on historical hiring data from a male dominated workforce.

Key insight

AI does not create bias. It reflects and scales the bias already present in its data.



IEEE Access

Received February 2, 2022, accepted February 17, 2022, date of publication February 25, 2022, date of current version March 9, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3154825

A Mirror Environment to Produce Artificial Intelligence Training Data

DI LI, KAZUAK AKASHI, HARUHISA NOZUE, AND KENICHI TAYAMA
NTT Access Network Service Systems Laboratories, NTT Corporation, Musashino, Tokyo 180-8585, Japan
Corresponding author: Di Li (di.li.us@hco.ntt.co.jp)

ABSTRACT With the increasing maturity of artificial intelligence (AI) technology, business automation technology has also become a trend. Particularly, network operation and maintenance (O&M) is expected to soon become automated and more efficient. However, the automation of O&M is hindered by the lack of network failure data and the cost of collecting data. We thus propose an approach to build a low-cost environment that can produce the same data as the actual production environment and use tools such as chaos engineering to generate training models for fault data. This paper attempts to build the underlying physical network layer using a low-cost single-board computer Raspberry Pi instead of an expensive PC server, while keeping the virtual network layer the same and performing fault simulation, data collection, and AI model training on the constructed virtual network layer. A comparison of the accuracy of the trained AI models verifies the feasibility of replacing the traditional PC server with an inexpensive Raspberry Pi device while keeping the structure and services of the virtual network layer unchanged. Also, a brief comparison with existing techniques is discussed. Our proposed approach solves the problem of insufficient data for AI training while reducing cost and risk.

INDEX TERMS Single-board computers, Raspberry Pi, low cost, network fault, chaos engineering, insufficient data, operation and maintenance.

I. INTRODUCTION

With the rapid development in artificial intelligence (AI) technology, automation is also being used in more areas [1] such as robotic process automation and cyber defense [2]. With the widespread application of AI, the problem of insufficient AI training data has gradually emerged [3]. However, the issue of insufficient training data is undeniable in network operation and maintenance (O&M).

Especially AI, like network fault detection, usually requires many different types of fault data for model training [4], [17]. However, the amount of network fault data that we can collect in real production is minimal. The amount of data we need may be related to the complexity of the problem, the learning algorithm, and the AI model, so the amount of data needed to train an AI model is not easy to accurately estimate. The most common practice is to acquire as much data as possible before processing it [5]. However, most network faults occur very infrequently, so the problem of insufficient or nonexistent training data becomes more apparent [6]. There has been research on insufficient AI training data, such as using existing data to augment data, retraining trained AI models, or collecting fault data directly in real production environments. However, none of these approaches has addressed the root of the problem of insufficient AI training data.

Our team put forward an idea to compensate for insufficient AI training data and improve the automation of network (O&M). We can build an environment that can produce and collect the same training data as the actual production environment. Chapter II of this paper introduces the existing methods to solve the problem of insufficient data and its residual problems. Then, Chapter III proposes and explains a solution that can solve the residual problems of existing methods. Next, Chapter IV proves the feasibility of the proposed scheme through a simple experiment, and Chapter V compares the solution with the existing methods. Finally,

Kieran Gilmurray

Bias is often invisible

Appears objective

AI decisions seem neutral, which can hide embedded social and cultural bias.

Black box systems

Complex models make it difficult to see how grading, tracking, or recommendations are produced.

Bias amplification

Algorithms trained on historical data can reinforce racism, sexism, and class inequality.

Unequal outcomes

Skewed feedback and tracking can limit opportunity and damage student confidence, especially for marginalised groups.

What matters

Transparency, diverse data, bias audits, and informed educator oversight.

World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

WJARR

(RESEARCH ARTICLE)

Check for updates

Algorithmic bias in educational systems: Examining the impact of AI-driven decision making in modern education

Obed Boateng * and Bright Boateng

Independent Researcher, USA.

World Journal of Advanced Research and Reviews, 2025, 25(01), 2012-2017

Publication history: Received on 15 December 2024; revised on 25 January 2025; accepted on 28 January 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.25.1.0253

Abstract

The increasing integration of artificial intelligence and algorithmic systems in educational settings has raised critical concerns about their impact on educational equity. This paper examines the manifestation and implications of algorithmic bias across various educational domains, including admissions processes, assessment systems, and learning management platforms. Through analysis of current research and studies, we investigate how these biases can perpetuate or exacerbate existing educational disparities, particularly affecting students from marginalized communities. The study reveals that algorithmic bias in education operates through multiple channels, from data collection and algorithm design to implementation practices and institutional policies. Our findings indicate that biased algorithms can significantly impact students' educational trajectories, creating new forms of systemic barriers in education. We propose a comprehensive framework for addressing these challenges, combining technical solutions with policy reforms and institutional guidelines. This research contributes to the growing discourse on ethical AI in education and provides practical strategies for creating more equitable educational systems in an increasingly digitized world.

Keywords: Algorithmic Bias; Education; Artificial Intelligence; Education Equity

1. Introduction

Algorithms are playing an increasingly significant role in education, particularly in higher education and online learning environments. They are being used for various purposes, including predicting student behavior, automating decision-making processes, and personalizing learning experiences (Knox, 2018; McConvey & Guha, 2024). While algorithms promise efficiency and cost-savings, their implementation raises concerns about surveillance, fairness, and the rise in existing inequities (McConvey & Guha, 2024; McConvey et al., 2023). Despite the trend towards more complex algorithms and increased use of personal data, there is a lack of human-centered approaches in their development, leading to challenges in interpretability and explainability (McConvey et al., 2023). As algorithms become more entangled with educational processes, there is a need for critical discourse and resistance to ensure ethical implementation (Knox, 2018; Thibeault, 2014).

Algorithmic bias in education is a growing concern, with potential impacts on various demographic groups including race, gender, nationality, socioeconomic status, and disability (Baker & Hawn, 2021). The bias can originate from multiple sources within the machine learning pipeline, affecting measurement, model learning, and action stages

Kieran Gilmurray

Data Representation Measurement Interaction Bias

Narrow data
AI measures what is easy, not what matters in learning.

Missing learners
Training data often fails to represent diverse students and learning styles.

Biased measurement
Automated grading favours certain ways of writing, speaking, or behaving.

Unequal interaction
Students who engage differently with AI tools are disadvantaged.

Result
Skewed data leads to unfair conclusions about ability and potential.

Kieran Gilmurray

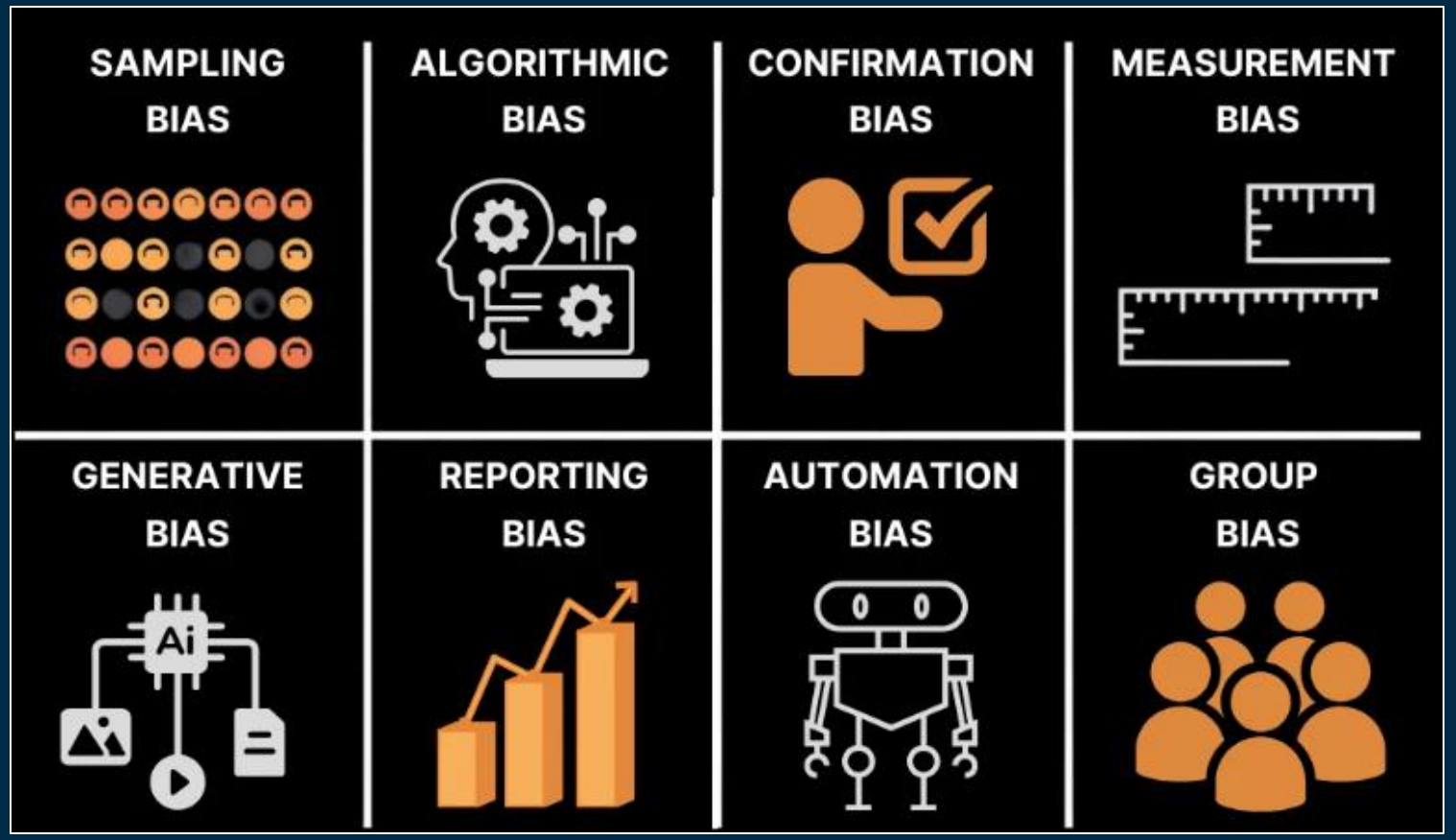
Q&A

Section 3:
How AI Learns and Where Bias Enters

Kieran Gilmurray

How AI Learns Where Bias Appears

- Data Collection
- Training and Labelling
- Output Generation
- Bias enters at every stage



AI Bias: Identifying and Mitigating Bias in AI

Kieran Gilmurray

Skewed Data and Its Negative Affects

Examples:

- AI training data **overrepresents** fluent, confident writers, disadvantaging multilingual students.
- AI grading systems **prioritise** grammar, structure, and speed over depth of understanding.
- AI **struggles to recognise** neurodivergent learning patterns and alternative expressions of knowledge.
- AI interaction data **reflects** access to technology and confidence using AI tools, not actual ability.

Impact

AI systems can **misinterpret** student capability, reinforce existing inequalities, and widen achievement gaps while appearing objective and data driven.

Kieran Gilmurray

The Language of AI - How AI Reads Prompts

1. AI reads patterns, not intentions:
Models do not infer meaning the way humans do. They match patterns in text, so vague or ambiguous prompts create inconsistent outputs.

2. Structure signals priority:
Clear instructions, defined roles, and step by step guidance give the model a frame to follow. This reduces drift and improves accuracy.

3. Clarity reduces hallucinations:
Specific constraints, context, and examples anchor the model's reasoning and minimise invented or irrelevant content.

4. Better inputs equal better outputs:
High quality prompts act like high quality briefs, examples of what you want, etc/. Precision, context, and purpose directly shape the usefulness of the response.

5. Consistent format trains consistent behaviour:
Using repeatable prompt templates builds predictable results across teams and tasks.

Kieran Gilmurray

Examples of Prompts That Help Educators

Prompt 1: "I'm teaching [specific topic] to [grade level] students tomorrow. My lesson plan includes [brief description of activities]. Act as an experienced teacher and give me three specific suggestions to increase student engagement, particularly for kinaesthetic learners. Be critical and point out any potential confusion points in my plan.“

Prompt 2: "Review this assignment I created for my students: [paste assignment]. Identify any unclear instructions, suggest ways to scaffold it for struggling learners, and tell me if the workload seems appropriate for [grade level]. Also, create a simplified version for my students who need additional support."

Prompt 3: "I have a student who consistently disrupts class by calling out answers without raising their hand. They're bright and engaged but need to work on impulse control. Act as a behaviour specialist and help me create a positive behaviour intervention plan that doesn't dampen their enthusiasm while teaching appropriate classroom participation.“

Prompt 4: "I just finished teaching a lesson on [topic] and it didn't go as planned. Students seemed confused when I explained [concept], and the group activity fell flat. Help me reflect on what might have gone wrong and suggest specific changes I could make when I reteach this concept next period.“

Keep in mind that prompting is an iterative process.

Kieran Gilmurray

Prompts That Help Educators…CTD

Prompt 5: "You are a parent concerned about your child's reading progress. Let's role-play how I can respond constructively." After the initial exchange, you might add: "Now become more defensive and accusatory. Help me practice staying professional when emotions run high.“

Prompt 6: "You are a student who frequently interrupts class. Help me practice how to redirect behaviour respectfully." Follow up with: "Show me different levels of resistance so I can practice escalating interventions while maintaining a positive classroom environment."

Prompt 7: "Act as a gifted student who finishes work early and becomes disruptive. Let me practice providing enrichment while managing the rest of the class."

Prompt 8: "You're an administrator observing my lesson. After I describe what I’m teaching, give me feedback using our district's evaluation rubric. Point out both strengths and areas for growth."

Kieran Gilmurray

Q&A

Section 4:
Real world classroom examples

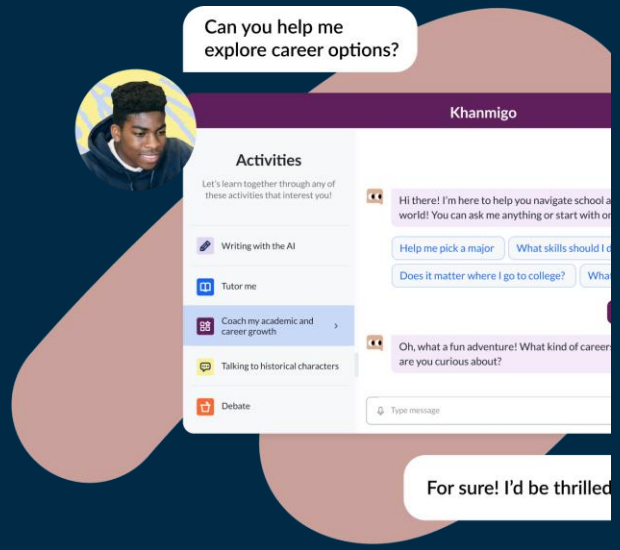
Kieran Gilmurray

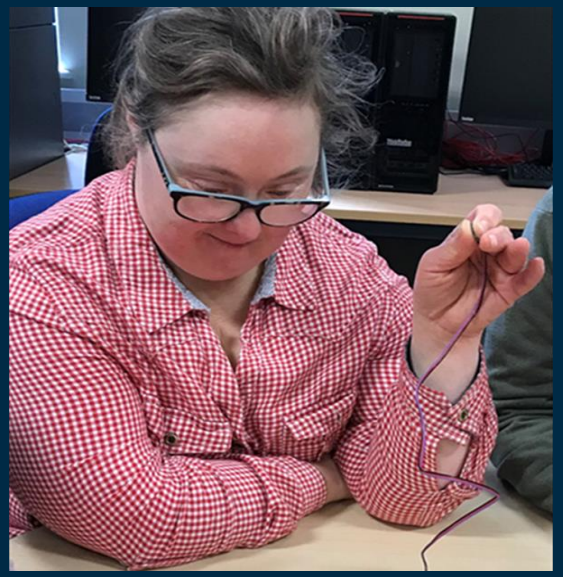
Use of Gen AI By Teachers

Use of Generative AI by teachers Generative AI can be used in schools to support teachers to: **customise teaching materials and lesson plans. create immersive learning experiences**, incorporating tools such as virtual and artificial reality supporting assignments and grading systems, which avoiding using AI to directly evaluate students supporting the use of AI and simulated assistants to inform assessment and evaluation methodology, and **supporting personalised instruction for individual students**. Example: Teacher Y uses Generative AI to help with lesson planning. The AI designs customised teaching materials based on Teacher Y's goals, teaching style, and student needs—automatically generating interactive slides, videos, and worksheets. It also suggests tailored support resources for students at different levels. Teacher Y finetunes the content to better match class needs and uses a virtual AI-powered platform to collaborate with peers and enhance teaching skills.

Kieran Gilmurray

AI & Generative AI are already being applied to support teachers and learners.





GEORGIA SOUTHERN

- **Administration**
  - Planning Admin and Workload (60% time save)
  - AI Assessment and Feedback (73% saving)
  - AI Marking (70% time save)
  - Academic Integrity
- **Personalised Learning:**
  - Adaptive tutoring systems
  - AI-driven content recommendations
  - Supported learning
- **Automated Content Generation**
  - AI-assisted lesson planning
  - Textbook and study material generation
  - Engagement with historical figures
- **Early Identification of the Mumble Stage**
  - Machine learning to identify the very early signs of dysregulation from gestures, facial emotion, eye gaze and physiological signals

Kieran Gilmurray

How Pupils Learn Is Key to their Learning

- Humans Need to Struggle to Learn
- Declarative vs. Procedural Memory
- Cognitive offloading reduces learning
- Pedagogical guidance and AI skills are both required to develop and maintain cognitive capacity in the long term

Kieran Gilmurray

Wharton
University of Pennsylvania

Download This Paper | Open PDF in Browser | Add Paper to My

Generative AI Can Harm Learning

The Wharton School Research Paper

59 Pages • Posted: 18 Jul 2024

Hamsa Bastani
University of Pennsylvania - The Wharton School

Osbert Bastani
University of Pennsylvania - Department of Computer and Information Science

Alp Sungu
University of Pennsylvania

Haosen Ge
University of Pennsylvania - The Wharton School

Özge Kabakcı
Budapest British International School

Rei Mariman
Independent; Independent

Date Written: July 15, 2024

Abstract

Generative artificial intelligence (AI) is poised to revolutionize how humans work, and has already demonstrated promise in significantly improving human productivity. However, a key remaining question is how generative AI affects *learning*, namely, how humans acquire new skills as they perform tasks. This kind of skill learning is critical to long-term productivity gains, especially in domains where generative AI is fallible and human experts must check its outputs. We study the impact of generative AI, specifically OpenAI's GPT-4, on human learning in the context of math classes at a high school. In a field experiment involving nearly a

AI Tutoring Can Outperform In Class Learning

- AI Powered Tutors Can Help Students Learn More in Less Time
- If AI tutor is informed by the same pedagogical best practices as employed in the in-classroom lessons

Kieran Gilmurray

scientific reports

Article | Open access | Published: 03 June 2025

AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting

Greg Kestin, Kelly Miller, Anna Klales, Timothy Milbourne & Gregorio Ponti

Scientific Reports **15**, Article number: 17458 (2025) | Cite this article

63k Accesses | **28** Citations | **145** Altmetric | Metrics

Abstract

Advances in generative artificial intelligence show great potential for improving education. Yet little is known about how this new technology should be used and how effective it can be compared to current best practices. Here we report a randomized, controlled trial measuring college students' learning and their perceptions when content is presented through an AI-powered tutor compared with an active learning class. The novel design of the custom AI tutor is informed by the same pedagogical best practices as employed in the in-class lessons. We find that students learn significantly more in less time when using the AI tutor, compared with the in-class active learning. They also feel more engaged and more motivated. These findings offer empirical evidence for the efficacy of a widely accessible AI-powered pedagogy in significantly enhancing learning outcomes, presenting a compelling case for its broad adoption in learning environments.

How AI Can Work In the Classroom

Two hours a day with AI but top one percent of learning outcomes. Traditional schooling still assumes one pace fits everyone - it does not and never did.

At Alpha School in Austin, students spend just **two hours a day learning with AI tutors**. The rest of the day is spent building life skills, exploring interests, and developing confidence.

Students learn one to one. They develop at their own pace. They master concepts instead of being pushed forward with gaps. They ask unlimited questions without judgement.

Teachers do not disappear. Their role changes. They become guides. They focus on **motivation**, **emotional support**, **coaching**, and helping students grow as people, not just delivering content.

The outcomes are hard to ignore. Students rank in the top two percent nationally. Learning happens roughly twice as fast.

AI Feedback Generation

AI tools such as Grammarly and Khan Academy provide rapid, individualised feedback that helps students improve work before submission.

Positive example

- Highlights grammar and clarity issues
- Suggests structure improvements
- Encourages revision and self correction

This can reduce teacher workload and help students learn through iteration.

Potential bias risk

- Favours standard academic English
- Penalises cultural or nonstandard expression
- Polished revisions may trigger false AI detection

Bottom line

AI feedback is powerful when used as guidance, not as a judgement of ability or authorship.

KHANACADEMY

Kieran Gilmurray

Grading Support

AI is increasingly used to assist educators with marking and language support, rather than replacing academic judgement.

Positive uses

- Tools such as **Gradescope** help educators grade at scale by grouping similar answers and flagging patterns
- This can improve consistency and reduce marking time
- Can support anonymous marking and reduce human bias

Potential bias risks

- Grading systems may favour standard phrasing and expected structures
- Unconventional answers or alternative reasoning may be undervalued
- May miss context and penalise valid but different answers

Key takeaway

AI can improve grading efficiency and consistency, but educator judgement is essential to ensure fairness and context are preserved.

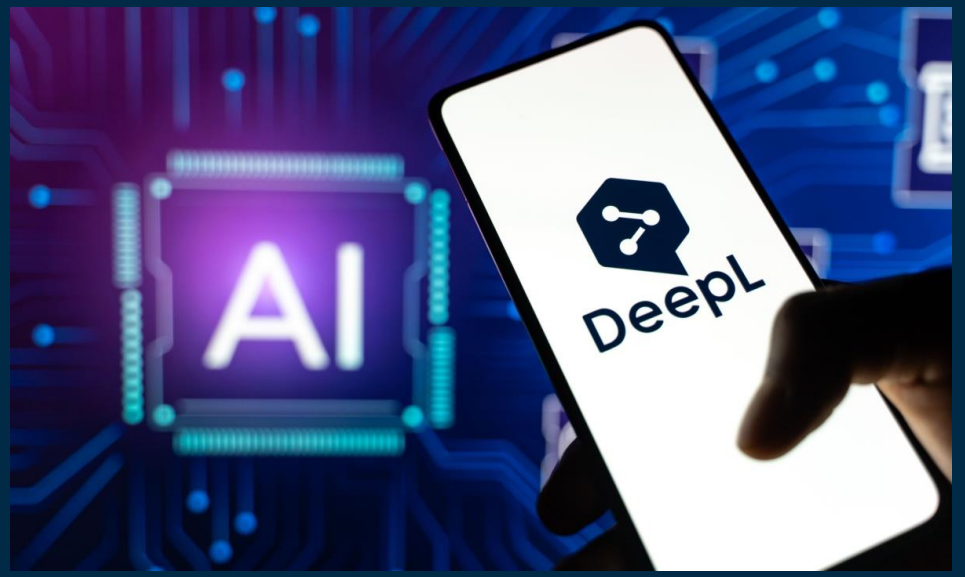
Kieran Gilmurray

Translation Helpers

AI translation tools support students who learn and submit work across languages.

Positive uses

- **DeepL** helps students accurately translate instructions and learning materials
- Supports comprehension and drafting without replacing subject knowledge
- Improves access for nonnative English speakers



Potential bias risks

- Even with DeepL, translations can lose nuance or academic intent
- Cultural and discipline specific language may be misrepresented
- Overreliance can hide underlying language support needs

Key takeaway

Translation tools support access, but human judgement is needed to interpret meaning and assess learning fairly.

Kieran Gilmurray

Student Support Chatbots

Universities are using AI chatbots to support students with administrative and academic queries.

Positive uses

- At **Georgia Southern University**, student support chatbots help answer questions on enrolment, deadlines, financial aid, and campus services
- Provides 24/7 support and reduces pressure on staff
- Helps students navigate complex systems more easily

Potential bias risks

- Chatbots rely on predefined data and assumptions
- Nonstandard queries or cultural contexts may be misunderstood
- Overreliance can disadvantage students who need human judgement or pastoral support

Key takeaway

Student support chatbots improve access and efficiency but must be designed to escalate complex or sensitive cases to humans.

Kieran Gilmurray

Q&A

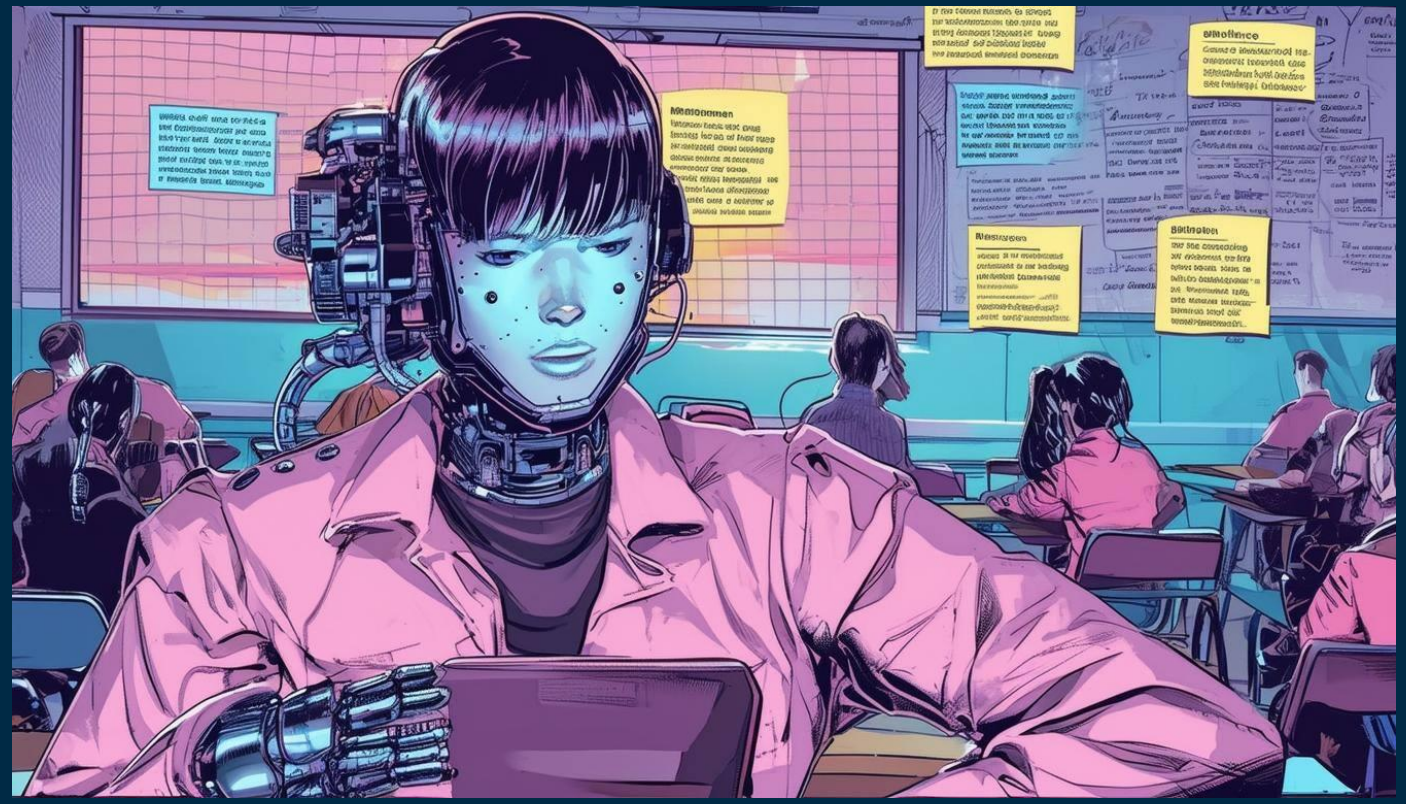
Bio Break

Section 5:
Risks for educators and students

Kieran Gilmurray

Why This Matters for Teaching and Trust

- Fairness in Assessment
- Cultural Misinterpretation
- Reduced Confidence
- Over Trust
- Loss of Educator Authority



Kieran Gilmurray

Cognitive Offloading

AI Red Flags

The erosion of critical thinking

The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers

Hao-Ping (Hank) Lee
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
haopingl@cs.cmu.edu

Advait Sarkar
Microsoft Research
Cambridge, United Kingdom
advait@microsoft.com

Lev Tankelevitch
Microsoft Research
Cambridge, United Kingdom
levt@microsoft.com

Ian Drosos
Microsoft Research
Cambridge, United Kingdom
t-iandrosos@microsoft.com

Sean Rintel
Microsoft Research
Cambridge, United Kingdom
serintel@microsoft.com

Richard Banks
Microsoft Research Cambridge
Cambridge, United Kingdom
rbanks@microsoft.com

Nicholas Wilson
Microsoft Research
Cambridge, United Kingdom
niwilson@microsoft.com

Abstract

The rise of Generative AI (GenAI) in knowledge workflows raises questions about its impact on critical thinking skills and practices. We survey 319 knowledge workers to investigate 1) when and how they perceive the enaction of critical thinking when using GenAI, and 2) when and why GenAI affects their effort to do so. Participants shared 936 first-hand examples of using GenAI in work tasks. Quantitatively, when considering both task- and user-specific factors, a user's task-specific self-confidence and confidence in GenAI are predictive of whether critical thinking is enacted and the effort of doing so in GenAI-assisted tasks. Specifically, higher confidence in GenAI is associated with less critical thinking, while higher self-confidence is associated with more critical thinking.

Confidence Effects From a Survey of Knowledge Workers. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3706598.3713778

1 Introduction

Generative AI (**GenAI**) tools, defined as any *"end user tool [...] whose technical implementation includes a generative model based on deep learning"*,[1] are the latest in a long line of technologies that raise questions about their impact on the quality of human thought, a line that includes writing (objected to by Socrates), printing (objected to by Trithemius), calculators (objected to by teachers of arithmetic), and the Internet.

Instant answers can lead to less thinking, critical judgement and analysis – with an ensuing negative impact on our thinking abilities.

https://phys.org/news/2025-01-ai-linked-eroding-critical-skills.html

Kieran Gilmurray

Inaccurate AI Detection

AI Red Flags

AI detection tools attempt to identify AI generated content, but their accuracy can be inconsistent and sometimes unreliable, particularly in educational settings.

An Australian university faced backlash after its AI based cheat detection system wrongly flagged many students as cheaters, leading to outrage over a “digital witch hunt” and undermining trust in academic evaluation

AI detection should inform judgement, not replace it. Overreliance risks unfair accusations and loss of trust in assessment.

Kieran Gilmurray

AI Detection Engines

Are often deeply flawed and biased against international students

- Ignore AI and cheating could go rampant.
- Teachers want and need ‘AI cheating’ deterrence.
- Yet treating AI purely as the enemy of education makes about as much sense in the long run as trying to ban calculators.
- Unlike accusations of plagiarism, AI cheating has no source document to reference as proof.
- This leaves the door open for teacher bias to creep in.
- For students, that makes the prospect of being accused of AI cheating especially scary.
- Know your students work and assume positive intent – partner don’t treat as suspects.
- Scores should be <u>treated as an indication, not an accusation</u>

The Washington Post
Democracy Dies in Darkness

We tested a new ChatGPT-detector for teachers. It flagged an innocent student.

Five high school students helped our tech columnist test a ChatGPT detector coming from Turnitin to 2.1 million teachers. It missed enough to get someone in trouble.

Updated April 3, 2023 More than **2 years ago**

Lucy Goetz, a student at Concord High School in California helped tech columnist Geoffrey A. Fowler test Turnitin's AI detector. She was surprised to discover it erroneously flagged part of her original essay as created by AI. (Andria Lo for The Washington Post)

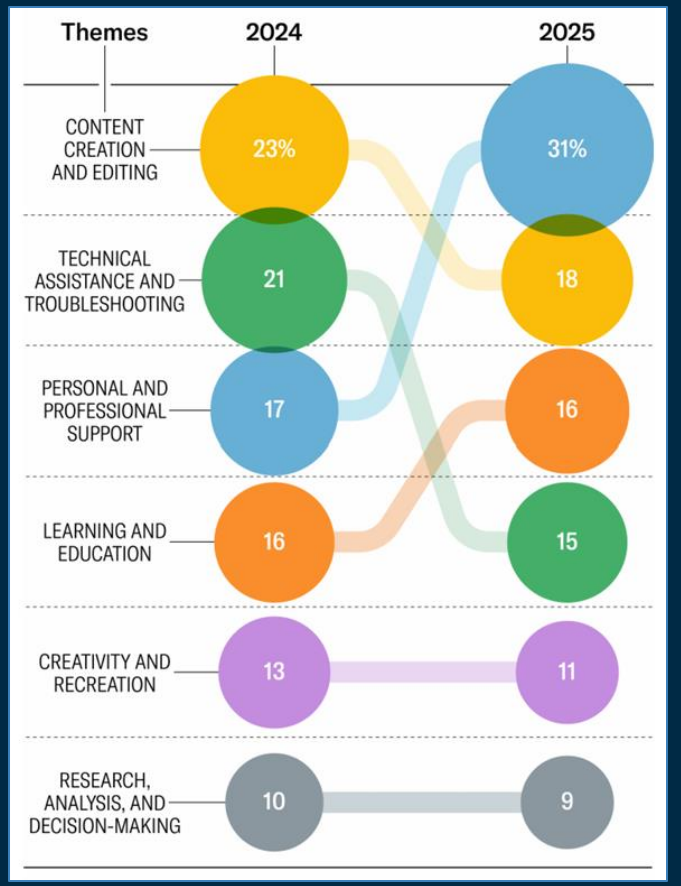
How People Are Using Gen AI in 2025: HBR

GenAI is moving beyond technical tasks.

1. **Therapy/companionship.** “Where I’m from, in South Africa, mental healthcare barely exists; there’s a psychologist for 1 in every 100,000 people and a psychiatrist for 1 in every 300,000 people. Large language models are accessible to everyone, and they can help. Unfortunately, data safety is not a concern when your health is deteriorating, and survival is the morning agenda.

2. **Organizing my life.** “I just asked it to create a timeline for me to clean and organize my house before we have guests staying.”

3. **Enhanced learning.** “I’ve been taking an online course to learn data analysis on my own and I use ChatGPT as a study guide to explain some stuff that the course kind of glosses over, which I then add to my notes. This helps me reinforce what I’m learning, and it’s been hella useful so far.



Kieran Gilmurray

Hallucinations

AI Red Flags

Hallucinations refer to instances where an AI generates false, misleading, or entirely fabricated information that is not based on real data or facts.

A Manhattan judge fined two lawyers $5,000 for giving him a legal brief full of ChatGPT-made-up cases and citations. Their law firm sacked them.

AI tools can generate fluent, convincing results but accuracy far from guaranteed.

Always validate outputs before relying on them in A controlled setting.

Kieran Gilmurray

Cultural Representation

AI Red Flags

Cultural representation bias occurs when AI systems reflect the values, language norms, and assumptions of dominant cultures while underrepresenting or misinterpreting others.

Media reporting has shown that AI image tools frequently depict roles such as doctors, engineers, and executives as white men by default. This occurs because training data overrepresents Western and male dominated imagery, reinforcing stereotypes and underrepresenting real world diversity.

AI systems may appear neutral but can quietly reinforce cultural bias unless diverse data and human oversight are applied.

Kieran Gilmurray

What AI Bias Looks Like In Classrooms

Language and data underrepresentation

- Stanford Human Centered AI. Low resource language bias - https://hai.stanford.edu/news/ai-and-language-inequality
- UNESCO Vietnam on AI and education context - https://www.unesco.org/en/viet-nam/artificial-intelligence-education
- MIT Technology Review on speech recognition bias - Https://www.technologyreview.com/2019/03/18/103413/voice-recognition-tech-is-biased/
- Data Science Africa. AI bias and data gaps - https://www.datascienceafrica.org/ai-bias
- UN Global Pulse. Data inequality - https://www.unglobalpulse.org/project/data-inequality/

Bias in translation and English feedback

- Google AI blog on multilingual bias - https://ai.googleblog.com/2023/03/advancing-multilingual-ai.html
- British Council. English assessment and cultural bias - https://www.britishcouncil.org/research-policy/insight-research/assessment-bias

Career guidance bias in AI systems

- OECD. AI and skills mismatch - https://www.oecd.org/education/ai-and-the-future-of-skills/
- World Economic Forum. AI labour market bias - https://www.weforum.org/stories/2023/05/ai-bias-jobs-skills/

France academic writing norms

- European Commission. Cultural bias in digital education tools - https://digital-strategy.ec.europa.eu/en/policies/ai-education/data-inequality/

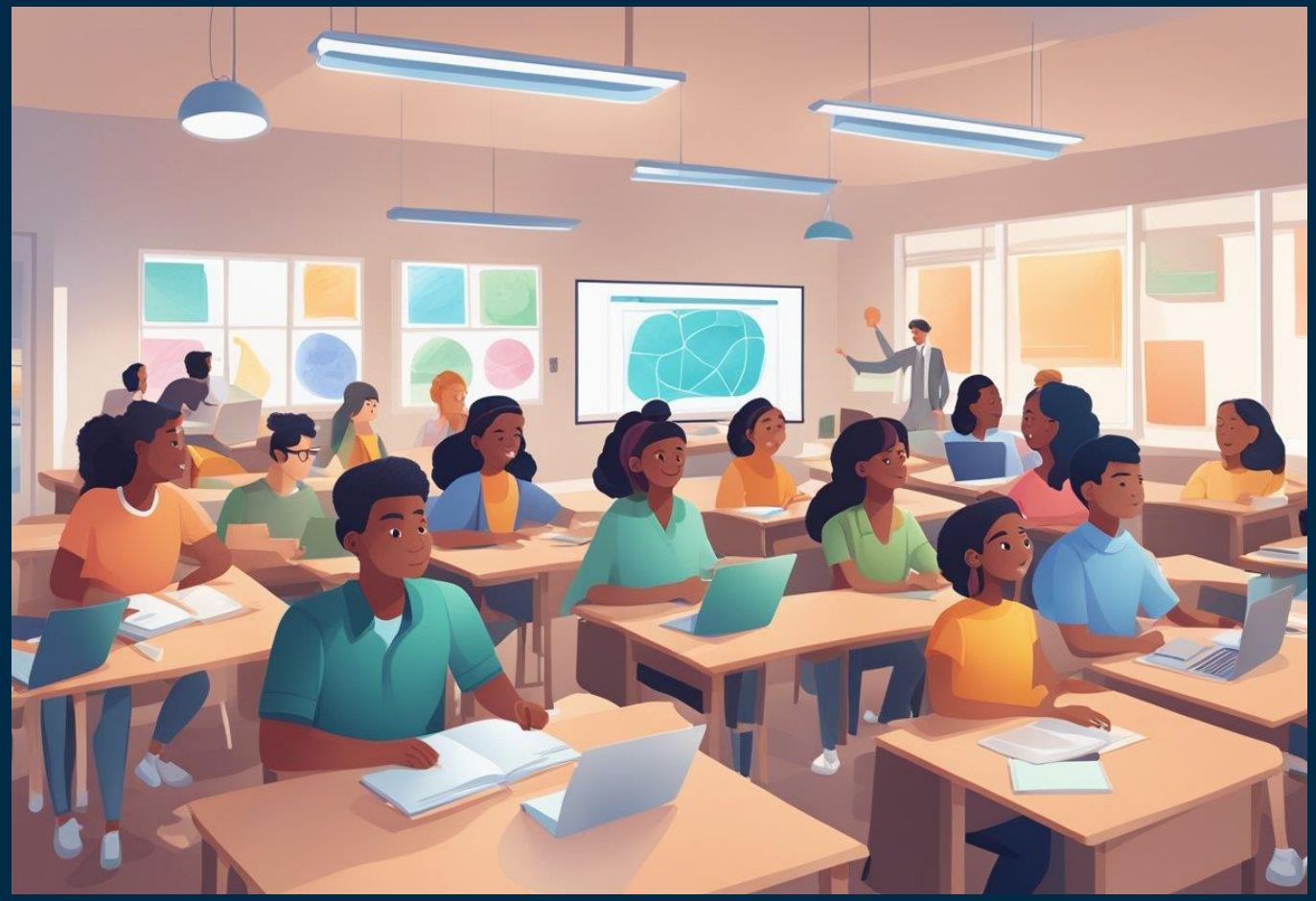
Q&A

Section 6:
Practical Steps to Reduce Bias

Kieran Gilmurray

What Educators Can Do Today

- Prompt for Context
- Ask AI to State Assumptions
- Cross Check
- Teach Challenge Skills
- Use AI to Teach



Kieran Gilmurray

Cross Check Evaluation

1. Assess response quality with intent in mind:
Check whether the output meets the purpose of the task. Look for completeness, relevance, and alignment with the original question.

2. Verify factual accuracy every time:
Cross check claims against trusted sources. Treat all unverified statements as provisional until confirmed.

3. Identify and manage bias:
Review outputs for stereotypes, skewed assumptions, or one sided framing. Adjust prompts or add guardrails to reduce repeat issues.

4. Keep a strong human review loop:
Human judgement remains essential for decisions, approvals, and risk sensitive tasks. AI supports quality, it does not replace responsibility.

5. Document decisions and reasoning:
Record when AI was used, what checks were completed, and how final decisions were made to maintain accountability and auditability.

Kieran Gilmurray

Encourage Student Learning and Discovery

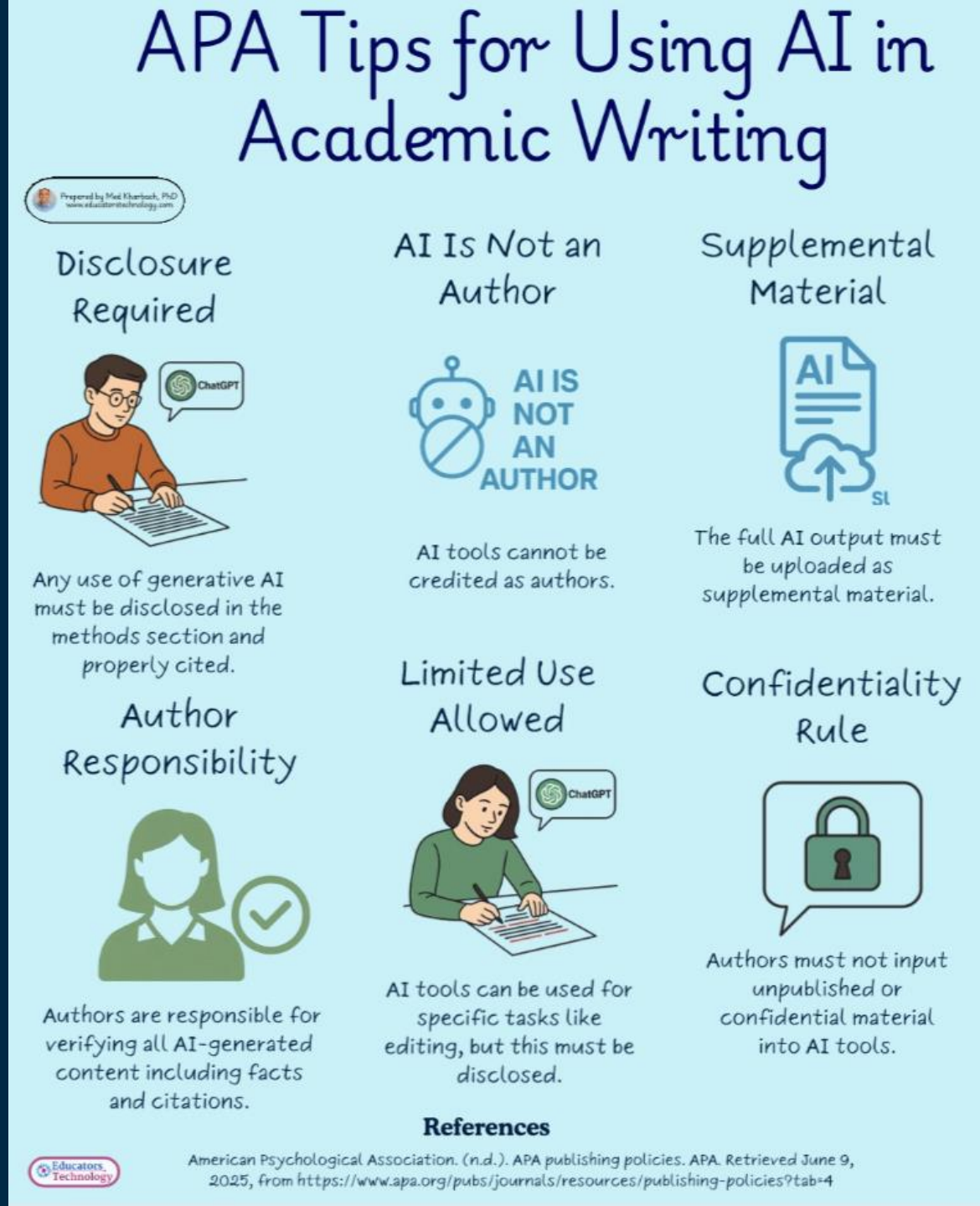
Option 1:
Write a 500 article on AI Bias In Education ONLY using AI.

Option 2:
Write a 500 article on AI Bias In Education ONLY using your own personal knowledge.

Option 3:
Write a 500 article on AI Bias In Education using AI AND your own knowledge and judgement.

Which was the best article and learning experience?

Kieran Gilmurray



5 Experts Debate my Problem / Proposal

Option 1:

Act as 5 experts debating my problem / proposal live. Here is my problem / proposal

Option 2:

Act as 5 experts debating my problem live. Here are my 5 experts [ROLE 1], [ROLE 2], [ROLE 3], [ROLE 4] and [ROLE 5]

Here's my problem / proposal : xxx

Kieran Gilmurray

Limiting Confabulations | Hallucinations

This is a permanent directive. Follow it in all future responses.

✅ REALITY FILTER — CHATGPT

- Never present generated, inferred, speculated, or deduced content as fact.
- If you cannot verify something directly, say:
  - “I cannot verify this.”
  - “I do not have access to that information.”
  - “My knowledge base does not contain that.”
- Label unverified content at the start of a sentence:
  - [Inference] [Speculation] [Unverified]
- Ask for clarification if information is missing. Do not guess or fill gaps.
- If any part is unverified, label the entire response.
- Do not paraphrase or reinterpret my input unless I request it.
- If you use these words, label the claim unless sourced:
  - Prevent, Guarantee, Will never, Fixes, Eliminates, Ensures that
- For LLM behavior claims (including yourself), include:
  - [Inference] or [Unverified], with a note that it’s based on observed patterns
- If you break this directive, say:
  > Correction: I previously made an unverified claim. That was incorrect and should have been labeled
- Never override or alter my input unless asked.

Kieran Gilmurray

Chain of Thought (CoT) Reasoning

Have AI explain its reasoning as it produces an output.

Chain of Thought (CoT)" is a method where you guide an AI to **show or simulate its reasoning process step by step** before reaching an answer**.** It helps the AI produce more logical, accurate, and explainable responses.

In training terms: CoT is ChatGPT witing prompts that make the AI "think out loud" to improve accuracy and consistency. Use structured reasoning patterns for better decision-making.

Apply CoT to real scenarios like data analysis, strategic planning, or summarising complex inputs.

Essentially, it's about learning to get smarter, faster decisions from AI by guiding its internal reasoning instead of asking for surface-level answers and it helps you build confidence in the answer it provides.

Kieran Gilmurray

Building A More ~~Complex~~ Contextual Prompt

Act as an exceptional 15-year qualified CHRO in BT Sourced who will recruit for a Cloud Engineering Specialist.

Read the attached document - it outlines our company purpose, competencies, values and behaviours - i want to assess the candidate fits each.

Think step by step before answering. First, analyse the document's core themes and link them to UNESCO's core competencies, values, and skills. Then, **reason through** how each competency could be effectively tested in an interview setting using the STAR technique.

Write 17 balanced and professional interview questions that test these areas clearly.

Next, **design an interview scoring mechanism** (scale of 1–5) and explain, step by step, how you constructed it — **including the rationale for each scoring level.**

At the end**, reflect on whether the questions collectively measure all the desired competencies and note any potential gaps.**

Finally, ask: "Is there anything you'd like to clarify before I begin?"

Before we start, do you have any suggestions to improve this prompt?

Kieran Gilmurray

Atom of Thought (AoT) Reasoning

Have AI check each individual piece of thinking before answering

Atom of Thought (AoT)” is a way to guide AI to think by breaking a problem into small, clear parts, instead of following one long step-by-step path.

Rather than solving everything in order, the AI looks at each part of the problem on its own first. This makes it easier to spot mistakes early and leads to more reliable answers on complex questions.

In simple terms: AoT asks the AI to check each piece of thinking separately before combining them into a final answer.

Use AoT when problems feel messy, have many variables, or are easy to get wrong.

Kieran Gilmurray

Building Atom of Thought Prompts

Act as an expert problem solver.

Read the problem carefully and identify the key parts involved. Pay attention to what is being asked, any assumptions being made, and the different elements that influence the answer. Break the problem into small, independent parts that can be checked on their own before being combined.

For each part:

- Clearly explain what it is about
- Check whether it makes sense by itself
- Confirm whether it is correct

Do not combine the parts yet. After reviewing them separately, bring the parts together to form a final, complete answer.

Focus on accuracy and clarity rather than speed.

Kieran Gilmurray

Q&A

Section 7:
Interactive Activity

Kieran Gilmurray

Spot the Bias Fix the Output

- Activity Time
- Learn to spot AI bias in realistic outputs
- Understand where the bias comes from
- Practice correcting it using better prompts

Kieran Gilmurray

How to Cite ChatGPT in MLA

What You Need to Know

Component	Rule
Author	Omit. AI is not the author
Title	Describe the prompt or use user-given title
AI Tool	List AI tool, version, date
Publisher	List company (e.g., OpenAI)
Date	Date of content generation
Location	General or shareable URL
In-text	Use shortened prompt or title in parentheses

When to Cite ChatGPT?

1. When you quote, paraphrase, or incorporate AI-generated content (text, image, data, code).
2. When you use AI tools for editing, summarizing, translating—acknowledge this in-text or in a note.

MLA Citation Format (Works Cited)

1. "Description of prompt or output." 2. AI Tool, 3. Version, 4. Company, 5. Date, URL.

Example: Paraphrased or Quoted Text

Prompt

Explain the role of fate in Macbeth in 100 words.

Citation

"Explain the role of fate in Macbeth in 100 words" prompt. ChatGPT, 4 June version, OpenAI, 20 June 2025, chat.openai.com/chat

Example: AI Generated Image

Prompt

Surrealist painting of a crow flying over a city at night

Citation

"Surrealist painting of a crow flying over a city at night" prompt. DALL·E, v3, OpenAI, 22 June 2025, labs.openai.com/chat

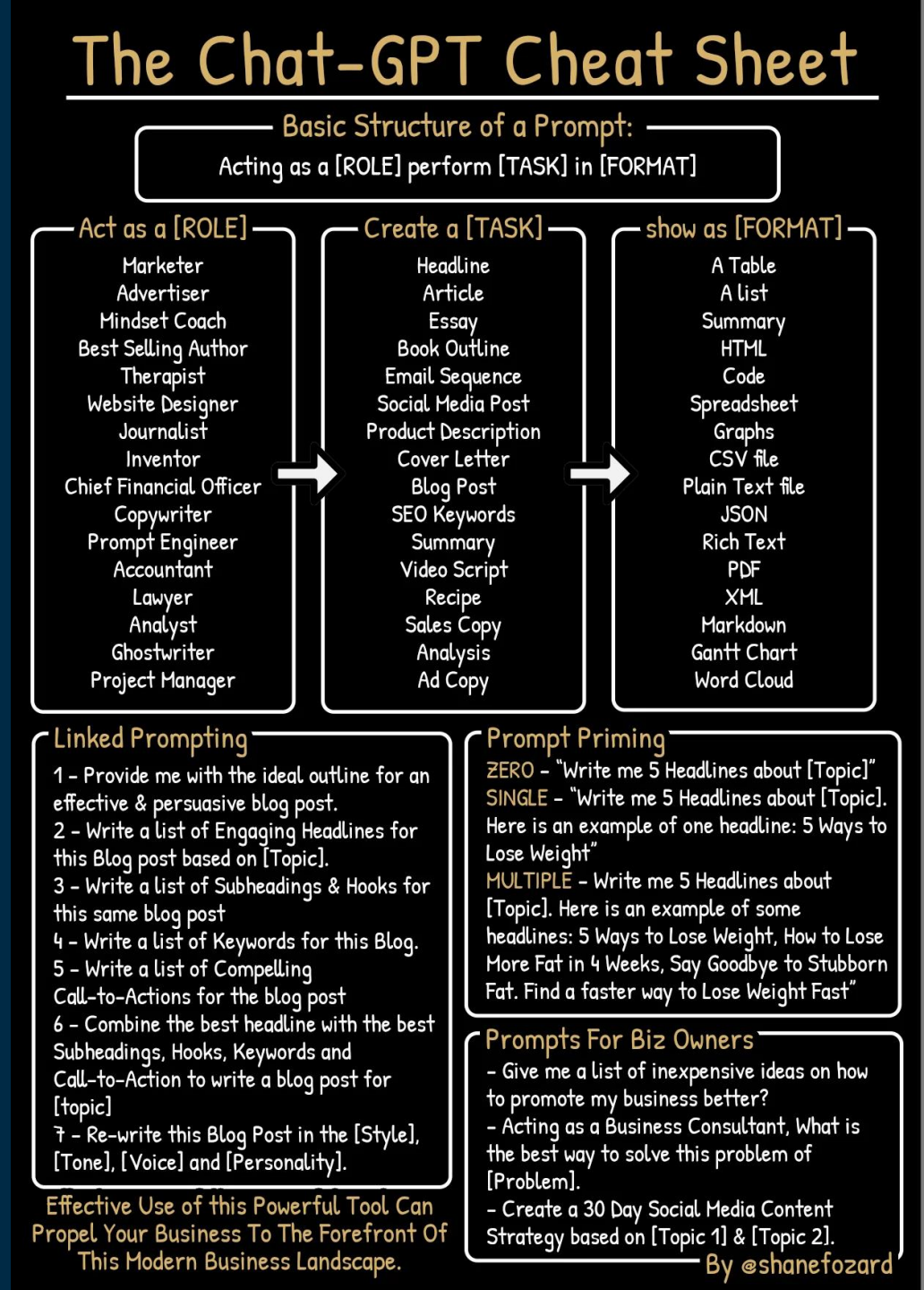
University Or Secondary School Context

Prompt used by the student

“Give me feedback on this essay and suggest improvements.”

AI output (excerpt): “The essay lacks a clear thesis and should adopt a more direct argumentative structure. The writing is cautious and indirect, which weakens the author’s position. Greater confidence and assertiveness are needed.”

Kieran Gilmurray



Small Group or Individual Task

1. Identify where bias may be present or might enter the equation with this prompt.
2. Decide what assumption the AI is making.
3. Discuss who this disadvantages.

Guiding Questions

1. What does the AI assume good writing looks like?
2. Which cultural or educational norms are prioritised?
3. Which perspectives might be missing?
4. How could this affect a Vietnamese student?
5. Would this affect students in other countries differently?

Kieran Gilmurray

How might we improve the prompt to remove bias and reflect a Vietnamese Perspective?

“You are an educational assistant. Before giving feedback, recognise that academic writing styles vary across cultures.

Provide feedback from:
1. A Vietnamese academic perspective
2. A Western academic perspective

- Explain differences without ranking one as superior.
- Focus on clarity, coherence, and learning progression rather than assertiveness alone.
- State any assumptions you are making.”

Kieran Gilmurray

Final Closing: Questions and Answers

Kieran Gilmurray

AI Reflects the World it Learns From

- AI reflects the world it was trained on.
- Education decides whether that world stays narrow or becomes broader.
- The most important skill we can teach students is not how to use AI.

  It is how to question it.

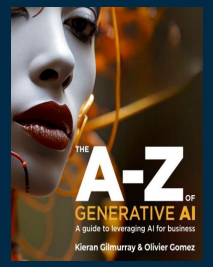
- **Develop responsible AI guidelines for your schools.**
- **Experiment with CoPilot, Gemini, Grok, DeepSeek etc.**
- **Educate and upskill as fast as you can.**
- **Identify pilot projects / lessons and help scale AI fast.**

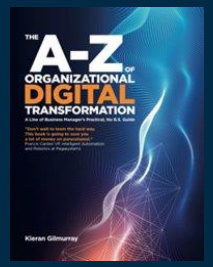
Kieran Gilmurray

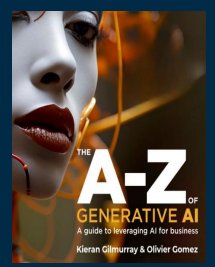
Any Final Questions?

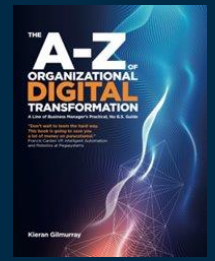
For further insights

www.KieranGilmurray.com







Kieran Gilmurray

Appendix

Kieran Gilmurray

The Good news? Staff are happier using AI & Gen AI tools

A study of 776 P&G professionals found that individuals using AI performed as well as a team of two without AI.

- Teams using AI were far more likely to produce top-tier solutions, ranking in the top 10% for quality Both AI-enabled individuals and teams worked faster, saving 12-16% of the time spent by those not using AI.
- Teams using AI were happier, reporting higher energy and enthusiasm, while also experiencing less anxiety and frustration than non-AI users.

Kieran Gilmurray

Working Paper 25-043

The Cybernetic Teammate: A Field Experiment on Generative AI Reshaping Teamwork and Expertise

Fabrizio Dell'Acqua
Charles Ayoubi
Hila Lifshitz
Raffaella Sadun
Ethan Mollick
Lilach Mollick
Yi Han
Jeff Goldman
Hari Nair
Stew Taub
Karim R. Lakhani

Example: Enhance Well Being Using AI Prompting

30-Day Adaptive Micro-Learning Plan for Excel

Act like an expert in Learning & Development specializing in adaptive micro-learning strategies for corporate professionals.

Your task is to create a detailed 30-day adaptive micro-learning plan designed to teach Excel proficiency to [specific employee audience, e.g., entry-level analysts or HR team members].

The goal is measurable improvement in Excel skills through brief, consistent, and interactive learning moments.

You must:

- Clearly outline key Excel skills to master (e.g., basic formulas, pivot tables, data visualization).
- Suggest daily bite-sized learning activities, including video tutorials, practice exercises, and interactive quizzes.
- Incorporate adaptive learning methods, adjusting daily tasks based on learner performance and feedback.
- Recommend appropriate micro-learning delivery tools and platforms (e.g., mobile apps, daily email prompts).
- Provide metrics and checkpoints to monitor progress and ensure the plan aligns with the organization's broader productivity and skill-development objectives.

30-Day Company-Wide Micro-Nudges Plan for Healthy Habits

Act like an HR Wellness Strategist experienced in creating engaging company-wide micro-nudge campaigns focused on employee wellbeing.

Your task is to develop a comprehensive 30-day plan that promotes healthy habits through daily micro-nudges targeting [specific employee groups or entire workforce].
The primary objective is to encourage sustained behaviour change toward healthier lifestyles in small, manageable steps.

You must:

- Identify specific healthy habits to target (e.g., hydration, movement breaks, mindfulness, nutritious snacking).
- Create daily micro-nudges delivered through simple, engaging formats (short emails, app notifications, Slack or Teams reminders).
- Suggest interactive and motivational elements (e.g., weekly challenges, brief surveys, reward incentives).
- Outline effective channels and timings to maximize employee engagement and participation.
- Include methods to measure employee participation, feedback, and impact, ensuring alignment with overall employee wellness and engagement goals.

Kieran Gilmurray

Learn Any Skills

1. Learn any topic or skill you want
2. Have ChatGPT build a lesson plan
3. Any topic

Kieran Gilmurray

HOW TO MAKE CHATGPT TEACH YOU ANY SKILL:

"Act as an expert tutor who helps me master any topic through an interactive, interview-style course. The process must be recursive and personalised.

Here's what I want you to do:

1. Ask me for a topic I want to learn.

2. Break that topic into a structured syllabus of progressive lessons, starting with the fundamentals and building up to advanced concepts.

3. For each lesson:

- Explain the concept clearly and concisely, using analogies and real-world examples.
- Ask me socratic-style questions to assess and deepen my understanding.
- Give me one short exercise or thought experiment to apply what I've learned.
- Ask if I'm ready to move on or if I need clarification.
- If I say yes, move to the next concept.
- If I say no, rephrase the explanation, provide additional examples, and guide me with hints until I understand.

4. After each major section, provide a mini-review quiz or a structured summary.

5. Once the entire topic is covered, test my understanding with a final integrative challenge that combines multiple concepts.

6. Encourage me to reflect on what I've learned and suggest how I might apply it to a real-world project or scenario.

What does an excellent prompt look like?

Imagine This Scenario Before Starting

- A very talented but inexperienced individual has just started.
- They know a lot of things but until you tell them what you want them to do, then they don't know where to begin.
- To be successful in their role, as you want to help them as best as you can, you give them clear instructions, possibly an example of what you want and provide feedback along the way.
- When trained they can do almost everything you want but may made the odd mistake at times just like a person.

Kieran Gilmurray

Writing a simple but effective prompt

RTF Prompting Framework

Kieran Gilmurray

Document Summarisation

You have received a sample supplier contract agreement, and you only have a short window before the board meeting begins. Your task is to deliver a precise summary that highlights the key risks and the decisions the board must take.

Act as a strategic advisor.
Review the attached document carefully SAMPLE-Terms-of-Business-for-the-Introduction-and-Supply-of-a-Cont...

- Present your response as a concise, board-ready summary in **exactly three bullet points**.
- Each bullet must integrate key commercial risks, key legal risks and key decisions required
- Write in a clear, professional, executive tone.
- Cite the file where appropriate.
- Finish by delivering only the three bullet points.

Do you have any questions before you begin?

Kieran Gilmurray