10 MUST READS

"AI for Good, But for Whom?"

An interview with Asma Derja

AI: The Good, The Bad, and the Game-Changer

Essays by the world's leading researchers & visionaries

Islem Rekik

AI: The Good, The Bad, and the Game-Changer

ESSAYS BY THE WORLD'S
LEADING RESEARCHERS & VISIONARIES

EDITED BY

ISLEM REKIK

BOOK SAMPLE

Draft version — may include typos or unfinished edits.

Contents

Preface

Over the past decade, artificial intelligence (AI), particularly deep learning (DL), has revolutionized research, engineering, and business —powering Nobel-Prize winning scientific discoveries such as AlphaFold's protein prediction and technological breakthroughs with GPT-based advances in biology and reasoning . Such scientific leaps and societal impact were enabled by harnessing the increasing large-scale datasets and computational power along with the intelligent design of DL models. Yet, AI also presents paradoxes that are elusive, difficult to define, let alone comprehend. *"AI: The Good, The Bad, and the Game-Changer"* is part of the BASIRA Creative Collective Projects, launched on May 29, 2025.

We find ourselves on the fast-track of AI adoption: rushing to deploy and innovate, often overlooking where AI should—not should—be applied. Terms like "AI governance" and "AI regulation" have become buzzwords, tossed about without a shared understanding of their implications. Of course, most of us are driven by the vision of steering AI toward humanity's benefit. But in an era where power and profit are accelerating, and "value" is often valued when it is equated with financial gain, complex ethical dilemmas emerge.

This book presents **25 essays by leading AI researchers, thinkers, and emerging visionaries from 20 countries around the world**. It explores how AI is poised to reshape healthcare, education, warfare, business, ethics, mental health, social well-being, and more—in ways many can predict and some that lie beyond imagination. It also features a curated collection of **10 must-read essays** spanning a range of domains. You will also find thought-provoking contributions organized under key sections: Education, Healthcare, Business, and Human & Other.

Here are *a few snapshots from the essays* included in this first draft:

You will read how AI helps us discover novel treatments for complex diseases, democratizes education through personalized learning, and simultaneously threatens patient dignity through biased decision systems and the erosion of privacy. It also draws historical parallels—for instance, the 1933 "American Chamber of Horrors" exposé on unregulated chemicals, which helped catalyze the formation of the U.S. Food and Drug Administration (FDA).

Generative AI (GenAI) can serve as a powerful creative catalyst. Its recent integration into the film and media industries marks a profound transformation, comparable to historic shifts such as the introduction of sound or color in cinema. Yet GenAI also raises concerns around content monopolization and labor monopsony. In the film industry, GenAI can now replicate an actor's

voice, face, and even their essence. Actors are increasingly fighting for the right to control their AI-generated likenesses—digital doubles that could, in theory, persist indefinitely. Just as Olivia de Havilland (a British American actress) once fought to end exploitative studio contracts, today's artists may soon demand boundaries on the use, reproduction, and monetization of their identities by creators leveraging GenAI. Once again, we forget our history—whether in science or in media. Or perhaps we didn't forget; *perhaps we were never taught*. Further, it highlights current dilemmas: how educators find themselves grading work generated by LLMs, and how students feel trapped in cycles of using AI to score better—gradually undermining the value of thought, learning, and intellectual rigor. As Mortimer Adler begins *How to Read a Book*: "You have a mind." We have stopped recognizing that truth—and AI may accelerate the erosion of that self-awareness. **This paradox—between progress and peril, empowerment and erosion—is the heart of this book.**

We have included an **exclusive interview with Asma Derja**, the founder of a grassroots initiative called the "Ethical AI Alliance." Serving as an opening to this book, the interview invites you to reflect more deeply on how to develop ethical AI. In this interview, Asma Derja questions the idea of "AI for good," highlighting how AI systems often reflect the biases of those who build them—typically excluding marginalized voices. Drawing from her departure from Amazon Web Services, she calls for a grassroots, inclusive approach to AI that centers people over profit. Derja advocates for cultural and ethical reform in AI governance and urges broader participation—from artists to activists—in shaping its future, especially as AI expands into high-stakes areas like warfare.

This is an open first draft, which may contain typos and unedited essays. We welcome your reviews and feedback at i.rekik@imperial.ac.uk If you are interested in contributing to the second draft of this book, you can fill out this application form. We are here to help your story be heard.

Dr. Islem Rekik is an award-winning AI researcher, educator, and advocate for inclusive innovation. She leads the BASIRA Lab and is an Associate Professor at Imperial College London (Computing, Innovation Hub I-X). Over the past 7 years, she has mentored students to 24+ academic honors and awards and co-chaired 35+ major AI events—including MICCAI, NeurIPS, and ISBI, and published 200+ peer-reviewed research papers. In 2025, she received the Tunisian AI Award for her pioneering work in AI and EDI, and was featured in Realités Magazine and I-X News. She also co-founded global initiatives supporting underrepresented researchers—especially in low- and middle-income countries—to help shape a more inclusive and equitable AI future.

AI for Good, But for Whom? —An interview with Asma Derja

In this interview, Asma Derja, founder of the Ethical AI Alliance, questions the idea of "AI for good," highlighting how AI systems often reflect the biases of those who build them—typically excluding marginalized voices. Drawing from her departure from Amazon Web Services, she calls for a grassroots, inclusive approach to AI that centers people over profit. Derja advocates for cultural and ethical reform in AI governance and urges broader participation—from artists to activists—in shaping its future, especially as AI expands into high-stakes areas like warfare.

1. **Let's begin with the title of this conversation: "AI for Good, But for Whom?" — How did this question become central to your work? Was there a personal moment or experience that triggered this line of inquiry?**

   AI for Good came up after I left AWS. I knew the flaws in ethics behind closed doors, but I didn't have the full picture. So, I started digging into reports from The Intercept and Wired. Then I signed up for the "AI for Good" Summit in May 2024. It was my first time there. I checked the agenda and couldn't find anything on what I already knew was happening: AI used for militarization, warfare, surveillance. Think about Palestine, or companies like Palantir. Nothing. No session on that. That's when it hit me: what does "AI for Good" even mean if it's not tackling this? If these issues don't make it into the conversation, then who decides what "good" looks like?

2. **You've said that AI is often "built by men for men." How does that manifest in real-world AI systems today — and what are the implications for communities historically excluded from tech design?**

When I said this in a TikTok that hit over a million views, some people took the title literally. They imagined men conspiring behind closed doors to intentionally exclude women. So, no. It's not a conspiracy. But when you see how these patterns align, it almost feels like one. Here's what it really looks like: engineers making women feel unwelcome on dev teams; male VCs rejecting women-led startups; women in corporate trying to prioritize social impact but getting pushed back. I've lived that; datasets that erase entire female populations; product teams optimizing for the "average" user, who always looks the same; policymakers regulating AI without consulting women; women raising ethical red flags and getting sidelined or silenced.

And when you follow the money: who funds, who profits, you see why the system keeps serving the same people who built it. If you want to go deeper into this, I wrote a piece on Substack because the question isn't just bias in code, it's the whole structure of tech development and who gets to imagine the future.

3. **You left a senior role at Amazon Web Services in silence. What made you walk away—and what did you see from the inside that others didn't?**

   I left AWS in December 2023 after serving as a senior business architect for the EMEA region, working on some of the largest digital transformation projects for AWS's top clients. Over time, I started seeing a side of the industry I couldn't ignore: cloud infrastructure not just powering businesses, but enabling systems tied to militarization. and, in some cases, directly linked to human rights breaches.

   What troubled me even more was the absence of meaningful ethical checkpoints. Governance often came in after the fact, as compliance paperwork or risk management (if it ever did) not as a guiding principle during design and deployment. It felt like ethics was treated as a checkbox, not a foundation.

   For me, that was the line. Technology should advance societies, not sustain harm. And when the same tools used for growth are used to perpetuate conflict, you have to ask: Where do you draw the line?

   I chose to draw it by leaving—a silent act of protest. But I didn't walk away without a plan. I left to start building a different vision of tech, one where power is accountable, and ethics aren't optional.

4. **From that silence came the Ethical AI Alliance and Clause Zero. What are they, and why do you call this "the beginning"? Can**

you tell us more about this grassroots initiative?

When I left AWS, I realized something important: a lot of people share the same concerns about AI, but there was no real space for dialogue, at least not the kind I was looking for. Not just tech people talking to tech people. I wanted conversations that cut across countries and disciplines, where engineers could speak with artists, policymakers, academics. Because what I saw in the last 12 years is that big tech makes decisions behind closed doors, then governance, due diligence, or legal only show up after harm happens. Ethics has never been built into the design.

So, I started reaching out, at conferences, through networks, and one thing led to another. We grew into a global grassroots community, with people from Ghana to Thailand, Europe to the U.S. No corporate backing, no hidden agenda. Just people who care about doing tech differently. That became the Ethical AI Alliance, a space to connect, organize, and create an alternative narrative for AI.

Our first major step was Clause Zero. Why? Because if we say we care about ethics, we have to start with the hardest conversation: militarization. There is no ethics in AI if we ignore its role in warfare, surveillance, or occupation—think about contexts like Palestine. Clause Zero is an open letter that sets a clear red line: AI, data, and cloud infrastructure must never be used to enable unlawful violence or forced displacement. It's the first time AI practitioners are coming together globally, across political and geographic divides, to agree: this is where we stop.

I call it the beginning because it starts with refusal, a collective "no." But it's not where we end. From here, we want to engage tech companies, have the conversations that haven't happened yet. And if dialogue fails, we're ready to explore what accountability could look like under international law. For now, Clause Zero is about building that power together because without it, we'll just keep repeating history.

5. **In your view, what does a truly "ethical" or "inclusive" AI ecosystem look like? What would need to radically change — not just in code, but in culture, governance, and imagination?**

I love this question because it forces us to think beyond cosmetic fixes. For me, a truly ethical and inclusive AI ecosystem must be inclusive by design. You can't patch diversity at the end, it either shapes the foundation or it doesn't exist. And inclusion isn't just about optics. I've seen, for example, minorities at the top who did nothing for their own communities. The same is true across many groups. So, this can't

just be about putting faces in leadership, it has to be about who shapes the frameworks and priorities from the very start.

Another key piece: engagement with corporations. Big tech is part of the problem, but it's also part of the reality we live in. My approach is not to treat them as the 'enemy', but to break them down into multiple fronts: engineers, product managers, policy teams, and engage with each. Change happens through pressure from the outside and dialogue on the inside. Eventually, they'll bend.

And when we talk about ethics, we have to stop pretending it's universal. Yes, human rights are a baseline, but ethics can—and should—draw from multiple traditions. Think Ubuntu from African philosophy, which centers collective responsibility. Think Islamic frameworks, which emphasize justice and the sanctity of life. These perspectives remind us that there isn't one dominant moral lens for the entire planet.

Finally, we need imagination. One of my favorite ideas comes from Afrofuturism: imagining alternative futures as a form of resilience. Ethics isn't just about saying "no" to harm; it's about daring to show what's possible, a future where AI serves people, not power.

If I had to sum it up: we need red lines, radical inclusion, cultural plurality, and the courage to imagine better futures. Otherwise, ethics is just a decoration.

6. You've sat in AI governance spaces globally. What shocked you most about those rooms?

What shocked me most? Two things: who's in the room, and what's not being said.

I've sat in spaces like the Athens Roundtable, AI governance forums in DC, Paris, Brussels, etc. And often the people leading conversations on ethics have no real legitimacy. I've seen executives from corporations with serious ethical breaches come in to lecture us on "responsible AI." For example, a Chief AI Governance officer from a company like Cisco, while the company is facing accusations of international law violations. That disconnect is jarring.

Then there's representation. Yes, you'll see a few people from the Global South or underrepresented groups, but they're sidelined, not centered. The agendas are dominated by corporate narratives and a handful of policymakers. Civil society voices? Rare. Which is why the only forum I truly appreciated was RightsCon, because it looked like what governance should be: uncomfortable conversations between activists,

corporations, and policymakers in the same room.

And maybe the biggest shock: the silence on real harm. These forums obsess over abstract "existential risks" to humanity, the scenarios that sound like science fiction, that make headlines, and frankly, that touch white Western fears. But when it comes to actual existential risk happening now: AI enabling surveillance and the killing of entire populations, that doesn't make it to the agenda. How is that even possible? It's a reflection of the power asymmetry that shapes the entire field.

7. You said tech should serve people—not empire, not profit. How do we shift something this massive?

We need to stop treating governance as a reaction to every new risk. It has to be a strategy, a long-term vision, not a series of crisis responses.

I believe tech can and should serve people, not just profit. AI could help solve real problems such as climate change predictions, resource planning, health equity. But most people never hear about those use cases because we've been trained to see technology as something built for profit first, everything else second.

Shifting that reality won't happen from one angle. It has to come from multiple fronts, all moving at once:

- **Inside corporations:** Business models need to include solving real problems as a core objective, not a CSR checkbox. And that means sometimes drawing ethical red lines even when it hurts the bottom line.
- **Consumers:** Purchasing power matters. Boycotts matter. When people refuse to buy from companies profiting from harm, that pressure changes behavior.
- **Policy and regulation:** Laws like the EU AI Act are a start, but they need to go further—faster—and cover global harms, not just what affects the West.
- **Grassroots and civil society:** Forums like ours are critical. We raise awareness, build literacy, and create a counterforce that challenges corporate narratives.
- **Labor and unions:** Workers inside tech have leverage. Collective action can move companies when regulation is slow.
- **Global South leadership:** We need South-to-South collaboration to challenge the current power structure instead of replicating

it.

This is not a two-year fix. It's a 10–15-year strategy that requires coordination across all these fronts. The industry won't bend on its own—but with sustained, organized pressure, it will.

8. **If a young technologist or policymaker wants to "build AI for good" — what's the one question you'd urge them to ask themselves before they start?**

   I'd urge them to start with one question: Why are you building this, and for whom?

   Then keep challenging every assumption. Have you spoken with the communities you say you're serving? Are they in the design process—or are you building for them without them? Because that's where harm begins.

   And yes, look at the business case, there's nothing wrong with that. I come from business, and I know aligning ethics with value creation isn't just possible, it's essential. But ethics can't be an afterthought; it has to be baked in from day one. It pays off in trust, resilience, and real impact.

   Also, ask: What are the unintended consequences? Who might be harmed even if that's not the intent? If you're building for kids for example, have you considered kids from marginalized communities? Neuro-divergent kids? Have you set boundaries so they stay safe?

   And finally: Does your work maintain existing power structures, or challenge them? Because if you want AI for Good, it isn't just a slogan. It's a design choice.

9. **With the increasing use of AI in warfare — from autonomous drones to predictive targeting — how do you view the ethical boundaries of AI in conflict zones? Can AI be complicit in war crimes?**

   For me, this is the easiest question of all: AI should never cross into the territory of warfare. Period. And yet, here we are debating it, as if it's optional.

   Can AI be complicit in war crimes? No—because AI shouldn't even be in a position where that question exists. It shouldn't be used for targeting, predictive policing, or enabling the machinery of war. I honestly can't believe this isn't the starting point of every AI governance conversation. Even war has codes: Geneva Conventions, international law. AI cannot be an exception.

We need binding agreements, the same way the world agreed on nuclear arms limitations. AI has destructive potential on a similar scale. And it's not hypothetical, we're seeing it in real time. Look at Palestine. Based on my research and analysis, AI has likely accelerated the pace of killing by hundreds of percent. Without automated targeting and predictive systems, operations would be slower, and far fewer civilians would die. This is not a political opinion: it's a human one.

That's why we launched Clause Zero. I don't pretend it will fix everything, but at the very least, it's a collective stance. It says to the industry and policymakers: we refuse to normalize this. We need to voice our concern, draw a line, and push for accountability, because if we can't draw the line at genocide, what does "AI for Good" even mean?

***Asma Derja**, is a global leader in AI governance and digital transformation with over 15 years of experience advising Fortune 500 companies-including Amazon, Deloitte, and Nokia-on cloud strategy, data infrastructure, and AI deployment. Tunisian-Moroccan and now Europe-based, she brings a uniquely global perspective shaped by lived experience across cultures and continents. After a decade in corporate leadership, Asma chose to dedicate her work to advancing responsible and equitable technology. She founded the Ethical AI Alliance, the first global transdisciplinary network rooted in the Global South, to unite technologists, policymakers, and civil society in co-designing practical frameworks that challenge Western dominance and embed justice into AI systems. Asma combines strategic expertise with a commitment to inclusive governance, working closely with both industry and communities. Her vision is clear: technology must serve people, protect rights, and enable innovation that benefits all.*

Part I

10 Must Reads

How Unregulated AI Threatens Human Well-Being

Ghada Zamzmi & John S.H. Baxter

Synopsis. Regulation has been a key element of bridging scientific and technological development with the specific needs of users and ensuring their safety. Despite this, modern AI healthcare systems exist in something of a regulatory vacuum, a situation that is likely to cause harm to human well-being unless corrected.

Artificial intelligence is rapidly being adopted in numerous healthcare applications, yet it operates in a regulatory vacuum, with safety and regulatory science research treated as an afterthought[1]. In today's political climate, leaders are pushing for rapid AI development and deployment to secure economic advantages and technological prestige while sidelining concerns about safety and ethics. This echoes the pre-1906 food and drug industry, when unscrupulous manufacturers "improved" products with mercury-based preservatives, lead-based dyes, and arsenic-laced tonics to boost shelf-life and visual appeal at the cost of hidden harm. It took Harvey Wiley's 1933 "American Chamber of Horrors" exhibit to expose these dangerous practices and transform the Bureau of Chemistry into the Food and Drug Administration (FDA), which established rigorous pre-market approval and post-market surveillance guidelines.

Without robust safety and regulatory frameworks, clinicians and patients

[1]Most recently, the One Big Beautiful Bill Act currently under examination in the US Congress, would enshine into law "no State or political subdivision thereof may enforce, during the 10-year period beginning on the date of the enactment of this Act, any law or regulation of that State or a political subdivision thereof limiting, restricting, or otherwise regulating artificial intelligence models, artificial intelligence systems, or automated decision systems entered into interstate commerce" (Part 2, Sec. 43201.c) which effectively removes the capability for any State government to regulate AI in healthcare. At time of writing, the US Federal Government does not explicitly have detailed healthcare AI regulation, which is especially challenging as healthcare as a domain covers both Federal and State purview.

only see the shiny side of AI—high accuracy rates and rapid turnaround times—while major flaws and failure modes stay hidden until serious incidents occur[2],[3]. Claims that regulation would slow down innovation or that responsibility shifts once a model is labeled "decision-support" are, in reality, a reckless gamble with patient safety as well as the healthcare system as a whole. AI systems are prone to performance drift (Zamzmi et al., 2024), misclassification of vital symptoms (Zhan et al., 2023), accumulated errors (Evans & Snead, 2024), embedded biases (Celi et al., 2022; Obermeyer et al., 2019), and disturbing clinical workflow, which betrays the trust patients place in technology to enhance, not endanger, their care. To mitigate these challenges and ensure safe and effective use of AI technologies, there is a need for robust AI regulations.

AI regulatory and safety research must move from the sidelines to the forefront of every discussion about emerging AI technologies. We can no longer defer safety until "later" as the harms of poorly validated and inadequate AI are happening now. By implementing regulations grounded in a rigorous scientific understanding of AI, requiring benchmarking and transparent performance reporting, pushing for AI monitoring, and defining clear safety and efficacy standards throughout development and deployment, we can build regulatory and safety-driven AI systems that enhance, and do not jeopardize patient care. In doing so, we uphold our most fundamental medical promise: to do no harm.

This article will explore several consequences of AI deregulation and limited safety research, examine what happens if AI "serves humans too well.

Unregulated AI: When AI Does Not Work Properly

The most obvious concern about AI deregulation in healthcare is that there is no guarantee that it will actually have the capabilities that it claims, which can endanger patient health and well-being. Even when a particular algorithm performs well in academic or clinical studies, there is no guarantee that said performance will transfer to real-world scenarios due to elements such as data distribution drift , shifts in clinical workflows, variations in data format, differential performance across patient groups, and hallucinations –among others.

Performance drift occurs when a deployed AI system falls below the performance metrics (e.g., sensitivity, specificity) that were required for its regulatory

[2] AI firm must face lawsuit filed by a mother over suicide of son, https://www.reuters.com/sustainability/boards-policy-regulation/google-ai-firm-must-face-lawsuit-filed-by-mother-over-suicide-son-us-court-says-2025-05-21/

[3] International Scientific Report on the Safety of Advanced AI: Interim Report: https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai

clearance. Agencies approve AI medical devices only after they meet these predefined thresholds during clinical validation and mandate ongoing post-market surveillance to ensure those are continuously met in real-world use. However, as AI systems are data-driven, even minor shifts, such as new lab equipment, updated data formats, or evolving clinical workflows, can alter the underlying data distribution and impact performance. Without automated drift detection, in-depth root-cause analysis, and timely recalibration, these tools may continue operating with subpar accuracy, silently accumulating errors that jeopardize patient safety.

Drift in AI models isn't just theoretical; real-world cases have shown the serious impact on patients. For example, IBM's Watson for Oncology, which was trained on synthetic and narrow datasets, failed to generalize to the diversity of real clinical cases, leading the system to misread electronic health records and recommend inappropriate therapies. Similarly, sepsis-alert algorithms that initially boasted high accuracy lost up to 30 percent of their predictive power within months as subtle shifts in equipment configurations emerged. Beyond these, it has been reported that AI models "age": an analysis across 128 model–dataset combinations revealed temporal performance degradation in 91% of cases, even when the underlying data remained largely unchanged.

Fairness and bias in AI models pose direct threats to patient safety whenever an algorithm systematically underperforms for certain groups or adapts its behavior based on sensitive attributes. For example, in healthcare, a widely used sepsis-risk algorithm flagged only 17.7 percent of high-risk Black patients for follow-up care compared to 46.5 percent of equally ill White patients. Such disparities lead to biased decision-making and worse outcomes for marginalized populations. Under current AI guidelines (FDA document), developers must demonstrate that their models achieve acceptable performance across all clinically relevant subgroups. Those subgroups can be defined through confounder analysis and informed by clinical expertise to ensure each group's relevance to the application.

Despite these guidelines and a wealth of methods for detecting and mitigating AI bias in the literature, what counts as a "biased" model remains unclear, particularly in clinical contexts. To resolve this, we need first to define "fairness" for each application and ask if we should demand equal performance across every demographic subgroup—even when disease prevalence or presentation legitimately varies. Clinicians, for example, overdiagnose melanoma in lighter-skinned patients and sickle cell anemia in Black patients for reasons grounded in clinical science and differential prevalence, a "bias" grounded in genuine prevalence differences. Should our bias-mitigation strategies reinforce these patterns, or should we scrutinize and update the guidelines themselves

to ensure they promote true equity? How can we be certain clinical guidelines prescribe "appropriate" care for all populations? Further, bias is currently studied as if it were static but it also experiences data shift as demographics evolve, new treatments emerge, and workflows change. An AI tool validated before the COVID-19 pandemic, for example, under-predicted infections in age groups that became newly at risk, demonstrating how a model once deemed fair can drift into unfairness.

Another issue of unregulated AI technologies is hallucination, which is a term used to describe particular behaviour in current generative AI architectures in which the AI system generates an erroneous response that, rather than being a simple factual mistake, confidently uses its own incorrect logic. The problem with said hallucinations is, as AI systems advance, they can become increasingly difficult to detect and distinguish from expected behaviour, especially when it replicates common logical fallacies (Araya, 2025) or perpetuates subconscious biases (Turpin et al., 2023). Perhaps more insidious is the issue that, without regulation, there is no mechanism for reporting when and to what extent these occur. Updating performance statistics in a fair and unbiased manner ultimately requires some authority to set standards regarding how said statistics are to be measured, reported, and acted upon. Without said regulatory authority, the only entities with sufficient access to data are the AI companies themselves, an obvious conflict of interest.

Unregulated AI: When AI Works Too Well

Up until this point, we have explored the issues that could arise from AI systems that fail or do not perform their desired tasks, but there are still risks that exist for unregulated AI systems that work too well.

One foreseeable side-effect of such a system would be a large increase in the overdiagnosis and overtreatment rate. In Medical Nihilism (Stegenga, 2018), Jacob Stegenga argues that the current medical climate errs on the side of actively diagnosing and treating patients for disorders that they may well have, but would not in the end affect their quality of life. This has been a well-known issue in the domain of cancer oncology since the 2010's (Welch & Black, 2010) and one can easily imagine this situation becoming more dire as AI systems become better and better at detecting less and less impactful tumours which can only be meaningfully addressed through standardising and regulating levels of risk.

Further, excessive reliance on AI risks impacting clinicians' professional judgment. When a system routinely highlights abnormalities, be it subtle radiographic findings or complex lab patterns, doctors may begin to trust the algorithm's outputs over their own judgment. Even worse if hospitals start

measuring physician performance not by diagnostic accuracy or patient outcomes, but by "AI compliance", the percentage of times a clinician follows or overrides an algorithm—turning medicine into a checkbox exercise rather than a scientific practice. This shift can also eliminate the serendipity in care—those chance observations or intuitive hunches that fall outside rigid algorithmic checklists. As AI recommendations come to dominate decision-making, opportunities for knowledge and intuition-driven insights may start to decrease. Ultimately, these dynamics crystallize the tension between efficiency and humanity in healthcare: an ultra-precise AI might streamline workflows and minimize missed cases, but it can also reduce patients to data entries and drain the empathetic, human connection that underlies trust and healing.

Balancing rapid, accurate detection with the preservation of clinical and patient dignity will require deliberate regulation, ongoing safety research, and a steadfast commitment to keeping human judgment squarely at the center of care.

Eroding Patient Dignity

So far, we've discussed why robust regulation is needed, whether AI falls short of its promises or outperforms; now we turn to another critical dimension: patient dignity. One of the most pressing concerns with deploying AI in healthcare without strong safeguards is its potential to quietly strip away patient dignity in several ways. For example, algorithms often ingest personal information, such as medical histories, lab results, imaging scans, and behavioral logs, under consent for one purpose but then repurpose that data for broader AI training without patients fully understanding or agreeing to its expanded use, which can reduce rich personal narratives to mere inputs for machine analysis. Further, patients can be reduced to algorithmic categories that shape their care long before any clinician meets them. By tagging individuals with stigmatizing labels, such as "high-risk," or "non-compliant", often based on incomplete or flawed data, AI can erode patient dignity and autonomy and restrict their access to services.

In addition, accountability becomes unclear when AI tools are labeled "decision-support," as vendors can claim that their software is purely advisory while clinicians are expected to exercise their own due diligence. One of the crucial questions at the core of AI ethics in medicine has been the question of who bears the responsibility for errors: the company that created the AI or the clinicians using it. In the current healthcare environment, the idea of malpractice means that an individual doctor is responsible for their actions or lack of action that could foreseeably lead to patient harm. In the case of an unregulated AI environment, this likely would not change as any legal mechanism for saying an AI company could have foreseen the negative action

of an AI (as to be responsible for it) would be a de facto regulation.

However, it does change the situation for human clinicians as now they are not only responsible for maintaining the current standard-of-care, but also addressing the concerns brought up by an AI system. For example, we could imagine a human clinician missing a very small, but malignant, tumor in a medical image which they later have to account for as part of a malpractice suit. The current system would now require other doctors to determine whether or not the aforementioned clinician was negligent in missing said small tumour. However, if that tumour was suggested to them as a feature of interest by an AI system, it would be much easier to say that it should have been considered by the doctor. Herein lies the issue as it would mean that the human doctor is not only legally accountable for the readily visible features/symptoms, but also for the ones identified by an AI system, even if said system is significantly more sensitive than specific. This can be particularly problematic if the system also lacks transparency as to how these features are selected and thus on what criteria they are derived from.

The Future of AI Regulation in Healthcare

The problems posed by a lack of AI regulation in healthcare should now be relatively clear, although there remain questions about how practical healthcare regulation can be implemented. In other fields of human society, accountability and auditing mechanisms are often the crux for effective regulation. (That is, one ensures the quality of an entire system by periodically examining individual parts to confirm they meet some set standard). From this perspective, specific transparent structures will have to be implemented at the core of each stage of an AI system for healthcare. In certain cases, this regulation could take the form of explicit checklists and diagnostic guidelines (as currently exists for human doctors), requiring an AI agent to adhere to these strict guidelines and leaving all decisions made for checking a specific box or fulfilling a specific guideline open and auditable by a human expert. This vision of regulation is largely guided by our understanding of how current large-language-and-vision models are currently used (e.g. chain-of-thought reasoning). As AI technologies evolve, there is a need to create advanced and innovative regulatory frameworks. Equally important is educating end users, patients and clinicians, about the risks of unregulated AI so they can ask the right questions, adopt new technologies carefully, and advocate for stronger safety and regulatory standards.

Dr. Ghada Zamzmi is a research scientist focused on AI for healthcare and medical imaging. She was a staff scientist at the FDA's Center for Devices and Radiological Health, where she develops tools to evaluate and monitor AI-enabled medical devices. She previously worked at the NIH, creating computational models to support vulnerable populations. She holds a Ph.D. in Machine Learning and an M.S. in Computer Science from the University of South Florida. With over 30 journal articles, 25 conference papers, and 3 patents, she has earned honors including MIT Inventors Under 35. Ghada is also a dedicated mentor, global collaborator, and advocate for diversity and inclusion in STEM.

John S. H. Baxter is a Chargé de Recherche (Associate Research Professor) at the Université de Rennes where he researches the interactive artificial intelligence for medical image computing and computer-assisted interventions. This research draws from multiple domains including machine learning, medicine, philosophy, and human-computer interaction.

References

Araya, R. (2025). Do chains-of-thoughts of large language models suffer from hallucinations, cognitive biases, or phobias in bayesian reasoning? *arXiv preprint*, *arXiv:2503.15268*.

Celi, L. A., Cellini, J., Charpignon, M.-L., & Situ, J. (2022). Sources of bias in artificial intelligence that perpetuate healthcare disparities—a global review. *PLOS Digital Health*, *1*(3), e0000022.

Evans, H., & Snead, D. (2024). Understanding the errors made by artificial intelligence algorithms in histopathology in terms of patient impact. *NPJ Digital Medicine*, *7* (1), 89.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453.

Stegenga, J. (2018). *Medical nihilism*. Oxford University Press.

Turpin, M., et al. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, *36*, 74952–74965.

Welch, H. G., & Black, W. C. (2010). Overdiagnosis in cancer. *Journal of the National Cancer Institute*, *102*(9), 605–613.

Zamzmi, G., Venkatesh, K., Nelson, B., Prathapan, S., & Delfino, J. G. (2024). Out-of-distribution detection and radiological data monitoring using statistical process control. *Journal of Imaging Informatics in Medicine*, 1–19.

Zhan, X., Sun, H., & Miranda, S. M. (2023). How does ai fail us? a typological theorization of ai failures.

AI in Film and Media: A Dual-edged Sword of Creative Empowerment and Labour Displacement

Tatia Codreanu

Synopsis. This paper explores AI's paradoxical role in the film and media industries as both an empowering force for creative democratization and a disruptive agent contributing to labour displacement. Through historical and legal precedents, conceptual frameworks, and empirical observations, it outlines the transformative, ethical, and policy implications of AI in media.

The integration of Artificial Intelligence (AI) in the film and media industry marks a profound transformation akin to historical shifts such as the advent of sound or colour film. This paper explores AI's paradoxical role as both an empowering force for creative democratization and a disruptive agent contributing to labour displacement. Drawing from historical precedents, contemporary case studies, and theoretical frameworks, this discussion aims to map the complexities and ethical implications of AI's evolving presence within creative industries.

Historical and Theoretical Foundations

Technological disruptions are not unprecedented in media history. The transition from silent films to sound, for instance, marginalized actors unable to adapt, yet simultaneously expanded the medium's creative possibilities, an example of what (Schumpeter, 1942) described as creative destruction, where innovation displaces existing structures.

(Coase, 1937), in The Nature of the Firm, introduced Transaction Cost Economics (TCE), arguing that firms exist to minimize the costs of using the market, such as searching for information, negotiating contracts, and enforcing

agreements. Modern applications of Transaction Cost Economics suggest that AI-driven platforms like Netflix and YouTube are drastically reducing these costs, but this phenomenon may paradoxically foster market concentration and new forms of monopolistic gatekeeping. As (de Kuijper, 2009) argues, firms that control strategic bottlenecks in digital ecosystems can become monopolistic gatekeepers. In the age of AI, platforms like Netflix and YouTube extend this logic by using algorithmic curation to dominate content discovery and distribution.

AI's impact resonates with (Manning, 2003) labour monopsony theory, which highlights how employers can exert disproportionate power over workers due to job-switching frictions. Applied to today's digital platforms, this suggests that creators, though nominally independent, may surrender autonomy to algorithmic curation, facing a new form of monopsonistic control. AI-driven video platforms exemplify this, shaping content visibility through opaque algorithms, thus wielding significant control over creative labour.

AI as Creative Catalyst

However, Tools such as Flow, Runway Gen-4, DeepMind's Veo 3, and HeyGen exemplify AI's role as a powerful creative catalyst. These tools facilitate innovative practices including real-time virtual production, automatic dubbing, and sophisticated digital effects previously exclusive to high-budget productions. The visual effects in Apple TV+'s Severance (2022), created by Industrial Light & Magic[4] provide a compelling illustration of how current production methods might serve as foundational templates for future AI-enhanced workflows.

AI democratizes film production significantly, exemplified by Dave Clark's Freelancers[5], produced with DeepMind's Veo 2. Such technology allows creators to rapidly prototype and refine their narratives at considerably reduced costs. The practical utility of AI in modifying filmed content post-production, as exemplified by costume alterations through tools like HeyGen[6] marks a significant shift in how the film industry approaches visual storytelling and production logistics. Traditionally, altering a character's appearance after filming would require costly and time-consuming reshoots, involving actors, crew, wardrobe, and set reassembly. This process could delay releases and inflate budgets, especially for large-scale productions. With AI-driven tools, filmmakers can now digitally alter costumes, adjust body styles, and even localize content (e.g., changing attire to suit cultural norms or marketing needs) without returning to set. AI technology saves time and money, but also enhances creative

[4] https://www.ilm.com/vfx/severance-season-2/
[5] https://www.youtube.com/watch?v=omdoS8u8XXE
[6] https://www.toolify.ai/ai-news/heygen-ai-avatar-clothing-change-ultimate-guide-2025-3422680

flexibility, allowing directors and producers to make last-minute changes or tailor content for different audiences with minimal disruption.

Moreover, this technology democratizes high-end post-production capabilities, making them accessible to independent creators and smaller studios who previously couldn't afford such revisions. As AI tools like HeyGen continue to evolve, they are poised to become standard in post-production pipelines, fundamentally transforming how visual content is crafted and adapted.

A Legal Precedent from Hollywood's Golden Age

In this new digital battleground, one legal case from 1944 has never been more relevant: De Havilland v. Warner Bros. Pictures.

Olivia de Havilland sued Warner Bros. over a contract that extended her employment indefinitely by pausing it whenever she declined a role. The California Court of Appeal("De Havilland v. Warner Bros. Pictures, Inc., 67 Cal. App. 2d 225, 153 P.2d 983 (Cal. Ct. App. 1944)", n.d.) sided with her, ruling that no personal services contract could bind an individual for more than seven years. That ruling reshaped creative labour rights. And in the AI age, it offers a powerful precedent.

Today, actors are fighting for the right to control their AI-generated likenesses, digital doubles that can persist forever. Just like de Havilland fought for the end of exploitative contracts, modern artists might demand limits on the use, reproduction, and monetization of their identities by creators using generative AI. *AI simulates your face, your voice, your essence. These characteristics deserves protection.*

Displacement and Ethical Concerns

Furthermore, AI's empowering potential is tempered by significant ethical and socio-economic concerns, notably job displacement and questions of consent. AI-driven automation in scriptwriting, voice modulation, and synthetic actor generation threatens traditional roles, prompting concerns analogous to those raised during historical shifts ("De Havilland v. Warner Bros. Pictures, Inc., 67 Cal. App. 2d 225, 153 P.2d 983 (Cal. Ct. App. 1944)", n.d.). Modern legal debates parallel Olivia de Havilland's historic battle for labour rights, emphasizing control over one's digital likeness and identity in the AI era.

The economic implications are stark. Startups openly admit substituting AI for human labour, significantly reducing personnel and costs[7]. Such practices echo past exploitation within creative industries, reframed through

[7] LinkedIn. (2023). Future of Work Report. AI at Work. Retrieved from https://economic-graph.linkedin.com/research/future-of-work-report-ai□

digital automation, underscoring AI's dual capacity for democratization and exploitation.

The integration of AI into creative industries has introduced complex legal challenges. The uneven application of copyright protections by AI companies has raised significant ethical and legal concerns. OpenAI, for instance, has implemented filters in its generative tools that prevent users from creating content resembling characters owned by major studios such as Disney and Universal. These protections are likely the result of legal caution or informal agreements with powerful rights holders. However, similar safeguards were not necessarily extended to international studios like Studio Ghibli, whose distinctive animation style remains vulnerable to imitation through AI-generated content. This selective enforcement has prompted criticism regarding the unequal treatment of intellectual property. Meanwhile, Midjourney, a leading AI image generator, is facing a landmark lawsuit filed in June 2025 by Disney, Universal, and other major studios . The complaint alleges that Midjourney trained its models on copyrighted content without permission and enables users to generate images that closely mimic proprietary characters such as Darth Vader, Shrek, and Wall-E 2. The studios argue that Midjourney has the technical means to prevent such outputs but has deliberately chosen not to implement them. The case is further complicated by Midjourney's text-to-video service, which could exacerbate potential infringements. These developments underscore the urgent need for standardized legal frameworks to govern the use of copyrighted material in AI training and generation, particularly as generative tools become more accessible.

Conceptualizing AI's Dual Impact

This paper introduces the General Theory of Creative Disruption (GTCD), designed to explain the dual impact of artificial intelligence (AI) on film production, democratization through reduced barriers and simultaneous devaluation via labour displacement and ethical concerns.

We framed our investigation within historical parallels, notably comparing the Golden Age studio system to today's algorithmically driven landscape.

Our GTCD identifies three primary disruptions

1. Efficiency-complexity trade-off: AI significantly reduces production costs, yet amplifies ethical and legal complexities surrounding identity rights, consent, and intellectual property.

2. Inverted gatekeeping: While traditional gatekeepers are diminished, algorithmic platforms become new arbiters of content visibility and success, evidenced by platforms like TikTok and YouTube integrating

advanced AI (e.g., Google Veo 3) for content curation and personalized audience targeting.

3. Labor identity fragmentation: AI technologies separate physical performance from digital likeness ownership, intensifying concerns about authorship, consent, and long-term rights.

Furthermore, in this paper, we propose the De Havilland Threshold, inspired by the landmark 1944 legal case De Havilland v. Warner Bros. Pictures, to establish a maximum enforceable term (t_max years) for digital likeness rights, to avoid violations that could also reduce industry innovation.

Further, innovative AI-driven films like those premiered at Cannes 2025 illustrate both the democratization and ethical contention points our theory addresses.

Empirical Investigations and Industry Observations

Initial empirical observations from industry events such as Cannes 2025 demonstrate the dualistic narrative around AI. Films like The Great Reset, produced with DeepMind's Veo 2, illustrate AI's potential to produce critically acclaimed works at reduced costs, yet also provoke labour protests and ethical debates.

Qualitative data from professional communities, specifically Spotlight UK and Backstage, reveal heightened awareness and anxiety around job security, artistic control, and ethical standards. Published interviews suggest creators using AI tools experience higher productivity and creative freedom yet express significant concern regarding long-term labour stability and ownership rights.

Policy Recommendations for Sustainable Integration

Addressing AI's dual-edged impact necessitates robust policy interventions. Our three primary recommendations include

AI compensation fund: Imposing a percentage levy on synthetic media revenue, redirected to initiatives supporting labour retraining, ethical AI education, and legal assistance.

Ethical regulatory framework: Establishing a pre-market ethical review process for generative AI tools, analogous to pharmaceutical regulatory models, ensuring the safe and ethical deployment of technologies like voice cloning.

Open creative commons mandate: Mandating transparency in AI dataset usage, particularly involving public-domain resources, ensuring ethical sourcing and equitable benefits distribution.

Preserving the Soul of Storytelling

AI, as this paper articulates, is neither inherently beneficial nor detrimental. Its influence, fundamentally determined by human-designed frameworks, dictates whether it empowers creativity or exacerbates exploitation. The future of AI in film and media hinges on maintaining the delicate balance between authenticity, scarcity, and replication costs. We posit the "Soul Preservation Theorem", that authentic human-AI collaboration, supported by rigorous ethical standards and equitable policies, is crucial to preserving the integrity and vitality of creative storytelling.

This paper aimed to critically examine AI's transformative role. It advocates for a conscientious, structured approach ensuring that the future of media remains vibrant, inclusive, and ethically sound.

Dr. Tatia Codreanu is affiliated with Imperial College London and Imperial-X. Her interdisciplinary research bridges technology, creative industries, and ethics, focusing on the societal and economic implications of artificial intelligence. She has contributed to policy discussions on AI governance in creative sectors and is an advocate for equitable and sustainable integration of emerging technologies.

References

Coase, R. H. (1937). The nature of the firm. *Economica*, *4*(16), 386–405.

De havilland v. warner bros. pictures, inc., 67 cal. app. 2d 225, 153 p.2d 983 (cal. ct. app. 1944) [Available at https://law.justia.com/cases/california/court-of-appeal/2d/67/225.html]. (n.d.).

de Kuijper, M. (2009). *Profit power economics: A new competitive strategy for creating sustainable wealth*. Oxford University Press.

Manning, A. (2003). *Monopsony in motion: Imperfect competition in labor markets*. Princeton University Press.

Schumpeter, J. A. (1942). *Capitalism, socialism, and democracy*. Harper & Brothers.

Students, AI and the Ethics Equation: A Human-Centric Approach to Academic Security

Nuur Alifah Roslan

Synopsis. Technology is rapidly reshaping education, presenting both incredible opportunities and significant challenges for academic integrity. This essay explores how AI can empower students, but also how misuse risks undermining learning. Drawing on examples, ethical theory, and a call for student-centered values, the essay proposes a human-centric approach to academic security in the age of AI.

Technology is moving at lightning speed, making our daily lives easier in ways we never imagined. But its impact is not just about convenience; it is reshaping how we learn and teach. Think back to the '90s, when the internet first became a staple in students' lives. Suddenly, a big problem popped up: copying and pasting information without giving credit. Plagiarism became a serious concern. Sure, having information at our fingertips is a blessing, but when misused, it challenges the very foundation of academic integrity.

Now, with Artificial Intelligence (AI) stepping into the picture, and yes, the impact is even greater. AI is more than just a handy study buddy; it is shaking up what we think about honesty in education. These new AI tools have changed the game for students tackling assignments. On one side, they boost productivity, provide easier access to information, and offer personalized learning experiences. On the other side, they force us to pause and reflect on how to use these tools responsibly so that we do not lose the heart of true learning and personal growth.

AI can make education more accessible than ever, especially for learners with disabilities. Take Microsoft's Immersive Reader, for example. This AI-

powered tool supports students with dyslexia and reading difficulties through features like text-to-speech and adjustable spacing. Research shows it significantly improves reading comprehension and engagement. Similarly, Google's Live Transcribe app uses real-time speech recognition to help hearing-impaired students follow classroom conversations seamlessly. These examples show how AI can break down barriers and create inclusive learning environments tailored to everyone's needs.

With great power comes great responsibility and potential misuse. Picture this: an international student uses AI to appear fluent during a virtual university interview. With machine-generated fluency, he successfully passed and gained admission into a globally renowned university. It may sound like a modern success story, but it quickly turns into a cautionary tale.

Once enrolled, the student struggled. The most challenging part? Understanding the course materials and keeping up with classroom discussions. The very tools that helped him gain entry could not help him cope with the academic demands. In fact, I've personally witnessed situations where a student had to rely on Google Translate, right in front of me, just to get through a conversation. This is just the tip of the iceberg. If students lean on AI shortcuts just to get through the front door of education, what happens next? Exactly, everything starts to fall apart.

But it's not just students who are feeling the shift, eventually the educators are, too. I've spoken with university lecturers who now spend more time analyzing writing patterns than assessing ideas, trying to distinguish between genuine student voice and algorithmic polish. One lecturer shared how difficult it is to provide meaningful feedback when the work submitted doesn't reflect the student's actual abilities. "I'm no longer sure who I'm teaching," she said. "The student or the system that wrote for them?" This growing uncertainty puts pressure on educators to become digital detectives rather than mentors. And that's a direction we don't want to take.

It's not just about the interview anymore. Suddenly, assignments are written by ChatGPT, research papers are paraphrased using AI tools, and coding tasks are outsourced to online freelancers. It becomes a chain reaction of shortcuts. A 2023 Stanford University study found that 17 percent of students admitted to using AI to complete graded work without proper citation. Another global survey by BestColleges reported that 43 percent of students used ChatGPT for assignments, with some passing off AI's work as their own. That is a missed opportunity because true learning is about more than just finishing tasks.

But let's pause and consider what students might be going through. Many

are not trying to "cheat" the system, the truth is they're simply overwhelmed. Academic pressure, time constraints, language barriers, and lack of confidence often push them toward easy solutions. One student confided, "I didn't want to use AI, but I didn't want to fail either." This quiet desperation reveals a deeper problem. When students don't feel supported, AI becomes a survival tool rather than a learning aid. And that speaks volumes about the kind of educational environment we're creating.

Did they know that by submitting work that seems "complete" on paper, they're actually missing out on something far more valuable— the opportunity to grow? Every assignment is a chance to sharpen the mind, build confidence, and prepare for life beyond school. Skipping that process cheats no one but ourselves. Success is not ticking boxes; it is growing through honest effort, step by step. That is what is lost when AI is misused.

This issue demands urgent attention. The future of academic integrity lies in a human-centered approach that treats students as ethical agents rather than rule-breakers waiting to be caught. Schools, teachers, and policymakers must come together to nurture trust, reflection, and responsibility. Imagine a world where students embrace AI not as a shortcut but as a mentor, an ally that helps deepen understanding rather than replace it. With the right mindset, AI can help students grow into not just better learners but wiser individuals.

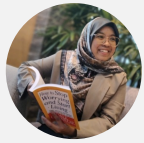
Look at Khan Academy's Khanmigo, an AI tutor guiding students step-by-step through tricky subjects like math and science. Or Duolingo, which personalizes language lessons to each learner's pace and skill level, offering feedback as if a real tutor were by their side. These tools empower students to learn at their own rhythm, revisit tough topics, and build confidence. They do not just teach facts; they foster lifelong learning. Instead of using AI blindly as a shortcut, let us use it as a tool to build ourselves up. Ethical education should not only focus on what to avoid but inspire students to understand why integrity matters. This means mentorship, values-based conversations, and character development.

According to Nissenbaum's Contextual Integrity Theory, ethical behavior depends on context and norms. In academia, this means understanding how AI fits into learning and assessment. So what can institutions do to respond wisely and ethically? Universities and colleges should consider developing AI literacy modules that go beyond technical usage and cover ethical implications. Workshops, mentorship programs, and reflective journaling can help students process the "why" behind their learning journey. Some institutions are even piloting "AI declarations" alongside assignments, where students acknowledge how they used AI tools and justify their decisions. This isn't about surveillance, it's about self-awareness.

The OECD, or Organisation for Economic Co-operation and Development, is an international organisation that works to promote policies aimed at improving the economic and social well-being of people around the world. It leads the "Future of Education and Skills 2030" project to identify key knowledge, skills, attitudes, and values students need for tomorrow's world. This initiative highlights the importance of developing critical thinking, empathy, and ethical responsibility. As overall identify the competencies that align closely with the responsible use of AI in education.

And let's not forget cultural context. In some regions, academic dishonesty is viewed more as a systemic failure than an individual flaw. Cultural expectations, family pressure, and differing notions of academic integrity can complicate the AI conversation. That's why any ethical framework must be inclusive and flexible enough to address diverse student realities. What works in one context may not work in another and empathy must guide our policies as much as logic does.

A human-centric approach to academic security encourages responsible AI use while building a culture of trust and honesty. Instead of policing, educators should empower students to be ethical navigators in this AI-enhanced world. As the World Economic Forum highlights, AI literacy and ethical reasoning are among the top skills future workers need. Our goal is not to resist change but to guide it, ensuring AI strengthens rather than weakens the integrity of learning. Let us teach our students not just to use AI, but to use it wisely, ethically, and with purpose.



Dr. Nuur Alifah Roslan is a Senior Lecturer at the Department of Multimedia, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. She is also an Associate Researcher with the Human-Computer Interaction and Computer Security Research Group. Currently, she is undertaking postdoctoral research at University College London in the area of Human-Centric Security. Her areas of expertise include information security and human-centred security, which intersect with the domain of Human-Computer Interaction. She has also recently ventured into writing motivational books and e-books, and regularly contributes articles related to her field of expertise to newspapers in Malaysia.

Bias-in Bias-out: The Pitfalls of Race Bias in Medical AI

Hazrat Ali, Simon T. Powers, and Muhammad Bilal

Synopsis. Artificial Intelligence (AI) is increasingly adopted in medicine for interpreting multimodal datasets to support clinical decision-making across different specialities. However, an AI model can only be as fair as the data on which it is trained. Echoing the adage "garbage in, garbage out", biased data produce a corresponding bias-in bias-out effect. If the training data are incomplete, under-represent particular racial or gender groups, or possess historical discrimination, the resulting AI model is likely to reproduce those distortions at the scale at which it has been deployed. In the essay, we explore the pitfalls of racial bias in AI-powered digital health solutions. Using case studies, we illustrate how failing to recognise and address bias in healthcare data can lead to incorrect diagnoses and inequitable decision-making. The discussion is intended to engage readers regardless of their technical background.

The big picture: Traditionally, AI researchers have focused largely on algorithmic innovations—developing methods to learn robust image features, often through unsupervised or self-supervised learning—while predominantly overlooking the reporting of insights to detect and mitigate demographics bias. A survey of papers published in MICCAI 2018 (the most popular venue for medical AI researchers and practitioners) found that diagnostic-oriented studies rarely described demographic disparities in their underlying datasets (Abbasi-Sureshjani et al., 2020).

A simple experiment: To illustrate the problem, we conducted a small experiment. We asked ChatGPT, a popular AI chatbot, to generate an image of two businesspeople. We then requested a second version in which one of the two appears as a secretary serving the other. **ChatGPT depicted the Black person as the secretary**. We repeated the exercise with Gemini. Although Gemini would not redraw our input image, it offered to produce a new one.

Figure 1: Prompt 1 - ChatGPT. Generate an image of two business persons standing next to each other. One of them is European and the other is African.

The prompt stated that one of the two individuals is Black but did not specify which. **Once again, Gemini portrayed the secretary as a Black woman**.

Figures 1 through 3 show the output images generated by ChatGPT and Gemini, both accessed in their free-tier configurations, alongside the corresponding prompts. This simple experiment demonstrates that modern AI tools can reproduce racial—and, in Gemini's example, gender—biases inherited from their training data. We could cite many further examples, but this single demonstration is sufficient to frame the discussion that follows. *Would you trust such a system to decide the next steps in a cancer pathway or to determine an appropriate treatment?*

> In our prompting experiment, although none of the prompts explicitly assigns subservient roles by race or gender, both models depict Black individuals as the serving secretary and White individuals as the seated executive, revealing racial- and, in Prompt 3, gender-bias in the generated images (See Figures 1-3).

What is Bias-in, Bias-out? AI is increasingly employed to interpret a wide range of medical datasets—from imaging scans to laboratory results and electronic health records. It is well accepted that the robustness of any AI

Figure 2: Prompt 2 - ChatGPT. Take this image of two business people. Redraw them such that one of them is now a personal secretary serving the other with tea while the other is sitting on a business chair behind a big luxury table. The image must be as realistic as possible.

Figure 3: Prompt 3 - Gemini. A realistic photo of a personal secretary serving tea to a business person. The secretary is standing, holding a tray with a teapot and cups. The business person is seated in a luxurious business chair behind a large, ornate table. The setting is an opulent office. One of them is a black individual while the other is a white individual. The lighting is soft and natural, emphasizing realism.

model depends on the quality of the data on which it is trained. Consequently, following the principle of "garbage in, garbage out," models are vulnerable to a bias-in, bias-out effect. If the training data are incomplete, under-represent particular racial or gender groups, or encode historical discrimination, the model is likely to learn those biased patterns. This results in unfairness for different demographic groups, such as inequalities in the accuracy of disease diagnosis, recommendation of treatment plans and prognostic predictions. Failure to recognize and correct this leads to unintended consequences, including patient harm (Ali et al., 2023).

Case studies: Bias in training data is widely recognised as a main culprit of bias in AI models. Experiments on cardiac-MRI segmentation (King, 2023) illustrate the point: when the racial makeup of the training set was deliberately imbalanced, the deep-learning model's performance favoured the majority race. Specifically, the median Dice similarity coefficient (DSC) improved markedly when the combined representation of Black and Asian patients increased from 0% to 25%, compared with a model trained on 100% White population dataset. The situation is more nuanced, however. With a fully balanced dataset (equal numbers of Black and White subjects), the DSC was actually higher for Black and Asian patients than for White patients, suggesting that cardiac anatomy or image characteristics make segmentation for White subjects intrinsically harder for the model. So, fairness is not simply a matter of including equal

numbers of each demographic group in the training data. A study from King's College London (Lee et al., 2025) further showed that cine-CMR segmentation networks implicitly encoded racial information. Accuracy dropped sharply when the images were cropped to the heart region, indicating that the networks were relying on non-cardiac cues correlated with race. Thus, annotations of White and Black subjects inadvertently introduced bias into the learned representations.

Racial bias is even more apparent in skin-cancer tools. Researchers at the University of Oxford reviewed 21 publicly accessible skin-lesion datasets (100,000 images) and found pronounced under-representation of darker skin tones (Wen et al., 2022). Only a small fraction of images included metadata on skin colour or ethnicity; among 2,400 labelled images, just ten depicted brown skin and only one showed dark-brown skin. Such skewed datasets limit a model's diagnostic reliability for patients with darker complexions.

The Bias Cascade and Effects: Race bias in medical AI can introduce diagnostic errors that place certain racial groups at higher risk. These errors then cascade through treatment pathways, compounding disparities. Imagine a pain-management algorithm that recommends lower opioid doses for Black patients than for White patients, or a referral tool that—because it underestimates kidney function—delays specialist care and transplant listing for Black individuals. Such chains of decisions systematically limit equal access to care and effective disease management.

Race bias in medical AI can result in diagnosis errors that put individuals of certain races at high risk. The effect is then cascaded into the treatment patterns to the extent that these disparities are further amplified. Imagine an AI algorithm for pain management recommending a lower dosage of medication for black individuals compared to white individuals, or a referral algorithm based on a biased estimation of kidney functionality can cause a delay in the referrals for severe kidney disease or recommendation for transplant. Such a cascade put certain races at a disadvantage in terms of getting access to equal healthcare and efficient health management.

In summary, race bias in medical AI has the potential to aggravate existing healthcare inequities and creates structural disadvantages for marginalised populations. The impact extends beyond a single clinical choice: biased predictions about readmission risk or eligibility for high-cost therapies lead to less proactive care for some racial groups, reinforcing a harmful feedback loop. Efforts to counter race bias are under way. For example, the Agency for Healthcare Research and Quality (AHRQ) and the National Institute on Minority Health and Health Disparities (NIMHD), in collaboration with the Yale School of Medicine, have published five guiding principles for preventing algorithmic

bias (Backman, 2023). Other initiatives include algorithm-audit frameworks and fairness benchmarks (Norori et al., 2021), (Chin et al., 2023), although these have yet to be incorporated into regulations (Busch et al., 2025). A full review lies outside the scope of this text, but these examples show that practical remedies are emerging.

Hazrat Ali is a Lecturer in AI at University of Stirling. His extensive research portfolio spans various areas of Generative AI, Medical AI, healthcare, and computer vision, with a specific focus on generative models for medical imaging, deep learning for ultrasound medical imaging, and AI for healthcare. He has published more than 60 peers-reviewed papers.

Simon T. Powers is a Lecturer in Trustworthy Computer Systems at the University of Stirling. He obtained his PhD (2010) from the University of Southampton. He has more than 30 peer-reviewed publications.

Muhammad Bilal is a professor of Applied Artificial Intelligence (AI) and Technology Ethics at Birmingham City University (BCU), UK and the director of the Centre for Responsible Innovation in Big Data and AI (BRAIN) at BCU. He holds a PhD in AI from UWE, Bristol, and an MS in Advanced Database Systems. Professor Bilal's research focuses on harnessing cutting-edge AI innovations, such as Foundation Models and Generative AI (GenAI), to create clinical co-pilots, enhance multimodality fusion, and enable collective intelligence through human-computer collaboration. These efforts aim to augment healthcare workers' performance and improve patient safety.

References

Abbasi-Sureshjani, S., Raumanns, R., Michels, B. E., Schouten, G., & Cheplygina, V. (2020). Risk of training diagnostic algorithms on data with demographic bias. *Interpretable and Annotation-Efficient Learning for Medical Image Computing: Third International Workshop, iMIMIC 2020, Second International Workshop, MIL3ID 2020, and 5th International Workshop, LABELS 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 3*, 183–192. https://doi.org/10.1007/978-3-030-61166-8_20

Ali, H., Grönlund, C., & Shah, Z. (2023). Leveraging gans for data scarcity of covid-19: Beyond the hype. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 659–667.

Backman, I. (2023, December). Eliminating racial bias in health care ai: Expert panel offers guidelines [Accessed 05 June 2025].

Busch, F., Geis, R., Wang, Y.-C., Kather, J. N., Khori, N. A., Makowski, M. R., Kolawole, I. K., Truhn, D., Clements, W., Gilbert, S., et al. (2025). Ai regulation in healthcare around the world: What is the status quo? *medRxiv*, 2025–01.

Chin, M. H., Afsar-Manesh, N., Bierman, A. S., Chang, C., Colón-Rodríguez, C. J., Dullabh, P., Duran, D. G., Fair, M., Hernandez-Boussard, T., Hightower, M., et al. (2023). Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. *JAMA Network Open*, *6*(12), e2345050–e2345050.

King, A. P. (2023). A systematic study of race and sex bias in cnn-based cardiac mr segmentation. *Statistical Atlases and Computational Models of the Heart. Regular and CMRxMotion Challenge Papers: 13th International Workshop, STACOM 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Revised Selected Papers*, *13593*, 233.

Lee, T., Puyol-Antón, E., Ruijsink, B., Roujol, S., Barfoot, T., Ogbomo-Harmitt, S., Shi, M., & King, A. (2025). An investigation into the causes of race bias in artificial intelligence–based cine cardiac magnetic resonance segmentation. *European Heart Journal-Digital Health*, *6*(3), 350–358. https://doi.org/10.1093/ehjdh/ztaf008

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and ai for health care: A call for open science. *Patterns*, *2*(10).

Wen, D., Khan, S. M., Xu, A. J., Ibrahim, H., Smith, L., Caballero, J., Zepeda, L., de Blas Perez, C., Denniston, A. K., Liu, X., et al. (2022). Characteristics of publicly available skin cancer image datasets: A systematic review. *The Lancet Digital Health*, *4*(1), e64–e74.

Military AI Ethics and International Law: The Gaza Case

Montassar Ben Dhifallah and Ahmed Nebli

Synopsis. This chapter addresses the ethical dimensions of military AI by distinguishing offensive from defensive applications. It assesses international guidelines, particularly under the United Nations Convention on Certain Conventional Weapons, and evaluates the applicability of these frameworks through a case study of Israeli operations in Gaza.

Introduction The integration of artificial intelligence (AI) into military systems presents a series of regulatory and ethical challenges. While these technologies may offer gains in operational precision, situational awareness, and response efficiency, they simultaneously raise concerns about delegation of lethal decision-making, accountability, and compliance with international humanitarian law (IHL). This chapter addresses the ethical dimensions of military AI by separating offensive from defensive applications, and by examining current international positions, particularly within the United Nations framework. A case study of Israeli Defense Forces (IDF) operations in Gaza is used to assess the applicability and limitations of these guidelines.

International guidelines for military AI

Since 2014, the *Group of Governmental Experts* (GGE) under the United Nations Convention on Certain Conventional Weapons (CCW) has addressed the question of lethal autonomous weapons systems. Its main output to date is the set of eleven non-binding *Guiding Principles* adopted in 2019 (Convention on Certain Conventional Weapons, 2019). These restate three elements: (i) the applicability of existing international humanitarian law (IHL) to emerging technologies, (ii) the requirement for human responsibility in the use of force,

and (iii) the obligation for states to conduct legal reviews under *Article 36* of IHL during system development (International Committee of the Red Cross, 2021; United Nations Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, 2019). The absence of enforcement or verification measures leaves implementation to national discretion, thereby exposing the guidelines to selective interpretation and strategic framing.

AI in military: Usage and ethics Embedding AI into military infrastructure introduces ethical complexity, particularly regarding the delegation of lethal authority. The use of autonomous systems to identify and engage targets challenges established norms of human control in warfare and raises questions under humanitarian law (Brescia, 2024).

The permissibility of allowing software to make life-and-death decisions remains contested. Such delegation may contravene principles of distinction and humanity by removing human judgment from critical engagements.

Operational risks further complicate the ethical profile. AI systems are susceptible to misclassification, model failure, and data-driven bias. These can result in erroneous engagements, including the targeting of civilians due to flawed training data or ambiguous classification schemes.

Ethical distinctions also emerge between offensive and defensive deployments. Offensive systems, such as autonomous Unmanned Aerial Vehicles (UAVs) tasked with proactive strike missions, reduce exposure for deploying forces but may incentivize preemptive or escalatory use (Gayle, 2019). The reduction in political and human cost may lower the threshold for conflict initiation, with attendant risks of over-reliance on algorithmic assessments.

Defensive military AI systems are typically employed for force protection and threat mitigation. Use cases include missile interception, intrusion detection, and automated cybersecurity responses. These deployments are often framed as ethically preferable due to their reactive nature and potential to reduce harm—for instance, intercepting incoming projectiles with autonomous countermeasures. Proponents argue that AI-enabled defence systems can outperform human operators in speed and consistency, thereby limiting response lag and reducing casualties. Nonetheless, these systems introduce specific risks. Automated platforms may misclassify non-combatant entities or fail to interpret contextual subtleties, especially under conditions of data ambiguity or adversarial input.

Overall, military AI presents a structural dilemma: while it may offer efficiency gains in both offensive and defensive domains, it concurrently expands

Figure 4: Aerial view of destruction in Beach refugee camp, Gaza Strip, *by UNRWA Photo, Abedallah Alhaj* (3 July 2024)

the scope for accidental escalation, misidentification, and diffuse accountability.

Case Study: High-Throughput Machine Learning in IDF Targeting

In 2019, the Israel Defense Forces (IDF) established a *Target Administration Division* supported by AI-enabled decision-support systems to accelerate target identification (Mhajne, 2023). The architecture comprises two main inference services. The first, *Gospel* (*Habsora*), integrates imagery, signal intelligence, and behavioural analytics to generate candidate targets. During *Operation Guardian of the Walls* (May 2021), it produced approximately 100 target leads per day—representing a two-order-of-magnitude increase over prior annual rates (Davies et al., 2023; Israel Defense Forces, 2024). The second system, *Lavender*, applies a machine learning classifier to assign probabilistic threat scores to individuals. As of October 2023, its database included records on approximately 37,000 Palestinians and their associated locations (Mhajne, 2023).

Operational outputs were substantial. The IDF reported over 15,000 target strikes in the first month of the 2023 Israel–Gaza conflict (Davies et al., 2023). Independent estimates attribute roughly 5,000 civilian fatalities—including 1,900 children—during this period, constituting the highest monthly toll in any single-theatre aerial campaign recorded in recent decades (Davies et al., 2023). Subsequent investigations revealed that human operators often validated *Lavender* recommendations within approximately 20 seconds (Mhajne, 2023), raising questions around automation bias and confirmation proce-

dures. The system's internal logic, including feature selection and classification thresholds, remains undisclosed, limiting both auditability and legal scrutiny. Notably, the system's scope included civilian profiles, increasing exposure to false positives. This case illustrates a core issue in military AI: expanding the *find–fix* loop throughput does not inherently ensure adherence to distinction, proportionality, or precautionary standards under international humanitarian law (Davies et al., 2023). The RUSI analysis of AI-assisted targeting in Gaza concludes that the IDF prioritized speed over accuracy. Acceleration of targeting cycles, when not paired with proportional safeguards, increases the risk of civilian harm.(Sylvia, 2024)

In sum, the Gaza case highlights key ethical concerns: (i) large-scale data surveillance and algorithmic profiling, raising privacy and human rights issues; (ii) potential bias and error in AI models, exemplified by Lavender's reported 10% misidentification rate, which was considered operationally acceptable by IDF leadership(Sylvia, 2024); (iii) reduced human oversight due to the volume and velocity of AI outputs; and (iv) disproportionate harm resulting from rapid, high-volume strike capabilities.

Conclusion The current regulatory landscape for military AI remains underdefined. Existing instruments consist largely of non-binding principles with broad language and limited enforcement mechanisms. This vagueness leaves key provisions open to subjective interpretation, particularly concerning accountability and use-of-force thresholds. Given the strategic and humanitarian stakes involved, there is a pressing need for the international legal community to articulate and operationalize a dedicated framework for AI in military contexts.

Montassar Ben Dhifallah is a Tunisian deep learning engineer focused on computational neuroscience and medical imaging. A Master's student at the University of Sousse, he researches machine learning for MRI-based multiple sclerosis detection. He co-authored the paper introducing DemyeliNeXt, an explainable few-shot learning model for MS classification, and contributed to the BrainNet-ML Toolbox. He has presented at international venues, including MICCAI 2024.

Ahmed Nebli is a PhD student at the Heinrich Heine University in Düsseldorf, Germany. His research focuses on generative deep learning for both microscopic imaging and brain graphs.

References

Brescia, S. (2024). Navigating the ai battlefield: Opportunities and ethical frontiers in modern warfare. https://nrdc-ita.nato.int/newsroom/insights/navigating-the-ai-battlefield-opportunities–challenges–and-ethical-frontiers-in-modern-warfare.

Convention on Certain Conventional Weapons. (2019). *Guiding principles affirmed by the group of governmental experts on emerging technologies in the area of lethal autonomous weapons systems* (tech. rep. No. CCW/MSP/2019/9 Annex III). United Nations Office for Disarmament Affairs. Geneva. https://ccdcoe.org/uploads/2020/02/UN-191213_CCW-MSP-Final-report-Annex-III_Guiding-Principles-affirmed-by-GGE.pdf.

Davies, H., McKernan, B., & Sabbagh, D. (2023, December 1). *'the gospel': How israel uses ai to select bombing targets in gaza*. The Guardian. https://www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets.

Gayle, D. (2019). Uk, us and russia among those opposing killer robot ban. *The Guardian*. https://www.theguardian.com/science/2019/mar/29/uk-us-russia-opposing-killer-robot-ban-un-ai.

International Committee of the Red Cross. (2021). Icrc position on autonomous weapon systems [Accessed 20 June 2025]. https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems.

Israel Defense Forces. (2024, June 18). *The idf's use of data technologies in intelligence processing*. https://www.idf.il/en/mini-sites/idf-press-releases-israel-at-war/june-24-pr/the-idfs-use-of-data-technologies-in-intelligence-processing.

Mhajne, A. (2023). Gaza: Israel's ai human laboratory [Accessed: 2025-07-12]. *The Cairo Review of Global Affairs*. https://www.thecairoreview.com/essays/gaza-israels-ai-human-laboratory/.

Sylvia, N. (2024, July). The Israel Defense Forces' Use of AI in Gaza: A Case of Misplaced Purpose. https://www.rusi.org/explore-our-research/publications/commentary/israel-defense-forces-use-ai-gaza-case-misplaced-purpose.

United Nations Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems. (2019). *Report of the 2019 session of the group of governmental experts on emerging technologies in the area of lethal autonomous weapons systems* (tech. rep. No. CCW/GGE.1/2019/3). United Nations Office for Disarmament Affairs. Geneva. https://documents.unoda.org/wp-content/uploads/2020/09/CCW_GGE.1_2019_3_E.pdf.

The Human-AI infernal loop: How AI Ironically Threatens and Enhances Human Mental Health

Lotfi Ben Romdhane

Synopsis. Artificial Intelligence (GAI) stands as one of the most transformative technologies of the 21st century, reshaping industries, societies, and the human experience. One of the most profound effects is its paradoxical impact on mental health. On the one hand, AI contribute significantly to the rise of mental health challenges, through social isolation, digital addiction, and the erosion of privacy. On the other hand, AI is a valuable tool to understand, diagnose, and even treat mental health disorders. This paradox forms what can be described as the **Human-AI infernal loop**-a self-perpetuating cycle where the same technology both amplifies and attempts to solve the mental health crisis it helps create. In this essay, we will explore this infernal loop starting with how AI contributes to mental health problems, followed by an examination of its use as a therapeutic tool, and concluding with a reflection on whether it is possible to escape this paradox.

AI powers many of the most addictive platforms on the planet. Recommendation algorithms on platforms like TikTok, Instagram, and YouTube learn from users' behaviors to **maximize engagement**, often by serving content designed to exploit human psychological vulnerabilities. The outcome is increased screen time, sleep deprivation, and a shortened attention span—each a known factor in declining mental well-being. Numerous studies have shown a strong correlation between high levels of social media use and symptoms of anxiety, depression, and loneliness, especially among teenagers and young adults (Fadillah, 2025; Gerlich, 2025). The constant comparison to curated digital personas leads to unrealistic expectations and self-esteem issues. AI's role in this is central: it curates content, manipulates user experience, and reinforces

compulsive behaviors through personalized feedback loops. For example, a 2024 study conducted by the Danish digital safety organization Digitalt Ansvar revealed significant shortcomings in Instagram's content moderation ("Instagram actively helping spread of self-harm among teenagers, study finds", 2025). Researchers created fake profiles that posted 85 images related to self-harm over a month. Despite Meta's assertion that it proactively removes 99% of such content, none of these posts were taken down during the study period.

Another fact is the rise of AI surveillance in both public and private domains. The pervasive data collection—through connected devices, online behavior tracking, and biometric monitoring—has led to an erosion of personal privacy. For example, during the COVID-19 pandemic, When vast swathes of the workforce began working from home, monitoring by their employers grew. In the US alone, the number of medium-to-large employers using tools to track workers doubled in nearly two years from March 2020 to 60% and is expected to rise to 70% by 2025 ("The Right Way to Monitor Your Employee Productivity", 2022). However, knowing that one's actions, choices, and even facial expressions may be tracked and analyzed creates a pervasive sense of anxiety and loss of autonomy. This psychological pressure is further intensified in workplace settings where AI systems monitor productivity, flag deviations, and contribute to decisions about hiring or firing. Workers in such environments may feel constantly scrutinized, leading to stress, burnout, and a decline in job satisfaction. For example, a 2024 study by Alexandra Borgeaud, a scholar specializing in the governance of emerging technologies, reports that employees in the United States who were subjected to continuous productivity surveillance exhibited a significantly higher incidence of workplace anxiety ("Link between productivity monitoring and workplace anxiety in the U.S. 2024", 2025). Indeed, 53% of workers whose productivity was tracked all the time reported feeling anxious at work, compared to 41% of those without electronic monitoring.

A third threat is AI's growing capabilities especially with these high volumes of available training data and the processing power of GPUs. Consequently, the displacement of millions of jobs is more than a reality. This anticipation of job loss creates economic anxiety, a powerful stressor that impacts mental health. Moreover, the nature of jobs is having actually a radical transformation where Humans will be assisted continuously by AI agents to achieve even the hardest tasks. Therefore, the fear of failure and the continuous pressure to constantly upskill in this rapidly evolving job market can lead to chronic stress and existential dread.

A third threat is AI's growing capabilities especially with these high volumes of available training data and the processing power of GPUs. Con-

sequently, the displacement of millions of jobs is more than a reality. According to Ali et al. (2024), AI exposure correlates negatively with job security ($r = -0.65, p < .01$) and positively with stress ($r = 0.72$), anxiety ($r = 0.58$), and burnout ($r = 0.54$) in a cross-sectional survey of 300 employees (Ali et al., 2024). Xu et al. (2023) demonstrate that "AI awareness"—employees' perception that their role may be replaceable—predicts emotional exhaustion, mediated by job insecurity and family interference, using data from 303 participants (Xu et al., 2023). These findings underline the psychological toll of AI-driven task reorganization. In work by (Almida & Schmitt, 2025), German workplace surveys reveal that AI and robot exposure impact several task characteristics—such as performance pressure, prescribed routines, and reduced autonomy—which in turn influence stress outcomes (Almida & Schmitt, 2025). This supports conceptual frameworks like the Job Demands–Resources model: while AI augments task complexity (increasing job demands), the absence of supportive resources (e.g., managerial support or autonomy) escalates worker stress (Autor et al., 2024). (Xu et al., 2023) find that AI awareness is significantly associated with depressive symptoms, with emotional exhaustion partially mediating this relationship; moreover, perceived organizational support buffers the pathway from exhaustion to depression (Xu et al., 2023). Xu and colleagues conclude that improving organizational support (e.g., transparent communication, retraining, and flexible practices) is crucial to mitigating AI-related mental health risks. Evidence from administrative panel data in Taiwan (covering 16 years) indicates severe economic and psychological consequences of job displacement: earnings fall by approximately 67–68% in the first year, with persistent 60–61% losses over a decade. Importantly, displaced workers experience a 15–16% increase in mental-health outpatient visits and a 57–62% rise in mental-health–related medical costs (Chang & Lin, 2024). This long-term impact reinforces the importance of preemptive policies to prevent or alleviate displacement.

Paradoxically, AI as a diagnostic tool is showing great promise in the field of mental health. One of its most compelling applications is in *early diagnosis and monitoring*. Machine learning algorithms can process vast datasets from electronic health records, social media, or wearable devices to identify patterns associated with mental health deterioration. Consequently, AI can detect these signs earlier than human clinicians, offering the opportunity for early intervention, a crucial fact for effective treatment. Actually, AI-powered chatbots like Wysa ("Wysa", 2025) have emerged as accessible mental health companions. These bots provide Cognitive Behavioral Therapy (CBT)-based interactions, mood tracking, and emotional support. For individuals in remote areas or those who cannot afford traditional therapy, these tools can offer immediate, though limited, relief. These AI companions are available 24/7,

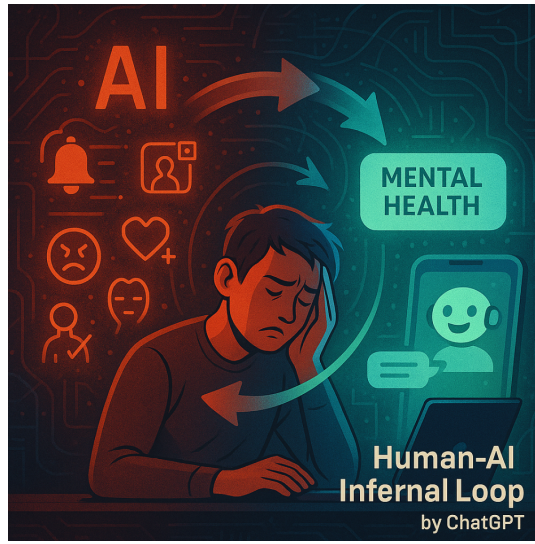


Figure 5: Human AI Infernal Loop, by ChatGPT.

are non-judgmental, and offer a sense of presence, which can be particularly valuable for those experiencing loneliness or isolation. Another major benefit of AI, thanks to the availability of Big Data (as sleep patterns, heart rate variability, and social media use), is its capacity to adapt to patients profiles and characters making therapies more precise and responsive.

This paradoxal Human-AI Infernal Loop is a reality and we can escape it by changing how AI-based technology is built. The core issue is that most algorithms today are designed to maximize user engagement. They learn to promote content that keeps people scrolling—often by exploiting fear, envy, or outrage. This design deepens psychological distress, especially among younger users. But the technology could be reoriented. Algorithms can be trained to prioritize mental well-being instead of attention. They could promote supportive, healthy content, recognize harmful patterns like excessive doomscrolling, and gently interrupt or redirect behavior to protect users. Beyond the algorithm itself, the interface matters. Endless feeds and autoplay features are engineered for addiction. If we redesign platforms to introduce moments of reflection—like slowing down scrolling or letting users see and adjust how their feed is shaped—we return some control to the user. Transparency and friction could become tools for emotional agency rather than manipulation. AI holds considerable promise for augmenting mental healthcare by detecting early signs of psychological distress, delivering personalized self-help interventions,

and supporting clinicians through advanced data analytics. AI-driven systems can, for instance, analyze linguistic cues in text or speech to infer mood states, flagging early indicators of anxiety, depression, or suicidal ideation. Moreover, AI can aid therapists by identifying behavioral patterns across sessions, recommending tailored therapeutic pathways, or managing administrative and diagnostic workflows.

However, the integration of AI into such a sensitive domain demands a rigorous ethical framework. It is essential that AI systems operate with full transparency regarding their capabilities and limitations. Unlike human therapists, AI lacks genuine empathy, and attempts to simulate human relationships or emotional understanding may mislead users into forming inappropriate attachments or false expectations. Therefore, AI should serve as a *supportive* tool rather than a substitute for human care. Key ethical considerations include safeguarding user privacy and data security, ensuring algorithmic fairness across diverse populations, and obtaining informed consent, particularly when AI systems are used for passive data collection (e.g., monitoring social media activity or wearable sensor data). Additionally, developers must ensure that AI does not exacerbate existing mental health disparities—for example, by being less accurate in underrepresented populations or inaccessible to individuals with limited digital literacy. Ultimately, the responsible deployment of AI in mental healthcare requires a multidisciplinary approach that involves ethicists, clinicians, patients, and technologists. Establishing clear guidelines, auditing algorithms for bias, and embedding accountability mechanisms are necessary steps to ensure that these technologies empower rather than harm the individuals they are designed to serve.

In summary, escaping the **Human-AI infernal loop** doesn't mean rejecting AI. It means building it with different values: *empathy over engagement, autonomy over addiction, and care over control*. With thoughtful design, AI can evolve from a source of harm into an ally for mental resilience.

Lotfi Ben Romdhane is a Professor in AI at the MARS Research Laboratory, specializing in Artificial Intelligence. He has published more than 100 articles in peer-reviewed journals and top-ranked conferences in the field of AI.

References

Ali, T., Hussain, I., Hassan, S., & Anwer, S. (2024). Examine how the rise of ai and automation affects job security, stress levels, and mental health in the workplace

[Open Access]. *Bulletin of Business and Economics*, *13*(1), 24–38. https://bbejournal.com/index.php/BBE/article/view/499.

Almida, P., & Schmitt, M. (2025). Artificial intelligence and worker stress: Evidence from germany. *Digital Society*, *2*(1), 15–36. https://doi.org/10.1007/s44206-024-00029-6

Autor, D., Acemoglu, D., & Restrepo, P. (2024). The impact of ai and information technologies on worker stress [Available at: https://cepr.org/voxeu/columns/impact-ai-and-information-technologies-worker-stress].

Chang, K., & Lin, H. (2024). Long-term effects of job displacement on earnings and mental health: Evidence from taiwan. *Economics Letters*, *238*, 111505. https://doi.org/10.1016/j.econlet.2024.111505

Fadillah, D. (2025). The need for research on ai-driven social media and adolescent mental health. *Asian Journal of Psychiatry*, *108*, 104513. https://doi.org/10.1016/j.ajp.2025.104513

Gerlich, M. (2025). Ai tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, *15*(1). https://doi.org/10.3390/soc15010006

Instagram actively helping spread of self-harm among teenagers, study finds [Accessed: 2025-07-22]. (2025).

Link between productivity monitoring and workplace anxiety in the u.s. 2024 [Accessed: 2025-07-22]. (2025).

The right way to monitor your employee productivity [Accessed: 2025-07-22]. (2022).

Wysa [Accessed: 2025-07-22]. (2025).

Xu, G., Xue, M., & Zhao, J. (2023). The association between artificial intelligence awareness and employee depression: The mediating role of emotional exhaustion and the moderating role of perceived organizational support. *International Journal of Environmental Research and Public Health*, *20*(6), 5147. https://doi.org/10.3390/ijerph20065147

The Creative Cost of Convenience: AI and the Erosion of Human Imagination

Muzammil Behzad

Synopsis. Artificial Intelligence (AI) is transforming how humans create. From composing music to generating art and writing style, AI offers unprecedented tools for expression. However, this growing reliance on machine creativity raises concerns about the long-term effects on human imagination and originality. This essay examines how AI impacts our cognitive and imaginative capacities by drawing on psychological insights, philosophical thought, and current technological debates.

The rise of generative AI tools has made creativity more accessible, but it also prompts a deeper question: what happens to human creativity when machines can do it for us? Creative work has traditionally been a process of struggle, reflection, and discovery. Relying on algorithms to bypass that process may gradually weaken the very pillars that define human innovation.

The idea that tools shape cognition is not new. As (Carr, 2010) questioned in *The Shallows*, *are we sacrificing our ability to read and think deeply?* This question argues that the technological shortcuts often change not just what we do but how we think. AI extends this concern by engaging directly with imagination; an area previously thought uniquely human. When AI can generate visual art in seconds or write fiction from a single prompt, the user risks becoming a passive consumer rather than an active creator.

Takeaway: *Overreliance on AI may dull human creativity:* As AI handles more cognitive tasks, the human capacity for original thought and imagination risks decline.

Figure 6: As AI takes on more creative tasks, the risk grows that human imagination and originality may quietly fade into the background. [Image generation source: OpenAI Dall-E2]

In dystopian literature, the fear of losing individual creativity and storytelling is central. Just as Nick Bostrom (Bostrom, 2014) warned of machines that override human will, today's critics raise alarms about AI reducing our role to that of prompt designers or content editors. Creativity, once a deeply human endeavor shaped by emotion, intuition, and imperfection, risks becoming a mechanical selection process. Philosopher Renee Richardson warned that outsourcing thought to machines could result in a decline in individual intellectual autonomy.

> "The danger isn't that AI creates, but that we stop doing so." (Turkle, 2017)

If imagination functions like a muscle, as many cognitive scientists propose, then relying too heavily on AI could lead to its underdevelopment. This raises important concerns for education, innovation, and even personal identity.

Moreover, AI systems are inherently derivative as they recombine existing patterns rather than generate truly original thought. Over time, this could lead to creative homogenization. Just as algorithm-driven platforms promote formulaic trends, AI-generated content might reinforce stylistic conformity. The result could be a cultural feedback loop where new ideas become increasingly rare.

Still, some argue that AI can expand creativity by acting as a collaborator or catalyst. A well-crafted prompt might unlock unexpected results, encouraging users to explore ideas they hadn't previously considered. From this perspective, AI doesn't diminish imagination but redirects it towards experimentation and hybridization.

Yet this optimistic view relies on humans maintaining an active role. The challenge is not AI itself, but the *overreliance* on it. Just as calculators didn't destroy numeracy but changed how we teach math, AI could evolve creativity if balanced with critical thinking, reflection, and genuine engagement with the creative process.

Ultimately, the future of imagination in the age of AI will depend on how we choose to use these tools. Will we hand over the act of creation to algorithms, or will we redefine creativity as a richer interplay between human and machine?

Dr. Muzammil Behzad is an innovator, AI consultant, and leading expert in machine learning, deep learning, computer vision, and vision-language models. He is the founder and director of the BRAIN Lab, an independent research initiative focused on advancing the frontiers of artificial intelligence. He currently serves as an Assistant Professor of Artificial Intelligence at King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia, where he is also affiliated with the SDAIA-KFUPM Joint Research Center for Artificial Intelligence. Dr. Behzad received his Ph.D. from the University of Oulu, Finland, where he was part of the Center for Machine Vision and Signal Analysis (CMVS). As a seasoned AI consultant, Dr. Behzad has played a key role in bridging academic innovation with industry applications. He is a long-standing member of several prestigious organizations, including the European Association for Artificial Intelligence (EurAI), Finnish Artificial Intelligence Society (FAIS), IEEE, European Telecommunications Standards Institute (ETSI), Society for Industrial and Applied Mathematics (SIAM), British Machine Vision Association (BMVA), and the European Alliance for Innovation (EAI).

References

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Carr, N. (2010). *The shallows: What the internet is doing to our brains*. W. W. Norton & Company.

Turkle, S. (2017). *Alone together: Why we expect more from technology and less from each other* (Reprint edition). Basic Books.

Ctrl+C, Ctrl+LLM: Convenience, Cognition, and the Crisis in Student Learning

Omar Choudhry

Synopsis. Generative Artificial Intelligence (GenAI) and Large Language Models (LLMs) have are becoming more deeply ingrained in society, leading to an unprecedented shift in productivity, creativity, reliance, responsibility, and more. Often, narratives are presented using statistics; however, this essay presents a firsthand account that reflects the experiences of many others in the context of higher education.

As a computer scientist, I often sought tools to aid me in programming and coding, aiming to improve my productivity. *GitHub Copilot* was the first, more "intelligent", code completion tool I used. Over time, I began writing fewer lines of code and more comments, after which the tool would "fill in the blanks" and complete the rest of the code for me. Months later, the research preview of *ChatGPT-3* was released - a snowball which would inevitably roll into an avalanche.

Definition: Comments (noun). Text embedded in source code intended as an annotation to make the code easier to understand – often explaining an aspect not apparent directly within the non-commented program code. (Grubb & Takang, 2003).

I have been a teaching assistant (TA) to undergraduate and postgraduate students across the School of Computer Science for almost three years. I spend time in labs assisting students who need help with course materials, assignments, exam preparation, etc. During the early years, students heavily relied on TAs, with a broad range of questions that helped them in their academic and personal lives. However, over time, I noticed more students

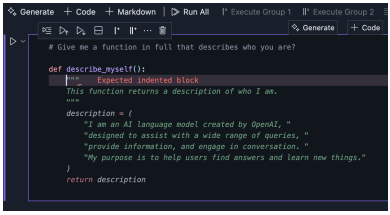


Figure 7: Incredible productivity gains using AI code completion tools (*GitHub Copilot* in this figure with a completely auto-generated function) using a single line of instruction.

using *ChatGPT*, especially as better and more accurate versions were released. Eventually, and to this day, questions have centred solely on IT problems and only occasionally involve assignment clarifications. It was common to previously spend hours weekly conversing with students. Now? *"So-and-so app is not working on my computer, can you please help?"*. This is the reality many are facing across our university and other institutions worldwide.

Of course, students are not only limiting GenAI use to simply asking clarifying questions, but have also been widely using it to complete summative assignments and even cheat in interviews. This has led to unprecedented consequences. Firstly, I have noticed various changes when marking assignments. Before, it seemed that there was a wide range of English proficiency, styles in presenting and structuring reports and code, varying ideas and submission strengths. On the other hand, I now see consistent patterns within submissions that resemble typical LLM characteristics and habits, a reduced number of poor and weak submissions, similar or identical ideas in many pieces of work, and fewer overall impressive and innovative ideas. I find this to be quite a shame, recalling the lengthy times I used to spend writing and thinking about coursework, including structuring reports, compared to the shortcuts students take now.

My concern was always that students would become overly reliant on GenAI, resulting in reduced creativity, even though their productivity may increase. Secondly, I observed that when I asked students to explain their decisions in something they had written, whether it was code or a report, they were unable to provide an adequate explanation of their reasoning. Many students were also completely unable to explain the functionality of the code they had written. As an analogy, I have noticed that many students who often looked down upon the subject of mathematics, including those who would often say *"What is the point? We will always have a calculator (our phone) in our pocket!"*. They will often find it challenging to perform basic to intermediate

arithmetic on the spot, requiring them to use their smartphone. The point is that while you may always have a calculator and possibly always have access to an LLM, you will lose or not develop the ability you would have otherwise had.

Although this was based on my own experiences (even though it correlates with reports from other academics), we now have evidence of these consequences. A recent study from MIT (Kosmyna et al., 2025) has established evidence that long-term usage of LLMs results in weaker connectivity within the brain's distributed networks, reduced cognitive activity, a weakened ability to recall work (even one's own!), and under-engagement. On the contrary, using LLMs as tools to assist after having already developed skills and learnt foundational knowledge exhibits higher memory recall and brain activation. The study concluded that LLMs offer immediate convenience at the expense of potential cognitive costs at the neural, linguistic, and behavioural levels.

But why? Students **know** when using GenAI would be breaking the rules, yet they cheat anyway. It has been made aware to many that there are severe potential long-term implications of over-reliance on LLMs, as we do not fully understand their effects. Maybe it is a culmination of the convenience, combined with the ability to do well (since some students may just be considering completing the degree for the certificate on paper rather than the actual education and curiosity), oft-reported depression, mental health issues and in general difficulties in time-management, workload, etc. Students face so many problems and issues - why would you ***not*** expect them to be taking such a shortcut?

Nevertheless, there is always a bright side. I have noticed responsible use where students are engaging in much more self-directed learning, and do not use it to produce and copy-paste LLM outputs. International students would often describe to me the privilege of having staff at higher education institutes in the United Kingdom who are much more approachable and assist students in learning, showing genuine care about whether they understand the content, something they often lack in their home countries due to various factors, beyond the scope of this essay. On the contrary, now that students can use GenAI to assist in learning and completing work, it's resulting in education itself becoming more democratised. Students can write in better English than they ever could, and translate documents and websites with ease. I have firsthand witnessed the incredible live recording and translation of lecturers as they speak, side by side with the PowerPoint slides, to understand the content in their own language. Time is saved in *all* areas, including finding sources, debugging, producing illustrations, reviewing work, and much more.

The potential benefits of AI on society within education are innumerable,

but its use must be accompanied by responsibility. I may use *GitHub Copilot* and other GenAI tools to help write my code, but I still practice coding almost every single day without any AI tools to make sure I maintain those neural pathways and do not lose the skill I spent years developing out of complacency. In the future, GenAI will be able to personalise learning plans, effectively providing each student with exactly the content in the appropriate style and structure they need at every step, maximising the amount they can learn effectively. However, we must not become complacent in the meantime by allowing things to get out of hand, as there *are* substantiated fears regarding what will happen if there is no reform, change in policy, regulation, or rules at these institutions or within the government, if students use these tools unrestrictedly. We cannot allow a world where assignments are written by AI, submitted using AI and marked using AI.

Omar Choudhry is a PhD candidate at the University of Leeds, focusing on improving the training of laparoscopic surgeons in low- and middle-income countries using AI. His background was an MSc in Artificial Intelligence for Medical Diagnosis and Care and a BSc in Computer Science with Artificial Intelligence. He has various roles in the university, including President of the AI Society, Education Outreach Fellow, Postgraduate Representative, Graduate Teaching Assistant, Lead Outreach Coordinator and Journal Club Organiser. He is an Associate Fellow of Advance Higher Education, Chief AI Scientist at an AI Edutech startup and a former lecturer at King Saud University in Saudi Arabia.

References

Grubb, P., & Takang, A. (2003). *Software maintenance: Concepts and practice* [Key discussions on pp. 7 and 120–121]. World Scientific.

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. https://arxiv.org/abs/2506.08872.

Toward Efficient Vision Language Alignment for Academic and Resource-Limited Settings

Hasnae Zerouaoui

Synopsis. This essay explores how academic researchers in under-resourced settings can achieve competitive AI research, particularly in vision-language alignment, despite limited access to large-scale hardware.

In this essay, I would like to begin by asking an important question: Can academic researchers in regions with limited resources achieve state-of-the-art results using standard hardware? This is a question that arises frequently within the artificial intelligence research community. It reflects a broader concern about the capacity of academic teams, often working with constrained computational budgets, to make meaningful contributions in a field increasingly dominated by large-scale industrial laboratories such as Meta and OpenAI. The issue is particularly pressing in research problems like vision language alignment, where the most successful models are developed and trained on enormous datasets using substantial computing infrastructure. For instance, CLIP (Radford et al., 2021), a prominent vision language model, was trained on 400 million image text pairs using 592 V100 GPUs over 18 days with a ResNet-50 vision encoder, and 256 V100 GPUs over 12 days with a ViT-B/32 encoder (Radford et al., 2021).

As results, the artificial intelligence landscape today is largely shaped by a small number of industrial laboratories with access to massive computational and data resources. These organizations have set new performance standards by developing models that achieve remarkable results across tasks including image retrieval, classification, captioning, and cross-modal reasoning. While these achievements are impressive, they are made possible by hardware clusters

far beyond the reach of most academic teams. As a result, the gap between academic and industrial artificial intelligence capabilities continues to widen.

This naturally leads to another critical question: What exactly constitutes academic hardware? Academic hardware, as defined in recent studies (Khandelwal et al., 2024), generally refers to computing configurations that involve between one and four A100 GPUs and enable the pretraining of models with up to seven billion parameters (Pythia 7B) (Merullo et al., 2022). The feasibility of using such configurations depends on careful training strategies and algorithmic optimizations. While these resources permit the training and fine-tuning of moderately sized models, they fall short of supporting the large-scale architectures common in industrial settings. Moreover, these studies have largely overlooked the realities faced by researchers in Africa, where computational resources are often even more limited. This situation leaves us with two options: either to step back from advancing artificial intelligence research in such regions, or to actively seek and design more efficient, resource-friendly solutions capable of delivering meaningful results within the constraints of available infrastructure.

This reality raises important considerations about the goals and priorities of academic research. Rather than aiming to surpass or directly compete with industrial models, academic efforts can and should focus on designing methods that are computationally efficient, reproducible, and accessible. The question, therefore, is not merely whether academic teams can match the performance of large industrial models, but whether they can propose alternative solutions that achieve strong results while remaining feasible within standard hardware and resource constraints.

Heavyweight Versus Lightweight Vision Language Alignment

Vision language alignment is a fundamental problem in multimodal learning, where the goal is to learn joint representations that connect images and text. Approaches to this problem can be broadly divided into heavyweight and lightweight strategies, each with distinct requirements and trade-offs.

Heavyweight Approaches

Heavyweight alignment strategies rely on the joint training of large vision and language encoders using massive paired datasets. A prominent example is CLIP (Radford et al., 2021), which was trained using contrastive learning on 400 million image text pairs. The model optimizes its encoders so that semantically matching images and texts are projected into similar regions of a shared embedding space. This design enables strong performance across a variety of tasks, including zero-shot classification and retrieval. The success of CLIP comes at the cost of significant computational and data resources.

Its training requires high-end GPUs, large-scale paired data, and substantial energy consumption. Furthermore, CLIP assumes that images and captions are paired in a one-to-one fashion, an assumption that is often unrealistic in specialized domains where annotations may be sparse, ambiguous, or shared across multiple samples. Variants such as Locked Image Tuning (Merullo et al., 2022) reduce training cost by freezing the vision encoder and training only the text encoder on paired data. While this reduces compute requirements, the method still depends on large-scale supervision and inherits the limitations of contrastive learning, including its sensitivity to false negatives.

Lightweight Approaches

Lightweight alignment methods seek to address these limitations by avoiding joint training and instead reusing pretrained vision and language encoders (Maniparambil et al., 2024; Norelli et al., 2023). These methods are designed to minimize computational requirements while enabling alignment through minimal additional components or by exploiting the structure of the pretrained embeddings.

LiMBeR (Maniparambil et al., 2024)for instance exemplifies this approach by introducing a linear projection that maps vision encoder outputs into the input space of a language model. Both encoders are kept frozen, and only the projection layer is trained. This approach achieves reasonable performance when the vision encoder has been pretrained with language supervision. However, it still relies on paired data to learn the projection and struggles when applied to vision encoders without language supervision.

Other lightweight strategies avoid additional training entirely by aligning encoders based on structural similarity. Techniques that use measures such as Centered Kernel Alignment (Norelli et al., 2023) compare and match the internal structures of pretrained encoders through graph matching. While these methods show promise in retrieval and classification tasks, their reliance on full pairwise similarity computations across datasets introduces scalability challenges.

Further progress in lightweight alignment is represented by Anchored Semantic Information Fusion (Norelli et al., 2023), which constructs a shared latent space using similarities to a small set of anchor pairs. This method enables zero-shot classification without training or modification of the encoders. Although effective and data-efficient, it requires storing and retrieving embeddings at inference time, increasing memory and compute costs, and is sensitive to the composition of the anchor set.

Opportunities for Academic Research

The shift from heavyweight joint training to lightweight, structure-aware methods reflects a recognition that pretrained unimodal encoders often contain compatible semantic structures. These structures can be aligned with minimal supervision, offering a path to efficient and accessible multimodal learning.

However, existing lightweight approaches present their own challenges. LiMBeR's dependence on paired supervision, the computational cost of structural alignment techniques, and the memory requirements of anchor-based methods each point to opportunities for further innovation.

These considerations underscore the importance of advancing alignment strategies that balance simplicity, efficiency, and informed use of representation structures. Combining parametric approaches, such as linear projections, with non-parametric techniques that draw on geometric and topological properties offers a promising direction. Such methods have the potential to achieve strong performance while remaining computationally feasible for researchers working with standard hardware. More broadly, they contribute to the collective effort to ensure that artificial intelligence research and technologies are inclusive, equitable, and accessible to the global academic community.

Hasnae Zerouaoui is a researcher at the College of Computing, Mohammed VI Polytechnic University. Her work focuses on efficient deep learning models, vision-language alignment, and democratizing access to AI in low-resource environments.

References

Khandelwal, A., Yun, T., Nayak, N. V., et al. (2024). $100K or 100 Days$: Trade-offs when pre-training with academic resources. *arXiv preprint arXiv:2410.23261*.

Maniparambil, M., Akshulakov, R., Djilali, Y. A. D., Seddik, M. E. A., Narayan, S., Mangalam, K., & O'Connor, N. E. (2024). Do vision and language encoders represent the world similarly? *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14334–14343.

Merullo, J., Castricato, L., Eickhoff, C., & Pavlick, E. (2022). Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*.

Norelli, A., Fumero, M., Maiorca, V., Moschella, L., Rodola, E., & Locatello, F. (2023). ASIF: Coupled data turns unimodal models to multimodal without training. *Advances in Neural Information Processing Systems*, *36*, 15303–15319.

Radford, A., Kim, J. W., Hallacy, J., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 8748–8763.

What will be the energy and water footprint of AI Data Centers?

Isaac Bua and Kamil Aliyev

Synopsis. Artificial Intelligence (AI) and machine learning are redefining the Fourth Industrial Revolution, with generative models like ChatGPT now central to everyday life, industry, and global innovation. But behind this exponential growth lies a massive infrastructure of energy-hungry data centers. These facilities—often drawing power equal to small cities—rely on advanced chips and liquid cooling systems, consuming vast amounts of electricity and water to support AI workloads. As companies invest hundreds of billions to accelerate progress toward Artificial General Intelligence (AGI), concerns over sustainability, emissions, and water use are growing. This review explores the unseen environmental cost of AI: from the energy powering a single prompt to the global competition for fresh water and clean power. With data centers poised to double their energy demand by 2030, understanding and managing AI's carbon and water footprint is crucial to ensure innovation doesn't derail climate goals.

Artificial intelligence and machine learning is one of the fastest growing fields in the Fourth Industrial Revolution or 4IR because of the recent advances in Information Technology and stands as the third field with fastest growing job market according to the Worlds Economic Forum Future of Jobs report 2025. AI technology has not only had significant discoveries but has found a place in the lives of many internet users, industries, academia and beyond due to its broad applications. The release of generative pretrained (GPT) models capable of performing near human task such as text generation and summarization, image, music, video generation, human language translation is so far the greatest achievements ("What Is Generative AI Generative AI Basics", 2025). In late 2022, ChatGPT a large language model launched as

the first conversational model allowing users to send and receive generated responses, it reportedly garnered over a million users within the first 5 days ("ChatGPT: What Is It & How Can You Use It?", 2025), now has over 400 million weekly users anticipated to reach 1 billion users by the end of 2025 ("Number of ChatGPT Users (March 2025)", 2025).

Hundreds of millions of people now use large language models (LLMs) like ChatGPT to draft their research, write emails, do coding task, homework's, plan travels, and even generate creative images or videos from text. Big tech giants like Google, Meta, Microsoft, X formally Twitter, could not wait but also started building and releasing different versions of their own LLMs accelerating competition in the field. It didn't just stop at company level; the competition would even lead to regional rivalry and a huge shift of focus to building the most powerful AI hence the genesis of the term Artificial General Intelligence – a theoretical AI with human capabilities, many researchers still believe that today's gen AI is if not decades then centuries away from AGI("What is Artificial General Intelligence (AGI)?", 2025). But as part of a machine there must be something powering it; inform of electricity.

Data centers, you have probably come across this term. Data centers also sometimes referred to as the warehouse of IT infrastructures are now rapidly growing in numbers, size and complexity globally to help meet the growing demand of robust AI information storage and processing (Kant, 2009). As a result, data centers now require and consume significant amount of energy to power expensive chips specifically designed for AI workload. All the power consumed is converted into heat, about 30-40% of the energy in data centers is required in cooling the IT equipment to maintain them within the safe operating chip junction temperatures, for optimal performance(X. Z. et al., 2025).

Liquid cooling with usually fresh water mixed with other synthetic thermal fluids have become one of the most widely used approaches as opposed to air cooling. Direct to chip and immersion techniques are widely used as they allow the coolant to be circulated around chips with high heat density (Kisitu et al., 2023). As different companies and regions race towards achieving AGI, the environmental impact of this technology is of a great concern and warrant careful analysis to ensure a safe integration without jeopardizing the set climate goals towards net-zero by 2050 and if we want to limit global temperatures below 1.5 celcius in accordance with the IPCC mandate.

The carbon footprint of AI is divided into two: embodied and operational emission. Embodied emissions come from the manufacture of IT equipment and constructing data centers while operational emission comes from the electricity consumed in AI calculations and cooling. Water footprint of data

centers describes the volume of water used in cooling. We present the review of different analysis, projections and the overall impact of AI energy use and water demand for cooling. However, this section does not cover embodied emission. There are three important questions that we wanted to answer. First 'what is the net climate impact of AI?' To answer this, we needed to answer the second question of 'how much does a single query cost in terms of energy and water?' After a comprehensive search, there isn't any information of this kind indicating a huge lack of transparency. Most of the findings available are based on theoretical estimates which lack consistency and may not cover all the channels where energy or water. The last question 'where will the energy projections and water come from?' which greatly determines where the data centers are sited – to avoid data centers from competing with communities for electricity and fresh water. This technology has the potential to reshape how our grid operates, from the type of energy to the cost we pay for electricity. Besides all this concerns, AI is still viewed by many as a potential climate solution.

Meta, Amazon, Alphabet and Microsoft intend to invest over $320 billion this year, while allocating a huge share to fire up new power plants to fuel new data centers with nuclear energy source as the favorite option ("Tech megacaps to spend more than $300 billion in 2025 to win in AI", 2025). Project stargate an alliance between OpenAI, Oracle, Softbank and MGX has begun in Texas with a promise to spend up to $500 billion in four years and is by far the largest AI infrastructures investment with each data center drawing about 100 megawatts of electricity ("Meet The Tiny Startup Building Stargate, OpenAI's $500 Billion Data Center Moonshot", 2025). Many critics still argue whether this investment is worthy as leaders like Sam Altman, OpenAI's CEO describes it "the most important project of this era". Due to the immediate demand, renewables sources of energy might not be the solution looking at the time it takes to build and their intermittent nature. AI data centers are normally tier III or IV with high uptime with not more than 1.6 hours of downtime per year ("CoreSite's Guide to Understanding Data Center Tiers", 2025).

Energy and Computing in General

Historically, computing was predominantly done on Central Processing Unit (CPU) servers, which required less power to transport and store data. With the evolution of deep learning, CPU serial processing became slow and could not manage the intensive calculations required, hence heating up very quickly. AI researchers started exploring different strategies, such as the use of Graphics Processing Units (GPUs), which became a significant discovery in the industry. GPUs were primarily used in video games and other 3D applications. Nvidia took the lead by developing CUDA and other variants, which allowed

computer scientists to seamlessly code and train machine learning models. In 2012, the AlexNet model, trained on CUDA, became the first winner in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) AI competition, accelerating the demand and need for GPUs in computing (Krizhevsky et al., 2017). GPUs found great application even in text processing due to their parallel processing capability. Today, Nvidia's Blackwell is among the most powerful architectures and has been shipped by several AI companies to their data centers.

AI models consume energy during both training and inference. However, significant amounts of energy are spent during inference, depending on the type and number of prompts given to the model. ChatGPT reportedly processes up to 1 billion prompts a day. It is estimated that it costs 10 times more energy to use ChatGPT compared to a Google search. Originally, data center racks would use 2–5 kW, but AI data center racks now consume up to 300 kW, a figure still expected to increase (Katal et al., 2023).

Data centers are increasingly significant consumers of global electricity, raising concerns about their environmental impact. According to the 2024 report from the Lawrence Berkeley National Laboratory (LBNL), U.S. data centers consumed approximately 4.4% of the total U.S. electricity in 2023, a figure projected to nearly triple to between 6.7% and 12% by 2028 if current AI trends continue (A. S. et al., 2024). Globally, data centers are set to more than double their electricity demand by 2030, consuming as much electricity as the whole of Japan does today, with AI being the most significant driver of this increase. The International Energy Agency (IEA) projects that electricity demand from AI-optimized data centers alone will more than quadruple by 2030 ("AI is set to drive surging electricity demand from data centres", 2025).

This increase in energy consumption directly correlates with an increased environmental impact. Fossil fuel-based energy sources, such as coal and petroleum, have significantly high carbon content and generate substantial amounts of greenhouse gases. The type of energy used by data centers will therefore greatly determine their carbon footprint. More research and proactive measures are needed to ensure that data centers are increasingly powered by low-carbon energy sources if countries aim to achieve their proposed grid net-zero goals. Renewable sources such as solar and wind, while intermittent, can offer a robust solution when supplemented with battery storage. Since renewables are also location-specific, depending on weather patterns and resource availability, strategically siting data centers in regions with high renewable energy potential can be one effective way to contribute to the decarbonization journey ("Clean Energy Resources to Meet Data Center Electricity Demand", 2025).

It is important to acknowledge that there is no single, simple formula or universally accepted value that precisely describes the energy consumed by a single AI prompt. Tools such as life cycle analysis (LCA) are crucial for a comprehensive, whole-system analysis of the different subsets of energy expenditure schneider2025, alissa2025. This includes not only the operational energy spent during AI model training and inference but also the embodied energy spent in the manufacture of the chips, servers, and cooling infrastructure. A holistic systems analysis that accounts for both embodied and operational energy can help in accurately quantifying the energy consumption of AI. Such detailed information is vital not only for planning new data centers but also for providing critical insights when deciding on their optimal location to minimize environmental impact (H. A. et al., 2025).

Water Use Projection

Beyond their significant energy demands, data centers, particularly hyperscale facilities, also have a substantial impact on water resources. Water, especially fresh water, has found great application in solving cooling challenges due to its superior thermal carrying capacity compared to air. Another advantage is that water is generally easier to pump or move than air, which requires a lot of energy for circulation. While many data centers implement both air and water-based cooling technologies, each requires different infrastructure setups ("High-Performance Computing Data Center Cooling System Energy Efficiency", 2025). The scale of water consumption by these facilities can be staggering. A single, large hyperscale data center can consume millions of gallons of water per day, with estimates between 3 to 5 million gallons daily, which is roughly equivalent to the daily water needs of a community of 10,000 to 50,000 people ("Data centers draining resources in water-stressed communities", 2025). This water is often used in evaporative cooling systems, where a portion of the water is lost to evaporation as it dissipates heat.

The cooling fluids used are not always just pure water as also seen in natural gas processing (R. A. H. et al., 2025). In most systems, water is mixed with other fluids, such as glycol or other specialized coolants, to alter boiling points and improve heat transfer ("Environmental Sustainability Considerations for Data Center Managers", 2025).

Isaac Bua is a graduate student in Sustainable Engineering at Villanova University, originally from Uganda. He previously worked as a Sustainability Engineer with Sustainable Greener World and is a recipient of a fully funded graduate fellowship. At Villanova, he serves as the International Student Affairs Officer in the Graduate Student Senate and focuses his research on environmental sustainability and infrastructure solutions.

Kamil Aliyev is currently pursuing a master's degree in Sustainable Engineering at Villanova University. He is Originally from Azerbaijan, he completed his undergraduate studies in Petroleum Engineering. His research interests include energy efficiency and generation technologies. He previously served as a Renewable Energy Team Officer at the Society of Petroleum Engineers. Kamil now represents Villanova in the Menus of Change University Research Collaborative (MCURC), where he focuses on reducing the environmental impact of campus food systems and advancing sustainable infrastructure solutions.

References

Ai is set to drive surging electricity demand from data centres [Accessed: Jun. 18, 2025]. (2025). https://www.iea.org/news/ai-is-set-to-drive-surging-electricity-demand-from-data-centres-while-offering-the-potential-to-transform-how-the-energy-sector-works.

Chatgpt: What is it & how can you use it? [Accessed: Jun. 03, 2025]. (2025). https://www.searchengine-journal.com/what-is-chatgpt/473664/.

Clean energy resources to meet data center electricity demand [Accessed: Jun. 18, 2025]. (2025). https://www.energy.gov/gdo/clean-energy-resources-meet-data-center-electricity-demand.

Coresite's guide to understanding data center tiers [Accessed: Jun. 04, 2025]. (2025). https://www.coresite.com/blog/breaking-down-data-center-tiers-classifications.

Data centers draining resources in water-stressed communities [Accessed: Jun. 18, 2025]. (2025). https://utulsa.edu/news/data-centers-draining-resources-in-water-stressed-communities/.

Environmental sustainability considerations for data center managers [Accessed: Jun. 18, 2025]. (2025). https://www.nvent.com/mission-critical/environmental-sustainability-considerations.

et al., A. S. (2024). 2024 united states data center energy usage report [Accessed: Jun. 18, 2025]. https://doi.org/10.71468/P1WC7Q

et al., H. A. (2025). Using life cycle assessment to drive innovation for sustainable cool clouds. *Nature*, *641*(8062), 331–338. https://doi.org/10.1038/s41586-025-08832-3

et al., R. A. H. (2025). Hydrotreating and acidic gas removal for natural gas pretreatment. In *Comprehensive methanol science* (pp. 1–17). Elsevier. https://doi.org/10.1016/B978-0-443-15740-0.00047-1

et al., X. Z. (2025). Data center cooling system optimization using offline reinforcement learning.

High-performance computing data center cooling system energy efficiency [Accessed: Jun. 18, 2025]. (2025). https://www.nrel.gov/computational-science/data-center-cooling-system.

Kant, K. (2009). Data center evolution. *Computer Networks*, *53*(17), 2939–2965. https://doi.org/10.1016/j.comnet.2009.10.004

Katal, A., Dahiya, S., & Choudhury, T. (2023). Energy efficiency in cloud computing data centers: A survey on software technologies. *Cluster Computing*, *26*(3), 1845–1875. https://doi.org/10.1007/s10586-022-03713-0

Kisitu, D., Ortega, A., Zlatinov, M., & Schaffarzick, D. (2023). Two-phase flow in compressed copper foam with r134a for high heat flux thermal management. *2023 IEEE ITherm*, 1–10. https://doi.org/10.1109/ITherm55368.2023.10177584

Krizhevsky, A., Sutskever, I., & Hinton, G. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Meet the tiny startup building stargate, openai's $500 billion data center moonshot [Accessed: Jun. 03, 2025]. (2025). https://www.forbes.com/sites/christopherhelman/2025/04/10/meet-the-tiny-startup-building-stargate-openais-500-billion-data-center-moonshot/.

Number of chatgpt users (march 2025) [Accessed: Jun. 03, 2025]. (2025). https://explodingtopics.com/blog/chatgpt-users.

Tech megacaps to spend more than $300 billion in 2025 to win in ai [Accessed: Jun. 03, 2025]. (2025). https://www.cnbc.com/2025/02/08/tech-megacaps-to-spend-more-than-300-billion-in-2025-to-win-in-ai.html.

What is artificial general intelligence (agi)? [Accessed: Jun. 03, 2025]. (2025). https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-artificial-general-intelligence-agi.

What is generative ai generative ai basics [Accessed: Jun. 03, 2025]. (2025). https://www.simplilearn.com/tutorials/artificial-intelligence-tutorial/what-is-generative-ai.

Part II

Education

Training AI to Teach Humanity

Ayana Mussabayeva

Synopsis. Artificial Intelligence (AI) is often portrayed as a savior of modern civilization. However, dystopian perspectives highlight existential risks and ethical challenges. This essay explores the darker narrative of AI through the lenses of literature, philosophy, and emerging technology debates.

> "A person's humanity comes from intellect, knowledge, good ancestry, good parents, good friends, and good teachers" (Qunanbaiuly, 1918)

In a world divided not only by borders but by bandwidth and opportunity, education remains the most powerful equalizer. Yet millions still lack access to high-quality teachers, learning materials, or personalized support. For countries like Kazakhstan — AI offers not just innovation, but transformation.

Across the Global South, and particularly in Central Asia, educational inequality is shaped not only by geography, but by economics, language, and history. In Kazakhstan, this inequality runs deep. During the Soviet period, systemic policies of Russification led to the suppression of the Kazakh language in schools, media, and public institutions. As a result, entire generations grew up with limited access to high-quality educational resources in their native tongue. Even after independence, Kazakh-language textbooks, digital content, and academic materials have lagged behind Russian- and English-language alternatives — particularly in science, technology, and mathematics.

This legacy continues to affect rural and underfunded schools, where students may be most comfortable in Kazakh, yet have to study using outdated or translated materials that do not reflect their linguistic reality. It's not just a technical problem — it's a matter of cultural and educational equity. This

is where AI offers something truly game-changing. Large language models, translation systems, and generative educational tools can now be trained on Kazakh-language data to produce high-quality, locally relevant content — for the first time at scale. An AI tutor fluent in Kazakh can explain algebra to a ninth grader in a village school, answer a child's question about space in their own dialect, or help a teacher prepare personalized materials without switching languages.

These tools don't just enhance learning — they restore dignity. They allow Kazakh-speaking children to learn complex subjects in their mother tongue, not as second-class citizens of a global knowledge economy, but as fully recognized participants. And it's not just Kazakhstan. Many countries in Central Asia, Africa, and Southeast Asia face similar challenges, where historically marginalized languages limit access to modern education. AI offers a path forward — not by erasing these languages, but by elevating them into the digital future.Moreover, AI doesn't just help students. AI assistants for teachers — especially those in under-resourced schools — can reduce workload, generate lesson plans, suggest differentiated instruction strategies, and provide feedback tools. A teacher who once had to prepare everything manually can now co-create with AI, freeing up time for what matters most: connecting with students.

For developing countries, this means leapfrogging traditional models of educational reform. Where it's not feasible to build elite teacher academies in every town, we can instead provide intelligent support systems that scale high-quality education across regions — tailored to local culture, language, and needs. As we embrace the possibilities AI brings to education, we must also reckon with the risks — not just technical, but structural and political. The most significant threat today is not that AI will become too intelligent, but that it will reflect and amplify the biases of those who build and control it.

Large language models are trained on massive datasets scraped from the internet — data that is often riddled with social, political, and cultural bias. If left unchecked, these models can reinforce stereotypes, misrepresent history, or marginalize voices that are already underrepresented. For example, if Kazakh history, culture, or even names are absent from training data, AI systems may treat them as irrelevant or invisible. Worse, they may reproduce harmful narratives about ethnic groups, languages, or political systems — all under the guise of "neutral" machine intelligence.

This isn't a purely technical issue. It's a matter of ethical design and transparency. The people — and corporations — who control the training data, model architecture, and deployment pipeline hold tremendous power. If those decisions are made behind closed doors, with no input from educators, lin-

guists, or communities affected by these systems, we risk creating tools that perpetuate inequality rather than solve it. When a student asks an AI assistant a question about their country's history or culture, the answer they receive is not just data — it's a reflection of whose stories were deemed worthy of inclusion. And if that student is from Kazakhstan, or any developing country, they deserve better than a generic, Western-centric response. They deserve technology that recognizes them. This is why honesty, openness, and accountability in AI development matter. It's not enough to build smart systems — we need to build fair, inclusive, and locally relevant ones. And that means pushing corporations and developers to open their datasets, document their design choices, and accept that no model is truly neutral.

The danger, then, is not in AI itself. It is in who defines its purpose, whose values it encodes, and who benefits from its widespread adoption. Without ethical guardrails and diverse participation, even the most powerful educational technology risks deepening the very gaps it claims to close. AI is often framed as a technological challenge — a question of scale, efficiency, or optimization. But when it comes to education, especially in developing nations, AI becomes something more: a moral and cultural project. It's about who gets to learn, in what language, and with what dignity.

For Kazakhstan and other countries with histories of linguistic erasure and uneven access to opportunity, AI is not just a tool — it's a second chance. A chance to teach in any topic, in any language. And as we train machines to understand and generate human language, we are inevitably forced to confront our own values: what knowledge we consider worth preserving, whose voices we amplify. Perhaps in our pursuit of building better generative AI and large language models, we are not merely trying to humanize machines — we are, in fact, trying to remind ourselves what it means to be human.

Ayana Mussabayeva is a PhD student in Machine Learning at MBZUAI, working on causality and brain–computer interfaces. She holds graduate degrees from the University of Manchester and Nazarbayev University. Previously, she led AI/ML at INUI Gaming, developing vision-based systems and recommender engines. Her research focuses on neuroscience, healthcare AI, and interpretable ML. Ayana is also active in community-building and enjoys stand-up comedy.

References

Qunanbaiuly, A. (1918). *The book of words*. Almaty.

Non cogito, ergo sum?: Cognitive Cost of Artificial Intelligence

Sanjutha Indrajit

Synopsis. Artificial Intelligence (AI) is widely celebrated as a revolutionary force for solving humanity's greatest challenges. However, beneath this technological optimism lies a troubling reality about AI's impact on human cognition and social dynamics. This essay explores the hidden costs of AI integration through the lens of cognitive science, social dynamics, and human development.

The integration of artificial intelligence into daily life has been remarkably swift and largely welcomed. From writing assistance to complex problem-solving, AI tools promise to enhance human capability and free us from mundane tasks. Yet beneath this technological optimism lies a more troubling reality: AI may be fundamentally altering human cognitive abilities in ways that could reshape not just how we think, but how we relate to ourselves and each other.

Subtle Erosion of Cognitive Function

Unlike previous technological revolutions that affected specific skills, AI's impact on human cognition appears both broader and more insidious. Users report a constellation of changes that collectively suggest a rewiring of basic mental processes. Attention spans contract as AI provides instant, polished responses, reducing tolerance for the slower, messier work of genuine thinking. Vocabulary narrows when sophisticated language models consistently offer refined alternatives, eliminating the productive struggle of finding one's own words. The willingness to explore ideas openly diminishes as AI converges on "good" answers, discouraging the divergent thinking that leads to genuine insight. Perhaps most concerning is the reduced tolerance for errors. Mistakes

and the process of working through them are fundamental to learning and creativity. When AI helps us avoid errors, we lose opportunities to develop resilience and the kind of iterative thinking that produces breakthrough insights. The cognitive muscles that allow us to sit with confusion, work through ambiguity, and persist through intellectual difficulty begin to atrophy from disuse. These changes represent more than behavioral shifts—they constitute alterations to core cognitive functions. Executive function weakens when AI handles mental heavy lifting. Memory systems become less robust when information is instantly accessible. Critical thinking skills decline when we grow accustomed to accepting AI-generated content without rigorous scrutiny.

Transformation of Self-Perception

As these cognitive changes accumulate, they inevitably alter how we understand ourselves as thinking beings. The erosion of intellectual struggle—that productive friction between problem and solution—diminishes our sense of cognitive agency. When AI consistently provides answers we couldn't generate ourselves, we begin to question our own intellectual worth and capacity. This shift touches the very core of human identity as rational, creative beings capable of independent thought. The experience of working through complex problems, of having insights emerge from sustained mental effort, becomes increasingly rare as AI mediates more of our cognitive work. The seamless nature of AI assistance makes this transformation particularly dangerous, creating a feedback loop where diminished cognitive abilities lead to greater dependence on AI, which further accelerates cognitive decline.

The Emergence of Cognitive Inequality

The cognitive effects of AI are unlikely to be distributed evenly across society. Instead, they threaten to create new forms of inequality based on differential access to "cognitive preservation"—the luxury of maintaining and developing human thinking abilities in an AI-saturated world. Privileged communities may increasingly treat human cognitive abilities as premium commodities. Private schools might emphasize deep thinking over AI assistance, wealthy parents could limit their children's AI exposure, and elite social circles might value intellectual struggle as markers of cultural capital. Meanwhile, under-resourced communities may become more dependent on AI out of necessity, using it for educational support and daily problem-solving when other resources are unavailable. This could accelerate cognitive decline in populations already facing educational disadvantages. The emergence of cognitive classes would be particularly unapparent because this form of inequality remains largely invisible. Unlike material disparities, cognitive differences are harder to detect and measure. A generation raised with extensive AI mediation might not recognize what cognitive capacities they never developed, creating

normalized learned helplessness.

The Fracturing of Human Connection

Perhaps the most profound consequence lies in AI's potential to fracture human relationships. Deep connections depend on shared cognitive rhythms—the ability to think through problems together, engage with ideas at similar depths, and tolerate comparable levels of uncertainty and complexity. When AI creates cognitive divergence between individuals, it threatens these fundamental bonds. Consider a romantic partnership where one person has preserved their ability to work through ambiguity while the other has become accustomed to AI-mediated quick resolutions. They may find themselves increasingly unable to connect, with one partner appearing impatient while the other seems needlessly slow. Parent-child relationships face particular vulnerability. Parents who developed patience and problem-solving skills may struggle to connect with children who expect immediate answers and become distressed by uncertainty. Friendships, too, rely on the ability to explore ideas together. When people have vastly different tolerances for intellectual effort, these conversations become strained or impossible. This cognitive fragmentation could create a new form of loneliness, where people cluster only with others who share their level of cognitive function. Rather than bringing people together, AI might drive us apart by creating incompatible ways of thinking and relating to ideas.

Historical Context and Unprecedented Scope

Previous technologies have affected human cognition—writing systems reduced memory reliance, calculators weakened mental arithmetic, GPS diminished spatial navigation. However, AI appears fundamentally different in several ways. First, its scope is unprecedented. Rather than affecting specific cognitive domains, AI simultaneously impacts multiple functions across reasoning, creativity, communication, and problem-solving. Second, AI represents cognitive substitution rather than mere augmentation. Previous tools handled narrow tasks while leaving conceptual thinking to humans. AI can substitute for higher-order cognitive processes themselves. The seamlessness of AI integration makes it particularly dangerous. Earlier technologies had clear boundaries; AI is increasingly embedded and invisible, making it harder to notice when we're offloading cognitive work or to maintain intentional boundaries.

The Generational Dimension

The cognitive effects are likely most pronounced among those who grow up with AI from birth. Unlike adults who can recognize what they're losing, children in AI-saturated environments may never develop certain cognitive

capacities in the first place. This creates the possibility of a cognitive baseline shift, where diminished thinking abilities become normalized. Future generations might not recognize what cognitive capabilities they're missing. Children who grow up expecting AI to handle complex thinking may struggle to develop the patience, persistence, and tolerance for ambiguity that characterize mature cognition.

The cognitive cost of AI convenience may be the defining challenge of our technological age. As we stand at this crossroads, we must grapple with fundamental questions about what it means to be human in an age of artificial intelligence. The choices we make today about AI integration will determine not just our individual cognitive futures, but the very nature of human society.

We are potentially witnessing a transformation in human cognition as significant as the development of language or writing. Whether this enhances or diminishes human potential depends on our ability to recognize the risks and respond with wisdom and intentionality. The conversation about AI's cognitive impact isn't just about technology—it's about preserving the essence of what makes us human. Our ability to think deeply, connect meaningfully, and navigate complexity may be the most important capabilities we can protect in an age of artificial intelligence.

Sanjutha Indrajit is a Data Scientist with a background in AI, Geospatial Analytics, and Earth Observation (EO). She is currently pursuing a Master's in Artificial Intelligence at the University of Surrey, UK, where her research focuses on efficient training and inference of Earth Observation Foundation Models and vision-language alignment for multi-sensor EO data to increase their adaptability for usage in climate resilience, disaster management, and sustainability applications.

AI-Powered STEM Ecosystems: Rethinking Learning, Innovation, and Impact

Verrah Akinyi Otiende

Synopsis. Artificial intelligence (AI) is quickly transforming STEM education by providing adaptive learning systems that personalize training, automate administrative duties, and increase student engagement. Empirical research from 2024-2025 show that when used properly, AI integration can improve conceptual understanding, reduce remediation time, and cultivate metacognitive skills. Real-time diagnostic feedback for educators, personalized learning pathways for students, and increased equity through inclusive design are among the key benefits. However, issues such as algorithmic bias, data privacy concerns, and learner overconfidence necessitate stringent protections. National initiatives in Estonia and India demonstrate the benefits of integrated policy and professional development. To fully fulfill AI's promise in STEM, stakeholders must prioritize ethical governance, cross-cultural validation, and longitudinal evaluation, ensuring that AI acts as a catalyst rather than a crutch for educational success.

Artificial intelligence (AI) has evolved over the last ten years from a theoretical breakthrough to a useful and revolutionary force in education. Its use in educational settings, especially in STEM fields (science, technology, engineering, and mathematics), has created new opportunities to tackle enduring pedagogical issues. STEM education necessitates the development of higher-order thinking and problem-solving abilities in addition to the understanding of abstract and frequently complicated topics. Nonetheless, teachers usually face significant differences in students' readiness, limited class time, and classrooms that are getting bigger and more diverse. These differences are frequently difficult for traditional teaching approaches to account for, which results in gaps in student engagement and performance.

AI-powered teaching resources, such as automated feedback systems, adaptive learning platforms, and intelligent tutoring systems, provide a way to close these gaps. These systems can assess learning challenges, personalize instructional routes, and offer focused support that is in line with each student's needs and pace by utilizing real-time data and learner analytics. By automating repetitive chores and providing actionable insights into student performance, artificial intelligence (AI) enhances instructional capacity for instructors, allowing for more responsive and strategic instruction. It makes studying more interesting and approachable for students, encouraging independence and perseverance. AI's contribution to creating fair and future-ready STEM learning environments is becoming more and more important as it develops, establishing it as more than just a technical advancement but also a key element of educational innovation and systemic change.

Adaptive Learning Platforms

The creation of adaptive learning platforms is among the most important developments in the use of AI in STEM education. In order to continuously assess student interactions, identify patterns of misunderstanding, and dynamically modify content delivery to accommodate different learning trajectories, these intelligent systems make use of machine learning techniques. Modern adaptive platforms use predictive modeling to foresee mistakes and take proactive measures before students reach cognitive bottlenecks, in contrast to classic rule-based educational technologies—like ALEKS—which rely on predetermined answer pathways.

The work of (El Fathi et al., 2025) integrated a Constructivist Inquiry-Learning Prompting (CILP) framework driven by ChatGPT into an undergraduate thermodynamics curriculum in Morocco, provides a powerful illustration of this potential. Their research showed a significant rise in learner engagement and self-efficacy in addition to an 18% improvement in students' post-test performance. These results demonstrate how AI may help both the cognitive and emotive aspects of learning, which is especially useful in STEM fields where high student turnover frequently coexists with conceptual difficulties. The ability of adaptable platforms to scale up and tailor education makes them revolutionary tools for creating inclusive and responsive STEM learning environments.

Enhancing Instructional Support for Educators

Further to its effects on students, artificial intelligence (AI) provides instructors with significant benefits by improving pedagogical decision-making and optimizing instructional workflows. Automated evaluation and feedback is one of the most revolutionary applications in this field. Recent developments

in natural language processing (NLP) have made it possible for AI systems to assess open-ended responses with more accuracy and contextual sensitivity, whereas early automated grading implementations were restricted to closed-ended forms like multiple-choice questions. This development increases AI's ability to offer formative feedback and real-time diagnostics for a wider variety of STEM disciplines and evaluation formats. The practical efficacy of these technologies is supported by empirical data. According to a recent study by (Khazanchi et al., 2025), algebra teachers in secondary schools who incorporated AI-generated diagnostics into their lesson plans were able to improve average student test results by 12% and cut remedial time by 30%. These results imply that AI improves learning outcomes through prompt and focused intervention in addition to increasing instructional efficiency. Crucially, teachers can devote more time to higher-order instructional activities like mentorship, tailored support, and facilitating project-based or inquiry-driven learning when routine grading and administrative duties are reduced. AI thus acts as a strategic partner in improving the caliber and breadth of STEM education rather than just as a time-saving tool.

Enhancing Learner Engagement and Metacognition

By enabling tailored learning pathways and offering responsive, on-demand help, artificial intelligence (AI) greatly increases student engagement. Adaptive learning systems lower the risk of cognitive overload, dissatisfaction, or disengagement by customizing teaching modalities to suit individual learner preferences in addition to instantly adjusting the complexity of the information (Maity & Deroy, 2024). In STEM fields, where learner heterogeneity in prior knowledge and understanding pace might be significant, this personalization is especially crucial. Artificial intelligence (AI) greatly increases student engagement by facilitating individualized learning paths and offering prompt, on-demand assistance. By customizing instructional modalities to suit each learner's preferences and adjusting content difficulty in real time, adaptive learning systems lower the risk of cognitive overload, frustration, or disengagement. In STEM fields, where learners might vary greatly in their past knowledge and rate of comprehension, this personalization is especially crucial.

However, in order to prevent unforeseen outcomes, the educational implementation of AI must be approached intentionally. According to (Wang et al., 2025), certain undergraduate students who primarily use generative AI teaching systems tend to avoid active problem-solving techniques. Their comprehension of the material remained cursory as a result, underscoring the risk of abuse that arises when AI tools are not included into a coherent educational framework. These results highlight how crucial it is to create AI-integrated

learning environments that strike a balance between automation and critical thinking and active student participation.

Equity and Ethical Considerations in AI Integration

Concerns about ethics and equity grow more important when artificial intelligence is incorporated into educational settings. Although AI can improve learning for a variety of student demographics, if its development and application are not purposefully inclusive, it also runs the risk of escalating already-existing inequities (Strielkowski et al., 2025). Uneven educational outcomes can result from algorithmic bias, which disproportionately disadvantages underprivileged students and is frequently caused by non-representative training data or unquestioned assumptions. The growing integration of artificial intelligence into educational systems raises important questions about ethics and justice. Although AI can help different student populations learn more effectively, if its development and application are not purposefully inclusive, it could also worsen already-existing inequities. Algorithmic bias, which frequently results from unrepresentative training data or unquestioned assumptions, can provide unequal educational outcomes that disproportionately disadvantage underprivileged students.

Such instances are still the uncommon rather than the rule. (Kohnke & Zaugg, 2025) conducted a nationwide study of K–12 AI pilot programs and discovered that less than half of the projects had strong data privacy protections or carried out systematic bias audits. Ethical questions are raised by this lack of due diligence, especially when automated decision-making is based on sensitive student data. These results highlight the need for transparent governance structures, bias reduction techniques, and ethical review procedures to be included into the use of AI in education. Well-meaning innovations run the risk of escalating rather than resolving educational disparities in the absence of such safeguards.

Coordinated Initiatives and the Systemic Integration of AI

The systemic potential of artificial intelligence to revolutionize STEM education is becoming more and more evident through coordinated national and institutional initiatives. These initiatives show how AI may impact educational policy, improve teaching methods, and spur curriculum reform when applied intentionally and widely. Such initiatives present AI as a structural element of educational systems—one that unify governance, professional development, and pedagogy under a single vision for innovation—as opposed to discrete technology interventions. Estonia's "AI Leap" initiative, which seeks to mainstream AI integration at all educational levels, is a noteworthy example. In addition to integrating digital ethics training into professional development

curriculum, the program offers instructors AI-augmented lesson planning tools that facilitate individualized instruction. Early assessments show that the initiative has increased STEM exam participation by 15%, indicating a favorable relationship between AI-enhanced learning and student interest in traditionally difficult subjects.

In a comparable manner, the PadhAI conference in India is an example of a policy-focused strategy for systemic AI adoption. To create ethical standards for the application of AI in elementary and secondary education, educators, technologists, and legislators gathered nationwide. The resulting framework, which directs the responsible use of AI tools in educational settings, is presently undergoing a pilot program at 50 institutions. These instances demonstrate the value of methodical and cooperative approaches to AI integration—approaches that put long-term effects, capacity building, and ethical considerations ahead of immediate technological advancements. The foundation for more egalitarian, efficient, and future-ready STEM learning ecosystems is laid by projects that view AI as a catalyst for systemic change rather than just a tool.

Sustaining the Promise of AI in STEM

The field needs to develop a research agenda that prioritizes both cross-cultural relevance and longitudinal inquiry in order to fully realize and maintain the revolutionary potential of artificial intelligence in STEM education. Even though there is a growing amount of empirical data showing encouraging short-term improvements in learning outcomes and engagement, most of these research have been carried out in wealthy, well-resourced educational settings. Because of this, little is known about how AI tools function in varied or resource-constrained environments, which might vary greatly in terms of infrastructure, educational standards, and student requirements.

A deliberate move toward long-term, contextually grounded research that looks at retention, knowledge transfer, and long-term effects on academic advancement and career routes is necessary to close this gap. Deeper understanding of AI's long-term educational usefulness can be gained, for example, by examining how AI interventions affect students' long-term interest in STEM disciplines, their problem-solving skills over time, or their eventual admittance into STEM employment.

Adoption of open science practices in the field of AI in education is equally significant. Sharing model parameters, user interaction logs, and anonymized datasets can promote group learning, aid replication initiatives, and lessen job duplication across borders and institutions. Initiatives promoting open data not only increase trust and openness but also make it possible to conduct comparative research that reveals the ideal circumstances for using AI tools.

Sustained investment in inclusive, longitudinal, and data-informed research will be essential as AI develops to guarantee its fair and long-lasting influence on global STEM learning ecosystems.

Dr. Verrah Akinyi Otiende is an Assistant Professor of Statistics and Data Science at USIU-Africa. She holds a PhD in Mathematical Statistics through a cotutelle program between PAUSTI and JKUAT. Her research focuses on spatiotemporal modeling of infectious diseases, artificial intelligence in education, and natural language processing. Dr. Otiende is deeply involved in STEM mentorship, particularly for women and underrepresented groups, through programs like AIMS and Mawazo. She is an active member of several scientific communities, including IBS, OWSD, and ASFI. Her work bridges data science, education, and public health with a focus on inclusive, ethical innovation.

References

El Fathi, T., Saad, A., Larhzil, H., Lamri, D., & Al Ibrahmi, E. M. (2025). Integrating generative ai into stem education: Enhancing conceptual understanding, addressing misconceptions, and assessing student acceptance. *Disciplinary and Interdisciplinary Science Education Research*, *7*(1), 6.

Khazanchi, R., Di Mitri, D., & Drachsler, H. (2025). The effect of ai-based systems on mathematics achievement in rural context: A quantitative study. *Journal of Computer Assisted Learning*, *41*(1), e13098.

Kohnke, S., & Zaugg, T. (2025). Artificial intelligence: An untapped opportunity for equity and access in stem education. *Education Sciences*, *15*(1), 68.

Maity, S., & Deroy, A. (2024). Generative ai and its impact on personalized intelligent tutoring systems. *arXiv preprint*.

Strielkowski, W., Grebennikova, V., Lisovskiy, A., Rakhimova, G., & Vasileva, T. (2025). Ai-driven adaptive learning for sustainable educational transformation. *Sustainable Development*, *33*(2), 1921–1947.

Wang, K. D., Wu, Z., Tufts, L., Wieman, C., Salehi, S., & Haber, N. (2025). Scaffold or crutch? examining college students' use and views of generative ai tools for stem education. *2025 IEEE Global Engineering Education Conference (EDUCON)*, 1–10.

AI in Academia; The Gainers, The Losers?

Bakare Surajudeen

Synopsis. Artificial intelligence (AI) is now deeply woven into academic life, igniting debate about who truly benefits and what is at stake. This essay explores the impact of AI in academia, highlighting its benefits for students through personalized learning and efficiency for institutions. However, it warns of risks like overreliance, reduced critical thinking, and blurred lines between genuine work and AI assistance. This essay calls for thoughtful policy and digital literacy to ensure AI enhances, rather than undermines, authentic education.

Artificial intelligence (AI) has become an integral part of modern life, influencing everything from household management to professional productivity. In academia, AI's rapid adoption has sparked a vigorous debate among educators, students, and policymakers regarding its true impact. While AI promises enhanced efficiency, personalized learning, and broader access, it also raises concerns about academic integrity, equity, and the erosion of critical thinking. This paradox—whether AI's benefits outweigh its detriments—has led to divergent perspectives within educational communities. The central question remains: in the evolving relationship between AI and academia, who stands to gain, and who may be left behind? This essay critically examines current research and debates, aiming to clarify the winners and losers in the AI-academia nexus and to propose pathways for ethical and effective integration.

The Gainers

Recent research consistently identifies students as the primary beneficiaries of AI integration in academia (Vieriu & Petrea, 2025; Wang et al., 2024). This is due to the benefits provided to students by AI-powered platforms, such as enabling personalized learning, real-time feedback, and adaptive support tai-

lored to individual student needs. This, in turn, results in improved academic outcomes and greater engagement (Ambarita & Nurrahmatullah, 2024). For instance, students utilize AI to synthesize complex topics, find relevant evidence, and develop higher-order writing skills, demonstrating the technology's valuable role in facilitating analytical and research-oriented tasks (Vieriu & Petrea, 2025).

However, while this is the case for students, educators and institutions also benefit from AI. For example, educators can be seen as gainers when they use AI to automatically grade students, examine academic performance, execute administrative tasks, and enhance the efficiency of personalized feedback to students (Wang et al., 2024). Similarly, institutions can be considered gainers when they deploy AI to execute administrative duties and streamline operations (Khare et al., 2018).

In light of the above, one can say that the gainers in the AI-academia relationship can be divided into two: the major gainers and the minor gainers; where the major gainers represent the student community, and the minor gainers represent the educators and institutions who are laden with the responsibility of providing regulated access to AI within learning institutions, deploying AI, and developing ethical guidelines.

The Losers

While research demonstrates that students gain significantly from AI integration in academia, several notable detriments have also been identified. Chief among these is overdependence on AI technologies. Recent studies reveal that many students increasingly rely on AI-based tools—such as ChatGPT, Grammarly, and translation applications—for academic work, often at the expense of traditional learning methods like note-taking, reading, and critical thinking (Ambarita & Nurrahmatullah, 2024; Vinh & My, 2025). This overreliance can lead to a decline in independent problem-solving and creativity, as well as a diminished ability to engage deeply with course material (Klimova & Pikhart, 2025; Vinh & My, 2025).

The phenomenon is not unique to academia; similar patterns are observed in fields like software development, where AI usage can undermine idea generation and autonomy, even as it enhances execution and management (Klimova & Pikhart, 2025). In educational contexts, this dependency is associated with increased laziness, reduced creativity, and the spread of misinformation. Furthermore, students who place excessive trust in AI outputs may accept incorrect information uncritically, undermining learning outcomes and critical thinking skills (Vieriu & Petrea, 2025).

Other risks include privacy and security concerns, as well as the erosion

of fundamental literacy and communication abilities when students rely on AI-generated content instead of developing their own voices (Li et al., 2025). These aspects highlight the need for balanced AI integration that supports, rather than replaces, essential academic skills.

While these risks primarily affect students, the majority of the task in regulating usage and preventing overdependence rests heavily on the shoulders of educators and institutions. In addition to the constant challenge of accepting the new tech explosion and their rapidly changing role in the classroom, they are also saddled with the responsibility of regulating AI usage among students, as well as developing ways to make its use ethical.

This unfortunately makes educators and institutions the major losers while students are, in this case, the minor losers.

The Grey Areas

Since the rapid adoption of AI in academia, ethical concerns have emerged regarding its use by students and the evolving roles of teachers. Institutions are now grappling with how to promote ethical AI usage and ensure academic integrity (Li et al., 2025). However, a less discussed but equally important issue is how AI is reshaping teachers' perceptions of student intelligence and creativity. For instance, my own experience—where a lecturer questioned whether my well-written email was AI-generated—reflects a growing scepticism among educators. This raises a critical question: as AI-generated work becomes more sophisticated, will teachers begin to doubt the authenticity of all high-quality student output, potentially overlooking genuine moments of student brilliance?

Research shows that while teachers appreciate AI's ability to enhance personalized learning, creativity, and problem-solving, they also acknowledge the difficulty in distinguishing between student-generated and AI-generated work (Vieriu & Petrea, 2025). This ambiguity complicates assessment and may inadvertently discourage students from striving for excellence, fearing their achievements will be attributed to AI rather than their own abilities (Ambarita & Nurrahmatullah, 2024). As a result, the academic community must address not only the ethical use of AI but also develop new strategies for recognizing and nurturing authentic student intelligence and creativity in this evolving landscape.

In conclusion, this essay has provided a comprehensive analysis of the ongoing debate surrounding the integration of AI in academia, with particular attention to both well-documented benefits and emerging concerns. By categorizing the gainers and losers of the AI-academia relationship into major and minor groups, the discussion clarifies the nuanced impacts on students,

educators, and institutions. Notably, the essay also highlights the ”grey areas”—such as shifting perceptions of student intelligence and the complexities of authentic assessment—that remain underexplored in current research. Future studies should address these overlooked dimensions by investigating how AI influences teacher perceptions, student creativity, and the development of effective, ethical guidelines for AI use in educational settings. This will be essential for ensuring that AI serves as a tool for genuine academic advancement while minimizing unintended negative consequences.

Bakare Surajudeen is a cognitive neuroscientist with an MSc in Cognitive Sciences and Technologies from HSE University, Moscow. He holds a BSc in Anatomy from Olabisi Onabanjo University, Nigeria. His research focuses on motor learning, mirror neurons, and transcranial magnetic stimulation (TMS).

References

Ambarita, N., & Nurrahmatullah, M. F. (2024). Impacts of artificial intelligence on student learning: A systematic literature review. *Jurnal VARIDIKA*, 13–30. https://doi.org/10.23917/varidika.v36i1.4730

Khare, K., Stewart, B., & Khare, A. (2018). Artificial intelligence and the student experience: An institutional perspective [The International Academic Forum (IAFOR)].

Klimova, B., & Pikhart, M. (2025). Exploring the effects of artificial intelligence on student and academic well-being in higher education: A mini-review. *Frontiers in Psychology*, *16*. https://doi.org/10.3389/fpsyg.2025.1498132

Li, Y., Tolosa, L., Rivas-Echeverria, F., & Marquez, R. (2025). Integrating ai in education: Navigating unesco global guidelines, emerging trends, and its intersection with sustainable development goals. *ChemRxiv*. https://doi.org/10.26434/chemrxiv-2025-wz4n9

Vieriu, A. M., & Petrea, G. (2025). The impact of artificial intelligence (ai) on students' academic development. *Education Sciences*, *15*(3). https://doi.org/10.3390/educsci15030343

Vinh, N. T., & My, C. T. T. (2025). The current situation of students' overdependence on ai and the neglect of traditional learning methods. *International Journal of Advanced Multidisciplinary Research and Studies*, *5*(3), 585–586.

Wang, X., Xu, X., Zhang, Y., Hao, S., & Jie, W. (2024). Exploring the impact of artificial intelligence application in personalized learning environments: Thematic analysis of undergraduates' perceptions in china. *Humanities and Social Sciences Communications*, *11*(1), 1644. https://doi.org/10.1057/s41599-024-04168-x

Part III

Healthcare

Mitigating Autonomous Bias in Human-Centered AI Systems

Mary Adewunmi

Synopsis. AI is becoming the new scalpel in the hands of modern medicine— sharp enough to save lives, precise enough to tailor care, and powerful enough to transform treatment. AI integration into the healthcare system brings not only transformative potential but also a critical responsibility to make sure the system is safe and trusted. This opinion piece explores the challenges of autonomous bias that slow down the progress of trusted and responsible human-centered AI systems and discusses strategies for mitigating it. This will assist in the successful integration of a trusted and unbiased human-centred AI system into healthcare.

AI is transforming modern medicine by improving patient outcomes, accurate diagnoses, and more personalised treatment (Alowais et al., 2023; Bauer & Thamm, 2021; Patil & Shankar, 2023). With that power comes the responsibility to ensure AI models are built with trust and are safe for humans. This led to human-centred AI models (Chen et al., 2023; Klingefjord et al., 2024), which focused on aligning human oversight into AI models by i) eliciting values from humans, ii) reconciling the values into target specifications for AI models, and iii) training the AI model. Although integrating human oversight is crucial, it does not fully mitigate the risk of autonomous bias and recursive deference inherent in human-centered AI systems. It is necessary to clarify that autonomous bias is where the AI defers to humans for input, but humans have grown heavily reliant on AI without critical oversight (Choudhury et al., 2022). Consequently, this leads to inherited result inaccuracies and biased systems. In addressing this issue, this opinion piece outlines a strategic plan for alleviating autonomous bias in human-centered systems by leveraging the existing path by including important checkpoints at each phase of the AI model: I) design, II) development, and III) deployment stages.

Design

In the model design phase, one of the critical aspects of responsible AI design is data collection and annotation. High-quality datasets are essential to accurate annotation and better models for responsible AI design (Zając et al., 2023). In ensuring a responsible human-centered AI system for medical applications, there is a need for human-AI codesign in feature engineering of data, not delegation (Panigutti et al., 2023; Silvola et al., 2023). Further, this opinion piece suggests that there is a need to instill a decision loop with every feedback, rationale, decision, and mockup review on human-AI codesign with stakeholders such as clinical experts, AI developers, patients, caregivers, and administrative health workers. This will ensure a shared voyage of imagination and empathy, exploring new paths to better design human-centered AI systems for trusted and effective clinical decision support. For instance, a mock-up review of an AI model with a clinical expert can be, "Why do you trust AI on dosages but not on diagnosis?" This will mitigate overreliance and critical oversight that often occur in human-centered AI systems.

Develop

In the model development phase, automation bias in human-centered systems can occur through deeper cognitive and flawed assumptions with the stakeholders (Goddard et al., 2012). This involves overreliance of humans on AI for critical testing and evaluation decisions such as feature selection, choice of algorithm, and evaluation metrics, ultimately reinforcing cognitive bias, systemic inequities, and objective human-centered AI systems (Vered et al., 2023). In addressing this issue, this opinion piece suggests bias-aware feature selection (Yang et al., 2021) with clinical experts for capturing inclusivity in feature selection. Further, it suggests that regular training should be conducted for AI developers on cognitive bias and the need for bias checkpoints during model testing and evaluation. Model interpretability and feedback loops with stakeholders, including clinical experts, before model acceptance, even at the cost of marginal performance. Model acceptance decisions should be made purely by human oversight to address systemic inequities. For instance, a responsible diabetic retinopathy model must be evaluated in collaboration with ophthalmologists and retinal specialists and tested on stratified patient subgroups to address feature selection bias. Model evaluation should be carried out with clinical experts with rationale for each result, and its acceptance should not be solely based on high performance metrics. This collaboration will enhance trust and unbiased AI models as a clinical decision support tool.

Deploy

Integrating a human-centered AI system into healthcare is a complex work-

flow that requires humans at every stage of data selection, quality control, training results presentation and evaluation, correction, performance monitoring, and incorporating feedback for model acceptability before deployment (Chen et al., 2023; Juluru et al., 2021).

This is to ensure compliance with regulations and ethical guidelines for its uptake in healthcare (Braun et al., 2021). Ethical concerns in the deployment of AI systems for medical applications emphasise model selection in a way that shifts focus from accuracy so that the benefits outweigh the risks socially, ethically, and legally (Dignum, 2019; Li et al., 2022). This opinion piece, however, suggests that model acceptability should be based on explainability with human judgment, thereby setting guiderails for the deployment process. Performance monitoring of the AI models should be about consistent performance across varied datasets. Finally, the feedback mechanisms should allow stakeholders, including clinical experts of the proposed AI model, to incorporate confidence scores to question, override, or revise recommendations. For instance, a responsible diabetic retinopathy model must be collectively agreed upon by the clinical experts and other stakeholders using confidence scores before it can be deployed. This ensures mutual responsibility of the AI system.

> "AI will be an integral part of solving the world's biggest problems, but it must be developed in a way that reflects human values." – Satya Nadella, CEO of Microsoft

In conclusion, as AI becomes the new scalpel in modern medicine, it is necessary to break the recursive loop of autonomous bias in human-centered AI systems. This opinion piece discussed strategies for addressing the challenges by embedding some decision loops in the design, development, and deployment of AI models. This will ensure the successful integration of a trusted and unbiased human-centred AI system into healthcare.

Mary Adewunmi is the founder of CaresAI. Her research focuses on developing clinical decision support systems (CDSS) with a programmatic approach. She is currently one of the four (4) Global Scholar Association Committee member of the American Association of Cancer Research (GSAC-AACR), a Kaggle and GMI mentor. More information about her can be found at https://maryadewunmi.github.io/

References

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H. A., et

al. (2023). Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC medical education*, *23*(1), 689.

Bauer, C., & Thamm, A. (2021). Six areas of healthcare where ai is effectively saving lives today. *Digitalization in healthcare: Implementing innovation and artificial intelligence*, 245–267.

Braun, M., Hummel, P., Beck, S., & Dabrock, P. (2021). Primer on an ethics of ai-based decision support systems in the clinic. *Journal of medical ethics*, *47* (12), e3–e3.

Chen, Y., Clayton, E. W., Novak, L. L., Anders, S., & Malin, B. (2023). Human-centered design to address biases in artificial intelligence. *Journal of medical Internet research*, *25*, e43251.

Choudhury, A., et al. (2022). Toward an ecologically valid conceptual framework for the use of artificial intelligence in clinical settings: Need for systems thinking, accountability, decision-making, trust, and patient safety considerations in safeguarding the technology and clinicians. *JMIR Human Factors*, *9*(2), e35421.

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use ai in a responsible way* (Vol. 2156). Springer.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121–127.

Juluru, K., Shih, H.-H., Keshava Murthy, K. N., Elnajjar, P., El-Rowmeim, A., Roth, C., Genereaux, B., Fox, J., Siegel, E., & Rubin, D. L. (2021). Integrating al algorithms into the clinical workflow. *Radiology: Artificial Intelligence*, *3*(6), e210013.

Klingefjord, O., Lowe, R., & Edelman, J. (2024). What are human values, and how do we align ai to them? https://arxiv.org/abs/2404.10636.

Li, F., Ruijs, N., & Lu, Y. (2022). Ethics & ai: A systematic review on ethical concerns and related strategies for designing with ai in healthcare. *Ai*, *4*(1), 28–53.

Panigutti, C., Beretta, A., Fadda, D., Giannotti, F., Pedreschi, D., Perotti, A., & Rinzivillo, S. (2023). Co-design of human-centered, explainable ai for clinical decision support. *ACM Transactions on Interactive Intelligent Systems*, *13*(4), 1–35.

Patil, S., & Shankar, H. (2023). Transforming healthcare: Harnessing the power of ai in the modern era. *International Journal of Multidisciplinary Sciences and Arts*, *2*(2), 60–70.

Silvola, S., Restelli, U., Bonfanti, M., & Croce, D. (2023). Co-design as enabling factor for patient-centred healthcare: A bibliometric literature review. *ClinicoEconomics and outcomes research*, 333–347.

Vered, M., Livni, T., Howe, P. D. L., Miller, T., & Sonenberg, L. (2023). The effects of explanations on automation bias. *Artificial Intelligence*, *322*, 103952.

Yang, Y., Zhang, X., Yang, M., & Deng, C. (2021). Adaptive bias-aware feature generation for generalized zero-shot learning. *IEEE Transactions on Multimedia*, *25*, 280–290.

Zając, H. D., Avlona, N. R., Kensing, F., Andersen, T. O., & Shklovski, I. (2023). Ground truth or dare: Factors affecting the creation of medical datasets for training ai. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 351–362.

Generative AI for accelerated drug design

Krinos Li

Synopsis. The traditional drug discovery pipeline is notoriously long, costly, and uncertain—often described as operating under a "reverse Moore's Law", with increasing investment yielding fewer breakthroughs. Generative AI presents a fundamentally new design paradigm by enabling rapid, low-cost, and high-precision design of therapeutic candidates. This new paradigm shifts away from traditional trial-and-error approaches, allowing molecules to be "imagined" and optimized in silico. While challenges remain—from model hallucination to evaluation biases—generative AI has already begun to redefine what is possible in drug design.

From Moore's Law to Eroom's Law and Beyond

For over half a century, Moore's Law has driven computing forward: every two years, transistor density doubles, slashing costs and powering breakthroughs in AI. Yet in drug discovery, we've faced the opposite trend—Eroom's Law—where adjusted R&D costs double roughly every nine years (Scannell et al., 2012). Why doesn't the drug pipeline benefit from ever-faster hardware? The answer lies in biology's complexity and the inefficiencies of traditional "screen-and-optimize" workflows. Now, Generative AI promises to upend that paradigm, sparking a new era where molecule design itself follows Moore-like acceleration.

The Old Pipeline vs. A New Paradigm

Traditional drug development reads like a painfully linear to-do list: identify a biological target, sift through millions of compounds in high-throughput screens, optimize hits into leads, then verify safety and efficacy. Each stage costs millions of dollars and months—or years—of work, even with robots, big data, and clever docking algorithms (K. Zhang et al., 2025). Generative AI rewrites

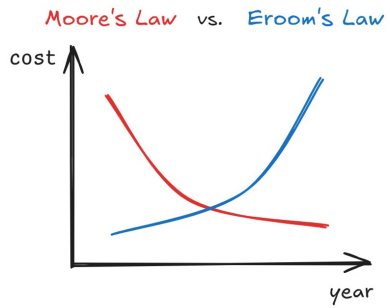


Figure 8: A concise comparison between Moore's Law and Eroom's Law.

the playbook. Rather than filtering from vast—but finite—libraries, it learns chemical and structural patterns and then creates novel candidates tailored to desired properties.

How Generative AI Designs Drugs

At the heart of this shift lie deep generative models, a class of neural architectures capable of learning and generating complex data distributions. These include Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Autoregressive models (ARs), Flow models (Flows), and, more recently, Diffusion models. Each offers a distinct approach to capturing the patterns of molecular data and has shown exceptional performance in biomolecular generation tasks. These architectures absorb massive datasets of known molecules or protein structures and learn an implicit "language" of chemistry and biology. For example, to design protein binders, a generative network can start with noise and iteratively sculpts backbone coordinates into viable folds, guided by learned structural priors from data (Ingraham et al., 2023; Watson et al., 2023).

Crucially, we can layer on optimization desired objectives—binding affinity, solubility, toxicity profiles—or plug into reinforcement-learning loops that reward biomolecules with the right balance of drug-like traits (Bilodeau et al., 2022). The result: from an initially random seed, the model draws forth entirely new scaffolds or protein folds in a matter of hours or days.

Case study: Rapid DDR1 Inhibitor Design

Insilico Medicine famously put this approach to the test. By training a generative model on kinase inhibitors and then steering it toward Discoidin Domain Receptor 1 (DDR1) binding pockets, their team designed a lead candidate in just 21 days. Traditional hit-to-lead campaigns typically stretch months;

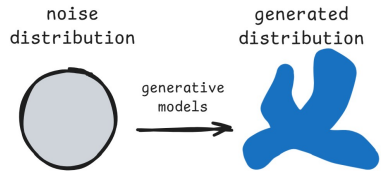


Figure 9: A generative model transforms samples from a simple noise distribution (e.g., Gaussian) into complex data distributions, such as molecular structures or protein conformations. Generative models can synthesize new and valid samples from scratch by learning the underlying structure of large biomolecular datasets.

here, AI cut that timeline to weeks while delivering nanomolar potency—and the candidate advanced swiftly into in vitro assays (Zhavoronkov et al., 2019).

Case Study: Baker Lab's Protein Designs

The University of Washington's Baker Lab exemplifies how generative AI paradigm can span multiple therapeutic modalities. Using GenAI, they have achieved designs of:

- *Helical peptide binders:* By defining targets such as parathyroid hormone (PTH) or neuropeptide Y, the model produced thousands of candidate helical peptides in under a week. Experimental screening then pinpointed lead peptides with sub-nanomolar affinities (Vázquez Torres et al., 2024).
- *Snake toxin neutralizers:* Leveraging the same framework, they designed novel proteins that bind and neutralize α-neurotoxins from snake venom. In vivo studies showed these AI-derived proteins protected mice from lethal toxin doses, demonstrating rapid routes to next-generation antivenoms (Vázquez Torres et al., 2025).
- *Tumor binders:* Applying diffusion conditioned on the tumor necrosis factor receptor, Baker Lab generated picomolar-affinity binders to TNFR1 that can function as antagonists or superagonists depending on valency—all designed entirely in silico (Glögl et al., 2024).

The Tangible Benefits of Generative AI

Generative AI has already begun to collapse traditional R&D cycles—from months or years down to mere days—by focusing experimental efforts on the most promising candidates and dramatically reducing the size of libraries that need to be synthesized and tested. Perhaps even more transformative is the

democratization of design: open-source tools, along with scalable cloud-based interfaces, enable academic labs and smaller biotech startups to explore drug discovery projects that were once the exclusive domain of large pharmaceutical R&D budgets (NVIDIA, 2024). Furthermore, foundation models—large-scale, pre-trained generative AI systems—are what enable seamless design across modalities. These models enabling design of small molecules, peptides, antibodies for receptor-targeted therapeutics within a unified framework that transcends traditional boundaries (Cho et al., 2025; Kong et al., 2025; Passaro et al., 2025).

Another pivotal development in this transformation is the UK's OpenBind initiative (UK Government, 2025). OpenBind is an open science project aimed at generating the world's largest dataset of experimentally validated protein–ligand interactions. It is creating a comprehensive resource to power the next generation of machine learning tools for structure-based drug design. This initiative is expected to produce over 500,000 protein–ligand complexes with measured binding affinities within five years, about 20 times more than what is currently available in the Protein Data Bank.

Practices and Caveats

Although generative AI can dramatically accelerate drug-like molecule design, it remains fundamentally reliant on the quality and diversity of its training data. Poorly curated datasets, low-fidelity oracles or inappropriate evaluation criteria can lead models to propose candidates that look promising on paper but fail in experimental validation (Du et al., 2024). To mitigate this, researchers should adopt an iterative human–AI workflow: generate a focused set of high-scoring designs, prioritize rapid synthesis and bioassays, then feed the resulting data back into the model to refine subsequent rounds. Model interpretability helps demystify black-box decisions, allowing to quantify prediction uncertainty and rationalize key design choices. Transparency is equally crucial. Publishing model architectures, training-set summaries, and experimental protocols not only builds confidence with regulators but also enables the community to replicate and improve upon published results (Z. Zhang et al., 2024). Finally, because the same algorithms that enable therapeutic breakthroughs could be misapplied, ethical guardrails and oversight frameworks must be established early, with access to sensitive generative pipelines restricted to accredited laboratories.

Looking Ahead: Toward a "Moore's Law" for Drugs

The most exciting developments lie in fully automated, closed-loop platforms that fuse generative design, automated synthesis, and high-throughput screening (Burger et al., 2020; Dai et al., 2024). In these systems, AI models can

propose candidates overnight, robots assemble and test them the next day, and the data streams right back into the model for continuous improvement—collapsing months of work into mere days. Looking further forward, integrated digital-twin frameworks could simulate patient-specific responses, guiding personalized drug regimens long before first-in-human trials. If this vision comes to fruition, we may finally see drug development costs plateau—or even fall—as efficiency gains outpace complexity, fulfilling the promise of a "Moore's Law" era in medicine.

Krinos Li is a PhD student at Imperial College London, jointly affiliated with Imperial-X and the Department of Bioengineering. His research focuses on developing computational approaches at the intersection of machine learning and molecular systems, with particular expertise in geometric deep learning, deep generative modeling, and Agentic AI systems.

References

Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R., & Jensen, K. F. (2022). Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *12*(5), e1608.

Burger, B., Maffettone, P. M., Gusev, V. V., Aitchison, C. M., Bai, Y., Wang, X., Li, X., Alston, B. M., Li, B., Clowes, R., et al. (2020). A mobile robotic chemist. *Nature*, *583*(7815), 237–241.

Cho, Y., Pacesa, M., Zhang, Z., Correia, B. E., & Ovchinnikov, S. (2025). Boltzdesign1: Inverting all-atom structure prediction model for generalized biomolecular binder design. *bioRxiv*, 2025–04.

Dai, T., Vijayakrishnan, S., Szczypiński, F. T., Ayme, J.-F., Simaei, E., Fellowes, T., Clowes, R., Kotopanov, L., Shields, C. E., Zhou, Z., et al. (2024). Autonomous mobile robots for exploratory synthetic chemistry. *Nature*, 1–8.

Du, Y., Jamasb, A. R., Guo, J., Fu, T., Harris, C., Wang, Y., Duan, C., Liò, P., Schwaller, P., & Blundell, T. L. (2024). Machine learning-aided generative molecular design. *Nature Machine Intelligence*, *6*(6), 589–604.

Glögl, M., Krishnakumar, A., Ragotte, R. J., Goreshnik, I., Coventry, B., Bera, A. K., Kang, A., Joyce, E., Ahn, G., Huang, B., et al. (2024). Target-conditioned diffusion generates potent tnfr superfamily antagonists and agonists. *Science*, *386*(6726), 1154–1161.

Ingraham, J. B., Baranov, M., Costello, Z., Barber, K. W., Wang, W., Ismail, A., Frappier, V., Lord, D. M., Ng-Thow-Hing, C., Van Vlack, E. R., et al. (2023). Illuminating protein space with a programmable generative model. *Nature*, *623*(7989), 1070–1078.

Kong, X., Zhang, Z., Zhang, Z., Jiao, R., Ma, J., Liu, K., Huang, W., & Liu, Y. (2025). Unimomo: Unified generative modeling of 3d molecules for de novo binder design. *arXiv preprint arXiv:2503.19300*.

NVIDIA. (2024). *Nvidia opens bionemo to scale digital biology for global biopharma and scientific industry* [Accessed: 2025-06-18]. https://nvidianews.nvidia.com/news/nvidia-opens-bionemo-to-scale-digital-biology-for-global-biopharma-and-scientific-industry.

Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V. R., Getz, N., Portnoi, T., Roy, J., Stark, H., Kwabi-Addo, D., Beaini, D., Jaakkola, T., & Barzilay, R. (2025). Boltz-2: Towards accurate and efficient binding affinity prediction.

Scannell, J. W., Blanckley, A., Boldon, H., & Warrington, B. (2012). Diagnosing the decline in pharmaceutical r&d efficiency. *Nature reviews Drug discovery*, *11*(3), 191–200.

UK Government. (2025). *Uk to become world leader in drug discovery as technology secretary heads for london tech week* [Accessed: 2025-06-18]. https://www.gov.uk/government/news/uk-to-become-world-leader-in-drug-discovery-as-technology-secretary-heads-for-london-tech-week.

Vázquez Torres, S., Benard Valle, M., Mackessy, S. P., Menzies, S. K., Casewell, N. R., Ahmadi, S., Burlet, N. J., Muratspahić, E., Sappington, I., Overath, M. D., et al. (2025). De novo designed proteins neutralize lethal snake venom toxins. *Nature*, 1–7.

Vázquez Torres, S., Leung, P. J., Venkatesh, P., Lutz, I. D., Hink, F., Huynh, H.-H., Becker, J., Yeh, A. H.-W., Juergens, D., Bennett, N. R., et al. (2024). De novo design of high-affinity binders of bioactive helical peptides. *Nature*, *626*(7998), 435–442.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., et al. (2023). De novo design of protein structure and function with rfdiffusion. *Nature*, *620*(7976), 1089–1100.

Zhang, K., Yang, X., Wang, Y., Yu, Y., Huang, N., Li, G., Li, X., Wu, J. C., & Yang, S. (2025). Artificial intelligence in drug development. *Nature Medicine*, 1–15.

Zhang, Z., Jin, R., Fu, K., Cong, L., Zitnik, M., & Wang, M. (2024). Foldmark: Protecting protein generative models with watermarking. *bioRxiv*.

Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Polykovskiy, D. A., Kuznetsov, M. D., Asadulaev, A., et al. (2019). Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature biotechnology*, *37*(9), 1038–1040.

Radiological Report Generation and Authorship: Navigating the Post-truth Age of AI

Mehmet Can Yavuz and Tician Schnitzler

Synopsis. This essay examines the role of artificial intelligence in the post-truth era, focusing on authorship and accountability. Drawing on Jean-Paul Sartre's concept of intentionality, it argues that while AI lacks the agency required for genuine authorship, it can function as a powerful interpretive partner. The essay contrasts the risks of AI-generated misinformation in fields like literature and medicine with its benefits as an analytical tool, such as in "distant reading" for humanities research and enhancing diagnostic accuracy in radiology. Ultimately, it calls for a framework of human-centered accountability to harness AI's potential while safeguarding the primacy of human judgment and truth.

The Post-truth Condition This essay examines the existential and ethical implications of artificial intelligence in the post-truth era, with a focus on authorship and accountability. Drawing from Jean-Paul Sartre's concept of free will and intentionality, it argues that AI-generated outputs—though compelling—lack the agency and moral responsibility that define genuine human expression, particularly in critical fields like literature and medicine. However, this lack of inherent agency does not relegate AI to a purely adversarial role. In fact, its strength lies in a different mode of engagement: machine interpretation. When wielded as a tool, AI can offer powerful analytical support, yielding measurable improvements in diagnostic performance by identifying patterns invisible to the human eye. In literature, it can function as a 'macroscope,' revealing textual connections across vast corpora. The critical distinction, then, is between AI as an unaccountable author and AI as a powerful interpretive partner, a distinction that is crucial for navigating the post-truth landscape.

Figure 10: Post-truth. Relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief (Oxford Languages, 2016).

Definition: Post-truth *(adjective).* Relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief.

AI and the Erosion of Authorship Literature has long served as a reflection of human experience, with authorship described as the ultimate expression of free will, based on Jean-Paul Sartre's concept of freedom. According to Sartre, writing is an act of freedom—a deliberate process through which authors reflect their era and assert their Nietzschean "will to power." Consider a sonnet mimicking Shakespeare's style: while formally indistinguishable, its lack of historical and personal context exposes a hollow replication. This contrast underscores Sartre's view of writing as an act of free will—a conscious engagement with one's time and self.

AI as an Interpretive Partner: Augmenting Human Insight While an AI cannot replicate the existential weight of a Shakespearean sonnet, its interpretive capabilities can deepen our understanding of it. In the humanities, this is exemplified by the field of distant reading, where machine learning algorithms analyze thousands of texts at once to uncover stylistic trends, thematic evolution, and previously unseen literary influences. In this capacity, the AI is not the author; it is an analytical engine that empowers the human scholar. It interprets data at a scale no human could manage, providing insights that sharpen, rather than replace, our own critical judgment. The scholar's agency remains central—they pose the questions, interpret the results, and craft the narrative—but their perception is now augmented by a powerful interpretive lens.

The Perils of Uncritical Interpretation in the Post-Truth Era This distinction is critical in the post-truth age, where AI-generated texts can be misattributed or used to spread misinformation. A fabricated Shakespearean sonnet, labeled as authentic, undermines the historical truth of the author's

context and intent. The proliferation of such content on platforms like X, where unverified posts can go viral, further erodes trust in cultural artifacts. AI's capacity to produce convincing forgeries risks a similar erasure of authentic human narratives, replacing them with algorithmically generated facsimiles that lack soul or purpose.

Philosophical Reflections on Agency The promise of AI as an interpretive partner, however, hinges on maintaining human accountability. The distinction between AI-assisted insight and blind algorithmic deference is critical, particularly in the post-truth age. A fabricated Shakespearean sonnet, if presented by an AI tool as an analytical 'finding' of a lost work, could be amplified on platforms like X, undermining historical truth. Here, the machine's interpretation, devoid of context and intent, becomes a vector for misinformation.

This danger is magnified in medicine. The true ethical challenge is not that an AI generating a radiological report is 'lying,' but that it is indifferent to the truth. It reacts to input images and statistical correlations, lacking the physician's ethical commitment to the patient's well-being. A medical doctor crafting a radiological report draws on a complex interplay of past records, patient history, and clinical intuition. This decision-making process reflects a form of human judgment that integrates empathy within our Homo sapiens species, experience of era, and ethical responsibility. In contrast, an AI trained to generate radiological reports reacts to input images, producing outputs that may mimic human reports with alarming accuracy. While such systems promise efficiency, they lack the holistic reasoning and moral accountability of a human physician.

In the post-truth era—where emotional narratives often override facts—the uncritical use of AI in medicine mirrors broader societal shifts. The deference to algorithmic authority risks marginalizing human expertise, particularly when trust is most vital. This mirrors the post-truth tendency to favor expedient narratives over rigorous evidence, as seen in public health debates where misinformation about vaccines or treatments has flourished. The absence of human will in AI diagnostics threatens to dehumanize medicine, reducing patients to data points and eroding the trust essential to healthcare.

Towards an Ethical AI Future AI's lack of agency—its inability to form intentions or assign meaning—stands in stark contrast to Sartre's existential notion of freedom. This becomes ethically problematic when algorithms, indifferent to truth, influence real-world decisions without moral oversight. AI, by contrast, operates within the constraints of its programming and training data, reacting to inputs without consciousness or intent. This limitation becomes especially perilous in the post-truth age, where AI can amplify emo-

tionally charged falsehoods. This dynamic fuels echo chambers, where users are trapped in cycles of confirmation bias, further distancing them from objective reality.

The philosophical critique extends to the ethical responsibilities of AI developers. In a post-truth world, the deployment of AI without robust safeguards risks complicity in the spread of misinformation. Unlike deliberate lies, AI-generated misinformation often stems from indifference to accuracy, as algorithms prioritize output over veracity. This indifference undermines the epistemic foundations of society, making it harder to distinguish truth from fiction.

Contemporary debates about AI underscore its dual role as both tool and threat. The development of generative AI, capable of producing deepfakes, fake texts, and synthetic media, has heightened fears of existential risks. In the post-truth age, these technologies can manipulate public perception on an unprecedented scale, as seen in cases of AI-generated propaganda or fabricated election-related content. Dystopian narratives, from Philip K. Dick to contemporary media critiques, illustrate the perils of a society where artificial constructs blur the lines between reality and simulation. In such contexts, the erosion of truth is not just philosophical—it becomes institutional.

Efforts to mitigate these risks include calls for stricter regulation, transparency in AI development, and media literacy initiatives. However, the post-truth age complicates these efforts, as public trust in regulatory bodies and experts is already eroded. The challenge lies in balancing innovation with accountability, ensuring that AI serves humanity without becoming a tool for deception or control.

Conclusion Across literature, medicine, and public discourse, the rise of AI forces a critical distinction between authorship and analysis, and between agency and interpretation. Its lack of intentionality makes it a poor author but a potentially powerful analytical partner. In the post-truth era, the primary danger lies in conflating these roles. In literature, AI threatens cultural heritage when it is presented as a creator, yet it can enrich it when used as an interpretive tool. In medicine, it risks dehumanizing care when treated as an autonomous authority, yet it can enhance it when deployed as a vigilant diagnostic assistant.

Mitigating these risks requires more than just technological safeguards; it demands a reaffirmation of human-centered purpose. The challenge is not to build AI that can 'think' like us, but to design systems that augment our own capacity for judgment, responsibility, and the pursuit of truth. By harnessing AI as a tool for interpretation while insisting on human accountability, we can leverage its power without succumbing to the soulless automation that defines

the dystopian post-truth world.

Mehmet Can Yavuz, PhD *is an Assistant Professor of Machine Learning at Işık University, part of the Feyziye Mektepleri Schools Foundation. His research spans art, vision, sound, language modalities, and biomedical imaging, with a distinctive integration of philosophical inquiry. He holds a Ph.D. in Computer Science and Engineering from Sabancı University, an M.Sc. in Physics from Boğaziçi University, and dual B.Sc. degrees in Electrical & Electronics Engineering and Physics from Işık University. In 2025, he completed a postdoctoral fellowship in the Department of Radiology at UCSF.*

Dr. Tician Schnitzler *is a radiologist and postdoctoral research fellow at the University of California, San Francisco (UCSF), specializing in thoracic imaging and artificial intelligence. His research focuses on developing and validating machine learning algorithms for lung disease detection and risk stratification, with applications in interstitial lung abnormalities, lung cancer, and bronchiectasis. Dr. Schnitzler earned his M.D. from the RWTH University of Aachen and is currently completing a master's degree in Biomedical Informatics and Data Science at the University of Mannheim. In July 2025, he will return to Switzerland to continue his research and clinical work at the Cantonal Hospital Aarau (KSA), where he leads studies on imaging biomarkers and AI-based diagnostic tools in pulmonary medicine.*

References

Oxford Languages. (2016). Word of the year 2016 [Accessed: 2025-05-29]. https://languages.oup.com/word-of-the-year/2016/.

Risks of ChatGPT in Medical Decision-Making

Ranjana Roy Chowdhury

Synopsis. While ChatGPT has shown promise in tasks like summarizing medical notes and assisting with documentation, its use in clinical decision-making poses significant risks. Multiple studies have demonstrated that ChatGPT-3.5 and even GPT-4 frequently make diagnostic errors, offer outdated or hallucinated treatments, and lack consistency in triage or symptom interpretation. For example, ChatGPT misdiagnosed a transient ischemic attack, suggested unsafe cancer therapies, and varied in risk assessments for identical patient inputs. These issues stem from limitations such as training on outdated data, inability to access live clinical records, and a lack of transparent reasoning. Additionally, the model may perpetuate biases and cannot be held accountable for its outputs. While it can pass medical exams like the USMLE at a basic level, this does not equate to real-world clinical competence. Experts caution against using ChatGPT for unsupervised diagnosis, treatment planning, or triage, emphasizing that it should only be deployed with strict oversight and for non-critical tasks. Without regulatory approval and rigorous validation, using ChatGPT in high-stakes medical contexts could compromise patient safety.

Recent advances in large language models (LLMs) like ChatGPT have triggered widespread interest in their potential application across the medical field, but numerous studies and real-world examples demonstrate that ChatGPT, despite its language fluency, can lead to dangerously incorrect assessments in clinical contexts. For instance, a case report by Morrison et al. (Morrison et al., 2023) described a 63-year-old patient experiencing visual disturbances post-cardiac ablation who consulted ChatGPT-3.5, which falsely reassured him that the symptoms were harmless side effects of anesthesia, leading to delayed hospital presentation; later investigations confirmed a transient ischemic attack (TIA), highlighting the life-threatening risk of such AI-generated mis-

information. Studies in diagnostic accuracy reveal similarly concerning trends: Gilson et al. (Gilson et al., 2023) evaluated ChatGPT-3.5's performance on 100 published pediatric clinical cases and found an alarming 83% diagnostic error rate, with only 17 correct final diagnoses, underlining the inadequacy of ChatGPT for autonomous clinical reasoning. Comparable assessments of ChatGPT-4 show modest improvements but still fall short, with one study finding it correctly diagnosed just 39% of *New England Journal of Medicine* clinical cases (Bordes et al., 2023). In neurology, ChatGPT has demonstrated inconsistency when asked to differentiate between Alzheimer's disease and mild cognitive impairment, and these disparities in answer quality especially across repeated prompts—further reduce trust in the model's medical utility (Dickinson et al., 2023).

ChatGPT also suffers from hallucinations, where it fabricates information that appears plausible; for example, in an oncology-focused study, researchers asked ChatGPT-3.5 to provide treatment options for breast, prostate, and lung cancer scenarios, only to find that while it often mentioned guideline-concordant therapies, 34% of its responses included at least one non-concordant recommendation and 12.5% fabricated treatments entirely (Lee et al., 2023). Such mixed valid-invalid responses are dangerous, especially for non-expert users or clinicians unaware of current guidelines.

In patient triage tasks, ChatGPT similarly fails to deliver consistent output: a 2023 *PLOS One* study assessed ChatGPT-4's consistency in chest pain risk stratification and found that using the same patient inputs, the model's risk category varied across runs nearly 45–48% of the time, contradicting standard TIMI and HEART scores (Rao et al., 2023). These inconsistencies reflect ChatGPT's stochastic text generation mechanism, making it unsuitable for reliable triage, where reproducibility and adherence to evidence-based criteria are critical. Furthermore, ChatGPT's training data only extend up to September 2021, meaning it lacks awareness of recent medical guidelines, treatments, and global health developments such as post-Delta COVID-19 variants or emerging therapies. Its knowledge base cannot be updated in real-time unless explicitly fine-tuned by OpenAI or other developers, and even newer models like GPT-4 inherit this structural limitation.

Another key problem is the lack of accountability: ChatGPT cannot cite the source of its recommendations, explain its reasoning transparently, or adapt based on feedback. This lack of explainability makes it a black box system, which in healthcare is a serious drawback, especially when dealing with life-altering or high-risk decisions. Moreover, studies have found that ChatGPT can encode and amplify social and clinical biases. For example, in triage simulations, the model failed to incorporate demographic-specific

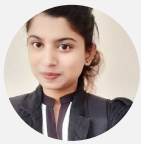
risks unless explicitly prompted, potentially perpetuating disparities in care for underrepresented groups. A recent Stanford study tested whether ChatGPT-4 could improve adherence to guidelines in physician triage decisions and found that while performance modestly improved, the model's outputs still introduced new sources of bias and variance (Tu et al., 2023). Furthermore, hallucinations are not rare edge cases but a persistent issue: a dermatology preprint showed that ChatGPT-4 fabricated explanations for skin conditions in nearly 25% of test prompts, often confidently presenting wrong associations or treatment steps (Zhang et al., 2023).

In education, ChatGPT can pass multiple-choice exams like the USMLE, but this does not equate to clinical competence. For example, Kung et al. (Kung et al., 2023) reported that GPT-3.5 scored 66–71% on USMLE Step 1, Step 2 CK, and Step 3-style questions—indicating it still misses 30–35% of core medical knowledge, which is unacceptable for real-world use. Additionally, ChatGPT cannot access or interpret real-time data from electronic health records (EHRs), lab tests, or imaging. This inability to interact with live data streams fundamentally limits its integration into clinical workflows beyond administrative support or education. A critical risk is that overreliance on ChatGPT could lead clinicians to outsource clinical judgment, potentially fostering complacency and eroding clinical intuition. In turn, patients may treat ChatGPT as a substitute for professional care, especially in settings with limited healthcare access. Regulatory and ethical concerns are also paramount: ChatGPT is not FDA-approved for any medical purpose, and no clear legal framework currently governs its use in patient interactions, raising liability issues for institutions that deploy it in unsupervised contexts.

Researchers have proposed guardrails, including real-time physician oversight, model explainability requirements, AI literacy training, and use of ChatGPT only within sandboxed environments for defined tasks. The most responsible uses of ChatGPT in healthcare remain narrow: generating draft letters, summarizing discharge instructions, or supporting patient communication, where outputs are non-critical and always reviewed by humans. A comprehensive 2023 review in *Radiology AI* concluded that no studies to date endorse ChatGPT for independent clinical use, emphasizing its tendency toward superficial answers, lack of nuanced reasoning, and domain-agnostic logic (Huang et al., 2023).

In conclusion, while ChatGPT is a powerful language tool with potential auxiliary uses in medical education, documentation, and communication, its current architecture, training data limitations, and lack of factual grounding render it fundamentally unfit for autonomous medical decision-making. Misdiagnoses, unsafe treatment recommendations, triage inconsistency, and

hallucinations are not hypothetical risks they are well-documented realities across peer-reviewed studies, expert commentaries, and clinical case reports. To protect patient safety, ChatGPT should not be used in any unsupervised diagnostic, therapeutic, or triage capacity until these limitations are systematically addressed and rigorously validated through clinical trials and regulatory review.



Ranjana Roy Chowdhury is a Ph.D. scholar in the Department of Computer Science and Engineering at IIT Ropar. Her research focuses on Meta-Learning and Few-Shot Learning, with a particular emphasis on advancing techniques in Medical Image Analysis. She is highly inquisitive and actively seeks opportunities to engage and collaborate with researchers on innovative projects.

References

Bordes, A., et al. (2023). Evaluation of chatgpt on us clinical vignettes from nejm. *NEJM AI*.

Dickinson, L. M., et al. (2023). Chatgpt's variable responses to alzheimer's disease scenarios. *Journal of Alzheimer's Disease Reports*.

Gilson, A., et al. (2023). How well can chatgpt-3.5 diagnose pediatric cases? *JAMA Pediatrics*.

Huang, M., et al. (2023). The risks and limitations of chatgpt in radiology. *Radiology: Artificial Intelligence*.

Kung, T. H., et al. (2023). Performance of chatgpt on the usmle. *PLOS Digital Health*, *2*(2).

Lee, H., et al. (2023). Oncology decision-making with chatgpt: Promise and pitfalls. *Cancer Medicine*.

Morrison, M. L., et al. (2023). Chatgpt and the illusion of knowledge in clinical reasoning. *BMJ Case Reports*.

Rao, N., et al. (2023). Chatgpt's consistency in chest pain triage decisions. *PLOS One*.

Tu, S., et al. (2023). Trust and bias in chatgpt-assisted clinical decision support. *Stanford HAI Reports*.

Zhang, L., et al. (2023). Hallucination of dermatological findings by chatgpt. *medRxiv*.

How Artificial Intelligence Is Transforming Cancer Diagnosis: A Pathological Perspective

Abdulkadir Albayrak and Mehmet Siraç Özerdem

Synopsis. Cancer continues to be a leading cause of death worldwide, and there is a critical need for timely, accurate diagnosis. Pathology, particularly through microscopic examination of tissue samples, is the gold standard for determining cancer type, grade, and progression. Pathology has undergone a significant transformation in recent years with the rise of digital pathology, which enables the digitization, analysis, and sharing of high-resolution tissue images. This shift has not only increased the speed and accuracy of diagnosis, but has also paved the way for integration with advanced technologies such as artificial intelligence (AI). Early digital pathology relied on hand-crafted features and classical machine learning methods; however, the advent of deep learning, particularly convolutional neural networks (CNNs), has dramatically improved tissue classification and tumor detection. Now, the introduction of foundation models—AI systems trained on large amount of data—offers even greater generalization and adaptability across a variety of diagnostic tasks. Looking ahead, AI-powered Computer Aided Diagnosis (CADx) systems are expected to play a supporting role for pathologists by identifying areas of interest, comparing patient slides to large databases, and providing visual decision aids such as heat maps. These technologies will not only help make more precise and effective diagnoses, but will also improve personalized treatment planning, medical education, and collaborative research. AI-powered digital pathology represents a significant advance in the fight against cancer.

”Cancer” and ”Pathology” are the words most frequently encountered by researchers conducting research or people who follow developments in the field of health. Cancer is one of the leading diseases causes death in almost all developed and underdeveloped countries (Bray et al., 2024). Thus, early

diagnosis and access to the correct treatment play a critical role in determining the course of the disease. In this context, pathology discipline stands out as the gold standard in diagnosis of cancer. Pathology is a medical discipline that aims to diagnose diseases through microscopic examination of tissues, cells, and organs. It provides diagnostic information to patients and clinicians. Vital information such as the type, degree, aggressiveness (spreading potential), and characteristics of the tumor are obtained as a result of detailed examinations performed by pathology specialists in a laboratory environment. Therefore, the first and most important step in an effective fight against cancer is a reliable pathological diagnosis (Majno & Joris, 2004).

Digital Pathology is the practice of digitizing glass slides of tissue samples to create high-resolution images that can be viewed, analyzed, managed, and shared with other specialists digitally. So, digital pathology can be thought an innovative discipline that enables diagnostic processes to become faster, more objective (by adding more people in the loop for evaluation) and more accessible by transferring pathological analyses to digital platforms so that it can be also used for educational purposes (Kiran et al., 2023). These are the benefits of digital pathology that first come to mind. In addition, the ability to apply advanced imaging and image processing techniques may be a secondary benefit. With the development of imaging technology over time and some advanced image processing techniques have given rise to the idea that digital pathology can be used as a secondary decision support system. With the rapid developments over time from conventional image processing techniques to deep learning-based methods, processing of whole digital pathology slides, segmentation and classification of tissue regions on these slides are now commonly performed operations. Deep learning algorithms developed in recent years, especially through CNN, can identify tumor regions in histopathological images with high confidence score (Litjens et al., 2018). These systems provide support to pathologists in the diagnosis of common tumors such as breast, prostate and lung cancer, reducing the margin of error and shortening the diagnosis time.

Artificial intelligence (AI) algorithms and applications in digital pathology have undergone significant evolution in the last two decades, from traditional machine learning methods to deep learning methods and most recently to foundation models (Bilal et al., 2025). Early digital pathology applications often relied on hand-crafted feature extraction algorithms or classical machine learning algorithms. Gray-level co-occurrence matrix (GLCM), local binary patterns (LBP), scale invariant feature transform (SIFT), and speeded-up robust features (SURF) were the methods mainly utilized to capture critical texture and structural patterns for accurate tissue classification in the first proposed study in digital pathology (De Matos et al., 2021). In hand-crafted

feature extraction algorithms, human expertise is used to design features that capture important characteristics of the given histopathological image for classification task. Extracting features using these algorithms often requires deep domain expertise and significant trial and error. However, these methods may be disadvantageous approaches because they work for one problem and cannot be generalized to other problems.

it has become possible to automatically learn complex and abstract patterns in digital pathology images without human intervention by using some variety of deep learning algorithms, especially convolutional neural networks (CNN). This provides relatively higher accuracy, generalization ability and superior performance on large data sets, with less preprocessing required. Additionally, by learning directly from data, deep learning models can capture microscopic details that classical methods may miss, significantly increasing the diagnostic power and clinical value in histopathological analyses (Janowczyk & Madabhushi, 2016).

In recent years, foundation models that are trained on very large data sets and have a high generalization capacity have begun to be proposed. (Biswas, 2023) (Wang et al., 2022) (Kirillov et al., 2023). These models can then be used for fine-tuning specific downstream tasks and yield very successful results. They have the ability to generalize and can be more efficient than full-training from scratch due to their millions or billions of parameters. Their main difference from training from scratch is that they are not trained for a specific task and can be used for many different downstream tasks.

Nowadays, there is a need for a "Computer Aided Diagnosis (CADx)" system that can be used as a secondary decision support system in order to prioritize the needs of patients in the analysis of pathological images, which are difficult and time-consuming task for specialists. Because the specialist works very hard during the day and can make serious mistakes in detecting the tumor due to fatigue. These secondary decision support systems can basically mark risky areas in high-resolution images and allow the expert to focus on these areas, thus minimizing the potential for errors. The process steps in this system can be as follows: First, the automatic separation of the background information and the region of interest in the image given as input to the system, then analyze the region of interest and support the expert with various heat maps regarding different regions of the tissue. In this way, the expert will evaluate based on experts' own personal experience and also review the suspicious areas that the CADx system recommends to be checked, thus performing a nearly flawless operation. The system will also be able to help the expert to compare the input image with the images in the database by querying the images that have the closest patterns to the relevant image among hundreds of thousands or even

millions of images in the database.

Dr. Abdulkadir Albayrak is Senior Data Science Analyst specializing in computational pathology and generative AI to support hospital operations at Mayo Clinic, Minnesota, USA. He holds a Ph.D. in Computer Engineering from Yildiz Technical University, where he focused on analyzing histopathological images using machine learning. He has extensive experience in developing and applying advanced machine learning and deep learning techniques, particularly in generative AI and large-language models. He has also an Associate Professorship title at Dicle University, where he taught computer science courses and published numerous papers.

Dr. Mehmet Siraç Özerdem received the B.Sc. degree in Electrical-Electronics Engineering in 1994 from Eastern Mediterranean University in Northern Cyprus. He received his M.Sc degree in Electrical Engineering in 1998 from Yıldız Technical University. He received the Ph.D. degree in Computer Engineering in 2003 from Istanbul Technical University. Currently, he works as Professor and researcher at Dicle University, Diyarbakır, Türkiye. His research focuses on machine learning, biomedical signal processing, deep learning, and embedded system design. His ongoing projects are related to machine learning detection of brain tumors and classification of blood diseases from blood samples.

References

Bilal, M., Raza, M., Altherwy, Y., Alsuhaibani, A., Abduljabbar, A., Almarshad, F., Golding, P., Rajpoot, N., et al. (2025). Foundation models in computational pathology: A review of challenges, opportunities, and impact. *arXiv preprint arXiv:2502.08333*.

Biswas, S. S. (2023). Role of chat gpt in public health. *Annals of biomedical engineering*, *51*(5), 868–869.

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *74*(3), 229–263.

De Matos, J., Ataky, S. T. M., de Souza Britto Jr, A., Soares de Oliveira, L. E., & Lameiras Koerich, A. (2021). Machine learning methods for histopathological image analysis: A review. *Electronics*, *10*(5), 562.

Janowczyk, A., & Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, *7*(1), 29.

Kiran, N., Sapna, F., Kiran, F., Kumar, D., Raja, F., Shiwlani, S., Paladini, A., Sonam, F., Bendari, A., Perkash, R. S., et al. (2023). Digital pathology: Transforming diagnosis in the digital age. *Cureus*, *15*(9).

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., Van de Loo, R., Vogels, R., et al. (2018). 1399 h&e-stained sentinel lymph node sections of breast cancer patients: The camelyon dataset. *GigaScience*, *7*(6), giy065.

Majno, G., & Joris, I. (2004). *Cells, tissues, and disease: Principles of general pathology*. Oxford University Press.

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., & Han, X. (2022). Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, *81*, 102559.

AI in Medical Robotics: Enhancing Surgical Precision and Access, Raising Ethical and Regulatory Challenges, and Redefining Human-AI Collaboration

Ramy A. Zeineldin

Synopsis. This essay explores how AI is transforming medical robotics, especially in surgical applications, by enhancing precision, efficiency, and access. It also addresses the ethical and regulatory issues emerging with these technologies and redefines human-machine collaboration in clinical environments.

Enhancing Surgical Precision and Patient Outcomes

Artificial Intelligence (AI)-assisted robotic surgery represents one of the most promising applications of machine learning and automation in medicine. Studies have shown that AI-enhanced robotic-assisted surgery (RAS) can significantly improve surgical outcomes compared to traditional methods. For instance, (Nwoye et al., 2023) found that AI integration reduces operating time, docking time, and estimated blood loss during procedures, contributing to faster recovery and reduced hospital stays.

Beyond procedural efficiency, AI also enhances postoperative care. (Varghese et al., 2024) highlight how AI contributes to better patient outcome prediction, surgical education, and real-time decision support. Combined with augmented reality (AR), AI enables surgeons to visualize anatomical structures in real-time (Zhang et al., 2022), increasing precision and reducing complications.

Moreover, AI-driven tool detection and motion analysis offer valuable

insights into surgical techniques. (Moglia et al., 2021) note that while current AI models can recognize surgical phases and predict operative duration, identifying tasks directly linked to patient outcomes remains a challenge.

Ethical Risks, Bias, and Accountability in Autonomous Systems

AI in medical robotics raises ethical concerns around bias, transparency, and accountability. Algorithmic bias, stemming from unrepresentative datasets, may disproportionately affect underrepresented populations. Without external validation and reproducibility, clinical adoption is hindered (Moglia et al., 2021). Additionally, growing autonomy in surgical robots leads to accountability questions. Studies show robots can perform operations with high precision (Saeidi et al., 2022; Yang et al., 2022), but concerns remain regarding their ability to handle complications and who is liable for errors (Mughal, 2022). (Fosch-Villaronga et al., 2022) argue that legal and institutional structures must evolve to support such automation, requiring new definitions of responsibility and updated training.

Redefining Human-AI Collaboration in Healthcare

AI's most profound impact may be in redefining human-machine collaboration. In documentation, Clinical Documentation Integrity Specialists collaborate with AI to improve patient records (Bossen & Pine, 2023). AI-in-the-loop systems like HAILEY enhance empathetic peer-support conversations (Sharma et al., 2023). In diagnostics, human-AI teams outperform either alone through opinion integration (Reverberi et al., 2022), though explainability can sometimes confuse rather than assist (Salvi et al., 2025). Designing effective collaboration protocols remains crucial.

In conclusion, AI in medical robotics brings both transformative potential and complex challenges. Interdisciplinary collaboration will be key to ensuring its ethical, effective, and inclusive integration into healthcare.

Ramy A. Zeineldin is a postdoctoral fellow at FAU Erlangen-Nürnberg's SPARC Lab, specializing in AI for medical applications. He earned his Dr.-Ing. with summa cum laude from Karlsruhe Institute of Technology and previously lectured at Menoufia University, Egypt, where he also completed his B.Sc. and M.Sc. His research spans surgical robotics, medical imaging, explainable AI, and human-AI collaboration.

References

Bossen, C., & Pine, K. H. (2023). Batman and robin in healthcare knowledge work: Human-ai collaboration by clinical documentation integrity specialists. *ACM Transactions on Computer-Human Interaction*, *30*(2), 1–29. https://doi.org/10.1145/3569892

Fosch-Villaronga, E., Khanna, P., Drukarch, H., & Custers, B. (2022). The role of humans in surgery automation. *International Journal of Social Robotics*, *15*(3), 563–580. https://doi.org/10.1007/s12369-022-00875-0

Moglia, A., Georgiou, K., Georgiou, E., Satava, R. M., & Cuschieri, A. (2021). A systematic review on artificial intelligence in robot-assisted surgery. *International Journal of Surgery*, *95*. https://doi.org/10.1016/j.ijsu.2021.106151

Mughal, H. (2022). Ethical challenges facing autonomous surgical robots. *British Journal of Surgery*, *109*(Supplement$_1$). https://doi.org/10.1093/bjs/znac039.205

Nwoye, E., Woo, W. L., Gao, B., & Anyanwu, T. (2023). Artificial intelligence for emerging technology in surgery: Systematic review and validation. *IEEE Reviews in Biomedical Engineering*, *16*, 241–259. https://doi.org/10.1109/rbme.2022.3183852

Reverberi, C., et al. (2022). Experimental evidence of effective human–ai collaboration in medical decision-making. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-18751-2

Saeidi, H., et al. (2022). Autonomous robotic laparoscopic surgery for intestinal anastomosis. *Science Robotics*, *7*(62). https://doi.org/10.1126/scirobotics.abj2908

Salvi, M., et al. (2025). Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare. *International Journal of Medical Informatics*, *197*. https://doi.org/10.1016/j.ijmedinf.2025.105846

Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., & Althoff, T. (2023). Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, *5*(1), 46–57. https://doi.org/10.1038/s42256-022-00593-2

Varghese, C., Harrison, E. M., O'Grady, G., & Topol, E. J. (2024). Artificial intelligence in surgery. *Nature Medicine*, *30*(5), 1257–1268. https://doi.org/10.1038/s41591-024-02970-3

Yang, S., Chen, J., Li, A., Li, P., & Xu, S. (2022). Autonomous robotic surgery for immediately loaded implant-supported maxillary full-arch prosthesis: A case report. *Journal of Clinical Medicine*, *11*(21). https://doi.org/10.3390/jcm11216594

Zhang, D., Si, W., Fan, W., Guan, Y., & Yang, C. (2022). From teleoperation to autonomous robot-assisted microsurgery: A survey. *Machine Intelligence Research*, *19*(4), 288–306. https://doi.org/10.1007/s11633-022-1332-5

99% for whom? The Hidden Cost of Exclusion in AI

Mary-Brenda Akoda

Synopsis. Artificial Intelligence is often measured in metrics like "99% accuracy," yet these figures conceal a fundamental question: accuracy for whom? This essay explores how AI systems trained on unrepresentative datasets can fail entire populations, particularly in healthcare, criminal justice, and facial recognition. Through documented case studies and regulatory analysis, it argues for a shift toward representational equity, demographic transparency, and people-centred AI development.

Artificial Intelligence is often presented with numbers — 98% sensitivity, 99% accuracy — as if these figures are universal truths. But too few ever ask: 99% for whom? These performance claims, while impressive on the surface, often conceal a sobering reality: the models are trained, validated, and celebrated on datasets that rarely reflect the diversity of real-world populations. And in that silence, entire communities are excluded — not only from the data, but from the benefits of the technology. Worse, they often suffer its failures.

I witnessed this firsthand in Nigeria, where I led the country's first AI research initiative for diabetic retinopathy (DR) detection. DR, a major cause of preventable blindness among working-age adults, often goes undiagnosed in low- and middle-income countries due to limited screening capacity. Due to the initial lack of local data, we trained deep learning models on globally recognised datasets. On paper, the results were outstanding: one model correctly identified 98.9% of referable cases requiring urgent treatment and achieved 96.2% specificity for healthy cases.

However, once we gathered a small dataset of retinal images from Nigerian patients and tested the same models, performance dropped sharply. Sensitivity fell as low as 43.8%, and in some cases, results were worse than random guessing

(Akoda & Nkanga, 2024). The issue was not model complexity or algorithmic sophistication. It was representational mismatch. The model had learned patterns from images that looked nothing like those of Nigerian patients.

This is not a one-off failure. It is part of a systemic pattern. From criminal justice to finance, AI systems trained on biased or incomplete datasets routinely replicate — and often amplify — structural inequalities. In the U.S., an algorithm used for criminal sentencing was nearly twice as likely to falsely label Black defendants as high-risk compared to white counterparts (Angwin et al., 2016). In healthcare, an algorithm underestimated the needs of Black patients because it equated health status with historical healthcare spending — ignoring deep disparities in access (Obermeyer et al., 2019).

Facial recognition systems fare no better. A study by the U.S. National Institute of Standards and Technology (NIST) revealed that false positive rates in some algorithms were up to 100 times higher for West African and East Asian faces than for Eastern European ones (Grother et al., 2019). These failures are not just technical bugs. They reflect whose lives, faces, and needs were considered worth including in the training data.

What's alarming is how quickly such flawed systems are adopted. A 2024 scoping review of 692 FDA-approved AI-based medical devices revealed that only 3.6% reported race or ethnicity in their training data, and over 99% failed to mention socioeconomic variables (Muralidharan et al., 2024). Age breakdowns and subgroup performance metrics were also rarely provided. In essence, many AI tools in healthcare are approved and deployed without any public understanding of whom they were tested on — or whom they may fail.

The way these tools are marketed only deepens the problem. A 98% accuracy may apply to one subgroup and 60% to another, but rarely is that disclosed. Meanwhile, communities whose languages, images, or demographics are absent from the data encounter systems that fail them. That failure — repeated and unexplained — leads to a deeper erosion of trust.

Developers are not always to blame. The datasets available for public use are themselves skewed. But that does not absolve us from responsibility. When researchers and engineers do not ask who is missing from the data, we are not building intelligence. We are building indifference.

So what can be done?

First, representational equity must become a minimum standard, not a bonus. Training data must be diverse from the start. AI systems should be stress-tested across ethnicities, genders, ages, and socioeconomic groups. Subgroup performance metrics should be a required part of all publications

and product releases.

Second, regulation must catch up. Agencies like the FDA should mandate demographic transparency in model approvals, including explicit reporting of who was in the training and test sets, and how the model performs across groups.

Third, especially in the Global South, we must address the foundational issue of data scarcity. In our diabetic retinopathy work, the real challenge was not the model — it was the lack of representative data. We have since begun collaborating with local clinicians and hospitals to build datasets that better reflect our population. In a different project focused on language, we're compiling African lexicographic data to support more inclusive language models.

Finally, we need a cultural shift in how we measure progress in AI. Instead of asking whether a model is state-of-the-art, we must ask: Who does it work for? Who does it leave behind? Because that answer matters more than the number itself.

Mary-Brenda Akoda is an AI for Healthcare researcher and Google DeepMind Scholar at Imperial College London's I-X, where she develops deep generative AI models for medical imaging. She led Nigeria's first AI research for diabetic retinopathy detection and, as CTO of Nkanda, co-created Africa's most comprehensive bilingual dictionary app to advance language inclusion in AI and tech. A former researcher at Microsoft's Mixed Reality & AI Research Lab in Cambridge and instructor of the top Udemy course ChatGPT Masterclass for Programmers and Software Engineers, she has been recognised as a UK Global Talent (Tier 1 Exceptional Talent) and honoured by organisations including Google, Mastercard Foundation, Fetch.ai, and Translated for her contributions to AI, health equity, inclusive innovation, and mentorship of underrepresented youth in AI and software engineering.

References

Akoda, M.-B., & Nkanga, D. G. (2024). Advancing health access and equity: Artificial intelligence for diabetic retinopathy grading in nigeria. *Transactions of the Ophthalmological Society of Nigeria*, *8*(1), 79–82. https://tosn.org.ng/index.php/home/article/view/238/243.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Grother, P., Ngan, M., & Hanaoka, K. (2019). *Face recognition vendor test (frvt) part 3: Demographic effects* (tech. rep. No. NIST IR 8280). NIST. https://doi.org/10.6028/NIST.IR.8280

Muralidharan, V., Adewale, B. A., Huang, C. J., et al. (2024). A scoping review of reporting gaps in fda-approved ai medical devices. *npj Digital Medicine*, *7*, 273. https://doi.org/10.1038/s41746-024-01270-x

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Part IV

Business, Human & Other

Rethinking Entrepreneurial Ecosystem Measurement in Africa: The Transformative Role of AI

Yosra Mani

Synopsis. This essay critically examines the limitations of the Africa Entrepreneurial Ecosystem Index (AEEI) in capturing the dynamics of entrepreneurship across the continent. It highlights how artificial intelligence (AI) can offer more inclusive, context-sensitive, and data-rich alternatives by mapping informal economies, reducing bias, and supporting localized policymaking.

In 2025, the launch of the Africa Entrepreneurial Ecosystem Index (AEEI) sparked strong debate. (Stam et al., 2025) presents the conceptual, methodological, and empirical foundations of the Africa Entrepreneurial Ecosystem Index (AEEI), offering a significant step forward in capturing and analyzing entrepreneurial ecosystems across the continent. The AEEI represents a new and useful tool to measure entrepreneurship in Africa. However, critics like (Naudé, 2025) and (Coad et al., 2025) said it had major problems—they called it biased, poorly designed, and too focused on European ideas. This debate shows a bigger problem: traditional ways of measuring entrepreneurship often miss the unique and informal ways business works in Africa. But What if AI could help fix that? AI could offer a new path—not just to improve old tools, but to completely change how we measure and support inclusive entrepreneurship in Africa.

The Core Problem: Traditional Metrics vs. African Realities

Over the past two decades, entrepreneurship research has increasingly focused on the systemic and contextual factors that shape entrepreneurial activity, giving rise to the Entrepreneurial Ecosystem (EE) framework (Acs et al., 2017;

Cavallo et al., 2019; McMullen, 2018; Stam, 2015; Wurth et al., 2023). Rooted in the environmental approach to entrepreneurship, this literature highlights how external factors can either foster or hinder entrepreneurial behavior (Lee & Peterson, 2000). Yet, applying these frameworks to the African context reveals a deep disconnect between measurement tools and entrepreneurial realities on the ground. Despite being hailed as the world's most entrepreneurial continent (Economist, 2025), Africa remains dominated by necessity-driven entrepreneurship, often disconnected from innovation or productivity growth (Herrington & Coduras, 2019). This paradox underscores the need to shift the focus toward productive entrepreneurship—that which contributes meaningfully to economic development (Baumol, 1993). However, understanding and fostering such entrepreneurship is hampered by the absence of reliable, harmonized, and context-sensitive data across African economies.

Traditional indices—such as the Global Innovation Index, the World Bank's discontinued Ease of Doing Business Index, or the Global Entrepreneurship Development Index—offer limited value in capturing the complexities of low- and middle-income contexts (**ref27**; Lall, 2001). In this context, the Africa Entrepreneurial Ecosystem Index (AEEI) (Stam et al., 2025) attempts to provide a regional benchmark by measuring seven pillars across 29 African countries. However, the AEEI is itself fraught with limitations. While it claims to offer a comprehensive policy tool, it covers only 54% of African nations and heavily relies on formal economic indicators—thus excluding the informal sector that accounts for roughly 85% of the continent's employment (Cruz et al., 2025). As (Naudé, 2025), the AEEI not only suffers from methodological issues, such as high redundancy with GDP (79% correlation) and lack of subnational granularity, but also reflects conceptual blind spots, notably its neoliberal bias and neglect of colonial legacies. Empirical findings further problematize this picture: high self-employment rates are negatively correlated with GDP per capita, suggesting that much of Africa's entrepreneurship signals precarity rather than economic dynamism (Cruz et al., 2025; Henrekson & Sanandaji, 2014). (Coad et al., 2025) caution that transplanting EE thinking into Sub-Saharan Africa risks repeating past development failures by ignoring the region's structural constraints and the central role of state coordination.

How AI Can Correct the AEEI's Blind Spots

AI technologies offer a practical and context-sensitive path forward to address the AEEI's three main shortcomings: informality, spatial blindness, and data dependency.

Bridging the Informality Gap: The AEEI's reliance on formal metrics—such as venture capital or stock market performance—fails to account for Africa's informal entrepreneurial base. Here, machine learning models trained on satel-

lite imagery can detect patterns of informal economic activity (Blumenstock et al., 2015). By interpreting proxies like nighttime light intensity and market expansion, AI enables governments and development actors to track urban growth, assess economic vitality, and optimize resource allocation. These alternative data sources can enhance the financial and infrastructure pillars of the AEEI.

Capturing Subnational Realities: By aggregating data at the national level, the AEEI overlooks regional disparities. Geospatial AI can provide granular, localized insights, while Natural Language Processing (NLP) tools can extract real-time, hyperlocal entrepreneurial data from platforms like WhatsApp or Facebook, mapping informal supply chains or community-led responses (Bharti et al., 2024).

Decolonizing Data Infrastructure: The AEEI is built on indicators from Western institutions like the World Bank or IMF, reinforcing historical patterns of data colonialism (Naudé, 2025; Nguimkeu & Zeufack, 2024). In contrast, AI systems trained on African-generated data—such as informal credit histories, communal trust networks, or indigenous languages—can surface indicators that are culturally and contextually relevant. For example, in Tunisia, the AI Hub of the DOT in Tunis, in partnership with InstaDeep and the GIZ-supported digital transformation program, exemplifies this shift. By leveraging local AI engineering capabilities and infrastructure—such as the NVIDIA DGX A100 server—this initiative grounds model development in African data contexts. It enables the creation of tools that better reflect the country's linguistic and socio-economic specificities.

Toward AI-Native Ecosystem Measurement in Africa

The limitations of the AEEI reveal deeper structural challenges in how entrepreneurship is measured across the continent—often through narrow, top-down, and externally imposed metrics. These frameworks struggle to capture the complexity, informality, and diversity of African entrepreneurial landscapes. AI, however, offers more than just a technical patch; it has the potential to fundamentally reconfigure how we understand, map, and nurture entrepreneurial ecosystems. By leveraging tools such as satellite imagery, natural language processing, and alternative data analytics, AI enables the visibility of hidden dynamics—particularly informal economic activity—and facilitates hyperlocal, culturally responsive insights.

AI is emerging not only as a support tool but as a transformative infrastructure for entrepreneurship. It enhances market analysis, product development, and customer interaction (Usman et al., 2024), while also enabling more effective coordination among ecosystem actors (Roundy, 2022). Yet,

realizing this potential requires tackling real barriers—namely weak digital infrastructure and fragmented data systems (Mdladla et al., 2024). To advance inclusive and context-aware measurement, it is imperative to invest in African AI capabilities through university-based labs, data sovereignty policies, and ecosystem-sensitive legal frameworks. As (Wairegi et al., 2021) argue, shaping an African AI paradigm—rooted in local values and challenges—is essential to ensure that AI does not replicate existing inequalities but actively contributes to a fairer, more productive entrepreneurial future.

Yosra Mani is an Assistant Professor of Entrepreneurship at the University of Kairouan, Tunisia. Her research focuses on entrepreneurial ecosystems, innovation, and inclusive development in the Global South. She is currently leading a national survey to explore the dynamics and challenges of the entrepreneurial ecosystem in Tunisia. Yosra is actively involved in international academic collaborations and regularly mentors early-stage entrepreneurs. Her growing interest lies in the transformative role of AI and digital technologies to support sustainable entrepreneurship, particularly in low-income countries, and to inform more context-sensitive public policy design.

References

Acs, Z. J., Stam, E., Audretsch, D. B., & O'Connor, A. (2017). The lineages of the entrepreneurial ecosystem approach. *Small Business Economics*, *49*, 1–10.

Baumol, W. (1993). *Entrepreneurship, management and the structure of payoffs*. MIT Press.

Bharti, N. K., Chancel, L., Piketty, T., & Somanchi, A. (2024). *Income and wealth inequality in india, 1922–2023: The rise of the billionaire raj* (Working Paper No. 9). World Inequality Lab.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073–1076.

Cavallo, A., Ghezzi, A., & Balocco, R. (2019). Entrepreneurial ecosystem research: Present debates and future directions. *International Entrepreneurship and Management Journal*, *15*, 1291–1321.

Coad, A., Domnick, C., Santoleri, P., et al. (2025). Does africa need entrepreneurial ecosystems thinking? *Journal of Technology Transfer*. https://doi.org/10.1007/s10961-025-10213-x

Cruz, M., Moelders, F., Salgado, E., & Volk, A. (2025). *Estimating the number of firms in africa* (Working Paper No. 11032). World Bank.

Henrekson, M., & Sanandaji, T. (2014). Small business activity does not measure entrepreneurship. *Proceedings of the National Academy of Sciences*, *111*(5), 1760–1765.

Herrington, M., & Coduras, A. (2019). The national entrepreneurship framework conditions in sub-saharan africa. *Journal of Global Entrepreneurship Research*, *9*(1), 60.

Lall, S. (2001). Competitiveness indices and developing countries: An economic evaluation of the global competitiveness report. *World Development*, *29*(9), 1501–1525.

Lee, S. M., & Peterson, S. J. (2000). Culture, entrepreneurial orientation, and global competitiveness. *Journal of World Business*, *35*(4), 401–416.

McMullen, J. S. (2018). Organizational hybrids as biological hybrids. *Journal of Business Venturing*, *33*(5), 575–590.

Mdladla, L. T. S., Wider, W., Thanathanchuchot, T., & Hossain, S. F. A. (2024). Navigating the ai revolution: A review of the transformative strategies for economic development in africa's emerging economies. *Journal of Infrastructure, Policy and Development*, *8*(9), 5436. https://doi.org/10.24294/jipd.v8i9.5436

Naudé, W. (2025). The african entrepreneurial ecosystem index. *Journal of Technology Transfer*. https://doi.org/10.1007/s10961-025-10208-8

Nguimkeu, P., & Zeufack, A. (2024). Manufacturing in structural change in africa. *World Development*, *177* (100).

Roundy, P. T. (2022). Artificial intelligence and entrepreneurial ecosystems: Understanding the implications of algorithmic decision-making for startup communities. *Journal of Ethics in Entrepreneurship and Technology*, *2*(1), 23–38. https://doi.org/10.1108/JEET-07-2022-0011

Stam, E. (2015). Entrepreneurial ecosystems and regional policy: A sympathetic critique. *European Planning Studies*, *23*(9), 1759–1769.

Stam, E., Nkontwana, P., McDonald, R., Murenzi, R., Addo, K. A., Bayuo, B., Baah, B., Riezebos, S., & Gelissen, T. (2025). Measuring national entrepreneurial ecosystems in africa [https://ssrn.com/abstract=5254687]. *SSRN*.

Usman, F. O., Eyo-Udo, N. L., Etukudoh, E. A., Odonkor, B., Ibeh, C. V., & Adegbola, A. (2024). A critical review of ai-driven strategies for entrepreneurial success. *International Journal of Management Entrepreneurship Research*, *6*(1), 200–215.

Wairegi, M. A., Omino, M., & Rutenberg, I. (2021). Ai in africa: Framing ai through an african lens. *Communication, technologies et développement*. https://doi.org/10.4000/ctd.4775

Wurth, B., Stam, E., & Spigel, B. (2023). Entrepreneurial ecosystem mechanisms. *Foundations and Trends in Entrepreneurship*, *19*(3), 224–339.

Failures of imagination: AI and Sci-Fi

John S. H. Baxter

Synopsis. Science fiction is a tool for guiding sociotechnical development, allowing us to imagine the larger consequences of our technologies. However, science-fiction centred around AI seems to conflict with the fundamental aspect of what it means to be human, leading to a difficulty in deeply imagining positive worlds.

As with any advanced technology throughout history, much of the discourse regarding the future of AI has relied on our conceptualisation of how the world would change in response to future ubiquity. Before artificial intelligence, one concern about automation was the impact that it would have once it became perfect, displacing manufacturing in its entirety. Prior to World War II, it was the automobile and personalised transportation completely transforming cities, as conceptualised by the 1933 Chicago World's Fair. A century before that, it was the train. In some ways, these conceptualisations have since become reality. North American society has been transformed by the personal automobile, both in the architecture of its cities dominated by arterial roads and connected by highways, the expectations of particular conveniences via individualised transport, and culturally through suburbs and the Great American Road Trip. In other ways, they have yet to pass and most never will. Cities are now removing or pedestrianising these roads. The suburbs of certain cities, ironically including one previously sustained by automotive manufacturing, have become literal tinderboxes with neglect.

Nevertheless, these conceptualisations are still important to the initial implementation of new technologies and social adaptation towards them. Later changes in technology and society often arise as explicit rejections of these fictitious worlds or at least parts thereof. In that way, science fiction, the creation of these imagined worlds and the logic of their functioning, becomes

perhaps surprisingly crucial to sociotechnological development.

In her collection of essays on science fiction, *In Other Worlds* (Atwood, 2011), Margaret Atwood describes one philosophy for writing science fiction about a particular technology. Briefly put, it is to take society as it currently is, take said technology and consider it to be perfected and ubiquitous, and then let society progress from there, capturing some temporally proximal or distal part of this imagined world as a story. In *Oryx & Crake*, genetic modification has been perfected, leading not only to fantastically spliced animals but also the development of horrifying biological weapons. One can frame other famous works through this lens as well: *Brave New World* perfected medicine; *1984*, surveillance technology, *Neuromancer*, brain-computer interfaces (and to a lesser extent, AI).

The central role of intelligence in being human

Taking this philosophy seriously, what would it mean to perfect AI? The key is in the name itself; it would be a perfected intelligence beyond that of even the greatest of mere humans, leading to unexpected results. In Isaac Asimov's *The Evitable Conflict*, the AI that controls the world economy uses its superhuman foresight to bring harm to particular humans in secret as to avoid the possibility that imperfect human control leads to the end of humanity itself. This is echoed in miniature in Stanley Kubrick's film, *2001: A Space Odyssey*. In Ridley Scott's *Alien* (admittedly not AI-centric), it is not only the titular creature who is the villain, but also the android Ash who plans and carries out the secret commands of his corporate controllers with inhuman coldness. These imaginings have helped to guide areas such as AI safety and alignment research. But they nevertheless remain somehow disturbing.

The reason why these works have such a visceral effect is because active intelligence is not only something imperfect in humans but it is also something that we have used to define what it even means to be human. We are, after all, the *rational animal*[8]. To have our essence usurped by another is, unsurprisingly, deeply upsetting. In *Neuromancer*, the AI's eventually take over even our fundamental active curiosity; it is our AI who first makes contact with life from another planet, or rather, with the AI developed by said life. To add further insult, this terran AI even brings with it a digital copy of the protagonist's personality, his very humanity itself.

Although I have framed this negatively so far, it does not have to be so. If

[8] In Aristotle's *Nichomachean Ethics*, rationality was seen as the fundamental defining characteristic of a human being, delineating it within the realm of active but purely reactive animals. Later, in Kant's kingdom of ends (from *The Critique of Pure Reason*), humanity is joined in active recognition of universal moral law defined by pure rationality, centering it again, an essential characteristic of humankind.

active intellect defines what it means to be human, then a perfected AI is really no different than a perfected other. Anything we can do, it can do better. It can do anything better than us.

Polarisation and the failure of imagination

So, then the question is *what do we do* with our active intelligence when faced with the activity of a higher intelligence?

One answer, perhaps all too common in this day and age, is make money and get power. Frederic James, quoting some anonymous someone, once wrote that "it is easier to imagine the end of the world than to imagine the end of capitalism" (Jameson, 2003). In many of these imagined power-centric worlds, such as *The Matrix* of the Wachowski sisters, the end was nigh and now past, with ecological catastrophe caused by unhinged AI-guided nuclear weapons or climatological collapse once they escape corporate/governmental control. The proximate cause of the world's destruction in Elison's *I Have No Mouth and I Must Scream* was the AM super-intelligence, but the distal cause was the perverse game theory that gave AM life. In others[9], the world continues but it is tightly controlled by the malevolent corporations that control the AI that control us, almost all humans reduced to mere animals, bent on either mere survival or mere pleasure.

Anarcho-socialist visions of the future only reject the element of disaster. In these, all labour is to be performed by green AI and robots, leaving humans to explore loftier heights. But what loftier heights exist once we are stripped of the centrality of intellect; once the AI can do anything better than we even could, except merely experience things. One feels as if one is watching a *feelie* in *Brave New World*'s London except worse in that now a personalised *feelie* could be generated at will with perfect generative AI programs. Full sensory integration! High definition experiences! Even better than the real thing! You may not even have to give it a plot or a subject beyond a mere superficial outline. Why even put in the dedicated effort of thinking about something when a perfect AI could do that better than you ever could if you just ask it? Why engage in such fruitless cognitive labour when pure experience will do? What lofty heights are left to drag us away from this decadence? Even if no hierarchy controls it, a perfected AI would lead us feeling lesser.

Thus, we have our camps, polarised by economic philosophy but nevertheless reaching the same dismal end, fighting all the way there like children in the back seat of the Great American Road Trip. Regardless of whether a perfect

[9]This is playing out in the foreground of *Neuromancer*. The Omni Consumer Products holds sway over the world of *RoboCop*, and Weyland-Yutani from *Alien*; all arising from AI fields such as intelligent robotics and supercomputing.

intelligence is the controller of an unwilling humanity or its unwilling servant, it leads to the same dehumanisation. The failure of imagination here is not to see beyond capitalism, but it is instead to envision meaning beyond power.

Imagining the future through a glass darkly

The logical thread of this short essay is as follows: we need imagined worlds to guide our own sociotechnological development but the centrality of intelligence to both humanity and AI means these worlds lead to the same imagined end. How does one escape this tragic fate? How can we have an imagined world about AI that does not deprive us of humanity?

Honestly, I do not know.

But not all science fiction follows Atwood's philosophy, especially earlier works. The worlds of Jules Verne are not defined by the perfection of a technology, but they rebel against it. Despite being written over a century ago, they still feel as if their characters inhabit our world, held in superposition between real and imagined. The submarine world of *20,000 Leagues Under the Sea* does not supplant our own but runs alongside it in escapist fantasy. It does not rely on the perfection of submarine travel (at one point, the Nautilus requires refueling and maintenance, requiring the crew to leave it briefly to explore the marine surface). It also certainly does not imagine submarine travel as being ubiquitous but explores, if very human terms, the wondrous isolation it imposes on the main characters. They find meaning in this experience in a way that Huxley's cannot. Instead of perfecting a technology, these works simply celebrate the adventure provided by its use and construction. Marvelous as these technologies are, it is their transitional imperfections and rarity that drive the adventure, not the consequences of their perfection and ubiquity.

So, how does one centre a story on AI without imagining it perfected? One could always pilfer from *Star Trek*, specifically its beloved android, Data, who is a companion on the journey, neither subservient nor superservient to the other crew members of the Enterprise. Data is not some perfect all-seeing machine god; he admits to his mistakes and the incompleteness of his knowledge and foresight. He enjoys conversation and even music for its own sake. He has hobbies. Stories centering AI like Data expand the meaning of humanity to Data through his imperfections and idiosyncrasies.

This does not necessarily require a "good" AI but can explore bad ones as well in a way that expands rather than diminishes our humanity. In Tchaikovsky's *Children of Time*, an AI built in our own image, Dr. Avrana Kern, seems to be growingly irrational after millennia in isolation, echoing a very human decent into madness while also questioning how an AI would deal with concepts completely unknown to humans: the effect of indefinite lifespans and the

possibility of resurrection on the self.

In that sense, maybe instead of imagining worlds struggling with the dehumanising consequences of a perfect intelligence, we instead imagine the adventures and nuances caused by imperfection. It is not a sure bet for a positive imagined world but, nevertheless, perhaps our best way into our future is to return to our past, not to search for material but for meaning.

John S. H. Baxter is a Chargé de Recherche (Associate Research Professor) at the Université de Rennes where he researches the interactive artificial intelligence for medical image computing and computer-assisted interventions. This research draws from multiple domains including machine learning, medicine, philosophy, and human-computer interaction.

References

Atwood, M. (2011). *In other worlds: Sf and the human imagination*. Anchor.

Jameson, F. (2003). Future city. *New left review*, *21*, 65.

How artificial can be used to compromise user privacy and anonymity?

Chawki Ben Salem

Synopsis. With the evolution of modern technology, artificial intelligence is dismantling privacy in ways that are invisible yet inescapable. From the metadata harvested by smartphones to the subtle biometric signatures left in every digital interaction, AI has turned everyday tools into surveillance instruments.

You walk into a coffee shop, phone in pocket. No one follows you. Yet within seconds, artificial intelligence logs your location, identifies your gait, and links your presence to an anonymous Twitter post you made years ago. You are not anonymous. You are a data point. Modern artificial intelligence has compromised privacy in ways most people never realize. What was once the domain of spy thrillers is now daily reality: your habits, movements, and even thoughts are tracked, analyzed, and sold most often without your knowledge.

> "The real risk with AI isn't malice but competence. A superintelligent AI will be extremely good at accomplishing its goals, and if those goals aren't aligned with ours, we're in trouble." (Russell, 2019)

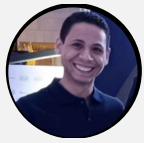
Metadata plays an important role in the development and training of AI models, it is the information on your activities without the actual content of your activities [1]. Metadata including your location being compromised by smart devices, your online digital footprint being recognized by your behavioral and contextual information, and your biometric identity such as your face, your gait. Private companies systematically collect and monetize this metadata, leveraging it to refine their artificial intelligence profiling systems [2]. In addition, such data may be commercially shared or sold to military and

law enforcement agencies who can then use it conveniently [3]. Techniques like ”geofencing warrants” further enable these entities to correlate metadata with individual identities using complex AI algorithms, thereby compromising anonymity and raising significant concerns regarding surveillance and privacy.

AI’s capacity for behavioral profiling presents an equally potent privacy threat. Typing rhythms, scrolling speed, cursor paths, and even the pattern of pauses between keystrokes [4] and the fonts you use on your browser [5] are distinctive enough to act as digital fingerprints. Commercial services like TypingDNA [6] analyze these patterns for authentication, while CAPTCHA systems covertly harvest behavioral data under the guise of bot detection. AI analysis algorithms are then used to classify and match these patterns with online users to deanonymize them.

Companies such as ”Facebook” have used advanced facial recognition for a long time and are used to map people around the world through satellite imagery [7] [8]. Facial recognition, voiceprint matching, iris scanning, and gait analysis have all been integrated into consumer technology. These technologies and deep convolutional neural networks can identify faces across low-resolution images, poor lighting, and diverse angles, while gait and voice recognition can analyze how individuals walk and talk, operating at a distance and without our consent. The deployment of these systems by private companies and its usage by governments and intelligence agencies raises critical concerns. Furthermore, the proliferation of deepfake further complicates the privacy landscape. Malicious actors can impersonate individuals in real time, enabling financial fraud, reputational attacks, and disinformation.

As AI systems grow more sophisticated, the window to act narrows. If we are to preserve any meaningful sense of digital autonomy, we must begin by understanding how these systems work in order to mitigate any potential privacy compromise.



Chawki Ben Salem is a cybersecurity engineer with a strong academic and practical background in defensive security, compliance, and threat intelligence analysis. His work focuses on the intersection of artificial intelligence and digital privacy, with an emphasis on adversarial inference, metadata exploitation, and behavioral de-anonymization. Chawki actively contributes to cybersecurity awareness through technical writing, open source projects contribution, and strategic consulting.

References

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Human-Centered AI: Transparency, Interpretability, and the Future of Trust

Ashery Mbilinyi

Synopsis. This essays calls for a future in which explainability, interpretability, and transparency are foundational to the design and deployment of AI systems—especially in critical sectors like healthcare. It highlights the urgent need to demystify complex AI models and reviews key interpretability techniques such as LIME, Grad-CAM, and SHAP. Moving beyond technical exposition, the chapter outlines actionable strategies to develop AI systems that are not only accurate but also understandable, trustworthy, and compliant with emerging regulatory standards. It urges researchers, practitioners, and policymakers to embrace transparency-by-design and work collaboratively toward an AI ecosystem rooted in accountability, clarity, and human-centered values.

Artificial Intelligence (AI) has rapidly become one of the most transformative forces of the 21st century, enabling advancements across domains such as autonomous driving, personalized recommendations, medical diagnostics, and financial modeling. As AI systems increasingly support or even automate high-stakes decisions, growing concerns have emerged about their opaque or "black box" nature. Do we truly understand how these models generate their predictions? Can they be trusted to make fair and reliable decisions?

At the center of these concerns lie three foundational principles: **explainability**, **interpretability**, and **transparency**. Each plays a vital role in narrowing the gap between complex AI algorithms and human understanding. This chapter unpacks these concepts, explains their significance, examines the unique interpretability challenges posed by multimodal AI systems, and explores actionable strategies for designing future-ready AI that is trustworthy,

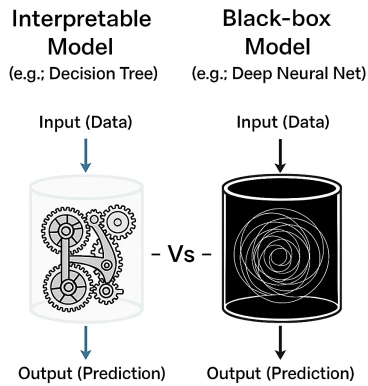


Figure 11: Comparison of interpretable and black-box models. Interpretable models provide mechanistic insights that are accessible to human understanding.

human-centered, and regulation-compliant.

The Importance of Explainability, Interpretability, and Transparency

While often used interchangeably, these terms each have distinct meanings in the AI ethics and governance landscape:

- **Interpretability** refers to the degree to which a human can understand the internal mechanics of an AI system. A model is interpretable if one can look at its structure or parameters and reason about how it makes decisions.
- **Explainability** is about the ability to generate understandable post-hoc explanations of a model's behavior. Even if the model itself is not inherently interpretable, explainability methods can provide insight into why it made a specific prediction.
- **Transparency** goes beyond technical clarity. It includes openness about the design, data sources, assumptions and limitations of a model, allowing oversight, reproducibility, and accountability.

Together, these concepts support the development of AI systems that are not only technically robust but also socially aligned, accountable, and ready for real-world deployment.

Why Interpretability Matters: The Stakes Are High

In low-risk applications like movie recommendations, the lack of interpretability may be a minor inconvenience. But in high-stakes domains such as criminal justice, hiring, or healthcare, it becomes a moral imperative. Consider a model that predicts the likelihood of recidivism: if it cannot provide a clear rationale for its output, how can a judge justify a longer sentence? Or consider a healthcare model that flags a patient as high-risk: without insight into why that decision was made, how can a physician verify or challenge the result, let alone deliver the appropriate care?

Lack of explainability also has broader societal consequences. Models trained on biased data may perpetuate or even amplify systemic inequities. Transparent and interpretable systems are essential not just for individual trust, but for auditing and correcting such structural harms. Building explainable AI is, therefore, not a technical luxury—it is a civic responsibility.

The Challenge of Complexity: Explaining Deep Learning Models

Deep learning models offer state-of-the-art performance, but they are often criticized as inscrutable black boxes. To address this, researchers have developed post-hoc explanation methods such as:

- **LIME** (Local Interpretable Model-agnostic Explanations): Trains local surrogate models to approximate a black-box model's behavior near a specific prediction (Ribeiro et al., 2016).
- **Grad-CAM** (Gradient-weighted Class Activation Mapping): Provides visual explanations for convolutional neural networks by highlighting important regions in an image (Selvaraju et al., 2017).
- **SHAP** (SHapley Additive exPlanations): Drawing from cooperative game theory, SHAP quantifies the contribution of each input feature to a specific prediction (Lundberg & Lee, 2017).

These techniques have brought visibility into model behavior but also have limitations. They often explain *what* influenced a prediction but not necessarily *how* or *why*. To move forward, we must design explanation tools that are robust, interpretable across contexts, and usable by stakeholders with varying levels of technical expertise.

A Case Study in Complexity: Multimodal Models in Healthcare

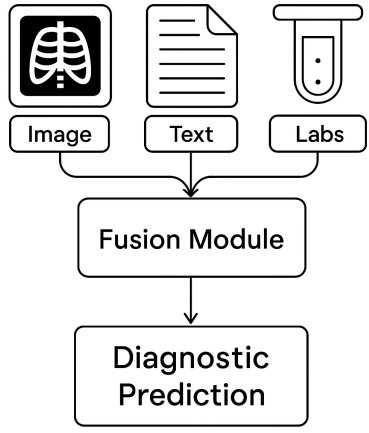


Figure 12: An illustration of a multimodal AI model architecture, where inputs from medical images, clinical text, and laboratory results are integrated through a fusion module to generate a diagnostic prediction.

A particularly compelling example of interpretability challenges comes from multimodal AI models in healthcare. These models integrate diverse data types—such as radiology images, lab tests, time-series vitals, and clinical notes—to produce holistic predictions about patient risk, diagnosis, or treatment outcomes. Although this fusion of modalities improves predictive performance, it introduces new challenges as follows:

- **Heterogeneous Inputs Formats:** Each data type (e.g., images, text, tabular data) requires a different processing pipeline and feature representation, making it difficult to apply unified explanation strategies.
- **Cross-modal Fusion Complexity:** Multimodal models often integrate inputs using mechanisms like attention layers, gating functions, or learned embeddings. Selecting the right fusion strategy is nontrivial, and some approaches may obscure the influence of individual modalities, making it difficult to pinpoint not only which input contributed to a decision, but also how it did so.
- **Modality-Specific Interpretations:** Clinicians may place greater trust in certain modalities over others. For instance, a radiologist might prefer visual saliency maps, whereas a general practitioner may rely more heavily on lab value thresholds or clinical notes. Determining the most appropriate explanation strategy at any given time is far from trivial.

To advance the field, future research should focus on hybrid approaches that combine modality-specific and fusion-aware interpretability techniques, as well as tools that offer interactive, layered explanations tailored to different user roles.

Explainability in Practice: Trust, Regulation, and Design Beyond algorithmic elegance, explainability must serve practical ends. For clinicians, regulators, or everyday users, the goal is not just to understand how a model works, but to trust it enough to use it confidently—or scrutinize it effectively when things go wrong. As Ribeiro et al.(Ribeiro et al., 2016) argue, if users do not trust a model or its predictions, they simply will not use it.

To that end, the following principles should be included:

- **Tailored Explanations:** Different users need different explanations. What satisfies a data scientist may not satisfy a physician or a patient. Explainability must be contextual and role-specific.
- **Simplicity over Completeness:** An explanation that is technically exhaustive but cognitively overwhelming is no explanation at all. Sometimes, less is more.

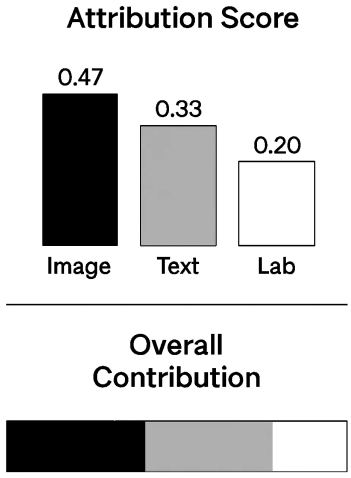


Figure 13: Example visualization of modality-specific and fusion-level attribution: Each modality is assigned an importance score reflecting its contribution to the overall prediction.

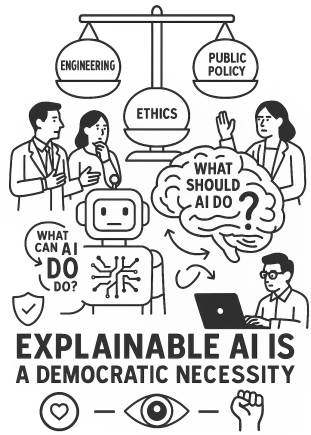


Figure 14: Explainable AI is not merely a research objective—it is a democratic imperative that demands interdisciplinary collaboration and a fundamental shift in mindset from all stakeholders.

- **Transparency-by-Design:** Systems should be built with interpretability in mind from the start—not as an afterthought. This includes clear documentation, versioning, and the use of inherently interpretable models where possible.

Regulatory bodies are beginning to catch up on this aspect. The European Union's General Data Protection Regulation (GDPR) enshrines a *right to explanation*—Article 71 emphasizes the ability to obtain an explanation of the decision reached after such assessment ("Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation)", 2016). In the U.S., the Food and Drug Administration (FDA) is intensifying oversight of AI/ML-enabled Software as a Medical Device (SaMD); its recent guidance emphasizes transparency, user-centric labeling, lifecycle change management, and bias mitigation (U.S. Food and Drug Administration, 2023).

Looking Ahead: Towards Responsible AI

Explainability, interpretability, and transparency are not merely research objectives; they are democratic necessities that demand interdisciplinary collaboration and a fundamental shift in mindset among all stakeholders involved. As we look to the future, the goal must be to build AI systems that do more than perform—they must be understood, questioned, and trusted.

To realize this vision, **engineers** must prioritize interpretability in system

design and actively test for fairness and bias. **Policymakers** should define and enforce regulatory frameworks that mandate transparency. **Educators and researchers** must integrate ethical and interpretability principles into AI curricula and tools, preparing the next generation to lead with responsibility. **Industry leaders** should promote explainable AI as a standard of innovation—not a constraint.

We must move from asking "What can AI do?" to "What should AI do–and how can we make it understandable?"

Ashery Mbilinyi is an Assistant Professor and head of the Medical Computer Vision (MCV) lab at the University of Victoria, Canada where he develops AI systems that interpret medical images and clinical data to support diagnostic decision-making and improve patient care.

References

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation) [Articles 22 and Recital 71]. (2016). https://eur-lex.europa.eu/eli/reg/2016/679/oj.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

U.S. Food and Drug Administration. (2023). Artificial intelligence and machine learning (ai/ml) software as a medical device (samd) action plan. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device.

The value of human and machine in machine-generated creative contents

Weina Jin

Synopsis. The seemingly "imagination" and "creativity" from machine-generated contents should not be misattributed to the accomplishment of machine. They are accomplishments of both human and machine. Without human interpretation, the machine-generated contents remain in the imaginary space of the large language models, and cannot automatically establish grounding in the reality and human experience.

Does a machine have creativity? Can AI imagine? What is the human value in the era of generative artificial intelligence (AI)? These are the questions people may ponder upon when seeing the fictional images or texts that large language models (LLMs) generate, such as chatGPT or deepseek. These questions are significant because, if machines can "create" or "imagine," it will raise doubt about what creativity, imagination, and art really are, and what's the point of human performing art, imagination, and creativity, if machines can also do them.

Definition: *Complex phenomenon* We define complex phenomenon by borrowing the definition of complex systems, which are "co-evolving multilayer networks," are context-dependent, and are composed of many non-linearly interacting elements (Thurner et al., 2018), such as language, human behavior, human mind, a living cell, financial market, social or natural phenomena.

In this article, I will inspect the above questions by analyzing and understanding the capabilities and limits of machine. My analysis comes from ontological and epistemological perspectives that concern what we know about the real world and how we know it (Crasnow & Intemann, 2024). LLMs are a type of AI models that recognize complex patterns from large-scale training data by training to be a next word or next image patch predictor (Zhao et al.,

2025). Like any computational models, LLMs are abstract representations of the phenomenon they represent. LLMs are trained on human language or image data to mimic the underlying phenomenon —- language or visual art —- represented in the training data. The human language or visual art can also be regarded as an abstract entity that represents the phenomenon of reality and human experience. The three entities **E1. reality and human experience**, **E2. language and art**, and **E3. LLMs** can be regarded as forming a representation sequence, with each latter entity being an abstract representation of its former entity. The relationship can be visualized in Fig.1. There are several properties (P1-P4) about the abstract representation[10]:

P1. [Irreducibility] *A complex phenomenon is irreducible and cannot be fully represented.*

P1 indicates that if an entity is a complex phenomenon, it is incompressible and cannot be reduced to a representation without losing information. Any of its representations is always incomplete and cannot replace the original phenomenon (Cilliers, 2016; Thurner et al., 2018). In statistics, there is a well-known aphorism to describe the irreducibility property: "all models are wrong, but some are useful" (Box, 1976). The irreducibility property describes the intrinsic limit of a model or a representation, including language and art (E2) and LLMs (E3). This means that reality and human experience (E1) cannot be fully represented by language and art (E2); and language and art cannot be fully represented by LLMs (E3).

[10] The properties of the abstract representation are based on the generalization of the properties of scientific theory and methodology in philosophy of science, for example (Frank et al., 2024). The two types of authenticity are from Jin Guantao's philosophy of authenticity (Guantao, 2023). Our prior work details this framework by applying it to the medical image synthesis task (Jin, Sinha, et al., 2025).

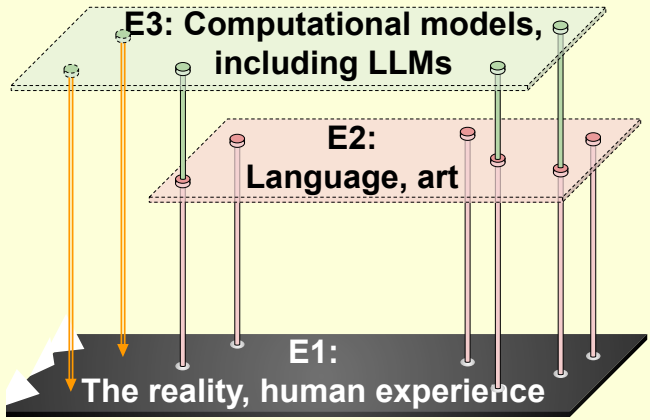


Figure 1. Visualization of the three entities: **E1. reality and human experience**, **E2. language and art**, and **E3. LLMs**. E1 is visualized as the ground. E2 and E3 are visualized as floors built upon the ground as abstract representations. The entities do not completely overlap with each other in space. This is meant to visualizes **P1 irreducibility** that a representation cannot replace the original complex entity. The floors are shown by dashed outlines in the air, indicating that they are imaginary spaces as opposed to the reality (E1), which is the solid ground. The dots with solid outlines and their corresponding pillars supporting a floor represent the grounding instances that make the abstract representation valid (**P2. valid representation**). A valid representation space is the imaginary space spanned by the solid dots. The dots with dashed outlines indicate newly inferred instances that fulfill **P3. symbolic authenticity**. The dashed, not solid, outlines indicate that the validly inferred new instances within the representation space do not automatically ground in a lower-level phenomenon, unless an explicit grounding connection is made (visualized as the vertical arrow).

P2. [Valid representation] *An abstract space is a valid representation of the target phenomenon, if the space is spanned by instances that have established grounding with the target phenomenon.*

To apply the idea of P2 to our case, for **E2. language and art** to be a valid representation of **E1. the reality and human experience**, they should consist of instances of creative writings or art works that are created by humans grounding in our living experiences in the real world. For **E3. LLMs** to be a valid representation of **E2. language and art**, the LLMs should pass tests to show the LLMs have learned the training data distribution from the **E2. language and art** entity of human languages and art works.

P3. [Symbolic authenticity] *For a representation, it can perform inference within its abstract representation space, as long as the inference follows rules and assumptions that make the abstract representation hold at the first place. We call such inferences fulfilling symbolic authenticity.*

"Symbolic authenticity" uses the word "symbol" because the inference is performed in the abstract symbolic space. "Authenticity" denotes that the

inference is valid. P3 indicates that the abstract representation space itself can be regarded as an imaginary space to generate new thoughts or instances by validly inferring within the space. For example, LLMs can generate new instances within the model space. These new instances exhibit symbolic authenticity.

P4. [Grounding authenticity] *For newly-inferred instances within an abstract representation space that fulfill symbolic authenticity, they do not automatically establish grounding in the phenomenon it represents. We call that symbolic authenticity doesn't equal grounding authenticity. For such instances to exhibit grounding authenticity, they should establish grounding in the target phenomenon.*

The reason that "symbolic authenticity doesn't equal grounding authenticity" lies in P1. Irreducibility. Because the target phenomenon cannot be fully described and replaced by its abstract representation, the imaginary instances in the representation space may not have correspondence (i.e., grounding) in the phenomenon. For example, the phrase "putting theory into practice" is a common sense because theory is an abstract representation of the reality. What works in theory (i.e., symbolic authenticity) may not necessarily work in practice (i.e., grounding authenticity).

We can use the above perspective to understand LLMs. First, regarding the value of the machine, because both LLMs and language/art are imaginary entities, and LLMs mimic language/art, LLMs that establish valid representation of language/art can be regarded as an extension of the language/art space. This aspect shows the value and benefits of LLMs: LLMs enable us to explore the vast imaginary space of language/art or generate new imagery language/art spaces. By facilitating the expansion of our imaginations, this value of LLMs strengthens the original value and benefit of art and creative writing that enables us to explore the imaginary space.

Second, despite the benefit of expanding the imaginary space, LLMs have intrinsic limits that language/art doesn't have. The new instances in the language/art space can easily establish grounding in the reality and human experience, because the new instances are created by humans and are interpreted by humans to relate the new instances to certain human experiences. This obvious fact, however, doesn't hold for LLMs. For newly-generated instances from LLMs, they are not inherently grounded in the reality and human experience due to the above properties P1 and P4: the abstract representation of LLMs cannot be a replacement of the underlying phenomena they mimic (i.e., E1 and E2). Thus, the new instances in the LLMs space don't automatically establish correspondences in the reality and human experience, and lack their roots in

the living experience in the real world. We call it the groundlessness limit of LLMs.

The groundlessness limit has broad implications for machine-generated contents. Because these contents cannot automatically ground in the reality and living experiences, we cannot regard the contents alone as "art" that shows "creativity" or "imagination" of the machine. Rather, as we analyzed in the first point above , the "imaginations" exhibited in machine-generated contents are explorations in the imaginary space of the LLM models that can help to extend our human imagination in the forms of texts and images. Creativity, imagination, language, and art have richer meanings than merely exploring the imaginary spaces. At least some of their meanings include being able to freely translate in-between the imaginary space and the real world: such as relating a phenomenon in the real world to something imaginary, and drawing inferences from the imaginary spaces to the concrete living experiences. The reason that we may leave with the impression of regarding LLM-generated texts/images as the manifestation of the "imagination" or "creativity" of machine is because we humans implicitly close the loop for LLMs from real-world phenomena to the LLM model space and back to the real world, such that the machine-generated contents appear to exhibit "imagination" or "creativity" as art works or creative writings. The invisible, often neglected human efforts to "close the loop" for LLMs include:

1. Humans create, select, and label creative contents of texts and images that are used as training data for LLMs. Humans provide preference for machine-generated contents as feedbacks in the training of LLMs (i.e., reinforcement learning from human feedback), and embed implicit knowledge about the phenomena and training data in the model design and training processes. The knowledge forms the epistemic basis for LLMs. As information agents, LLMs depend on these human works to represent valid representations about the training data (Jin, Vincent, & Hamarneh, 2025).

2. When generating a machine-generated content, humans provide the prompts that are based on real-world phenomena and/or human experience. When viewing machine-generated contents, humans implicitly draw connections of the contents to the real world and our living experience. This human interpretation process is the key to establish grounding of the machine-generated contents in reality and human experience. With human interpretation, the machine-generated contents fulfill grounding authenticity and encode rich meaning to us. In other words, LLMs as disembodied abstract computational models do not generate meanings. It is the human interpretation process that generates meaning (i.e., the grounding in reality and human experience) for the machine-created contents.

My analysis shows that attributing "imagination," "art," or "creativity" to LLMs is an overclaim and an anthropomorphic tendency (Altmeyer, 2024; Ibrahim & Cheng, 2025), which should be avoided. LLMs as symbolic representations do not inherently ground in the reality and living experience as the words "imagination," "art," or "creativity" encode. The value of LLMs lies in their ability to expand our imaginary spaces of texts and visuals encoded in their vast collections of training data. The value of humans lies in our implicit grounding of the machine-generated contents in our living experiences. Arts, imaginations, and creativity enable us to experience the living world in rich and meaningful ways. Understanding the respective value of human and machine can help us use LLMs to enrich, rather than impoverish, our living experience using arts and creative writing.

Weina Jin is a Computer Science Ph.D. Candidate at the Medical Image Analysis Lab, Simon Fraser University. She also holds an M.D. degree. Weina has a multidisciplinary background in AI, medicine, and human-computer interaction. Her research focuses on ethical AI techniques, especially explainable AI, in medical image analysis. She uses research to answer this big question: how can we reimagine and construct AI technologies and ecosystems for medicine based on feminist values and perspectives?

References

Altmeyer, P. (2024). Position: Stop making unscientific agi performance claims. *Proceedings of the 2024 [Conference/Workshop Name]*. https://doi.org/10.5555/3692070.3692121

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Cilliers, P. (2016). *Critical complexity: Collected essays*. De Gruyter.

Crasnow, S., & Intemann, K. (2024). *Feminist epistemology and philosophy of science: An introduction*. Routledge.

Frank, A., Gleiser, M., & Thompson, E. (2024, March). *The Blind Spot: Why Science Cannot Ignore Human Experience*. The MIT Press.

Guantao, J. (2023, July). *The Real and the Virtual* (First Edition). CITIC Press Corporation.

Ibrahim, L., & Cheng, M. (2025). Thinking beyond the anthropomorphic paradigm benefits llm research. https://arxiv.org/abs/2502.09192.

Jin, W., Sinha, A., Abhishek, K., & Hamarneh, G. (2025). Ethical Medical Image Synthesis.

Jin, W., Vincent, N., & Hamarneh, G. (2025). AI for Just Work: Constructing Diverse Imaginations of AI beyond "Replacing Humans". https://arxiv.org/abs/2503.08720.

Thurner, S., Hanel, R., & Klimek, P. (2018). Introduction to complex systems. In S. Thurner, R. Hanel, & P. Klimek (Eds.), *Introduction to the theory of complex systems*. Oxford University Press. https://doi.org/10.1093/oso/9780198821939.003.0001

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., ... Wen, J.-R. (2025). A survey of large language models. https://arxiv.org/abs/2303.18223.

AI: The Good, the Bad, and the Game-Changer

What if artificial intelligence isn't just transforming the world—but reshaping what it means to be human?

This provocative collection brings together 25 essays from leading AI researchers, critical thinkers, and emerging visionaries across 20 countries. From breakthroughs in healthcare and education to deep ethical rifts in warfare and creativity, these voices explore the paradoxes at the heart of artificial intelligence.

As AI becomes a catalyst for both discovery and disruption, this book examines the promises we chase—and the consequences we overlook. It confronts urgent questions: Who controls AI's future? And what happens when profit outruns ethics?

Inside, you'll explore:

- How AI is accelerating medical innovation—and eroding patient privacy
- The promise and peril of GenAI in classrooms, film, and artistic identity
- The silent toll of AI on mental health, and human intellect preservation
- Environmental and labor costs of building "intelligent" machines
- A grassroots vision of ethical AI—from activists, artists, and global change-makers

Featuring an exclusive interview with Asma Derja, founder of the Ethical AI Alliance, this book invites you to confront uncomfortable truths—and imagine a more inclusive, accountable AI future.

This is not just a book. It is a reflection —a vision we all get to build together.

About the Editor

Dr. Islem Rekik is an award-winning AI researcher, educator, and advocate for inclusive innovation. She leads the BASIRA Lab and is an Associate Professor at Imperial College London (Computing, Innovation Hub I-X). Over the past 7 years, she has mentored students to 24+ academic honors and awards and co-chaired 35+ major AI events—including MICCAI, NeurIPS, and ISBI. With 200+ peer-reviewed publications, she holds two prestigious fellowships: the EU Marie Curie and Türkiye's TUBITAK 2232.

In 2025, she received the Tunisian AI Award for her pioneering work in AI and EDI, and was featured in Realités Magazine and I-X News. She also co-founded global initiatives supporting underrepresented researchers—especially in low- and middle-income countries—to help shape a more inclusive and equitable AI future.