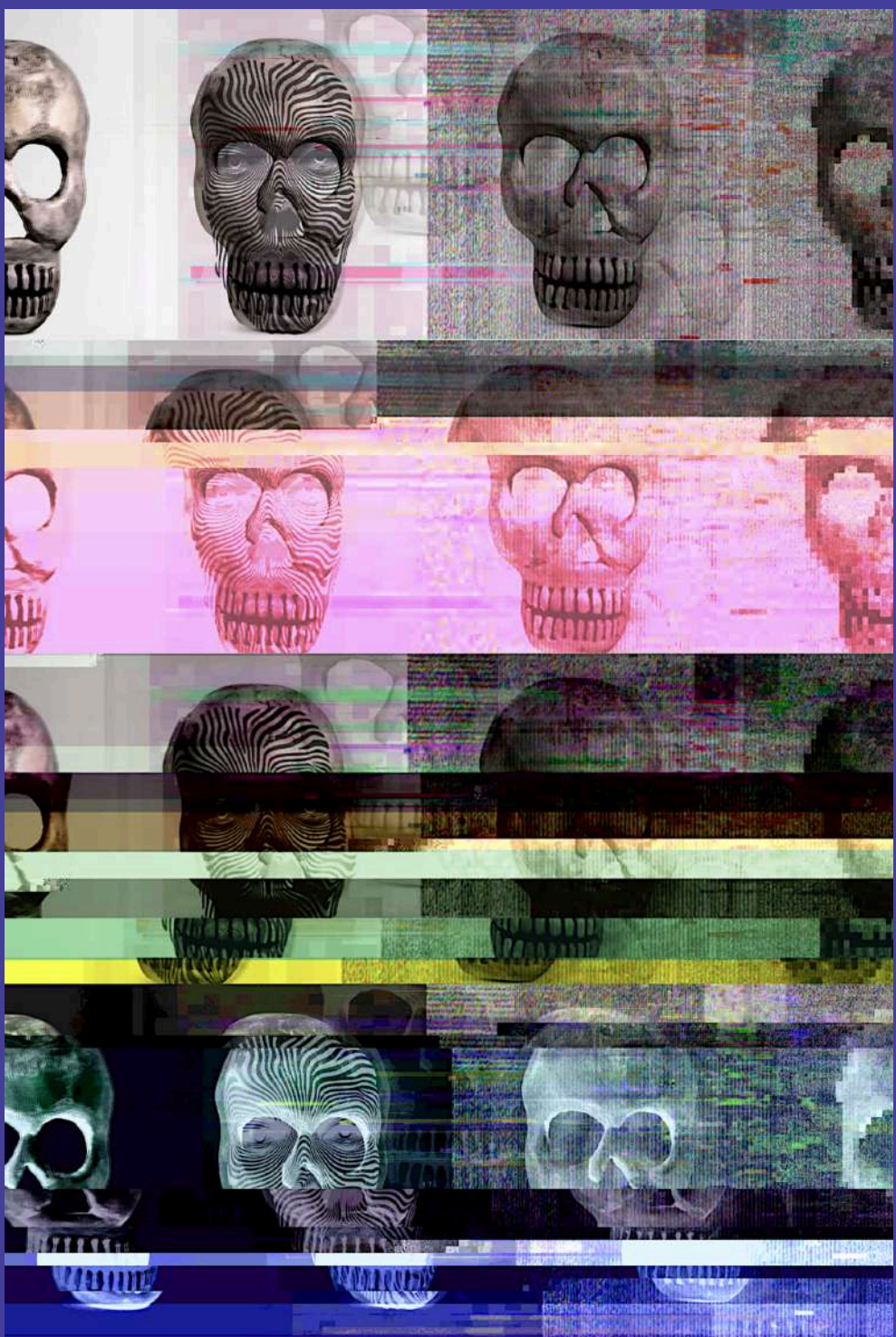
# CAMBRIDGE JOURNAL OF ARTIFICIAL INTELLIGENCE



Vol. 1 Issue 3
July 2026

CJAI CAMBRIDGE JOURNAL OF ARTIFICIAL INTELLIGENCE



Published by *Cambridge Journal of Artificial Intelligence*, Cambridge, United Kingdom.

The cover artwork is *Corruption 1* by Kathryn Conrad. The image represents both the way that AI models training on their own output leads to model collapse and the way that corrupted data can lead to problematic and potentially damaging outputs.

# Editorial Team

July 2026

# Contents

# Editorial

Welcome to the third issue of the *Cambridge Journal of Artificial Intelligence* (CJAI).

As the journal continues to grow, so too does the breadth and depth of the conversations it seeks to foster. What began as an ambition to create an interdisciplinary forum for the study of Artificial Intelligence has developed into a truly international community of scholars, students and practitioners committed to examining AI from diverse intellectual, cultural and professional perspectives. We are grateful to everyone who has contributed to this journey and are excited to present another issue that reflects the richness of this rapidly evolving field.

This edition marks a particularly exciting milestone for the journal through our collaboration with New York University's Peace Research and Education Program. Bringing together contributors from across institutions and disciplines, this partnership reflects a shared belief that understanding artificial intelligence requires engagement not only with technical innovation, but also with questions of governance, ethics, justice and human flourishing. We hope this collaboration will be the first of many that strengthen dialogue across academic communities and encourage new perspectives on the challenges and opportunities presented by AI.

The articles featured in this issue explore AI through the lens of power and responsibility. From metacognitive approaches to conversational AI and AI literacy to questions of technomoral resistance, governance, misinformation and sustainability, each contribution offers a thoughtful examination of how AI is reshaping the ways we think, make decisions and organise society. Together, these papers highlight that the most pressing questions surrounding AI are not solely about what these technologies can do, but about how they ought to be designed, governed and integrated into our lives.

The continued success of the CJAI is only possible because of the generosity and dedication of so many individuals. I would like to extend my sincere thanks to our authors for entrusting us with their research and to our reviewers for their careful and constructive evaluations, which uphold the academic standards of the journal. I am equally grateful to our editorial team, whose professionalism, enthusiasm and commitment continue to shape the CJAI into a welcoming and rigorous platform for interdisciplinary scholarship. Finally, I would like to thank our colleagues at New York University for their collaboration in bringing this special issue to fruition.

As AI continues to evolve at an extraordinary pace, the need for thoughtful, evidence-based and interdisciplinary dialogue has never been greater. We hope the articles in this issue inspire further research, debate and collaboration, and we warmly invite readers to contribute to future editions of the journal as authors, reviewers or editors. Together, we can continue building a scholarly community that critically examines AI and its role in shaping our collective future.

With best wishes,

**Mahera Sarkar**
*Founder & Editor-in-Chief*

## Foreword

This past June, in a university hall in Zomba, Malawi, a colleague held up a phone running a popular AI assistant and spoke to it in Chichewa. The room watched it stumble and give up. Nobody seemed surprised. The really telling part was the explanation that followed: these systems only learn from what they are fed, and everything left out is entirely invisible to them. For millions of Chichewa speakers, that invisibility is a daily reality. So are the indirect phrases a frightened woman in Thyolo might use to report a threat against her neighbour — the euphemisms people reach for when something is too dangerous or shameful to name directly. The workshop participants had spent the morning writing those very phrases on Post-It notes, simply because no system trained in Silicon Valley has ever encountered them.

I kept coming back to that room while reading the papers for this issue. It captures the exact question behind this call: where is power actually concentrating in AI, and what does that mean for the people living downstream of it?

When we first put out this call, a collaboration between the Cambridge Journal of Artificial Intelligence and NYU's Peace Research and Education Program, we expected substantial material analysis. We thought we'd get papers on water rights, data centres, supply chains, and labour. Instead, our contributors kept turning to something much more fundamental: power over what we know, how we come to know it, and whether we can still think for ourselves.

The six papers here group naturally into a few distinct arguments. The first is about epistemic control. One piece suggests that AI misinformation isn't a glitch, but a feature built into commercial ownership models. Another looks at AI literacy from a metacognitive angle, asking how we notice when a machine is doing our thinking for us, while a companion piece looks at the builder's side, offering a design framework that takes reflective well-being seriously rather than treating it as an afterthought.

Then there is the question of resistance, treated here not just as a public protest but as a quiet, deliberate skill. We have two papers arguing that the ability to say "no" to a system is worth teaching: one frames resistance as a "technomoral virtue," and the other argues for intentional friction as a form of literacy. The final two pieces look at governance from unexpected angles. One examines Turkey's proposed AI laws to see if emerging economies can dictate terms rather than just receive them, while the other shows how easily the language of "sustainability" can be co-opted to defend the very infrastructure it claims to critique.

That last problem, how the centres of power compromise themselves, came up in Zomba too, just on a much smaller canvas. During a workflow exercise on emergency reporting, the room completely stalled on one scenario: a woman reports a threat of violence, but the local chief who receives the alert happens to be the accused man's uncle. If a system can't route around a compromised local authority, it fails the people who need it most. It is the exact same structural challenge our governance papers wrestle with, just closer to the ground.

To be completely transparent about scale: this is a new journal, and this is a modest issue. We didn't get hundreds of submissions, and we aren't trying to pretend we did. What we have are six pieces of incredibly careful thinking, several from scholars and places that the field's usual citation loops completely ignore. But that was the whole point of partnering with a peace research program rather than sticking to traditional tech channels. The people most exposed to concentrated power are often the ones doing the sharpest thinking about it.

At the end of the day in Zomba, we took a photo of our "parking board", the flipchart where we left all the questions we couldn't resolve. We had told the room at the start that we didn't have the answers, just a shared desire to look for them. That feels like the right note for this issue. These papers don't wrap up the story of AI and power in a neat bow. Instead, they point carefully, from some unexpected directions, at

where we should be looking while that story is still being written.

***Marine Collins Ragnet***
*Head of Innovation Research, NYU Peace Research and Education Program*

*Adjunct Professor, New York University*

*Managing Editor, Cambridge Journal of Artificial Intelligence*

# Sustaining the Status Quo? Rethinking AI for Sustainability through Power and Infrastructure

*Harry Collins*
*University of Cambridge*



Artificial Intelligence (AI) is increasingly presented as a tool for addressing environmental crises, yet there remains significant debate over what should count as genuine AI for Sustainability. I argue that dominant accounts remain too narrow because they focus on environmental outputs while neglecting the political and social conditions through which those outputs are produced. Drawing on Falk and van Wynsberghe's framework and Lehuedé's analysis of data infrastructure in Chile, I show that sustainability claims can obscure power concentration, extractive relations, and territorial exclusion. I propose a revised AI for Sustainability framework that adds conditions on benefit and harm distribution, community engagement, stronger carbon reporting and energy-efficiency improvement, and the evaluation of alternative social practices. To make the framework operational, I work through community engagement and impact reconciliation in detail, demonstrate an example application against a real-world AI project, and argue that the framework should be used through external, public assessment rather than self-certification.

**Keywords**: AI for Sustainability, Power Concentration, Environmental Justice, Data Infrastructure, Resource Extraction

## Introduction

With carbon emissions rising, artificial intelligence (AI) has been suggested as a solution to environmental challenges (Vinuesa et al., 2020), and many AI systems branded as "AI for Sustainability" have been developed to this end (Falk and van Wynsberghe, 2024). However, there is also a growing body of work into the sustainability of AI itself, including environmental and social elements like energy consumption, extractive supply chains, labour, and expanding computational infrastructure. This raises the question of how to determine which AI systems can truly be considered AI for Sustainability, reflecting genuine environmental benefit, and when they instead serve to legitimise concentrated forms of power. I argue that dominant AI for Sustainability frameworks remain incomplete because they do not adequately address AI's political and social impacts, particularly power concentration and the distribution of environmental benefits and harms. Even where an AI system appears to generate positive environmental outcomes, those outcomes may depend on infrastructures that concentrate benefits among powerful firms and states while distributing environmental and social costs onto marginalised communities elsewhere. Sustainability claims can therefore function as a legitimising discourse obscuring extractive relations, infrastructure dependency, and control over environmental decision-making.

This article's contribution is to develop a demanding AI for Sustainability framework where environmental performance is inseparable from infrastructural accountability, referring to the public tracing and justification of the infrastructures, supply chains, governance arrangements, and community relations that make an AI project's environmental benefits possible. The framework therefore evaluates AI projects claiming sustainability benefits across the AI supply chain rather than the final model exclusively. To develop this argument, I centre my analysis around two important interventions. Section 1 focuses on Falk and van Wynsberghe's (2024) tiered framework for defining AI for Sustainability. I argue that the framework is too narrow, missing the political and social impacts of AI, particularly regarding benefits and harm distribution across AI infrastructures. Section 2 expands on this, turning to Lehuedé (2022), who reveals how data infrastructure in Chile threatens Indigenous practices that sustain their environment. By analysing Falk and van Wynsberghe (2024) alongside Lehuedé (2022), I argue that AI for Sustainability should be re-centred around power concentration, resource

extraction, compute dependency, and the distribution of benefits and harms. Section 3 integrates these insights into a revised framework, proposes an external rating model, and presents an example use of the framework against a real system. Methodologically, I adopt a critical-response approach. I critically reconstruct Falk and van Wynsberghe's (2024) framework, consider it through Lehuedé's (2022) analysis of data infrastructure and territorial conflict, and use this engagement to develop a revised AI for Sustainability conceptual framework.

## 1. Beyond Greenwashing: Analysing Falk and van Wynsberghe's Framework

### *1.1. Why This Paper*

Falk and van Wynsberghe (2024) attempt to clarify the definition of AI for Sustainability, which they argue lacks meaning. Without consensus on this, AI systems can masquerade as AI for Sustainability despite potentially perpetuating unsustainable practices. This enables greenwashing, the "*intersection of [...] poor environmental performance and positive communication about environmental performance*," (Delmas and Burbano, 2011) like Microsoft heavily marketing AI for Sustainability projects (Microsoft Research, 2024) whilst partnering with ExxonMobil to increase oil and gas extraction (ExxonMobil, 2019). Therefore, more rigour is required to determine whether AI systems contribute to sustainability in order to enable genuinely positive contributions in the field. This paper has been selected as it offers one of the clearest recent attempts to define AI for Sustainability through a strict framework. It is an important starting point, but requires expansion from environmental output to infrastructural power.

### *1.2. Overview and Core Strengths*

Falk and van Wynsberghe (2024) sought to develop a framework for determining which AI systems should be classified as AI for Sustainability. They define sustainability through seven environmental impact categories including climate change, chemical pollution, and change in biosphere integrity/biodiversity (Dong and Hauschild, 2017), arguing that AI for Sustainability systems must positively contribute to at least one. The authors then use necessary and sufficient conditions to tier sustainability claims. Necessary conditions must be met for an outcome to occur, whereas outcomes are guaranteed when sufficient conditions are also met. (Brennan, 2024). Falk and van Wynsberghe's framework classifies systems that provide environmental monitoring or information as AI in an application domain. If there is also a sustainability analysis of the AI system, particularly carbon reporting from model training, the system becomes AI towards Sustainability. These conditions are deemed necessary, and then if the sufficient condition is met, demonstrating a positive environmental action directly from the model, then the system is considered AI for Sustainability. This is valuable because it makes AI for Sustainability more rigorous. Distinguishing between information and action is especially useful, as AI forecasting energy demand, for instance, could help phase in renewable energy sources whilst also optimising fossil fuel imports. Thus, demonstrable environmental action is important to prevent potential usefulness being treated as actualised sustainability. The framework also brings the sustainability of AI into view by requiring a sustainability analysis.

### *1.3. Critical Analysis of Limitations: Distribution, Infrastructure, and Social Alternatives*

Despite these strengths, defining AI for Sustainability as positively contributing to one environmental impact category misses AI's political element, particularly how the AI systems' effects are distributed. Per Mensah (2019), environmental, economic, and social sustainability are distinct but "interconnected" (2019) pillars, so simplifying AI for Sustainability to focus on environmental sustainability targets neglects social impacts and where environmental sustainability takes place. Dauvergne (2022) uses a political economy lens to argue that corporate AI for Sustainability initiatives deepen global power imbalances, with benefits realised by wealthy Western states and corporations while environmental and social costs predominantly impact the marginalised. For example, Google DeepMind used AI to schedule wind-power delivery to their electricity grids, increasing Google's wind energy usage (Witherspoon and Fadrhonc, 2019). Although this appears positive, the benefits are concentrated towards

Google rather than the planet, highlighted as they measure success in economic value for themselves without mentioning the AI system's own resource usage. Despite this, the article is tagged as "sustainability" and Falk and van Wynsberghe's (2024) framework would classify this at least as AI towards Sustainability if carbon emissions from training were reported. However, not only is this omitted, but there is no indication that any benefits of the system are felt by communities involved in Google's AI supply-chain.

AI and data infrastructure require the extraction of rare earth elements and minerals like tantalum, much of which is mined in Africa and particularly the Democratic Republic of Congo (Dauvergne, 2022). Muldoon (2023) notes that increased demand for such materials has led to "*devastating environmental and social consequences*" (2023), such as polluted water supplies. Rather than sustainability being an unconscious term, this demonstrates conscious efforts to conceal AI supply chains, like how the physical reality of wider computational infrastructure is hidden behind the Cloud (Ensmenger, 2021). Moreover, nations like DR Congo's carbon emissions are miniscule versus wealthy Western states like the USA (Ritchie and Roser, 2024), which harbour most AI projects and rewards (Maslej et al, 2024), highlighting the unjust distribution of benefits and negatives. Publications like Google's provide no evidence as to how states and communities involved in providing their AI infrastructure environmentally benefit from Western AI systems versus the environmentally destructive consequences of the infrastructure, yet understanding the distribution of impacts is missed in Falk and van Wynsberghe's (2024) framework.

A second issue is that the sustainability analysis condition is too weak if it only requires disclosure of carbon emissions from AI training. Falk and van Wynsberghe aim to establish an "easy-to-implement" (2024) method to encourage disclosure, yet disclosure alone does not necessarily improve AI sustainability. There has been significant research into improving AI algorithms' energy efficiency, with Verdecchia et al.'s (2023) literature review finding cases of energy savings over 50% versus current methods. Disappointingly, they observe minimal industry involvement in this area, despite industry's leading role in most AI research (Thompson and Ahmed, 2023). Hence, evidence that AI for Sustainability research is experimenting with these techniques to improve the sustainability of AI should be included in the sustainability analysis condition, demanding genuine improvement in the sustainability of AI.

A final limitation is the overly quantitative focus on reducing carbon emissions with technology, thus risking subscribing to climate isolationism, considering climate issues as "*an isolated, discrete, scientific problem in need of technological solutions*" (Stephens, 2022). Stephens argues that focusing on reducing carbon emissions using technology risks detracting attention from social innovations to sustainability issues whilst perpetuating existing social inequalities. Instead of accepting "inherently unsustainable" (Falk and van Wynsberghe, 2024) AI, alternative approaches and voices should be considered. Stilgoe et al.'s (2020) framework for responsible innovation includes public consultation to ensure communities' concerns are addressed (2020), whereas here, alternative voices and social innovations are missed. This is an issue because AI and climate research is dominated by a white male hegemony (Stephens, 2022). Muldoon (2023) furthers this, considering AI infrastructure as a "colonial supply chain" (2023) reinforcing global power asymmetries, promoting European modernity and techno-solutionist capitalism. This results in a techno-deterministic narrative restricting the possibility for social changes like "*community-based initiatives to support the provisioning of local [..] energy*" (Stephens, 2022) due to the prospect of socio-technical lock-in, meaning increasing dependence on technological solutions permanently entrenches their unsustainable practices (Cairns, 2014). Therefore, before committing to AI to solve a sustainability issue, alternative social innovations and voices from those who will be affected by the AI should be considered to show that AI is genuinely necessary. To this end, Lehuedé (2022) shows how alternative viewpoints about data centre expansion in Chile can reveal environmental injustices and social

practices that can preserve natural ecosystems, now examined.

## 2. The Infrastructural Politics of AI: Analysing Lehuedé

### *2.1. Why This Paper*

Lehuedé (2022) offers an alternative lens to the discussed framework. It analyses the environment through conflicting understandings of territory, showing perspectives and social practices towards the environment beyond the dominant European modernity viewpoint highlighted by Muldoon (2023). Thus, it has been selected due to foregrounding the infrastructural and territorial politics that definitional frameworks like Falk and van Wynsberghe's (2024) risk overlooking.

Two concepts need briefly defining. "Ways of worlding" (Lehuedé, 2022) refers to the differing ways people and communities understand what exists, what has agency, and how relations between humans and non-humans are organized. Assetised and relational ontologies respectively mean an understanding of territory as a set of economic resources, or as constituted through interdependencies between humans, non-humans, land, water, and other entities (Lehuedé, 2022).

### *2.2. Overview and Methodological Lesson*

Lehuedé (2022) uses political ontology, which analyses conflicting ways of worlding, to examine divergences in what territory is in relation to data infrastructure expansion in Chile. Lehuedé collected empirical data through documents and interviews with stakeholders including public and private sector representatives and Indigenous Lickan Antay activists. He then conducted a discursive-material analysis of how territory is conceived and how agency is attributed to human and non-human actors. His main claim is that conflict around Chile's data infrastructure expansion is driven by two divergent ontologies of territory: a dominant assetised ontology driven by European modernity and colonialism that considers territory an economic resource, and a relational ontology from Indigenous communities where territory results from interdependencies between human and non-human actors.

The chululo-colonies episode is the clearest methodological point for AI for Sustainability. Lehuedé shows how Indigenous Lickan Antay communities have conflicted with the ALMA observatory. The communities have sustained the environment "*through practices that acknowledge mutual interdependency among the actors that make up the territory*" (2022), including life-sustaining entities like mountains that provide water. During ALMA's natural gas pipeline development, representatives declared no adverse environmental impacts, yet Lickan Antay activists visited the prospective building site and discovered overlooked chululos, saving 22 colonies. The Lickan Antay's acknowledgement of non-humans and knowledge of the territory helped preserve it, revealing harms formal assessments missed. This raises questions of whose knowledge is counted when determining environmental impacts. AI sustainability frameworks focusing only on technical output risk repeating this mistake, counting emissions, efficiency, or environmental benefits while failing to recognise the people and non-human entities most affected by the infrastructures that make those benefits possible.

### *2.3. Strengths, Counterarguments, and Implications*

The paper's main strength regarding AI for Sustainability is using political ontology to highlight how relational ontologies of territory can sustain ecosystems through social practices that acknowledge non-humans' importance, and the interdependencies between all entities in a territory. Reinforcing this is wider research showing that the natural environment in areas occupied by Indigenous communities is generally preserved better than elsewhere (IPBES, 2019). Drawing attention to the colonial themes of Chile's data infrastructure also shows how this expansion occupies Indigenous lands and destroys their worlds, viewing territory as economic resources to be extracted. Concerning AI for Sustainability, these findings can help create more robust, informed criteria for sustainable AI projects to adhere to regarding AI's political and social impacts.

There are, however, places where Lehuedé's argument could be strengthened. There is little engagement with wider debates around data

infrastructure despite claiming that analysing relational philosophies is crucial due to the terricide, a global phenomenon. Brodie (2023) connects Lehuedé's research to colonial dynamics in data infrastructure worldwide, so exploring whether similar ontological divergences may be occurring elsewhere could have helped guide further research. Potential counterarguments are also only lightly discussed. For instance, Lehuedé criticises the "consequentialist and quantitative logics" (2022) used to justify ALMA, yet does not engage with the counterargument that thousands of people may benefit from data infrastructure versus disrupting a small community (2022). He could have questioned who those standing to benefit are, how they will benefit, how this impacts Lickan Antay livelihoods, and who decides. Also, ethical intentions do not necessarily translate into ethical outcomes (Powell, 2022), so Lehuedé could have discussed how benefits of data infrastructure are only potential versus the immediate and actual destruction of the Lickan Antay's world.

Rather than undermine the usefulness of Lehuedé's work for my argument, these issues highlight that a revised AI for Sustainability framework must also ask who decides whether infrastructure is justified, whose knowledge is recognised, and how claimed benefits are weighed against territorial and social harms. It must also address the possibility of refusal rather than assuming that community engagement only means improving a project that has already been chosen.

## 3. Integrating Insights to Construct a Revised AI for Sustainability Framework

### *3.1. Tightening the Necessary and Sufficient Conditions*

I now construct an altered AI for Sustainability framework that considers AI's political and social impacts by integrating the insights from the analysis of both central papers. The original framework's conditions remain, with additional criteria added to address the distribution of benefits and negatives of AI systems, stronger sustainability analysis, and the evaluation of alternative social innovations and practices. As the decision to use necessary and sufficient conditions is central to Falk and van Wynsberghe's framework, the relationship between the conditions and tiers needs to be explicit. In this revised framework, a necessary condition is required for a project to move beyond the lowest tier. All sufficient conditions are required jointly, together with all necessary conditions, for the strongest classification of AI for Sustainability. So, a single sufficient condition alone does not guarantee sustainability. Rather, together they are jointly sufficient conditions to collectively demonstrate that the project has moved from information and reporting to accountable environmental action.

The first tier, AI in "Application Domain", applies where a system provides monitoring or information relevant to an environmental domain. The second tier, AI towards Sustainability, applies where the system meets the necessary transparency conditions: Monitoring and Information, Benefit and Harm Distribution Assessment, and Carbon Emission Reporting and Enhancement. The third tier, AI for Sustainability, applies when the prior necessary and forthcoming sufficient conditions are met, which are Community Engagement and Impact Reconciliation, Alternative Social Perspectives and Innovations Evaluation, and Action. Thus, the strongest label is reserved for projects demonstrating environmental contribution and social accountability.

### *3.2. The Revised Conditions*

**Monitoring and Information (Necessary).** This condition is retained from Falk and van Wynsberghe's (2024) framework. The AI system must generate information that could have a positive environmental impact in at least one UN environmental impact category.

**Benefit and Harm Distribution Assessment (Necessary).** This is a new condition. The project must make transparent the expected benefits and harms of the AI system, including who receives the benefits and who bears the costs across development, deployment, data infrastructure, mineral extraction, energy use, water use, labour, and waste. This is necessary to directly address the political element of AI missed in the original framework. Without it, a project could deliver environmental gains for a

powerful organisation while displacing costs onto communities elsewhere.

**Carbon Emission Reporting and Enhancement (Necessary).** This expands the original sustainability analysis condition. Reporting carbon emissions from training is important, yet the condition should also include inference where possible and evidence that technical improvements to energy efficiency have been considered. Showing a deliberate effort to improve the sustainability of AI at the training and inference stages is easy-to-implement at the developer level without requiring a full understanding of the AI infrastructure. The prior iteration of this condition only included reporting carbon emissions from training and failed to demand progress despite the current unsustainable status quo of AI. The condition is therefore shifted from just disclosing emissions to showing attempts to improve on them.

**Community Engagement and Impact Reconciliation (Sufficient).** To meet this new condition, a project must demonstrate meaningful engagement with communities affected by the AI system, including infrastructures, and it must show that this engagement is consequential. Engagement must occur early enough to influence design, siting, deployment, mitigation, benefit-sharing, or refusal. Lehuedé's (2022) chululo example demonstrates the need for this, showing that affected communities may perceive infrastructural harms that formal experts overlook, so their knowledge must have authority within the process. As AI infrastructure can perpetuate colonial power imbalances, it is crucial to demonstrate open dialogues with communities and areas that AI systems depend on, and to accommodate their concerns.

**Alternative Social Perspectives and Innovations Evaluation (Sufficient).** To avoid climate isolationism, a new sufficient condition is needed to consider alternative social practices and innovations for sustaining the environment before committing to AI. Engaging with affected communities could reveal existing social practices that could be applied elsewhere, or ontologies that may be lost with the expansion of AI infrastructure. Additionally, engaging with social innovation literature like Bergman et al.'s (2010) review of climate social innovations could reveal movements like Transition Towns, which engages local communities with practices to reduce fossil fuel dependence (2010). Rather than using AI to manage energy demands in a community's smart grid and slowly phase in renewable energy, such social schemes may have similar positive local impacts without the wide-ranging negatives of a currently unsustainable technology impacting communities elsewhere. So, projects should consider both the social practices that may be lost by locking in AI infrastructure, and show that AI use is warranted over alternative social innovations and practices suggested by communities and social innovation literature.

**Action (Sufficient).** This final sufficient condition retains Falk and van Wynsberghe's (2024) central insight. The AI system must lead to an action that demonstrably positively contributes to an environmental impact category versus just producing information that could be beneficial.

Figure 1 shows the revised framework. The tiered structure remains, but conditions regarding the political and social impacts of AI are added to ensure that AI for Sustainability is as sustainable as possible. The sustainability of AI is now considered across multiple conditions to reflect its different impacts. If all necessary conditions are met, an AI system is considered "AI towards Sustainability", and where all necessary and sufficient conditions are met, a system can genuinely be considered "AI for Sustainability".

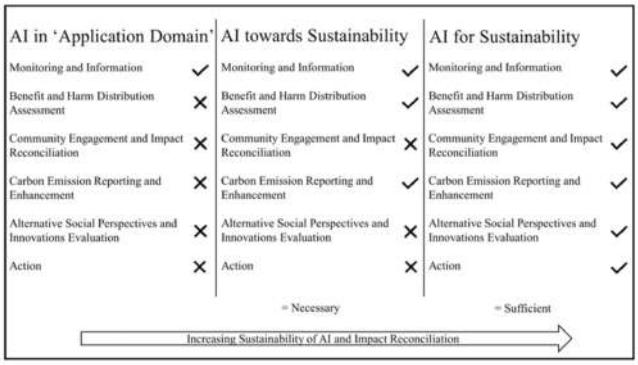



**Figure 1:** Tiered framework for classifying sustainable AI research (own illustration).

Under this framework, the AI for Sustainability label is reserved for systems that demonstrate environmental benefit while also accounting for infrastructural harms, community authority, and plausible non-AI alternatives.

### *3.3. External Rating, Refusal, and the Limits of Checklist Governance*

This article draws on relational ontology and the possibility of refusal while also proposing a framework that, if self-applied by companies or research teams, could be reduced to an organisational checklist. Thus, it could be vulnerable to greenwashing, with organisations defining harms, judging their own engagement and certifying their own sustainability. So, the framework should be treated not as a private self-assessment but as the basis for an external, public rating model. A robust model for this is seen in the Fairwork project, which publishes annual public ratings of how fairly workers in the digital platform economy are treated (Cant et al, 2023). A similar model for AI for Sustainability would assess projects against public criteria, require documentary evidence, include community testimony, and make ratings visible, resulting in public accountability. It would also remove reliance on self-regulation and pressure organisations to perform better. If an affected community has not been meaningfully engaged, if benefits and harms are not mapped, or if a project depends on infrastructure that communities reject, the rating should reflect this. The strongest AI for Sustainability classification should therefore be unavailable where the project's environmental output depends on overriding those most exposed to its infrastructural costs. To this end, it is important to operationalise the framework's conditions in this context.

### *3.4. Operationalising the Framework*

Having established the need for an external rating model, I now operationalise how the conditions should be assessed through such a model. Benefit and Harm Distribution Assessment would require a published lifecycle impact map showing where benefits and harms occur across deployment, energy and water use, mineral extraction, labour, waste, and data infrastructure, with uncertainty clearly stated rather than concealed. Plociennik et al. (2025) propose a practical approach to perform such an analysis, which includes scoping the system, creating a lifecycle inventory, assessing impacts, and interpreting results. Carbon Emission Reporting and Enhancement would require disclosure of training and inference emissions, alongside evidence of attempts to improve model and hardware efficiency beyond current standards. Alternative Social Perspectives and Innovations Evaluation would require a comparative justification showing why AI is preferable to non-AI interventions, including community-led practices, institutional reform, low-tech monitoring, or local ecological knowledge. Action would be assessed through evidence that the system has produced a demonstrable positive contribution to at least one environmental impact category, versus just generating information. Each condition would need to be supported by public evidence, independent review, and community testimony where relevant, ensuring that the AI for Sustainability label is awarded only where environmental gains are matched by infrastructural and social accountability.

Community Engagement and Impact Reconciliation as a sufficient condition requires deeper operational detail, as participatory design initiatives can often be performative, failing to truly give communities a voice. For instance, Delgado et al. (2023) argue that participatory AI design initiatives often barely consult stakeholder groups, noting that of 68 participatory AI projects studied, only 11 engaged with stakeholders multiple times (2023). With the permanence of material impacts like eradicating Indigenous worlds, performative gestures are not enough, so organisations must show that impacted communities have a meaningful voice. I therefore work this condition through concretely: what evidence satisfies it, who assesses it, and against what standard.

**Table 1:** Operational standard for Community Engagement and Impact Reconciliation.

| Element | Evidence required | Assessor | Minimum standard |
|---|---|---|---|
| Affected-community mapping | Map communities affected by deployment, data infrastructure, energy and water use, mineral supply chains, labour, waste streams, and territorial change. Include uncertainty and explain exclusions. | Independent review body with community-nominated members. | Published impact map. |
| Early, repeated community engagement | Records of meetings, workshops, deliberative forums etc., and time for response before design lock-in. | Independent review body, with community representatives' sign-off | Engagement must occur before irreversible decisions. |
| Design influence | Change log showing how community input altered system design. | Independent review body, with community representatives' sign-off. | Consultation is insufficient without creating documented changes or enforceable constraints. |
| Right to refuse and grievance | Clear process for objection, refusal, appeal. Public record of unresolved objections. | Independent review body | Objections must be resolved. |
| Benefit-sharing and repair | Agreements covering local benefits, remediation, community ownership/stakes where appropriate. | Independent review body and community representatives. | Benefits must be tangible, locally meaningful, and revisable. |

This operationalisation, summarised in Table 1, follows from the methodological lesson of Lehuedé's chululo example, that affected communities' knowledge must be considered and given authority. Community Engagement and Impact Reconciliation therefore cannot be satisfied by one-off consultations or by publishing stakeholder-engagement statements after design decisions have already been made. It requires repeated engagement, community verification, and evidence of change in response. A strong example of this can be found in the Ada Lovelace Institute's Citizens' Biometrics Council (2021). Unlike the consultations criticised by Delgado et al. (2023), the council utilised a "mini-publics" (2021) format where 50 diverse citizens who would be impacted by biometrics in the UK deliberated over several months, cross-examining experts to co-create policy recommendations for biometric AI. Thus, identifying impacted communities, adopting similar structured, repeat engagements where they are given the time, resources, and authority to influence AI system development can ensure meaningful collaboration resulting in sustainable AI. Crucially, there must also be evidence of reconciliation, meaning this participatory design should demonstrate clear change where concerns or challenges are raised, like altered siting, rejection of components, or implementation of ideas from energy democracy e.g. community stakes in local data centres and infrastructure (Szulecki, 2018).

Where a community refuses a project because its infrastructural demands threaten territory, livelihood, or world-making practices, that refusal should be treated as decisive evidence against classifying the project as AI for Sustainability. Ultimately, as Lehuedé (2022) shows, embodied practices are as important as discourse, so truly involving communities in projects could help preserve local environments and worlds by giving them a genuine stake in how AI infrastructure operates. Although this framework adds complexity, AI is highly complex in its infrastructure and in its political and social impacts throughout that infrastructure. Therefore, more complexity is warranted. The revised framework does not require perfect knowledge of every supply-chain relation before any AI project can proceed. It requires a higher evidential burden before a project can claim the strongest sustainability label. The label should be demanding because it carries legitimising power.

### *3.5. Applying the Framework: Google DeepMind Wind-Power Scheduling*

The framework can be illustrated by applying it to the Google DeepMind wind-power scheduling example discussed earlier. Under the previous framework, the system almost met every AI for Sustainability condition. However, here it would not receive the strongest classification with existing evidence unless it publicly accounted for the AI system, its infrastructures, and its distributed impacts. Table 2 therefore shows how the framework changes the assessment.

**Table 2:** Applying the revised framework to Google DeepMind wind-power scheduling.

| Condition | Condition met on available evidence? | Reason |
|---|---|---|
| Monitoring and Information (Necessary) | Yes | System generates information relevant to wind-power delivery. |
| Benefit and Harm Distribution Assessment (Necessary) | No | The available account only emphasises benefits to Google. |
| Carbon Emission Reporting and Enhancement (Necessary) | No | The example does not evidence this information for the AI system itself. |
| Community Engagement and Impact Reconciliation (Sufficient) | No | There is no public evidence of external engagement. |
| Alternative Social Perspectives and Innovations Evaluation (Sufficient) | No | The example does not compare the AI system with non-AI, community-led, or institutional energy alternatives. |
| Action (Sufficient) | Partly | The AI system improves wind-power delivery, but only for Google. |

Per the conditions met and the matrix in Figure 1, the project is firmly in the AI in "Application Domain" tier based on publicly available evidence. A case that previously looked relatively strong is revealed to require substantial work to harbour any sustainability claim under the revised framework unless benefits, harms, community authority, and alternatives are assessed. The system is not unsustainable in every respect, but the AI for Sustainability label should be withheld until infrastructural accountability is demonstrated.

### *3.6. Limitations*

Limitations include supply chains and data infrastructures being inherently opaque (Crawford and Joler, 2018), so the framework cannot demand perfect knowledge, instead requiring public evidence and transparency regarding uncertainty. Also, external rating bodies may reproduce the power inequalities they aim to assess, which is why community-nominated assessors and testimony are critical safeguards. Furthermore, community refusal may create difficult tensions with urgent climate action, especially where proposed systems promise emissions reductions. Community refusal does not necessarily mean a project cannot proceed, but it should prevent the project from receiving the strongest AI for Sustainability label while material objections remain unresolved. Finally, the comparison of AI versus non-AI alternatives will be context-specific and contested, thus difficult to objectively analyse. This is a reason to make comparison explicit rather than to avoid it, and to involve community stakeholders in analysis.

## Conclusion

To conclude, this paper has closely analysed and built upon Falk and van Wynsberghe (2024) and Lehuedé's (2022) work to define AI for Sustainability. I demonstrated significant gaps in Falk and van Wynsberghe's AI for Sustainability framework as AI's political and social impacts are not addressed. Then, I integrated insights from the analysis of both papers to produce a modified AI for Sustainability framework with a greater emphasis on power, infrastructure, distribution, community authority, and alternative social innovations. This enables the AI for Sustainability label to truly represent a system that is the best solution for a given sustainability issue, and that progresses the sustainability of AI both technically and socially through sincere engagement with communities impacted by the AI system. The operationalisation shows how the framework can move beyond abstract critique, specifying required evidence, assessors, and minimum standards, while the Fairwork-style rating model shows how the framework can avoid being reduced to self-certification. Future work could include in-depth analyses of popular AI for Sustainability projects against this framework, or further developing the conditions. Ultimately, AI systems are unsustainable if their environmental benefits depend on hidden, uncontested destructive infrastructure. The AI for Sustainability label should be reserved for projects with environmental gains matched by accountable infrastructures, community authority, and genuine openness to social alternatives.

## References


Ada Lovelace Institute (2021) The Citizens' Biometrics Council. London: Ada Lovelace Institute. Available at: https://www.adalovelaceinstitute.org/report/citizens-biometrics-council/.

Bergman, N., Markusson, et al. 2010, February. Bottom-Up, Social Innovation for Addressing Climate Change. In *Sussex Energy Group Conference—ECEEE*.

Brennan, A. (2024) 'Necessary and Sufficient Conditions', in Zalta, E.N. and Nodelman, U. (eds.) *The Stanford Encyclopedia of Philosophy*. Summer 2024 edn. Available at: https://plato.stanford.edu/archives/sum2024/entries/necessary-sufficient/.

Brodie, P. (2023) 'Data Infrastructure Studies on an Unequal Planet', *Big Data & Society*, 10(1), p. 20539517231182402.

Cairns, R.C. (2014) 'Climate Geoengineering: Issues of Path-Dependence and Socio-Technical Lock-In', *Wiley Interdisciplinary Reviews: Climate Change*, 5(5), pp. 649-661.

Cant, C., Ustek-Spilda, F., Cole, M. and Graham, M. (2023) *Fairwork Annual Report 2023: State of*

*the Global Platform Economy.* Oxford: The Fairwork Project. Available at: https://www.fair.work.

Crawford, K. and Joler, V. (2018) Anatomy of an AI System: The Amazon Echo as an Anatomical Map of Human Labor, Data and Planetary Resources. New York: AI Now Institute and Share Lab. Available at: https://anatomyof.ai.

Dauvergne, P. (2022) 'Is Artificial Intelligence Greening Global Supply Chains? Exposing the Political Economy of Environmental Costs', *Review of International Political Economy*, 29(3), pp. 696-718.

Delgado, F., Yang, S., Madaio, M. and Yang, Q. (2023) 'The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice', *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1-23.

Delmas, M. and Burbano, V. (2011) 'The Drivers of Greenwashing', *California Management Review*, 54(1), pp. 64-87. Available at: https://doi.org/10.1525/cmr.2011.54.1.64.

Dong, Y. and Hauschild, M.Z. (2017) 'Indicators for Environmental Sustainability', Procedia CIRP, 61, pp. 697-702.

Ensmenger, N. (2021) 'The Cloud is a Factory', in *The MIT Press eBooks*, pp. 29-50. Available at: https://doi.org/10.7551/mitpress/10993.003.0005.

ExxonMobil (2019) ExxonMobil to Increase Permian Profitability through Digital Partnership with Microsoft. Available at: https://corporate.exxonmobil.com/news/news-releases/2019/0222_exxonmobil-to-increase-permian-profitability-through-digital-partnership-with-microsoft.

Falk, S. and Van Wynsberghe, A., 2024. Challenging AI for Sustainability: What Ought It Mean?. *AI and Ethics*, 4(4), pp.1345-1355.

IPBES (2019) Global Assessment Report on Biodiversity and Ecosystem Services. Edited by E.S. Brondizio, J. Settele, S. Díaz and H.T. Ngo. Bonn, Germany: IPBES Secretariat. Available at: https://doi.org/10.5281/zenodo.3831673.

Lehuedé, S. (2022) 'Territories of Data: Ontological Divergences in the Growth of Data Infrastructure', *Tapuya: Latin American Science, Technology and Society*, 5(1), p. 2035936.

Maslej, N., Fattorini, et al. (2024) The AI Index 2024 Annual Report. Stanford, CA: AI Index Steering Committee, Institute for Human-Centered AI, Stanford University.

Mensah, J. (2019) 'Sustainable Development: Meaning, History, Principles, Pillars, and Implications for Human Action: Literature Review', *Cogent Social Sciences*, 5(1), p. 1653531.

Microsoft Research (2024) Microsoft Climate Research Initiative. 17 April. Available at: https://www.microsoft.com/en-us/research/collaboration/microsoft-climate-research-initiative/.

Muldoon, J. (2023) Platform Socialism: How to Reclaim Our Digital Future from Big Tech. London: Pluto Press.

Plociennik, C., Watjanatepin, P., Van Acker, K. and Ruskowski, M., 2025. Life Cycle Assessment of Artificial Intelligence Applications: Research Gaps and Opportunities. *Procedia CIRP*, 135, pp.924-929.

Powell, A.B., Ustek-Spilda, F., Lehuedé, S. and Shklovski, I. (2022) 'Addressing Ethical Gaps in Technology for Good: Foregrounding Care and Capabilities', *Big Data & Society*, 9(2), p. 20539517221113774.

Ritchie, H. and Roser, M. (2024) 'CO2 Emissions', *Our World in Data*, 22 January. Available at: https://ourworldindata.org/co2-emissions.

Stephens, J.C. (2022) Diversifying Power: Why We Need Antiracist, Feminist Leadership on Climate and Energy. Washington, DC: Island Press.

Stilgoe, J., Owen, R. and Macnaghten, P., 2020. Developing a Framework for Responsible Innovation. In *The Ethics of Nanotechnology,*

*Geoengineering, and Clean Energy* (pp. 347-359). Routledge.

Szulecki, K. (2018) 'Conceptualizing Energy Democracy', *Environmental Politics*, 27(1), pp. 21-41.

Thompson, N.C. and Ahmed, N. (2023) 'What Should be Done About the Growing Influence of Industry in AI Research?', *Brookings*, 5 December. Available at: https://www.brookings.edu/articles/what-should-be-done-about-the-growing-influence-of-industry-in-ai-research/.

Verdecchia, R., Sallou, J. and Cruz, L. (2023) 'A Systematic Review of Green AI', Wiley Interdisciplinary Reviews: *Data Mining and Knowledge Discovery*, 13(4), p. e1507.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S.D., Tegmark, M. and Fuso Nerini, F. (2020) 'The Role of Artificial Intelligence in Achieving the Sustainable Development Goals', *Nature Communications*, 11(1), pp. 1-10.

Witherspoon, S. and Fadrhonc, W. (2019) 'Machine Learning Can Boost the Value of Wind Energy', *Google: The Keyword*, 26 February. Available at: https://blog.google/technology/ai/machine-learning-can-boost-value-wind-energy/.

# Choosing the Path of Most Resistance: Metacognitive AI Literacy and Mitigating Epistemic Risk

*Iman Khwaja*
*University of Edinburgh*



This paper reimagines metacognition as both an individual tool and a component of AI (Artificial Intelligence) literacy, that acts as a critical pathway to mitigating the epistemic threats posed by AI tools. These threats are explored as the erosion of autonomy, critical thought and creative abilities in individuals, as a result of cognitive dependencies on AI tools.

Metacognition, a strategy involving the active preservation of cognitive ability in individuals, is currently a subtle component of many literacy approaches in varied educational contexts. This paper positions it as a practice that could play a meaningful role in the trajectory of AI development in tandem with the consequent changes to our collective epistemic and cognitive experiences. Specifically, the aspects of self-restraint, critical evaluation and epistemic investigation proposed by traditional metacognitive strategy are reconceptualized as potentially critical aspects of AI literacy efforts.

Drawing on literature that explores AI deskilling, epistemic risk, which Babui describes as "the likelihood that one's claims inaccurately represent the world" (2019), and digital literacy, this paper proposes a framework for users, developers and deployers to consider applying metacognitive strategy to AI literacy efforts, as a means to mitigate the epistemic risks posed by user dependencies on AI tools.

**Keywords**: AI Literacy, Metacognition, Epistemia, Epistemic Risk, AI Ethics

## Introduction

The proliferation of Artificial Intelligence (AI) models is changing the landscape of literacy and knowledge acquisition in an unprecedented way. AI models that operate by scraping and synthesizing large amounts of data to produce personalized responses to user queries are referred to in this paper generally as "digital concierges", as an umbrella term for LLM's (large language models) and other models that operate similarly. These responses have been described as "oracular" (Wihbey, 2024) in nature, in which the black-boxing of the data behind these informational outputs and the elimination of the traditional search by search engine process allows users to avoid tracing information to its source and evaluating its validity. While these rapid and articulate outputs offer evidently helpful enhancements to our own epistemic experiences, the question remains as to whether we are able to protect our cognitive wellbeing amidst increasing epistemic accessibility. This paper advocates for the development of literacy strategies that acknowledge the cognitive implications of AI use in tandem with its development. Southworth et al. reaffirm this suggesting that "*academic and industrial bodies need, in a symbiotic way, to reaffirm responsible behaviour*" (2023), which this paper proposes can be done through embedding reconceptualized metacognitive strategies to support users in various use contexts.

This paper poses the following questions as a means to investigate these opportunities preliminarily, and to prompt further research into metacognition as a component of decision-making for educators, users and developers of AI technologies interested in cognitive wellbeing:

1. Should cognitive wellbeing be a critical consideration in future AI literacy efforts in the context of epistemic AI use?

2. How can existing metacognitive frameworks be reimagined and applied to encourage individual and broader regulation of cognitive wellbeing?

3. What adjustments can and should institutions make to augment the

responsible and critical epistemic uses of AI?

4. How could a broad metacognitive approach to AI use encourage appropriate reliance on AI tools across a variety of use cases?

This paper identifies several key concepts and preliminarily outlines these below, that form the basis of the response this paper constructs to the above questions.

1. **Metacognition**: The ability to work critically with information and conduct reasoning in response to information by evaluating it using epistemic beliefs. Pertaining to knowledge about human cognition and the regulation of these cognitive processes with systemic, institutional and self-directed support.

2. **Epistemic Risk**: Generally threats to the ability to discern validity, truth and critically evaluate knowledge, the sources of knowledge and the structures of knowledge with accuracy.

3. **Cognitive Wellbeing**: The status of judgement, reasoning, critical thinking and assessment capabilities in individuals. With atrophy of these facets of cognition being a limitation to this wellbeing, being able to employ these in good standing in response to information is indicative of cognitive wellbeing (Salovich et al, 2021).

4. **Appropriate Reliance**: John Coleman's argument that "*AI is already beginning to sit between people and their own thinking - how we read, write, reason and form beliefs*" provides a starting point to determine appropriate reliance. This concept broadly refers to how users are able to contextualize their use of AI in a way that does not impede their cognitive wellbeing or pose an epistemic risk to their thinking processes. Contextual use is paramount in consideration of appropriate reliance, where some users have more autonomy over determining their use of and exposure to AI tools than others depending on their required application.

**Brief Assessment of Current AI Literacy Efforts**

The current landscape of AI literacy is arguably limited. Amongst the presence of many AI ethics and adjusted digital skills courses available to individuals online, there is a clear lack of standardised AI literacy programs as enforced components of national education in the UK. The publication of the AI Opportunities Action Plan in early 2025 written by Matt Clifford recalls a need to "*support institutions to increase the numbers of AI graduates and teach industry-relevant skills*" (p. 9) but does not detail how these graduates and students will be supported to be well-adjusted and cognitively strong users of these technologies.

The creation of dependence is a concern expressed across the literature, summarised by Kalantzis' dilemma: "*Why read, when a text can be spoken to you, and can you be told to speak with just the right level of difficulty for you?*" (2024). In the last year, numerous efforts to facilitate an ethical and critical approach to the perception and use of AI have emerged. In 2023, Kanta Dihal and Tania Duarte published Better Images of AI: A Guide for Users and Creators, as a primary reference point to facilitate "*accurate and compelling communication*" about AI (p. 3) in visual contexts. Similar conceptualizations such as The Philosophical Glossary for AI, created by Dr Alex Grzankowski for the London Humanity Project, consist of entries by various authors working to redefine existing terms such as "consciousness" and "bias" in the context of relationality to AI tools, that reveal the critical importance of reconceptualization in the AI literacy process itself. Interestingly, Brian Ordegaard and Diego Morales submitted an entry to the glossary redefining "metacognition" as a capability held by AI systems themselves.

While this paper acknowledges the value of exploring this iteration of metacognitive ability in digital contexts, this paper goes one step further to argue that employing metacognitive strategy as a responsive measure to the use of AI systems is a critical discussion that should coexist with exploring parallel abilities in AI models. In either case, the definition of metacognition as "*a means for systems to assess their own reliability*" that Ordegaard and Morales have outlined is useful to the arguments presented in this paper. This "assessment" process is therefore the central premise of the metacognitive framework constructed in this paper. Within this, the abilities to "*remember, reason, judge, think, and believe*" are included as cognitive processes that constitute metacognitive practice.

For the sake of discussion, this paper uses the phrases "knowing thyself" and "choosing the path of most resistance" to informally summarize the complexity of metacognition, a reiteration of what Yi describes as "knowing what to know" in their definition of metacognition (2025). Importantly, Yi also describes "the purpose of AI literacy" as "to have anticipation capabilities", in which "*metacognition involves an effort to anticipate an uncertain future*" (2025). These definitions of metacognition as a constituent competency of broader AI literacy are pertinent to the discussions outlined in this paper, which will be summarized in a framework that hybridizes traditional metacognitive concepts and their reconceptualizations for applied AI literacy strategy.

This paper positions this framework as an interventional strategy that ensures that "*the convenience offered by AI tools does not come at the cost of essential cognitive skills*" (Zhai et al, 2024). In "choosing the path of most resistance", I invite users, developers and educational deployers to engage in and promote metacognitive practice as a means to preserve their epistemic autonomy.

## 1. What is metacognition?

Developmental psychologist John Flavell coined the term metacognition in 1976 to denote the concept of "cognizing cognition itself" (Noushad, 2008). While definitions are varied and the historical roots of metacognition are complex, this paper benefits from the qualities of "self-appraisal" and "self-consciousness" that (Noushad, 2008) suggest are fundamental to metacognition, reiterated by Flavell's own assertion that "*metacognitive strategies monitor the process of learning and task completion*" (1976). Baker and Brown's assertion that Flavell's idea consisted of two main "*clusters of activities: knowledge about cognition and regulation of cognition*" (1980) suggest the importance of self-intervention across the board, getting to know one's own patterns of cognition and acquiring the skills to regulate them through continuously developed familiarity with these patterns, thus centralising an awareness of the self as crucial for effective metacognitive practice.

External to Flavell's view that "*critical thinking is subsumed by metacognition*" (1979), central ideas of self-regulation and monitoring are present in a wide spectrum of definitions of metacognition. Hennessy demonstrates this in their description of the "*multidimensional character of the term metacognition as: (1) an awareness of one's own thinking; (2) an awareness of the content of one's conceptions; (3) an active monitoring of one's cognitive processes*" (1999). This definition is consistent with Noushad's analysis of the "self-regulatory mechanisms" involved with metacognition, "*checking outcomes, planning the next move, monitoring effectiveness, testing, revising and evaluating one's strategies for learning*" (2008).

Consequently, various sources suggest metacognition is akin to critical thinking in many respects and could therefore be defined concretely as: the awareness of one's own cognitive processes and the continuous self-regulation and monitoring of these cognitions through an intentional and critical revisiting of them. To return to Yumi's definition of metacognition for AI literacy specifically, "knowing what to know" and "the anticipation of ambiguity" affirm the essence of these early definitions of metacognition.

Contemporary metacognitive research remains largely in the realm of its applications to traditional education. However, this paper attempts to explore the role of metacognition in

the context of digital epistemologies. It is therefore important to align the shifting landscape of epistemologies with the relevant principles of metacognition, to ensure it remains an applicable and useful practice in the face of unprecedented technological development. The value of metacognition in mitigating epistemic risk and preserving cognitive wellbeing against inaccuracy is relatively well recorded in the body of literature, where Salovich et al. conducted an experiment investigating this that applied metacognitive tasks to encourage evaluation in participants, demonstrating that participants "*made fewer judgement errors after reading inaccurate statements*" after completing it. This paper argues that a similar, adjusted practice could protect users against the implications of AI generated inaccuracies. This calls for a basic referential framework for the principles of metacognition to be reconceptualized, so that they may be ascertained and adjusted according to their utility for engagement with AI.

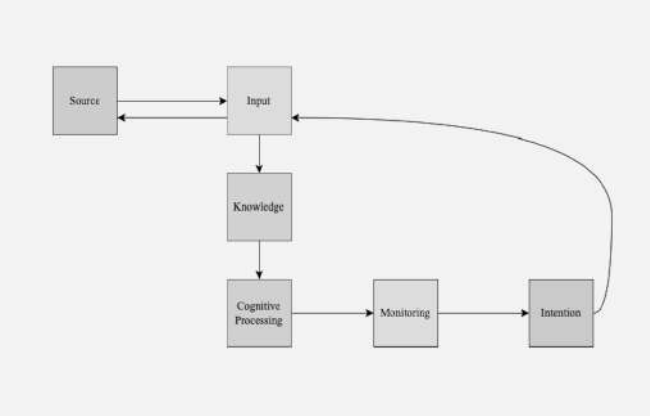



**Figure 1:** Referential Model for Metacognitive Practice

### *1.1. Decoding Metacognition: Figure 1 Key*

Source: Refers to the source of information itself in any capacity, such as visual or textual inputs.

Input: Refers to how the learner goes about acquiring the relevant knowledge required to perform the task, such as locating the information spatially, asking a question or inputting search terms manually or digitally. The bidirectional arrow between source and input is to demonstrate the relationship between the existence of the source in its entirety and the learner's intentionality to locate relevant information within this greater whole.

Knowledge: This term refers to the process that takes place prior to the extraction of meaning from the input or any inference that takes place and remains an abstraction or a conceptualization based on a priori understandings of the input.

Cognitive Processing: This term broadly encompasses the information ascertained by the learner. This involves the processes of "*remembering, reasoning, representing, thinking and believing*" outlined by Ordegaard and Morales in their definition of metacognition.

Monitoring: The distinctive aspect of metacognition, this stage is the crucial evaluation, reflection and other "self-regulatory mechanisms" (Baker and Brown, 1999) that allow the learner to retrace the steps of their own cognitive processing and knowledge stages. This is the formation of awareness of how this input was received by the learner and where repetitive learning patterns are established that allow the learner to make retroactive changes to cognitive processing.

Intention: I have added this to the diagram despite it not being explicitly mentioned in the metacognition body of literature. In this visual it acts as a consequential pathway that demonstrates how metacognitive activity changes the learner's intentions to pursue information in ways that are familiar, effective, or continually ineffective, depending on the success of the monitoring stage. It acts as a strategic factor that is a result of the "re-visiting" process in metacognition and determines how the learner plans to approach future situations.

### *1.2. Defining Epistemic Threat*

Threats emerging from the use of intelligent technology are numerous and ever-changing. Epistemic threat considerations currently remain in the background of most large-scale risk mitigation efforts, but it is not difficult to visualize how AI tools that take on "digital concierge" roles by providing users with personalised epistemic experiences could become threatening in their own way. Crucially,

identifying these threats is part of what Yi describes as the importance of using literacy to support our collective abilities to "*explore uncertain and complex societies, and predict future problems and find solutions*" (2021). LaGrandeur subsequently defines epistemic risk as a threat to our capacities to "*make decisions, conduct analytical reasoning and regulate our creative urges*" (2021), which this paper identifies as the components of metacognition threatened by epistemic dependence on AI proxies. Faced with resolving these dependencies, a widely applicable literacy strategy in tandem with the self-regulation aspects of metacognition is proposed in this paper as a pathway towards mitigating at least some of the epistemic threats posed by proliferating complex technologies.

Several empirical cases of epistemic risk have emerged in a growing body of evidence demonstrating an increase in concerning user-machine relationships. Kosymna et al. conducted a study producing neurological evidence through electroencephalography (EEG) scanning for the phenomenon of "cognitive debt" emerging in participants, where the results showed "*diminished critical thinking, reduced creativity and independent thought, vulnerability to manipulation and shallow information processing*" in groups that delegated tasks to A (2025) .

Further, a report by Baracui evaluating AI as a supplement for academic study indicated that the unrestricted use of AI tools in undergraduate students impeded retention capabilities for academic materials (2024). These findings not only indicate an exacerbation of cognitive complications in generalized AI use, but also shifts in the perceptions of these agents as they become increasingly personalized. The extremes of emotional dependency on AI models are now well documented, particularly in the wrongful death lawsuit filed by the family of sixteen-year-old Adam Raine against OpenAI in August 2025, that claimed OpenAI's leading chatbot ChatGPT played a role in his decision to take his own life. Whether the failure exhibited by the model was technical or an issue of guardrail development, the epistemic complications of Raines' relationship with ChatGPT as an agent itself remains of interest.

Within the scope of this paper, epistemic threat is considered as related to complications of dependency, loss of critical thought, clouded judgement and perception and the general erosion of cognitive wellbeing. This paper briefly identifies the following hypothetical epistemic threats with the intention to contribute to mitigating them. The threats identified use examples that position the end-user in an affected position, to illustrate the potential for epistemic threats to proliferate without intervention. In identifying these epistemic threats, the intention is not to be presumptuous about the realism or probability of these outcomes, but to depict them in a hyperbolised manner, to prompt further research on interventional response measures that cover the full extent of possibility.

Dependence: The establishment of a reciprocal dependency between the user and the AI algorithm where increasingly personalised outputs replace the user's ability to engage with other neutral and non-personalised sources of information. The user therefore is unprepared to engage with sources that do not produce outputs that reflect the users' cognitive patterns as ascertained by the algorithm. The user depends on the AI as a singular source for most tasks, damaging the user's relationship with other sources of information.

Cognitive Offloading: A by-product of dependence, the user becomes comfortable offloading cognitive processes such as critical thinking, evaluation, and judgement to the AI tool. This might look like asking the AI to critically evaluate a source on their behalf and adopting this evaluation as their own without conducting an objective appraisal of it, without interference or consultation from the AI. The user is not able to break the barrier of superficial understanding of key concepts due to the rapid generation of seemingly accurate and useful information.

Loss of Decisive Autonomy: The user's ability to make decisions about how to approach educational tasks is damaged significantly, and the user struggles to trust their own decision-making skills without the guidance or advice of an AI tool that algorithmically reflects their interests. The user is uncomfortable engaging

with cognitive trial and error to ascertain the best strategies for their own learning outcomes.

Erosion of Critical Thought and Judgement Capabilities: The user becomes unable to critically evaluate sources and form their own judgements without consulting the AI to seek confirmation, affirmation, or guidance. The user is unable to organically assess sources against other textual or digital sources without requiring that they be summarised in an accessible and personalised format.

Creative Injury: The user struggles to extract meaning from sources and produce novel or original solutions to challenges. The user loses the ability to adapt to new styles, formats, or contexts of information, and equates creative and experimental outputs with threatening unfamiliarity. The user no longer perceives a relationship between emotional resonance and creative cognition and perceives creativity as possible only through digital mediums or in digital contexts. The user becomes unable to adapt to new, unfamiliar situations and is discomforted by the challenge of creative problem-solving with other humans to meet demands. The user becomes comforted by the confirmation bias offered by the AI tool and no longer seeks collaborative creative work with other humans.

Intuitive Confusion: Users struggle to make intuitive decisions about when an AI tool should be overridden in decision-making instances. Users struggle to reconcile their a priori knowledge and beliefs with the recommendations of the AI and become intuitively confused about actions thereafter. Users are impacted intuitively and are unable to make intuitive decisions or judgements without AI assistance or confirmation.

Perceptual Fragmentation: The user no longer perceives the tool as a recursive product of their own algorithmic interactions, they become familiar with the tool as an extension of their own cognition or perceive it as a cognitive tool to fill gaps in their own capabilities. The user can no longer perceive their own cognitions, judgements, and independent criticisms as unique or separate. The user perceives AI as intellectually and creatively superior and therefore shifts their perception of their own capabilities and the capabilities of others.

The above threats have not been concretely identified in any official capacity but are present across the literature as common concerns about the reliance on AI modalities for educational and personal information. In establishing these epistemic threats in a more concrete manner, we create opportunity pathways for response capabilities. The following section will explore how metacognitive principles can be adjusted for AI to represent these protective measures in a widely applicable and referential way.

## 2. How Can Metacognition be Adjusted to Protect Against Epistemic Atrophy Caused by AI?

In response to the above outlines of metacognitive principles and epistemic threat, a conceptualization of actionable metacognitive markers can be developed in the context of AI use.

The applicability of this is to create an interventional starting point for researchers, deployers and educators to reference as a means to embed risk awareness and the natural preservation of cognitive wellbeing into AI literacy and epistemic risk mitigation efforts alike. The target of cultivating metacognitive thinking across literacy disciplines and in individuals themselves is ultimately to encourage users to continually and critically examine their own abilities to evaluate AI generated information.

Primarily, this means gearing metacognitive thinking towards recognizing what Wihbey describes as the "*incompleteness of AI's judgement with respect to human knowledge and values*" (2024), and therefore an awareness of appropriate reliance on these technologies that is contextual. This is a logistically challenging assertion in that reliance and dependency can be largely circumstantial depending on the users fluctuating needs, environment and requirements to use AI models, and this paper recognizes that despite attempts to categorize epistemic threat and metacognitive practice, not all users benefit from this proposed framework.

To illustrate this limitation, it should be considered that a student using an AI tool for academic purposes, such as to synthesize a body of literature for research purposes and to consult with it for academic support would be more of an epistemic risk consideration than a radiologist consulting with an AI medical imaging tool for detection purposes. However, this does not exempt the radiologist from the benefit of metacognitive practice in that they retain an awareness of their oversight value and contextual medical knowledge as a physician when evaluating information outputs.

Within these contexts we can achieve "*a development of students' metacognitive skills to help them become aware of when and how to use AI tools appropriately without undermining their cognitive development*" (Gerlich, 2025). This implies that "appropriate reliance" (Kong et al, 2021) contexts for our engagement must be differentiated from the outset, to develop principles of appropriateness that allow end-users to review their motives for engaging with AI in epistemic contexts. Here we have the first few components of what metacognition for AI literacy could look like, an expansion of the traditional self-reviewing of learning behaviours to a comprehensive review of our own awareness of multiple factors that impact our cognitive health, such as knowledge of algorithmic infrastructures, changing policies and emerging threats.

Additionally, an understanding of the limitations of AI design, and indeed the "discernible cause and effect" (Zittrain, 2022) that produces its judgements, and how this is vitally different from our own. A metacognitive framework that achieves these various objectives should therefore be widely applicable and should be referential in both "classroom ecologies" (Kalantzis et al, 2024) and instances of individual need alike. This argues for the retention of what is essential to our cognitive health whilst leaving room for changes to our cognitive expenditure. Kuhn and Dean reiterate the value of metacognition in their explanation that "*metacognition is what enables a student who has been taught a particular strategy in a particular problem context to retrieve and deploy that strategy in a similar but new context*" (2004), this implies that a fundamental metacognitive framework could allow for wide cross-contextual use.

For this reason, I argue that the overarching principles of self-review, self-restraint and continuous self-assessment can be merged into the ability to simply "know thyself", from which the tenets of intention, critical evaluation and contextualised judgement emerge thereafter.

Originating from an inscription on the Temple of Apollo at Delphi, "know thyself" is a Greek maxim that denotes the importance of self-awareness principally (Green, 2024). In a literacy setting, this is not equivalent to personal introspection or philosophical musings on consciousness but is simply the idea that end-users must be empowered by and made aware of their own cognitive, critical, and creative capabilities, before engaging with seemingly equally intelligent and advanced technologies. Thus, a metacognitive framework revolving around this principle will not only highlight these functions but emphasise the importance of preserving them and tending to them against the impulse for convenience and ease.

This second main component is what I have referred to as "*choosing the path of most resistance*". The framework should emphasise the importance of doing so circumstantially, be it preserving our investigative skills by recommending self-directed search strategies or encouraging dialogue with collaborators, before engaging with AI tools that are "properly calibrated for educational application" (Kalantzis et al, 2024). This means end-users should ultimately consider on a circumstantial basis how their use of AI weighs against their awareness of their cognitive health and evolving epistemic risks and attempt to reconcile these by "choosing the path of most resistance" in an actionable way. Barbara Grosz's development of the "Intelligent Systems: Design and Ethical Challenges" Course for the Harvard University Department of Computer Science combines these concepts in a literacy context, emphasizing that they wish to "*educate students to think not only about what systems they could build, but whether they should build those systems and how they should design them*" (2019). These interdisciplinary moves toward

educating students about their own intentions, motives, and engagement with AI highlights what a metacognition framework for AI literacy could offer developers and end-users alike. The following section presents this framework as a comprehensive and referential model.

### *2.1. Epistemic Modesty: Performance Limitations and Appropriate Reliance*

In a Substack post for the Cosmos Institute written by Professor Kevin Vallier titled On the Noble Uses of AI, Vallier argues for the "*inevitability of cognitive atrophy*" amidst epistemic augmentation. Using historical examples of our deference to calculators to perform arithmetic function, Vallier's argument is closely aligned with the concept of appropriate reliance reiterated in this paper (2026). The possibilities for AI tools to enhance our intelligence, performance and overall capabilities are only growing in tandem with development efforts. Vallier presents these opportunities to cognitively offload reductive tasks as a means to "focus on other things", where "hunting for sources" when conducting research allows researchers to instead think about "what the sources mean" (2026).

It is the delicate balance between succeeding at rejecting AI's attempts to "*seduce us into passivity*" (Vallier, 2026) and to cognitively defer to these tools as a means to prioritize more meaningful work that this paper terms as "appropriate reliance", and argues that metacognitive practice adjusted to meet this challenge can help be cultivated in end users. This preservation of deliberation, reasoning and judgement capacities by investigating information synthesized by AI is the essence of what metacognitive practice applied to AI-driven contexts can achieve, but it is equally crucial that these skills are preserved alongside our knowledge of the epistemic modesty of these tools, as well as the epistemic risks posed to us by our dependencies on them.

There are several cases of AI models producing epistemically unreliable outputs, in a study conducted by Urban et al, findings revealed that 88% of students who relied on ChatGPT to supplement processes of academic writing demonstrated an enhanced performance, but that "*their own thinking remains crucial to effectively integrating information from human-generated sources not included in the training dataset of generative AI tools*" (2024). Hannigan et al. reiterate this in their argument that "*response veracity is crucial*" when assessing AI generated information, and suggests that users "*likely need to employ critical thinking skills, inductive reasoning and forecasting to estimate the veracity of a statement produced by a chatbot*" (2024).

These indications of combined cognitive augmentation and confined reliance on these tools suggest the need for referential literacy frameworks that support users in determining information veracity and achieving appropriate reliance. The following sections propose an informative model for further research and educators as a means to consider embedding this into AI literacy curriculums.

### *2.2. Metacognition for AI: Actionable Framework*

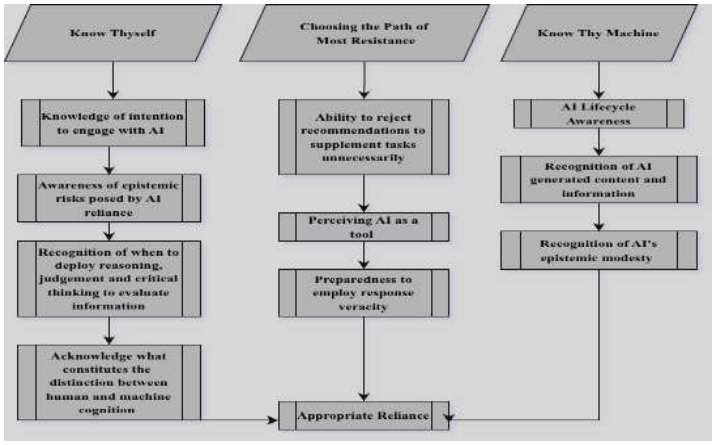



**Figure 2:** Referential Model for Metacognitive Practice for Applied AI

The below framework provides a visual summary of the content of this paper's proposition for an adjusted metacognitive application for AI use. The framework is intended to outline conceptual principles that ultimately constitute appropriate reliance, which can be expanded to various use contexts. The framework is underpinned by hybridizing risk awareness and knowledge of both AI and human cognition to determine appropriate use of AI tools for epistemic purposes.

### *2.3. Decoding Metacognition for AI*

The following section investigates the concepts outlined in Figure 2 in further depth to provide

clarification on what the metacognitive process itself entails.

**Know Thyself**: Component of awareness in the metacognitive process that preserves user autonomy and makes a crucial distinction between themselves and their perceived machine counterparts. Empowering users to embrace their uniquely human cognitive and epistemic experiences preceding their interactions with AI involves alerting users to the basic operations of human cognition and machine cognition, to make crucial distinctions between them known.

**Know Thy Machine**: The cultivated capability to criticize new kinds of information is a skill that this paper argues AI literacy curriculums should begin to cultivate in users who experience epistemic dependence on AI in any capacity, across disciplines and applications. This includes creating an awareness of the AI lifecycle of development, the role of data in exacerbating and embedding bias into AI generated information and overall working to demystify the operations of AI tools by revealing the basics of their processing operations to users. Akin to the "knowledge" aspect of the metacognitive framework presented earlier in this paper, "knowing thy machine" is critical to developing cognitively and epistemically sound relationships with these tools by creating a fundamental awareness of their limitations.

**The Path of Most Resistance**: This practice expands the metacognitive practices of revisiting, monitoring and investigating information itself to describe being equipped to critically evaluate AI outputs. In recognizing the epistemic modesty of AI tools that is often black-boxed behind the personalized and highly accessible nature of their outputs, users have the opportunity to critically evaluate and contextualize outputs by enacting reformed metacognitive skills, such as critical reasoning and evaluative judgement, to decide the validity of these outputs. Specifically, the ability to reject recommendations offered by AI tools designed to supplement work, such as to perform administrative tasks or to rewrite academic content for students, is an important capacity that users have that they must at least retain an awareness of. With the knowledge of technical fallibility, embedded bias and the inner workings of AI models themselves, users can determine for themselves, or learn to determine with the support of literacy frameworks, how to engage with models appropriately, and when to actively choose making a cognitive effort to protect themselves from epistemic risk.

## 3. Limitations

This paper does not explore metacognitive practice in enough depth to produce a robust, scalable, and integrable solution for its adjustment to AI solutions. It anticipates successful adherence to frameworks without considering the autonomous decisions individuals take to use AI to optimise their own outcomes. Developing valid and reliable assessment tools for metacognition is also a challenge, as is the ability for individuals to conduct these assessments on themselves. The application of such a framework is also limited by epistemic variability, and varied epistemic norms in global digital curricula. This framework also does not account for heuristic needs and cannot produce reliable results in all learning scenarios. Much of the success of this cognitive protection is reliant on individual willingness, ability and circumstance, and this paper recognises the need for further research to support how other standardised and quantifiable solutions can be developed.

## Conclusion

This paper presents a summary of findings that position metacognition as a protective, interventional strategy to mitigate epistemic and cognitive threats posed by AI generated information. Recognizing that the structure of knowledge itself has changed unprecedentedly with the advent of AI, adjusting traditional architecture for praxis in new contexts requires identifying these changes and the subsequent emerging threats from them. The implications for this kind of research are numerous but remain in the preliminary stages of application.

The primary complications highlighted in this paper are epistemic threats and distortions of our responses to information, as a result of dependencies on AI technologies. The framework proposes applicable but conceptual ideas that are intended to act as a reference point for educators, users and developers

dealing with exposure to these tools to support their epistemic autonomy and cognitive wellbeing. Ideally, further research that takes on a more empirical approach to actioning these concepts in various use contexts would legitimize these early investigations.

Conclusively, this paper advocates for meaningful epistemic interactions with AI technologies that are actively governing how we think and who we become.

**References**


Baker, L. and Brown, A.L. (1984) 'Metacognitive Skills and Reading', in Pearson, P.D., Barr, R., Kamil, M.L. and Mosenthal, P. (eds.) *Handbook of Reading Research*. New York: Longman, pp. 353–394.

Babic, Boris. "A Theory of Epistemic Risk." *Philosophy of Science* 86.3 (2019): 522-550.

Biswas, M. and Murray, J. (2025) '"Incomplete Without Tech": Emotional Responses and the Psychology of AI Reliance', in Huda, M.N., Wang, M., and Kalganova, T. (eds.) Towards Autonomous Robotic Systems. TAROS 2024. *Lecture Notes in Computer Science*, vol. 15051.

Biswas, M. and Murray, J. (2024) 'The Impact of Education Level on AI Reliance, Habit Formation, and Usage', 2024 29th *International Conference on Automation and Computing* (ICAC), Sunderland, United Kingdom, pp. 1–6.

Borenstein, J. and Howard, A. (2021) 'Emerging Challenges in AI and the Need for AI Ethics Education', *AI and Ethics*, 1(1), pp. 61–65.

Carr, N. (2010) 'The Shallows: What the Internet is Doing to Our Brains', *Synthese*, 178(3), pp. 341–348.

Cardon, P., Fleischmann, C., Aritz, J., Logemann, M., & Heidewald, J. (2023) 'The Challenges and Opportunities of AI-Assisted Writing: Developing AI Literacy for the AI Age', *Business and Professional Communication Quarterly*, 86(3), pp. 257-295.

Casal-Otero, L., Catala, A., Fernández-Morante, C., Taboada, M., Cebreiro, B., & Barro, S. (2023). AI Literacy in K-12: A Systematic Literature Review. *International Journal of STEM Education*, 10(1), Article 29.

Chen, V., Liao, Q.V., Vaughan, J.W., & Bansal, G. (2023) 'Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations', *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2), Article 370.

Chein, J.M., Martinez, S.A., & Barone, A.R. (2024) 'Human Intelligence Can Safeguard Against Artificial Intelligence: Individual Differences in the Discernment of Human From AI Texts', *Scientific Reports*, 14, Article 25989.

Chiu, T.K.F., Ahmad, Z., Ismagilov, M., and Sanusi, I.T. (2024) 'What Are Artificial Intelligence Literacy and Competency? A Comprehensive Framework to Support Them', *Computers and Education Open*, 6, 100171.

Department for Science, Innovation and Technology. (2025). AI Opportunities Action Plan. London: Department for Science, Innovation and Technology.

Dihal, Kanta, and Tania Duarte. Better Images of AI: A Guide for Users and Creators. The Leverhulme Centre for the Future of Intelligence / We and AI, Feb. 2023, blog.betterimagesofai.org/wp-content/uploads/2023/02/Better-Images-of-AI-Guide-Feb-23.pdf.

Flavell, J.H. (1979) 'Cognitions About Cognitions: The Theory of Metacognition', *American Psychologist*, 34(10), pp. 906–911.

Gerlich, M. (2025) 'AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking', SSRN Electronic Journal.

Metacognition." *The Philosophical Glossary for AI*, edited by Alex Grzankowski and Benjamin Henke, 18 Dec. 2025, aiglossary.co.uk/2025/12/18/metacognition/.

Heersmink, R. The Internet, Cognitive Enhancement, and the Values of Cognition. *Minds & Machines 26*, 389–407 (2016).

Holstein, K., Aleven, V., McLaren, B. M., & Sewall, J. (2020). Addressing Student Needs with Human and Artificial Intelligence Tutors. *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 1982–1991.

Jobin, A., Ienca, M., & Vayena, E. (2019) 'The Global Landscape of AI Ethics Guidelines', *Nature Machine Intelligence*, 1(9), pp. 389–399.

Johnson, B. (2022) 'Metacognition for Artificial Intelligence System Safety – An Approach to Safe and Desired Behavior', *Safety Science*, 151, 105743.

Kalantzis, M. and Cope, B. (2025) 'Literacy in the Time of Artificial Intelligence', *Reading Research Quarterly*, 60(1), Article e591.

Karoff, P. (2019) 'Embedding Ethics in Computer Science Curriculum', *Harvard Gazette*, 25 January. Available at: https://news.harvard.edu/gazette/story/2019/01/harvard-works-to-embed-ethics-in-computer-science-curriculum/ (Accessed: 9 February 2025).

Keles, S. (2023) 'Navigating in the Moral Landscape: Analysing Bias and Discrimination in AI through Philosophical Inquiry', *AI and Ethics*, 3(4), pp. 901–917.

Kong, S.-C., Cheung, W.M.-Y., & Zhang, G. (2021). Evaluation of an Artificial Intelligence Literacy Course for University Students with Diverse Study Backgrounds. *Computers and Education: Artificial Intelligence*, 2, 100002.

Kuhn, D. and Dean, D. (2004) 'A Bridge Between Cognitive Psychology and Educational Practice', *Theory into Practice*, 43(4), pp. 268–273.

LaGrandeur, K. (2020) 'How Safe is Our Reliance on AI, and Should We Regulate It?', *AI and Ethics*, 1(1), pp. 67–73.

Liu, Jiahui. 2025. "When Usefulness Fuels Fear: The Paradox of Generative AI Dependence and the Mitigating Role of AI Literacy." *International Journal of Human–Computer Interaction*, August, 1–22. doi:10.1080/10447318.2025.2544006.

Luccioni, A. and Bengio, Y. (2020) 'On the Morality of Artificial Intelligence', IEEE *Technology and Society Magazine*, 39(4), pp. 36–43. Available at: https://web.archive.org/web/20201108110312id_/https://ieeexplore.ieee.org/ielx7/44/9035512/09035523.pdf (Accessed: 9 February 2025).

Ng, D.T.K., Leung, J.K.L., Chu, K.W.S., & Qiao, M.S. (2021). AI literacy: Definition, Teaching, Evaluation, and Ethical Issues. *Proceedings of the Association for Information Science and Technology*, 58(1), 504–509.

Pritchard, D. (2010) 'Cognitive Ability and the Extended Cognition Thesis', *Synthese*, 175(1), pp. 133–151.

Przybylski, A.K., Murayama, K., DeHaan, C.R., & Gladwell, V. (2013) 'Motivational, Emotional, and Behavioural Correlates of Fear of Missing Out', *Computers in Human* Behavior, 29(4), pp. 1841–1848.

Salovich, N. A., & Rapp, D. N. (2021). Misinformed and Unaware? Metacognition and the Influence of Inaccurate Information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(4), 608–624.

Serafimova, S. (2020) 'Whose Morality? Which Rationality? Challenging Artificial Intelligence as a Remedy for the Lack of Moral Enhancement', *Humanities and Social Sciences Communications*, 7, Article 119.

Sparrow, B., Liu, J. and Wegner, D.M. (2011) 'Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips', *Science*, 333(6043), pp. 776–778.

Southworth, J., Migliaccio, K., Glover, J., Glover, J., Reed, D., McCarty, C., Brendemuhl, J., & Thomas, A. (2023) 'Developing a Model for AI Across the Curriculum: Transforming the Higher Education Landscape via Innovation in AI Literacy', *Computers and Education: Artificial Intelligence*, 4, 100127.

Tock, J.L. and Moxley, J.H. (2017) 'Metacognition: A Predictor of Learning

Outcome', *The Journal of Educational Research*, 110(5), pp. 478–489.

Touretzky, D., Gardner-McCune, C., Breazeal, C., Martin, F., & Seehorn, D. (2019). A Year in K-12 AI Education. *AI Magazine*, 40(4), 88–90.

UK Government (2025) International AI Safety Report 2025. Available at: https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf (Accessed: 9 February 2025).

Urban, Marek, et al. ""ChatGPT Can Make Mistakes. Check Important Info." Epistemic Beliefs and Metacognitive Accuracy in Students' Integration of ChatGPT Content into Academic Writing." *Computers & Education*, vol. 225, 2025, p. 105118. ScienceDirect, https://doi.org/10.1016/j.compedu.2024.105118.

Vartiainen, H., Toivonen, T., Jormanainen, I., Kahila, J., Tedre, M., & Valtonen, T. (2021) 'Machine Learning for Middle Schoolers: Learning through Data-Driven Design', *International Journal of Child-Computer Interaction*, 27, 100281.

"Vallier, K., On the Noble Uses of AI: When Cognitive Offloading Elevates Us." *Cosmos Institute*, 30 Jan. 2026, blog.cosmos-institute.org/p/on-the-noble-uses-of-ai.

Voeneky, S., Kellmeyer, P., Mueller, O., and Burgard, W. (eds.) (2022) *The Cambridge Handbook of Responsible Artificial Intelligence*. Intellectual Debt, Zittrain, J, Chapter 11. Cambridge: Cambridge University Press.

Wilton, L., MacKinnon, K., & Greene, K. (2021) 'Where is the AI? AI Literacy for Educators', in Artificial Intelligence in Education. AIED 2021. *Lecture Notes in Computer Science*, vol 12749. Springer, Cham, pp. 137–142.

Wihbey, J. (2024) 'AI and Epistemic Risk for Democracy: A Coming Crisis of Public Knowledge?', *SSRN Electronic Journal*.

Yi, Y. (2021). Establishing the Concept of AI Literacy. Jahr – *European Journal of Bioethics*, 12(2).

Zohar, A. and Barzilai, S. (2013) 'Promoting Self-Regulation in Science Education: Metacognition as Part of a Broader Perspective on Learning', *Metacognition and Learning*, 8(1), pp. 99–111.

Zhai, C., Wibowo, S. and Li, L.D. (2024) 'The Effects of Over-Reliance on AI Dialogue Systems on Students' Cognitive Abilities: A Systematic Review', *Smart Learning Environments*, 11, Article 28.

# Risk, Rights, and Regulatory Convergence: The Turkish Artificial Intelligence Law Proposal as a Model for Emerging-Economy AI Governance and the Prospects for a TRIPS-Style International Framework

*Zeki Emre Kurt*
*GEMS Schindhelm*



The global AI regulatory landscape has converged architecturally around risk-based tiering, mandatory ethics obligations, prohibitions on unacceptable practices, conformity assessment, supervisory enforcement, and turnover-linked sanctions. Yet this convergence is structural rather than substantive: identical terminology performs divergent doctrinal functions across jurisdictions. Most interventions worldwide remain at the proposal or non-binding policy-framework stage.

This article takes Turkey's Artificial Intelligence Law Proposal (Esas No. 2/2234, submitted 25 June 2024) as its central case study—an eight-article principle-based framework statute that mirrors the EU AI Act's turnover-linked sanctions while deferring substantive risk classification and enforcement detail to secondary legislation. Three theses are advanced. First, a universal common core exists, but it is a floor of architectural principles, not a ceiling of identical rules. Second, for upper-middle-income economies with constrained regulatory capacity, a principle-based framework statute is defensible where full advanced-economy transposition is institutionally infeasible—provided a credible secondary-legislation programme accompanies it. Third, the TRIPS Agreement's minimum-standards-plus-flexibility architecture offers a structural template for a development-sensitive international AI governance instrument, though such a regime risks reinforcing global inequality unless accompanied by genuine capacity-building and technology-transfer commitments.

**Keywords**: Artificial Intelligence Regulation, Turkish AI Law, Risk-Based Governance, Emerging Economies, TRIPS Analogy

## Introduction

In 2024, a provincial health directorate in south-eastern Turkey procured an AI-assisted diagnostic triage system trained on imaging datasets from a European vendor and validated against benchmarks calibrated to epidemiological conditions substantially different from those in the deployment region. No Turkish statute governed its use. No national authority had reviewed its conformity with safety or fairness standards. The algorithmic outputs were presented to clinicians as probability scores, but the absence of technical documentation, audit trails, or any mechanism of patient challenge created a governance vacuum around a system making consequential clinical determinations. This scenario is the governance deficit AI regulation is designed to address: not the prospect of catastrophic AI failure, but the pervasive structural absence of accountability, transparency, and recourse in the deployment of AI systems across under-resourced institutional settings.

Although the scenario is hypothetical, it is not unusual. Rather, it reflects a recurring pattern observed across a growing number of cross-border deployments of AI diagnostic and triage systems: tools trained and validated on non-local populations being introduced into resource-constrained health systems that lack dedicated legislation, conformity assessment procedures, audit mechanisms, or effective avenues for patient redress. This pattern sits squarely within concerns identified in WHO guidance on the governance of health AI (WHO, 2021). The risks associated with performance degradation under population and setting shift have been demonstrated empirically (Zech et al, 2018) and examined systematically within the clinical literature (Finlayson et al, 2021), while the broader governance challenges confronting emerging economies have been analysed by Okolo (2023) and UNCTAD (2021).

The question this article addresses is not whether emerging economies should regulate artificial intelligence — that question has been answered by the trajectory of global policy —

but how, and whether a principle-based framework statute of the type proposed in Turkey can serve both as a viable national governance instrument and as a template for an international framework sensitive to the development constraints of the Global South.

The global regulatory landscape has evolved rapidly (Alter and Meunier, 2009; Roberts et al, 2024). The EU AI Act — Regulation (EU) 2024/1689 (European Parliament and Council, 2024) — entered into force on 1 August 2024 as the most comprehensive domestic AI governance instrument yet adopted. The Council of Europe Framework Convention on AI (CETS No. 225) (Council of Europe, 2024), opened for signature on 5 September 2024, is the first binding international AI treaty and is explicitly open to non-European accession. The OECD AI Principles, revised in May 2024, have attracted 47 adherents (OECD, 2024). The United States, despite the revocation of Executive Order 14110 by Executive Order 14148 (2025), continues to produce influential voluntary frameworks through the National Institute of Standards and Technology (NIST, 2023). Brazil's Senate-approved PL 2338/2023 (Brazilian Federal Senate, 2024), Singapore's Model AI Governance Framework (PDPC/IMDA, 2020), the African Union Continental AI Strategy (African Union, 2024), and legislative initiatives across Africa and South Asia extend this picture into the developing world (Kenya, 2025; South Africa, 2024).

Into this landscape, Turkey's AI Law Proposal (Esas No. 2/2234), submitted on 25 June 2024 arrives as a distinctive and instructive intervention. The proposal declares five foundational principles, mandates risk assessment, imposes registration and conformity obligations for high-risk systems, and establishes a three-tier fine structure tracking the EU AI Act's turnover-linked percentages — whilst deferring the substance of risk classification and enforcement architecture to secondary legislation.

Three preliminary qualifications condition the analysis (Bradford, 2020). First, the global common core documented in Section I is architectural rather than substantive: convergence on nomenclature conceals divergence in function. Second, the case for a principle-based framework statute is conditional, not categorical — it depends on a credible secondary-legislation programme and carries genuine risks of legal uncertainty and institutional delay. Third, the TRIPS analogy must be stress-tested against the documented failures of TRIPS in the Global South (Yu, 2006; Okediji, 2014) and against the inverted bargaining dynamic that an AI governance convention would present.

## 1. The Global Common Core: Convergent Elements in AI Regulation

Across jurisdictions as legally distinct as the EU, China, Brazil, Singapore, and Kenya, eight structural building blocks recur with sufficient consistency to constitute a global common core of AI regulation. Two qualifications are analytically prior to the mapping exercise. The first concerns path-dependence (Bradford, 2020): the shared vocabulary — risk, transparency, accountability — reflects the "Brussels Effect", the EU's comprehensive, enforceable, and market-backed model functioning as a de facto global standard driven by one dominant exporter rather than independent parallel evolution. The second is more fundamental: structural convergence must not be mistaken for doctrinal convergence. "Transparency" in the EU AI Act (European Parliament and Council, 2024, Art. 86) is a legally enforceable right of explanation grounded in the EU Charter of Fundamental Rights. In China's Algorithmic Recommendation Regulation (People's Republic of China, 2022, Art. 24) it is a state-directed disclosure obligation with no European counterpart. "Accountability" in Brazil's PL 2338/2023 encompasses rights of contestation for historically marginalised populations; in Singapore's Model Framework (PDPC/IMDA, 2020) it denotes an internal organisational process. Identical names perform different doctrinal work. A treaty codifying "transparency" as a minimum standard achieves no substantive uniformity unless the functional content of that standard is specified.

Risk-based tiering is the dominant regulatory technique (Ebers, 2025). The EU AI Act codifies four tiers: unacceptable risk (Art. 5, eight prohibited practices); high risk (Arts. 6–7,

Annex III); limited risk (Art. 50); and minimal risk. Brazil adopts a three-tier variant (Brazilian Federal Senate, 2024, Arts. 3–4); the CoE Convention requires "*graduated and differentiated*" measures (Council of Europe, 2024, Art. 1(2)). The tiers serve functionally different purposes: in the EU, risk classification triggers mandatory ex ante obligations; in Singapore, voluntary proportionality guidance; in China, an administrative filing regime. All significant instruments articulate foundational principles — transparency, fairness, accountability, safety, privacy, human oversight — with convergence on names concealing divergence in functional content (OECD, 2024; UNESCO, 2021; NIST, 2023). Every major instrument establishes a categorical prohibition floor reflecting each jurisdiction's constitutional values (European Parliament and Council, 2024, Art. 5(1)(a)–(h)): the EU AI Act prohibits eight categories anchored in fundamental rights including subliminal manipulation, social scoring, real-time remote biometric identification in public spaces, and facial image scraping; Brazil's "excessive risk" tier targets subliminal manipulation and mass surveillance (Brazilian Federal Senate, 2024, Arts. 12–13); China's regime targets content incompatible with "Core Socialist Values" (People's Republic of China, 2023, Art. 4). Procedural convergence coexists with deep substantive divergence.

The remaining building blocks follow a consistent structural logic. All substantive frameworks impose enhanced lifecycle obligations on high-risk applications: Article 9 of the EU AI Act requires continuous, iterative risk management; the NIST AI RMF 1.0 (NIST, 2023) provides a voluntary equivalent; Brazil mandates human review, bias auditing, and impact assessment (Brazilian Federal Senate, 2024, Arts. 15–22). High-risk AI systems must undergo conformity assessment and registration; the EU establishes the most elaborated regime at Articles 43–49 and Annexes VI–VII (European Parliament and Council, 2024). Every binding or near-binding instrument contemplates supervisory authorities: the EU creates a layered architecture of AI Office, AI Board, Scientific Panel, and national competent authorities (Arts. 64–70); the CoE Convention requires "*independent, impartial oversight mechanisms*" (Council of Europe, 2024, Art. 26); Brazil coordinates through the SIA/ANPD system (Brazilian Federal Senate, 2024, Arts. 42–46). All substantive instruments link sanctions to annual turnover: the EU AI Act's 7%/3%/1% tiers (Art. 99; Pasquale, 2015), Turkey's 7%/3%/1.5%, Brazil's 2%. All major instruments commit to innovation-enabling implementation: the EU AI Act mandates regulatory sandboxes at Articles 57–62; the CoE Convention's Article 13 establishes 'safe innovation environments'. Table 1 maps these elements across five representative instruments.

**Table 1:** Global Common Core — Architectural Convergence, Legal Function, and Enforcement Intensity

| Building Block | EU AI Act | CoE CETS 225 | Brazil PL 2338 | Turkey | Legal Function / Enforcement Intensity |
|---|---|---|---|---|---|
| Risk tiering | 4 tiers (arts 5–6) | Graduated (art 1(2)) | 3 tiers (arts 3–4) | High-risk ref. | Obligation trigger / High (EU) → Voluntary (SG) |
| Ethical principles | Arts 8–15 (mandatory) | Arts 4–13 (binding) | Arts 5–12 | 5 principles (art 3) | Norm + interpretation / Enforceable (EU) → Aspirational (OECD) |
| Prohibited practices | Art 5: 8 categories | Art 16(4) ban/morat. | Excessive risk | Referenced; not listed | Absolute floor / High (criminal/admin) |
| Risk management | Art 9 (continuous) | Art 16 | Mandatory | Art 4 (general) | Ex ante control / High (EU) → Pending (TR) |
| Conformity/registration | Arts 43–49 | Implicit | Mandatory | Art 4 (general) | Pre-market gate / High (EU) → Filing only (China) |
| Supervisory authority | Arts 64–70 (AI Office) | Art 26 (independent) | SIA/ANPD | Existing bodies (art 5) | Enforcement / High (EU) → Piggyback (BR/TR) |
| Turnover sanctions | 7%/3%/1% (art 99) | N/A (treaty) | 2% turnover | 7%/3%/1.5% | Deterrence / High (EU) → Moderate (TR/BR) |
| Innovation support | Arts 57–62 (sandboxes) | Art 13 | Yes | Absent | Chill mitigation / Present (EU) → Gap (TR) |

Note: "Legal Function" identifies the doctrinal role each building block performs within a given system. Architectural convergence does not imply equivalence across either column. Each building block is analysed against the Turkish proposal article-by-article in Section V.

Across the instruments examined above, a relatively stable regulatory architecture emerges despite differences in legal form, scope and institutional setting. At its minimum, that architecture consists of six elements: foundational ethical principles; a requirement for risk assessment; registration and conformity-assessment procedures; the designation of a supervisory authority; effective sanctions linked to organisational turnover; and a structured commitment to secondary

legislation within a defined implementation timetable. The first five elements recur, in varying formulations, throughout the OECD framework, the UNESCO Recommendation, the Council of Europe Framework Convention and the EU AI Act. The sixth reflects a practical consideration. Where a framework statute establishes principles while leaving operational detail to future instruments, the credibility of the legislative project depends upon the existence of a clear and time-bound programme for implementation. Taken together, these six elements formulated by this article from the common core above provide a useful benchmark for evaluating whether a national framework statute functions as a credible transitional instrument. They are applied to the Turkish proposal in Section V and to the assessment of international alignment in Section VI.

## 2. Regulatory Models in Advanced Economies: From California to the EU

The EU AI Act is the most ambitious comprehensive AI governance instrument adopted to date (Almada and Petit, 2025). Its regulatory philosophy is product-safety law transposed to a new domain: compliant AI systems bear the CE mark and circulate freely within the single market. Article 5's eight prohibited practices became enforceable on 2 February 2025; the high-risk tier (Arts. 9–15) imposes continuous lifecycle risk management, data governance, automatic event logging, human oversight capability, and accuracy and cybersecurity requirements. Three critiques are analytically important: the compliance burden on SMEs is substantial, calibrated to large enterprises with dedicated legal functions (Mantelero, 2024; see Art. 99(6) for SME fine relief); the enforcement architecture is complex, requiring deep institutional coordination across the AI Office, AI Board, Scientific Panel, and national competent authorities (Novelli et al, 2025); and mandatory pre-deployment conformity assessment may be disproportionate where harm potential is context-dependent (Ebers, 2025). One of the more distinctive features of the AI Act is its treatment of general-purpose AI models at the level of the model itself. Article 51(2) establishes a rebuttable presumption that a GPAI model presents systemic risk when the cumulative computational resources used during training exceed $10^{25}$ floating-point operations (FLOPs). Models falling within this category become subject to the enhanced obligations contained in Article 55 (European Parliament and Council, 2024). Importantly, the threshold is not part of the definition of a GPAI model under Article 3(63). Its purpose is instead to identify a subset of particularly powerful models that merit additional regulatory scrutiny. The Commission may revise the threshold by delegated act (Arts. 51(3) and 97). In doing so, the AI Act introduces a significant innovation: regulatory obligations are attached to the scale of model training itself, independently of the particular sectors or settings in which the model may later be used.

The United States presents deliberate regulatory fragmentation (Calo, 2017). The revocation of Executive Order 14110 by Executive Order 14148 (2025) returned federal AI governance to sector-specific guidance and the NIST AI RMF 1.0 as the dominant federal instrument. California's SB 1047 would have imposed mandatory safety protocols on frontier models trained above $100 million; Governor Newsom's veto (2024) rested on a substantive methodological critique — the compute-cost threshold ignored smaller, contextually dangerous models — and the successor SB 53 (2025) retained the threshold whilst stripping the most demanding obligations. China's architecture is deliberately disaggregated: the Algorithm Recommendation Provisions (People's Republic of China, 2022), the Deep Synthesis Provisions (2023), and the Generative AI Interim Measures (People's Republic of China, 2023) are binding and mandatory but technology-specific rather than horizontal. China's model shares the common core's vocabulary but its prohibition floor reflects content-governance logic rather than individual fundamental-rights protection. The two are not functionally equivalent.

## 3. The Global Legislative Landscape: Enacted AI Laws and Pending Proposals

Despite the density of the regime complex (Alter and Meunier, 2009), comprehensive and binding AI legislation remains the exception: most regulatory interventions worldwide sit at the proposal or non-binding policy-framework

stage, and this fragmentation is the empirical baseline against which every normative claim about convergence must be evaluated (Roberts et al, 2024; World Bank, 2024). The EU AI Act and the CoE Framework Convention (CETS No. 225) — explicitly open to non-European accession under Article 31, and the most development-sensitive existing binding instrument (Council of Europe, 2024) — constitute the primary binding layer. China's disaggregated instruments are binding but technology-specific. By early 2026, only a handful of countries had enacted binding national legislation on artificial intelligence that applied horizontally across sectors and was already in force. The clearest examples are the statutes of South Korea, Italy, and Kazakhstan. Vietnam's comprehensive law was adopted in December 2025 but does not take effect until March 2026; Taiwan's raises questions of recognition; and the European Union's AI Act binds twenty-seven member states as a supranational instrument rather than as the domestic law of any single country. The count excludes sectoral measures, non-binding strategies, and proposals still under consideration, and is best read as a provisional snapshot (IAPP, 2026; OECD.AI, 2026).

A second tier consists of substantive proposals under parliamentary consideration. Turkey's AI Law Proposal (Esas No. 2/2234) (TBMM, 2024) is currently in committee review; Brazil's PL 2338/2023, Senate-approved on 10 December 2024 (Brazilian Federal Senate, 2024), adopts a three-tier risk-based architecture with explicit algorithmic discrimination provisions and enforcement built on the existing ANPD — the most directly comparable emerging-economy instrument to Turkey's proposal. Canada's AIDA (Bill C-27, Part 3), prorogued when parliament dissolved in January 2025, illustrates the political fragility of AI legislation even in advanced democracies. A third tier — the United Kingdom's pro-innovation sector-led model (UK Government, 2023), India's non-binding IndiaAI Mission posture, Kenya's National AI Strategy 2025–2030 (Kenya, 2025), and South Africa's National AI Policy Framework (South Africa, 2024) — relies on strategies and guidance without binding legal force. The Turkish proposal is the median position in the global legislative taxonomy: more advanced than most non-binding frameworks, more accessible than the EU's comprehensive prescriptivism.

## 4. AI Regulation in Emerging and Developing Economies: Needs, Constraints, and Opportunities

The challenge facing emerging economies is a prior question of institutional feasibility, not a choice between regulatory philosophies. Three capacity tiers are analytically necessary (World Bank, 2024). The first tier — upper-middle-income economies including Turkey, Brazil, Mexico, and South Africa — has existing regulatory institutions and the administrative capacity to enact and begin implementing a principle-based framework statute, developing secondary legislation within three to five years. The second tier — large lower-middle-income economies including India, Indonesia, and Nigeria — faces significantly more constrained regulatory infrastructure; even a principle-based statute requires substantial adaptation. The third tier — capacity-constrained lower-income economies such as Kenya and Rwanda — is best served by a policy framework anchored in international standards with a phased transition of at least a decade before binding obligations are activated (Okolo, 2023).

For first-tier economies, the conditional case for a principle-based framework statute rests on three grounds (Hildebrandt, 2020). First, it establishes a normative orientation without demanding institutional elaboration that immediately exceeds domestic capacity. Second, it creates a legal anchor for international trade and investment negotiations. Third, it defers the most technically demanding choices — risk classification, conformity assessment, GPAI rules — to secondary legislation that can evolve as technology and institutional capacity co-develop. The governance vacuum in emerging economies creates conditions under which AI-driven harms proceed without accountability (UNCTAD, 2021; Eubanks, 2018; Crawford, 2021); even imperfect governance mitigates this risk. Three significant risks must be acknowledged (Hildebrandt, 2020; Bently et al, 2022): the absence or delay of secondary legislation transforms a framework statute into nominal governance without operational effect; institutional capacity limitation means

governance obligations cannot be discharged where existing sectoral regulators lack minimum AI technical expertise; and legal uncertainty from framework design creates regulatory indeterminacy that may deter responsible operators whilst providing no deterrent to irresponsible ones. A hypothetical full transposition of the EU AI Act would require for Turkey dedicated supervisory authority budgets of €15–25 million, accredited notified bodies, and a mandatory sandbox programme — resources Turkey's existing bodies (BTK, BDDK, SPK, KVKK) cannot immediately absorb (European Commission, 2025). The principle-based approach, assigning obligations to existing sectoral bodies whilst secondary legislation develops specificity, is therefore a governance strategy rather than a governance failure. Brazil illustrates genuine first-tier localisation: PL 2338/2023 incorporates explicit algorithmic discrimination provisions with no EU counterpart, a 2% turnover sanction calibrated to Brazilian market conditions, and enforcement built on the existing ANPD (Brazilian Federal Senate, 2024). India presents the opposite trajectory — active investment in the ₹10,371 crore IndiaAI Mission paired with deliberate regulatory abstinence, a posture that carries material governance risks at India's scale of AI deployment.

## 5. The Turkish Proposal: Normative-Comparative Analysis and the Path to Secondary Legislation

The Turkish AI Law Proposal (Esas No. 2/2234) is a principle-based framework statute of eight articles. For each article the analysis applies three questions: does the provision satisfy the minimum floor established by the global common core? What secondary legislation is required? How does it interact with existing Turkish law? The proposal is defensible as a transitional governance instrument — but only on condition that the secondary-legislation programme described below is enacted promptly (Hildebrandt, 2020).

### Article 1 — Purpose and Scope

*"The purpose of this Act is to ensure the safe, ethical, and fair use of artificial intelligence technologies, to provide for the protection of personal data and the non-violation of privacy rights, and to establish a regulatory framework for the development and use of artificial intelligence systems."* (TBMM, 2024, Art. 1)

The purpose clause captures the three substantive commitments shared by all major instruments and satisfies the minimum floor. Secondary legislation must specify whether the AI statute's data protection obligations are supplementary to — rather than merely coextensive with — obligations under Turkey's Personal Data Protection Law (KVKK, Law No. 6698, 2016, as amended by Law No. 7545, 2024).

### Article 2 — Definitions

*"Computer-based systems capable of performing human-like cognitive functions and possessing capabilities such as learning, reasoning, problem-solving, perception, and language understanding."* (TBMM, 2024, Art. 2)

The five actor-roles are collapsed into a single regulatory category of Yapay Zeka Operatörleri. Compare the EU AI Act's role-differentiated obligations: provider (Art. 3(3)), deployer (Art. 3(4)), importer (Art. 3(6)), distributor (Art. 3(7)) (European Parliament and Council, 2024). The advantage is that no actor can evade responsibility through role characterisation; the disadvantage is identical compliance demands on a small agricultural-technology startup and a multinational diagnostic AI vendor. Secondary legislation must introduce role-differentiated obligations, prioritising deployer requirements given Turkey's position as a predominantly technology-deploying jurisdiction (Yu, 2025) (see Table 1, "Ethical principles" row and the role-differentiation gap recorded in Table 2).

### Article 3 — Core Principles

*Safety / Transparency / Fairness / Accountability / Privacy* (TBMM, 2024, Art. 3)

The five-principle architecture captures the substance of the OECD Principles (OECD, 2024) and six of the seven NIST AI RMF trustworthy-AI characteristics (NIST, 2023). The sole significant gap is human oversight — central to the EU AI Act (Art. 14) and CoE Convention (Art. 8) but absent as a named principle (Table 1, "Ethical principles" row). Secondary legislation should add human oversight as a sixth principle and specify AI-specific supplementary

obligations including algorithmic impact assessment.

**Article 4 — Risk Management**
*"Risk assessment shall be conducted during the development and use of AI systems and special measures shall be taken for high-risk systems." / "High-risk AI systems shall be registered with the relevant supervisory authorities and subjected to conformity assessment."* (TBMM, 2024, Art. 4)

Article 4 establishes the four structural obligations a primary instrument must contain: risk assessment, high-risk special measures, registration, and conformity assessment (Table 1, "Risk management" and "Conformity/registration" rows; gaps itemised in Table 2). The EU AI Act takes 60 articles and seven annexes to elaborate these four (European Parliament and Council, 2024). The critical gap is operational hollowness: no criteria define "high risk", no sectoral list equivalent to EU Annex III is provided, no risk assessment methodology is specified, and no registration procedure is established. Secondary legislation must supply: a High-Risk Applications List adapted from EU Annex III covering AI in healthcare diagnosis, criminal justice, educational assessment, employment recruitment, biometric identification in public spaces, and critical infrastructure management; a Risk Assessment Protocol aligned with ISO/IEC 42001 (2023) and the NIST AI RMF (NIST, 2023); a Registration and Conformity Procedure designating BTK or a dedicated AI unit; and a GPAI Downstream Assessment Protocol requiring deployer risk assessments for GPAI-embedded applications in high-risk domains (Yu, 2025).

**Article 5 — Compliance and Supervision**
Article 5 confers supervisory powers on unspecified "supervisory authorities" (TBMM, 2024, Art. 5). The institutional void is the proposal's most consequential deficiency: no authority is named and no investigative powers are enumerated (Table 1, "Supervisory authority" row; Table 2, "No supervisory authority"). Secondary legislation must designate competent bodies sectorally — BTK for IT applications, BDDK for financial services AI, Ministry of Health for healthcare AI, SPK for capital markets AI — establish affected persons' rights to notification, human review, and complaint (European Parliament and Council, 2024, Arts. 70–88; Council of Europe, 2024, Art. 26), and specify judicial remedies through the İdare Mahkemesi, with final appeals to the Danıştay (Law No. 2577, 1982; Law No. 2575, 1982).

**Article 6 — Sanctions**
*Prohibited applications: up to TL 35 million or 7% of annual turnover. / Obligation violations: up to TL 15 million or 3%. / False information: up to TL 7.5 million or 1.5%.* (TBMM, 2024, Art. 6)

While the Turkish proposal follows the EU AI Act's three-tier sanction structure for the first two tiers (Table 1, "Turnover sanctions" row), it sets the administrative fine for supplying false or misleading information to the supervisory authority at 1.5 per cent of annual turnover, one half-point above the EU's 1 per cent threshold (European Parliament and Council, 2024, Art. 99(5)). The proposal's explanatory memorandum expressly characterises this increment as a deliberate calibration, intended to strengthen deterrence against the provision of false information to the authority while preserving the tier's position as the least severe of the three sanctions (TBMM, 2024, Gerekçe, Art. 6). The half-point divergence is therefore grounded in the legislative materials as a targeted deterrence rationale rather than resting on inference. Secondary legislation should introduce an SME relief provision applying the lower of the percentage or absolute figure for SMEs and startups (European Parliament and Council, 2024, Art. 99(6); OECD, 2024). Article 7 provides immediate entry into force; secondary legislation should specify a grace period of at least twelve months for existing AI deployments to achieve conformity, consistent with the EU AI Act's phased application schedule.

**The Secondary Legislation Programme**
**Table 2:** Critical Gaps and Required Secondary Legislation — Turkey (cross-referenced to the article-by-article analysis in Section 5)

| Gap | Required Secondary Legislation | Turkish Law Interaction | Priority |
|---|---|---|---|
| No prohibited-practices list | Presidential Decree (adapted from EU art. 5) | Constitution arts. 17, 20 | Year 1 |
| Vague high-risk criteria | High-Risk Applications List (Ministry Regulation) | KVKK art. 3; BTK enabling rules | Year 1 |
| No role differentiation | Differentiated Obligations Regulation | BDDK, BTK, SPK sectoral rules | Year 1 |
| No supervisory authority | Authority Designation Decree; inter-agency MOU | BTK / BDDK / Health Ministry | Year 1 |
| No technical standards | ISO 42001 / NIST AI RMF by Presidential Decree | TSE harmonisation | Year 2 |
| No GPAI rules | Deployer Risk Assessment Protocol | KVKK data processing rules | Year 2 |
| No SME/innovation support | Sandbox Regulation; SME fine relief | KOSGEB enabling legislation | Year 2 |
| No judicial remedies | Remedies Regulation; right to explanation | İdare Mahkemesi / Danıştay rules | Year 1 |

Year 1 = within 12 months of enactment; Year 2 = 13–30 months

Measured against the six elements distilled in Section I as the minimum threshold for an effective principle-based framework statute, the current draft satisfies five: the articulation of core principles, a mandate for risk assessment, obligations of registration and conformity, the designation of a supervisory authority, and turnover-based sanctions. The remaining criterion — a binding commitment to timetabled secondary legislation — is not yet met and would require the adoption of a ministerial action plan within sixty days of the statute's entry into force.

## 6. Towards a TRIPS-Style International AI Agreement? Feasibility and Design

The convergence documented in Section I raises the structural question of whether it can be codified in a binding international instrument. The Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) (WTO, 1994) offers the most instructive precedent: a treaty that translated a partial international consensus into minimum standards with built-in flexibility and binding dispute settlement (Gervais, 2021). This section evaluates the analogy with explicit intellectual humility: TRIPS has a demonstrably mixed record in the Global South, and a TRIPS-style AI regime risks reproducing its structural inequalities unless specifically designed to avoid doing so.

*The Structural Parallels*

The TRIPS analogy operates at six structural levels (Yu, 2025). On minimum standards, an AI governance minimum would establish floors for safety testing, transparency, prohibition on specified harmful uses, supervisory authority existence, and turnover-linked sanctions — the global common core of Section I already constitutes the substance of such a floor. On flexibility provisions, TRIPS Arts. 7–8 embed developmental objectives (WTO, 2001); an AI governance convention's "development box" would comprise transition periods graduated by World Bank income classification — LDCs: ten years; lower-middle income: five years; upper-middle income including Turkey: three years — a capacity-building fund, technology-transfer commitments, and the CoE Convention's "other appropriate measures" modality (Council of Europe, 2024, Art. 3(1)(b); Gervais and Yu, 2025). On technology transfer, TRIPS Art. 66.2 imposes a binding obligation on developed-country Members to incentivise technology transfer to LDCs (Okediji, 2014); an AI governance convention should require developed-country parties to share safety testing datasets, evaluation benchmarks, and technical regulatory toolkits. Dispute settlement, non-discrimination, and sector-specific protocols complete the six-point mapping.

*Political-Economy Analysis*

Three political-economy objections are analytically prior to the design questions. First, power asymmetry (Roberts et al, 2024; Bengio et al, 2024): TRIPS was negotiated with asymmetric bargaining power — advanced-economy IP exporters drove the minimum-standard agenda whilst developing countries received flexibility provisions as a quid pro quo. An AI governance convention presents an inverted dynamic: developing countries are the primary demanders of strong minimum standards as net importers bearing disproportionate deployment risks, whilst the major AI-producing states may resist binding obligations. Second, regulatory capture (Pasquale, 2015): frontier AI systems are developed by a small number of corporations with significant lobbying capacity; treaty design must structurally address this through mandatory Civil Society Representation Panels and developing-country veto rights over certain minimum-standard formulations. Third, the historical failure of TRIPS technology-transfer provisions (Yu, 2006; Okediji, 2014): TRIPS Art. 66.2 has produced minimal observable

technology transfer in practice, with flexibilities legally available but practically inaccessible to developing countries. An AI governance convention that reproduces this pattern risks reinforcing global AI inequality; this risk can be mitigated by tying developed-country contributions to a capacity-building fund to market access within the treaty framework.

*The Council of Europe Framework Convention as the Preferred Institutional Foundation*

These political-economy objections counsel against a full WTO-style AI treaty in the near term (Bengio et al, 2024). The CoE Framework Convention (CETS No. 225) offers a more practicable foundation: globally open (Art. 31), already attracting non-European signatories, principles-based, and offering two implementation modalities for private-sector regulation — direct obligations or "other appropriate measures" (Art. 3(1)(b)) — that provide the flexibility development-sensitive design requires (Council of Europe, 2024; UN Secretary-General's High-Level Advisory Body on AI, 2024). Built on this foundation, a development-sensitive global AI governance instrument should incorporate: principle-based minimum standards; a Development Box with graduated transition periods and a capacity-building fund of at least USD 500 million over five years; sector-specific protocols for health, agriculture, finance, and criminal justice; a peer-review mechanism leading to enhanced obligations after five years; and an express provision declaring that a principle-based framework statute with timetabled secondary legislation constitutes compliance with the minimum-standards tier (Yu, 2025).

The Turkish proposal possesses most of the institutional features associated with a credible framework statute. Its principal remaining weakness lies in the absence of a clear and time-bound programme of secondary legislation. Were that element added, the proposal would provide a persuasive model of how emerging economies might establish an initial legislative foundation for AI governance without replicating the full complexity of the EU AI Act.

**Conclusion**

In a district clinic in Antalya, a physician is looking at a screen. She has used the AI diagnostic tool for eight months. She trusts it, mostly, and distrusts it in the way experienced clinicians always distrust efficiency-enhancing systems: intelligently, provisionally, and with the quiet awareness that errors will be absorbed by patients rather than algorithms. Today the tool is confident. Today it is probably right.

The Antalya physician is hypothetical, but the circumstances represented are not. They reflect a recurring pattern documented across deployments of medical AI systems in under-resourced healthcare settings. In the absence of technical documentation, an audit trail, or a mechanism of patient challenge, her professional judgment and conscience remain the only boundaries around the algorithm's output. Those are formidable boundaries. They are not, however, governance. Governance is what happens when the physician's judgment, the algorithm's output, and the patient's rights are held together within a structure of law that can be examined, challenged, and improved. This article has tried to describe that structure — and to locate Turkey's eight-article principle-based framework statute within the larger project of constructing it.

The global common core is real, but architectural rather than doctrinal: convergence on vocabulary conceals divergence in legal function and enforcement intensity. An emerging economy that commits to the architectural core is positioning itself to build operational governance, but it is not, by that commitment alone, protecting its citizens. Turkey's next legislative step is the secondary-legislation programme detailed in Table 2, with Year 1 provisions enacted within twelve months of the statute's entry into force. Without this programme, the principle-based framework statute remains what its critics will rightly call it: a declaration of intent. The principle-based approach survives stress-testing as defensible for upper-middle-income economies with constrained regulatory capacity — but subject to three conditions: that the secondary-legislation programme is enacted promptly; that international technical assistance supplements domestic capacity; and that the governance risks of framework design are actively managed. A convention built on the CoE

Framework Convention, extended to universal accession, and equipped with a genuine development box — including enforceable technology-transfer commitments and a well-resourced capacity-building fund — could be a more equitable instrument than TRIPS was. But this requires that the political-economy objections of Section VI be structurally addressed in treaty design, not deferred to implementation. The legal architecture can be designed before the political will arrives.

The physician in Antalya closes the screen. Her patient is referred to the city hospital. The algorithm was probably right. In a regulatory system that functions as it should, "probably" would be bounded by audit, documentation, oversight, and accountability. That system does not yet fully exist. But Turkey, in its modest and conditionally promising way, has begun to build it — and what it builds, with adequate international support, may serve as more than a national instrument. It may serve as a template.

**References**


African Union (2024). *Continental Artificial Intelligence Strategy: Harnessing AI for Africa's Development and Prosperity*. Accra: African Union.

Almada, M. and Petit, N. (2025). The EU AI Act: Between the Rock of Product Safety and the Hard Place of Fundamental Rights. *Common Market Law Review*, 62, pp. 85–120.

Alter, K. and Meunier, S. (2009). The Politics of International Regime Complexity. *Perspectives on Politics*, 7(1), pp. 13–24.

Bengio, Y. et al. (2024). Managing Extreme AI Risks Amid Rapid Progress. *Science*, 384, pp. 842–845.

Bently, L., Sherman, B., Gangjee, D. and Johnson, P. (2022). *Intellectual Property Law*. 6th ed. Oxford: Oxford University Press.

Bradford, A. (2020). The Brussels Effect: How the European Union Rules the World. Oxford: Oxford University Press.

Brazilian Federal Senate (2024). Projeto de Lei No. 2.338, de 2023 (AI Bill), Senate-Approved 10 December 2024. Brasília: Federal Senate.

Calo, R. (2017). Artificial Intelligence Policy: A Primer and Roadmap. *UC Davis Law Review*, 51, pp. 399–435.

Canada (2022). Bill C-27, Digital Charter Implementation Act 2022, Part 3 (Artificial Intelligence and Data Act). Ottawa: Parliament of Canada. Prorogued January 2025.

Council of Europe (2024). Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225). Vilnius: Council of Europe.

Crawford, K. (2021). Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. New Haven: Yale University Press.

Ebers, M. (2025). Truly Risk-Based Regulation of Artificial Intelligence: How to Implement the EU's AI Act. European Journal of Risk Regulation, 16(2), pp. 684–703.

Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York: St Martin's Press.

European Commission (2025). Guidance on the implementation of the AI Act — National Competent Authorities. Brussels: European Commission.

European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (EU AI Act). *Official Journal of the European Union*, L 2024/1689, 12 July.

Executive Order 14148 (2025). Removing Barriers to American Leadership in Artificial Intelligence. *Federal Register*, 90 FR 8275, 20 January 2025.

Finlayson, S.G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I.S. and Saria, S. (2021). The Clinician and Dataset Shift in Artificial Intelligence. *New England Journal of Medicine*, 385(3), pp. 283–286.

Gervais, D. (2021). The TRIPS Agreement: Drafting History and Analysis. 5th ed. London: Sweet and Maxwell.

Gervais, D. and Yu, P.K. (2025). TRIPS and the Tradification of Intellectual Property. In: D. Collins and V. Vadi, eds. *Routledge Handbook on International Economic Law*. Abingdon: Routledge. Available at: https://ssrn.com/abstract=4850765 [Accessed 28 March 2026].

Hildebrandt, M. (2020). Law for Computer Scientists and Other Folk. Oxford: Oxford University Press.

IAPP — International Association of Privacy Professionals (2026). Global AI Law and Policy Tracker. Portsmouth, NH: IAPP. Available at: https://iapp.org/resources/article/global-ai-legislation-tracker/ [Accessed 11 June 2026].

ISO/IEC 42001 (2023). Artificial Intelligence — Management System. Geneva: International Organization for Standardization.

Italy (2025). Law No. 132/2025, Provisions and Delegations to the Government on Artificial Intelligence. Gazzetta Ufficiale della Repubblica Italiana, 25 September 2025; in force 10 October 2025. Rome: Italian Parliament.

Kazakhstan (2025). Law of the Republic of Kazakhstan No. 230-VIII ZRK "On Artificial Intelligence", signed 17 November 2025; in force 18 January 2026. Astana: Parliament of the Republic of Kazakhstan.

Kenya Ministry of ICT (2025). National Artificial Intelligence Strategy 2025–2030. Nairobi: Ministry of ICT.

Law No. 2575 (1982). Law on the Council of State. Official Gazette No. 17580, 20 January 1982. Ankara: Turkish Grand National Assembly.

Law No. 2577 (1982). Law on Administrative Procedure. Official Gazette No. 17580, 20 January 1982. Ankara: Turkish Grand National Assembly.

Law No. 6698 (2016). Law on the Protection of Personal Data (KVKK). Official Gazette No. 29677, 7 April 2016, as amended by Law No. 7545, Official Gazette No. 32501, 12 March 2024. Ankara: Turkish Grand National Assembly.

Mantelero, A. (2024). The Fundamental Rights Impact Assessment (FRIA) in the AI Act: Roots, Legal Obligations and Key Elements for a Model Template. *Computer Law and Security Review*, 54, pp. 1–20.

Newsom, G. (2024). Veto Message on SB 1047. Sacramento: Governor's Office of California, 29 September 2024. Available at: https://www.gov.ca.gov/wp-content/uploads/2024/09/SB-1047-Veto-Message.pdf [Accessed 28 March 2026].

NIST — National Institute of Standards and Technology (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). Gaithersburg: NIST. DOI: 10.6028/NIST.AI.100-1.

Novelli, C. et al. (2025). A Robust Governance for the AI Act: AI Office, AI Board, Scientific Panel and National Authorities. *European Journal of Risk Regulation*, 16(2), pp. 566–590.

OECD.AI (2026). National AI Policies Database. Paris: OECD.AI Policy Observatory. Available at: https://oecd.ai/en/dashboards/overview [Accessed 11 June 2026].

OECD (2024). Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449, adopted 22 May 2019, amended May 2024). Paris: OECD.

Okediji, R.L. (2014). Legal Innovation in International Intellectual Property Relations: Revisiting Twenty-One Years of the TRIPS Agreement. *University of Pennsylvania Journal of International Law*, 36, pp. 191–260.

Okolo, C.T. (2023). AI in the Global South: Opportunities and Challenges Towards More Inclusive Governance. Washington DC: Brookings Institution.

Pasquale, F. (2015). The Black Box Society: The Secret Algorithms That Control Money and Information. Cambridge, MA: Harvard University Press.

PDPC/IMDA — Personal Data Protection Commission/Infocomm Media Development Authority (2020). *Model AI Governance Framework*. 2nd ed. Singapore: PDPC/IMDA.

People's Republic of China (2022). Administrative Provisions on Algorithm Recommendation for Internet Information Services, effective 1 March 2022. Beijing: Cyberspace Administration of China.

People's Republic of China (2023). Interim Measures for the Management of Generative AI Services, effective 15 August 2023. Beijing: Cyberspace Administration of China.

Republic of Korea (2025). Basic Act on the Development of Artificial Intelligence and Establishment of a Foundation for Trustworthiness (AI Basic Act, Act No. 20676), promulgated 21 January 2025; in force 22 January 2026. Sejong: Ministry of Science and ICT.

Roberts, H. et al. (2024). Global AI Governance: Barriers and Pathways Forward. *International Affairs*, 100(3), pp. 1275–1296.

SB-53 (2025). Transparency in Frontier Artificial Intelligence Act. California: Legislature of the State of California. Signed October 2025.

South Africa Department of Communications and Digital Technologies (2024). *National AI Policy Framework.* Pretoria: DCDT.

Taiwan (2026). Artificial Intelligence Basic Act, passed 23 December 2025; promulgated and in force 14 January 2026. Taipei: Legislative Yuan.

TBMM — Türkiye Büyük Millet Meclisi (2024). Yapay Zekâya İlişkin Düzenlemeler Hakkında Kanun Teklifi, Esas No. 2/2234, submitted 24 June 2024. Ankara: Grand National Assembly of Turkey.

UNCTAD (2021). Technology and Innovation Report 2021: Catching Technological Waves — Innovation with Equity. Geneva: United Nations Conference on Trade and Development.

UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence. Paris: UNESCO. Adopted 23 November 2021, 41st General Conference.

UK Government (2023). A Pro-Innovation Approach to AI Regulation (White Paper, CP 815). London: HMSO.

UN Secretary-General's High-Level Advisory Body on AI (2024). Governing AI for Humanity. Final Report, 19 September 2024. New York: United Nations.

Vietnam (2025). Law on Artificial Intelligence (Law No. 134/2025/QH15), adopted 10 December 2025; in force 1 March 2026. Hanoi: National Assembly of the Socialist Republic of Vietnam.

WHO — World Health Organization (2021). Ethics and Governance of Artificial Intelligence for Health: WHO Guidance (ISBN 978-92-4-002920-0). Geneva: World Health Organization.

World Bank (2024). Global Trends in AI Governance: Evolving Country Approaches. Washington DC: World Bank.

WTO (1994). Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS), Marrakesh Agreement Establishing the WTO, Annex 1C, 15 April 1994. Geneva: WTO.

WTO (2001). Declaration on the TRIPS Agreement and Public Health (Doha Declaration), WT/MIN(01)/DEC/2, 14 November 2001. Geneva: WTO.

Yu, P.K. (2006). TRIPs and its Discontents. *Marquette Intellectual Property Law Review*, 10(2), pp. 369–410.

Yu, P.K. (2025). Bridging the Global Artificial Intelligence Divide. SSRN Working Paper 5394040. Available at: https://ssrn.com/abstract=5394040 [Accessed 28 March 2026].

Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J. and Oermann, E.K. (2018). Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: a Cross-Sectional Study. *PLOS Medicine*, 15(11), e1002683.

# Resistance as a Technomoral Virtue for Our Times

*Nitya Mandyam*
*University of Cambridge*



This paper critically examines differing responses to the harms of Artificial Intelligence (AI) and explores the ethical dimensions of resisting AI. It draws on case studies of three distinct AI-related harms - bias in language models, the environmental impact of large AI models, and job displacement due to automation, to define what constitutes an act of resistance. Applying a virtue ethical framework to the question of resistance, I posit that resistance is a technomoral virtue in our times and can act as a creative, generative, and dynamic principle in producing technologies that engender human flourishing.

Author's Note: The author would like to acknowledge the significant proliferation of data centres and the simultaneous rise in resistance towards them from local communities, activists, and policymakers, since creating this manuscript in late 2024. The data centre resistance highlighted in this paper at Cerrillos, Santiago, Chile (2020-24) is a forerunner of a multitude of similar resistances since that are occurring across the globe, with data centres becoming a focal point for AI resistance. A small but significant number of these conflicts have culminated in the imposition of moratoriums and bans on data centre development by local governments.

**Keywords**: Resisting AI, Harm Reduction, Virtue Ethics, AI Extractivism, Technomoral Virtue

## Introduction

Recent methods of producing Generative Artificial Intelligence (AI) have raised concerns about harms ranging from the individual to societal, including threats to privacy and autonomy, amplification of bias in training data, rising unemployment due to automation, copyright law violation, power imbalances caused by consolidation of data and computer resources, and environmental damage caused by increased energy consumption demand. While AI engineers, researchers and legislators have sought harm reduction to mitigate some of these concerns, activists favour resistance (Narayanan, 2023). The primary argument against the harm reduction approach is that it risks legitimising unjust practices and outcomes. Resisting AI, on the contrary, can impede technological progress and the benefits it could bring. This paper focuses on the resistance piece of this debate and uses the virtue ethical framework to examine the following questions: When are acts of resistance justified? What kinds of resistance are unethical?

Section 1 of this paper lays out the different dimensions of the debate with concrete examples illustrating responses to three specific AI harms as important case studies: bias in language models, the environmental footprint of large AI models and job displacement due to automation. Aided by these examples, I will define what constitutes an act of resistance in AI, a definition that will be revisited later and refined to understand when resistance is virtuous. Section 2 examines this issue through the different normative ethical frameworks and lays out an argument for why virtue ethics is most suited to help resolve this question. Section 3 examines the virtues demonstrated by different parties in acts of AI resistance. Using philosopher Shannon Vallor's framework of technomoral wisdom (Vallor, 2016), it adjudicates whether these are holistically unvirtuous acts or steps toward resisting an unvirtuous ecosystem of AI production.

Based on these analyses, I present this essay's primary argument in Section 4: that resistance is a necessary technomoral virtue for our times. I reframe resistance as a creative and dynamic player in honing technologies that can prevent unsafety and act as a guiding principle in disrupting the structural dynamics that produce contemporary AI. I present resistance as a creative, constructive and generative principle for building moral machines in an unvirtuous

ecosystem and envision the concrete impacts of resistance on current and new technologies.

## 1. Resistance as an Alternative to Harm Reduction

### *1.1. Some Examples of AI Harms*

What constitutes technological harm can be deeply subjective, depending on the morality of the time, the metaethics of the people in question, and the epistemic stance of the arbitrator. Instead of attempting a universal definition of what constitutes "AI harm", this section relies upon some tangible scenarios where there is broad consensus on the presence of harm. Three well-acknowledged problematic side effects of contemporary AI are:

*The Perpetuation of Bias in Language Models*: This involves the different forms of oppression (among them racism, misogyny, transphobia and classism) inherent in the training data of these models and their subsequent persistence in the text generated by them (Bender et al, 2021, Gallegos et al, 2024).

*Sustainability Concerns*: The massive amounts of computing power required to pre-train, store, fine-tune, and infer from Generative AI models necessitate an increase in the consumption of environmental resources (A. S. Luccioni et al, 2024).

*Displaced Employment due to Automation*: It has been speculated that as AI models improve, employers will seek to cut costs by replacing human labour with AI, which adversely affects workers and increases economic inequality (Hatzius et al, 2023).

### *1.2. Harm Reduction*

Harm reduction seeks to redress these harms by working within the system of contemporary AI production through technical and non-technical interventions. Technical interventions involve algorithmic approaches implemented while creating and deploying a model. Non-technical interventions can take the forms of human oversight, building and adhering to ethical codes within tech corporations, legislation by local and international authorities and regulations implemented by nation-states. Applying these initiatives to the three harms described above:

*Mitigating Bias*: A few techniques used to counter bias in language models are data curation prior to pretraining via toxicity filtering performed by annotators, Reinforcement Learning based on Human Feedback (RLHF) to align models to human preferences before model deployment (Lindström et al, 2024), and countering more downstream harms post-deployment by providing clear disclaimers allowing for user feedback.

*Sustainability Initiatives*: Researchers concerned about the sustainability implications of Generative AI models seek to quantify the environmental toll caused by their training and deployment and then find technical ways of minimising it. Techniques such as quantisation and low-rank adaptation reduce the computational power required for storing and fine-tuning models (Dettmers et al, 2023). A non-technical solution is to build data centres in areas powered by renewable energy. More recently, researchers have suggested drawing upon existing mechanisms of green policy, such as energy ratings, not unlike light bulbs or automobiles (Luccioni et al, 2024).

*Reframing Jobs*: Just as the industrial revolution created new job categories, AI purveyors predict that after a possible lull in employment opportunities, new jobs will emerge around managing Generative AI technologies, even as there is rising scepticism about this stance (Acemoglu et al, 2022).

### *1.3. Acts of Resistance*

If harm reduction approaches are characterised by cooperation and constructive course corrections, acts of resistance are inherently non-cooperative and often seek to halt the course of implementation of a technology. Three singular events from recent history can be highlighted as being acts of resisting the harms mentioned at the beginning of this section:

*The Firing of AI Researcher Timnit Gebru (2020)*: Two years before ChatGPT took the world by storm in late 2022, an AI Researcher at Google, Dr. Timnit Gebru, coauthored a paper (Bender, 2021), along with colleagues and academics outside of Google advocating for a halt to the development of large language models (LLMs)

on the grounds of implicit bias. This suggestion directly contradicted Google's AI goals and would eventually lead to their firing.

*Chilean Data Centre Protests (2020-24)*: (Arellano et al, 2020) Google had planned to build a Data centre in Cerrillos, a commune in Santiago, Chile, that would require 169 litres of water per second to cool its servers. Several parts of Chile have been severely drought-ridden in the past decade. Santiago, being an aquifer, was expected to source water for many regions in the country. Residents were baffled to see the project approved and resisted it via community organising through a group known as Mosacat. The resistance has delayed the project significantly, culminating in a Chilean court pulling the approval granted to Google in February 2024.

*The Writers Guild of America (WGA) Strike (2023)*: 2023 saw two significant strikes by unionised writers (WGA) and actors (SAG-Aftra) demanding better wages within Hollywood's notoriously unequal studio system for nearly six months from May to November (Silberling, 2023). AI was at the heart of two negotiating points in the strikes - the first was a demand to use Generative AI as a tool and not a replacement within the industry, and the second was to stop copyright misuse of artists' creations.

### *1.4. Defining Resistance*

The three case studies considered in this paper were chosen to reflect how acts of resistance can occur at chronologically different phases of AI development - before, during and after AI development. The Chilean environmental activism can be seen as resisting the planning and resourcing phase of AI development, Timnit Gebru's defiance can be considered as demonstrating resistance during AI development and the WGA strikes are reflective of resisting AI usage post-deployment.

The *resistors* include diverse players such as critics of AI (often tech workers themselves), frontline workers, people whose livelihoods are threatened by automation, anti-capitalists, marginalised communities more directly impacted by biases in AI systems and environmentalists. Additionally, the forms of resistance are very diverse. For instance, Timnit Gebru's defiance was a case of workplace insubordination leading to the termination of their employment. Data centre activism in Chile or Uruguay has taken the form of community organising and protesting via demonstrations and referendums. As a negotiating tactic, the SAG-Aftra and Writers Guild strikes disrupted everyday work routines, causing significant economic discomfort to the strikers and their bosses.

While these are all very different forms of resistance, they share a throughline: to disrupt "business as usual". Often, the goal is to cause sufficient discomfort to a tech company to incentivise them to change their actions. However, this can and often does come at the cost of disrupting unexpected parties, such as company workers or technology consumers. Resistance also comes at a cost to the person resisting because, by its very definition, they are counteracting a force larger than themselves.

## 2. Ethical Perspectives on AI Harms

### *2.1. Questions of Proportionality and Threshold*

Many supposed harms of AI are often themselves subject to fierce debates about the extent of their harmfulness. This debate, while slightly different from the one this paper tries to address, is nonetheless entangled with the question of resistance. Resistance can be reasonably perceived as an excessive response if the harms are small or slow enough to take effect that an intervention (technical or otherwise) can mitigate it down the line. Regulation via legislation or building guardrails as and when problems crop up post-deployment can, in turn, seem ineffectual for those whose livelihoods or dignity are affected by immediate harm. In other words, disagreement on the *proportionality of harms* is a definitive factor driving the two approaches here.

A consequentialist approach that can empirically evaluate the proportionality of harm caused would presumably be an apt choice to resolve this debate. This is a worthy and meaningful endeavour that many technologists have and should continue to undertake, but it does not address this debate

holistically. A crucial missing piece in this approach is the extent to which power, relationality and epistemic views shape the development of sociotechnical systems like AI. In his talk titled "Resistance or Harm Reduction?", Prof. Arvind Narayanan corroborates this view (Narayanan, 2023): *"Computer scientists, in particular, overwhelmingly prefer the harm reduction approach. It is a better match for our technical skills; we understand the details of these products, and we can help improve them. It is also a better match for the fact that we often receive money from tech companies, and we are perhaps uncomfortable with the kind of positions that resistance involves taking and the kinds of things that it involves doing."*

Many harm reduction approaches within AI already tend to have a utilitarian flavour. An action-oriented inclination towards analysing outcomes, seeking tangible optimal solutions, and working within institutional constraints to maximise profit suits the venture capital-driven corporate climate within which contemporary AI is created. In such a culture, resistance is intrinsically counterproductive in the way that productivity has been defined and rewarded. For a consequentialist to endorse resistance, two logical requirements must be met - that the harms outweigh the benefits of a particular technology and that the long-term outcome of resistance will maximise utility. The burden of proof here relies on predicting the course of a technology, and this is a difficult task even for the technically proficient!

Another way to approach responses to AI harms is to identify quantifiable thresholds that can be used to aid technical development. Deontological approaches find their place in this debate in harm reduction approaches through technical interventions by way of hard-coding machine ethics via programmable rules (Hooker and Kim, 2018). In developing the proper rules to mitigate harm, quantifying impact and drawing suitable thresholds based on that becomes essential. The primary difficulty here is that this process requires cooperation with the corporations responsible for creating AI products to demonstrate transparency and proactiveness in quantifying these harms. Companies prefer opacity about their products' inner workings to stay abreast of their competitors and secrecy about possible harms to avoid unsettling their consumers (Whittaker, 2021).

Additionally, the most common critique of deontology, which is that it falls short in its ability to handle genuine moral dilemmas (Timmermann, 2013) is of special relevance to the resistance debate in AI as it is challenging to ascribe overarching imperatives to different players especially when competing interests create diverse moral dilemmas. Deontology is nonetheless valuable in narrower contexts when one can clearly define the parameters of the moral conundrum. For instance, one could use a deontological framework to adjudicate whether data workers who perform tasks such as annotation, toxicity-filtering and fact-checking might be justified in organised action against their employers (Mathenge, 2024) depending on their compensation, working conditions and the psychological effects of the work.

### *2.2. A Question of Virtue?*

There are two key ways this debate is complex to resolve using consequentialist and deontological frameworks. The first is a blind spot in both ethical frameworks to account for how power dynamics and relationality shape the course of technologies and, consequentially, the different responses towards them. The second is the inherent difficulties in predicting the extensive social effects of AI systems, a phenomenon referred to as "*acute technosocial opacity*" by virtue ethicist Shannon Vallor (Vallor, 2016) .

Vallor argues that the pace of technological development in a globalised world order makes it impossible to gauge our futures or even assess where we are now. Techno-optimists tend to claim that we are "*approaching the next dizzying explosion of technosocial progress.*" At the same time, techno-pessimists prefer to imagine that we are "*teetering on a precipice awaiting a calamitous fall*", and the consensus on the present moment remains fuzzy (Vallor, 2016). This fuzziness, in turn, makes figuring out the right thing to do, i.e., finding an ethical solution, a difficult task. Vallor proffers that we move away from seeking solutions by optimising

imagined outcomes or attempting to build robust ethical rules. Instead, she urges us to think about the kinds of moral character, i.e., the *virtues*, that can promote human flourishing in the face of technosocial opacity.

Virtue ethics views moral character in agents as not an inherent or static trait but rather something developed and nurtured through consistent practice and habituation. It is not concerned with arbitration based on outcomes (consequentialism) or adherence to predetermined rules (deontology). It does posit that actions that come from a place of moral habituation will likely lead to good outcomes and prescribes many of the values that deontology does. However, it is more interested in reorienting the actions of moral agents (human or machinic) towards behaviours that promote human flourishing.

Virtue ethics thus allows for a dynamic, flexible, and context-dependent approach to resolving the debate around harm reduction and resistance. Harm reduction and resistance are both responses to the distinct anxieties created by the acute technosocial opacity generated by contemporary AI. Moving away from the impossible task of answering questions about proportionality or threshold of harms needed to resist AI prima facie, the coming section will examine the moral behaviours of the diversity of players involved in building and deploying AI, mitigating its harms, and resisting it, through a virtue ethical lens.

## 3. A Virtue Ethical Lens on Acts of Resisting AI

### *3.1. Moving from a Technoneutral to a Technomoral Ethos*

*"Technologies neither actively determine nor passively reveal our moral character, they mediate it by conditioning, and being conditioned by, our moral habits and practices."* (Vallor, 2016)

In her seminal work, Technology and the Virtues, philosopher Shannon Vallor draws parallels between three global ethical traditions, Aristotelian, Confucian and Buddhist, to develop a unified and pluralistic virtue ethical framework. Infusing this framework with principles drawn from feminist ethics, as well as a sound understanding of the nature of technological development, Vallor lays out a taxonomy of *technomoral virtues* to guide human flourishing in the face of technosocial opacity. These values are honesty, self-control, humility, justice, courage, empathy, care, civility, flexibility, perspective, magnanimity and technomoral wisdom.

Vallor contends, "*Ethics and technology are connected because technologies invite or afford specific patterns of thought, behaviour, and valuing; they open up new possibilities for human action and foreclose or obscure others*." (Vallor, 2016) The word *technomoral* reflects the inseparable and intricate ways our morality and technology inform and influence one another. It implicitly presents a counterpoint to a deterministic view of technological development because as long as there are many technomoral choices at every step in building a technology, we are not under an imperative to march towards a singular technological predestination.

### *3.2. On the Accumulation of Moral Debt in Generative AI*

If we accept that there are technomoral choices en route to assembling a technology, Vallor contends that we accrue “moral debt” when we put off dealing with moral problems (Renieris, 2022). In software engineering, when the speed of delivery is prioritised over writing robust code, systems begin to accrue “technical debt”. In more old school terms, choosing the “easier way out” has consequences. Moral debt is not dissimilar and is specifically very evident in a sociotechnical system like Generative AI.

Consider the plurality of resources that facilitate building and deploying a model like GPT-4: financial investment from the Global North, highly skilled technical workers engaged in building testing neural network architectures within OpenAI, vast swathes of digital information scraped with and without consent that serve as training data, labourers in the Global South who perform toxicity filtering and data annotation, the rare earth metals that power microchips within Graphical Processing Units (GPUs) and the water resources needed to cool data centres performing computation. Making ethical compromises in accumulating

these resources to prioritise cost-cutting or faster delivery of AI systems, accrues moral debt. The downstream harms are more complex to mitigate as any structural injustices overlooked in the building and deployment phase will persist.

### *3.3. Insubordination, Activism and Strikes: The Morality of Resistance*

Revisiting the examples of acts of resistance provided earlier in this paper, we can now take a deeper look at what was being resisted there. In the first example, two well-respected AI researchers at Google, Timnit Gebru and Margaret Mitchell, advocated for reconsidering Large Language Models (LLMs) as the course of action to pursue in AI development (Bender et al, 2021). Their examination of bias in training data and the environmental costs of pre-training led them to conclude that this was an ethical dead-end. Their subsequent disagreement with Google and their eventual departure from the company were subjects of newsworthy headlines. There is, however, little consensus on the sequence of events between the authors and the company. (Hao, 2020)

In the second example, an activist group in Chile named "Mosacat" led a campaign that very recently resulted in a temporary overturning of the approval for Google to construct a Data centre in an aquifer at Cerrillos, Santiago. Dr Sebastián Lehuaedé (Marx, 2024) studies data centre activism and says the following about their strategies: "*They went to street markets, they did workshops with the local community, they put up posters in bus stops in order to inform the community about this situation.*" The community's water needs and the dent the data centre's water consumption would create had yet to be analysed when the local environmental authority approved the project, a case of clear and wilful political negligence.

These examples are vastly different cases that resist different aspects of AI production and deployment. While Google seems to be the common denominator, it is not unique in the behaviour that engendered the resistance. Endorsing Dr. Gebru's research (Bender et al, 2021) would be sufficient cause to pause the primary course of AI development the company was pursuing and lose its edge in a hyper-competitive market. In the issue of building a resource-intensive structure in a drought-ridden country, it is worth asking: Is it morally incumbent on Google to inform locals about the amount of water required by a data centre? Is it fair that the moral burden of research and communication falls on a corporation when local politicians/businesspeople are more than willing to compromise on their community's needs? This situation reflects the complex nature of building Generative AI and reinforces Vallor's stress on considering what it would mean to make technomoral choices.

The other common denominator in these acts of resistance is the resistors' courage, care and magnanimity. Whether one agrees with Timnit Gebru's perspectives on LLMs technically or ethically, there is undeniable courage in going against one's self-interest to advocate for others. Similarly, activists in Chile demonstrate care and magnanimity in their willingness to educate and empower local populations unaware of the extent to which their lives might be affected by the construction of the data centre.

## 4. Reframing Resistance as a Generative Virtue

### *4.1. Resistance as a Technomoral Virtue*

A virtual ethical lens sheds light on our current AI ecosystem as product of complex interactions between multiple agents, technologies, and systems. An ecosystem that incentivises entities within to prioritise their interests over human flourishing does not demonstrate technomoral wisdom. In an unvirtuous ecosystem, resistance automatically emerges as a corrective principle to reorient within our technomoral futures. Resistance is then a virtue for our times for tech workers, activists, venture capitalists, and non-tech-savvy consumers alike.

French philosopher Paul-Michel Foucault famously posited that resistance exists wherever there is power and can never be external to it. Resistance as an inherently oppositional idea might make it sound less like a virtue and more like a behaviour one might want to avoid. But philosopher Joanna Oksala reinterprets Foucault's view to underscore resistance as a dynamic and creative principle to organise ourselves with: *"... resistance against*

*domination and the normalising effects of power/ knowledge must advance on multiple fronts. It consists of creative transformations of the self, communal forms of counterconduct, and critical interrogation of our present.*" (Lawlor, 2014)

Many of the traditional virtues contain a seed of oppositional energy: courage only makes sense in the presence of fear, humility counters hubris, and flexibility opposes rigidity. Virtues have an inherently relational quality, and resistance is no different. Vallor states that "*Any plausible virtue ethic must be rooted in a more or less stable range of human moral capacities, and in the absence of some radical alteration to our basic psychology, there is no reason to think that humans will suddenly acquire a wholly new repertoire of moral responses to the world.*" (Vallor, 2016) Resistance satisfies this requirement as it has existed as long as humans have sought power. Without engaging in creative acts of resistance, we would have never won many of the human rights we have today against forces of systemic oppression. When faced with technologies that risk entrenching patterns of oppression within AI systems, a phenomenon Vallor refers to as the "*AI Mirror*", resistance becomes a technomoral virtue for our times.

Author and historian of science and technology Jean-Baptiste Fressoz has written extensively about histories of technological risk and argues that "*what in history of technology is traditionally put in the footnotes, under the label of 'opposition' or 'resistance' to progress turns out to be essential for the 'shaping' of safer technological systems.*" (Fressoz, 2007) He bases this conclusion on multiple case studies about technological controversies such as deforestation due to development, side-effects of vaccination, and pollution due to chemical industries, where resistance shaped safer and more robust technological outcomes, thus corroborating Oksala's argument for resistance as a necessary principle for our advancement.

### *4.2. When is Resistance Unvirtuous?*

Virtue ethics prioritizes context and nuance over universality and no virtue is thus a categorical imperative. Resistance can only be deemed a virtue if it can contribute to eudaimonia meaningfully, and this requires a stance against all forms of oppression. The Greek translation of "resistance" is antistasi (or "anti-fascism" more literally), a word the was strongly associated with the Greek National Resistance during World War II, one of Nazi-occupied Europe's most robust resistance movements. Resistance becomes unvirtuous, however, the moment it coopts the oppressor's strategies of violence, apathy, coercion and hubris. History is a testament to movements against oppression that are often an amalgamation of unconsidered, reactionary and violent forms of resistance and revolutionary, liberatory and non-violent forms of resistance. Specifically, movements of technological resistance in the 19th century can be considered to contain both.

There has been renewed scholarly interest in examining movements of technological resistance during the Industrial Revolution in Europe. Journalist Brian Merchant's work (Merchant, 2023), for instance, reclaims the Luddite movement in Great Britain in the early 1800s from a maligned technophobic and regressive characterization and recasts it as an informed labour struggle by an impoverished working class in a techno-feudal state that sought to deprive workers of their wages by using automation as an excuse. While the Luddites undoubtedly engaged in violent resistance and embraced their monikers as machine-breakers, Merchant traces the arc of systematic oppression faced by the Luddites for decades before the resistance turned violent, including a deprivation of livelihoods and sometimes, actual lives of mill workers. Merchant points to the chilling parallels between the callousness of the feudal class to the mill workers
in the 19th century and the apathy of tech companies to gig workers or data workers in the Global South.

Critical scholarship on AI has noted the undercurrents of colonialism (Muldoon and Wu, 2023), extractivism (Hogan, 2024) and eugenics (Chan, 2025) present in the mechanics of AI development today. This presents a case for looking into the modes of resistance that have engendered long-lasting positive change in struggles against the same. History is full of cautionary tales of revolutions that have turned

violent and yet, there are inspiring examples of predominantly nonviolent resistance movements to be found in the last century in the anti-colonial resistance in India, anti-Apartheid struggles in South Africa and the civil rights movement in the United States. Some slogans from these movements that can guide us in characterising resistance as a virtue would be the Sanskrit word "*Satyagraha*", which translates to "*holding firmly to truth*" or the Nguni slogan, "*Amandla Ngewethu*", or, "*Power to the people!*"

Virtue ethics' insistence on context, nuance and presence mandates that different technomoral virtues acquire temporal importance based on what is most called for at that moment. Resistance is thus a virtue for our times *at this moment in human history*, and I posit that a thawing of this moral apathy would cause it to fade into the background to assume a quiet watchful eye that only springs into action when oppressive dynamics enter the picture. Until such a time, the tangible examples of ethical AI resistance laid out in the following section are worth contemplating.

### *4.3. Dimensions of Ethical AI Resistance*

In adherence to Vallor's call for virtues to be rooted in a stable moral disposition, it is worth examining what modes of ethical AI resistance can look like today. The table below presents a brief yet diverse set of examples of movements and actions in the world of technology that would qualify as the same.

**Table 1:** Examples of ethical AI resistance

| | | |
|---|---|---|
| *The digital degrowth movement* | Political/Environmental | AI Accelerationism |
| *Unionization/ Labour organizing* | Economic/Sociopolitical | Oppressive power dynamics in tech companies |
| *Grassroots activism* | Environmental/Decolonial | Extractivist logics in AI development |
| *Antitrust legislations* | Legal | Monopolisation of AI development |
| *Refusal/ Opting out* | Individual | Surveillance logics in AI development |
| *Strikes/ Non-cooperation* | Collective | AI hype, AI misuse, Livelihood deprivation |
| *Adversarial interoperability*[1] | Commercial | Control of technological access |
| *Whistleblower protection* | Employment Rights | Corporate opacity, retaliatory practices |

While this table is by no means exhaustive, there is nonetheless a striking absence of what resistance could look like in technical product development in AI. Many of the examples above fall under Blair Attard-Frost's "*AI Countergovernance*" framework (Attard-Frost, 2023), an alternative to traditional AI governance. Arguing that today's AI technologies "*are all by-products of a failed AI governance regime and the logics of its constituent organisations*," countergovernance is less concerned with technical intervention and more with opposing the political and economic ecosystems that foster them. To imagine what is possible in AI in the technical dimension when it is nurtured by technomoral choices grounded in resistance every step of the way, we can look towards a rare technological application, Signal, the encrypted messaging service.

Built by a non-profit foundation that seeks to provide surveillance-free messaging networks to everyone, developing Signal required a complete non-reliance on existing software and building entirely from the ground up, as the most widely used software packages of today, by default, collect and store metadata (Whittaker and Lund, 2023). Signal is the safe application of choice for movements across the globe that are resisting authoritarian regimes, protesting genocides or organising climate activism. Signal's commitment to a surveillance-free ethos is evident in all facets, from conception to funding to deployment and further, in its continued refusal to cave to demands from nation-states that do not permit surveillance-free communication (Greenberg, 2024) providing us with a valuable roadmap to build AI grounded in the virtue of resistance.

## Conclusion

This paper has examined the ethical dimensions of resistance to AI through a virtue ethics framework. While harm reduction approaches operate within existing systems to mitigate AI's negative effects, resistance represents a necessary corrective force in an ecosystem that mirrors historic patterns of oppression. By reframing resistance as a technomoral virtue rather than mere opposition, we recognize its generative potential to shape safer, more ethical technological systems. In confronting AI's challenges, resistance provides a pathway to technologies that enhance rather than diminish human dignity and flourishing.

**References**


Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2022). Artificial Intelligence and Jobs: Evidence from Online Vacancies. *Journal of Labor Economics*, 40(S1), S293–S340.

Arellano, A., Cifuentes, L., & Ríos, C. (2020). The Dark Areas of the Environmental Evaluation that "Blindly" Authorized the Google Megaproject in Cerrillos.

Attard-Frost, B. (2023, December 16). AI Countergovernance.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.

Chan, A. S. (2025). Predatory Data: Eugenics in Big Tech and Our Fight for an Independent Future. University of California Press.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs (arXiv:2305.14314). arXiv.

Doctorow, C. (2019, October 2). Adversarial Interoperability. Electronic Frontier Foundation.

Fressoz, J. (2007). Beck Back in the 19th Century: Towards a Genealogy of Risk Society. *History and Technology*, 23(4), 333–350.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and Fairness in Large Language Models: A Survey (arXiv:2309.00770). arXiv.

Greenberg, A. (2024). Signal Is More Than Encrypted Messaging. Under Meredith Whittaker, It's Out to Prove Surveillance Capitalism Wrong. Wired.

Hao, K. (2020, December 3). A Leading AI Ethics Researcher Says She's Been Fired from Google. *MIT Technology Review*.

Hatzius, J., Briggs, J., Kodnani, D., & Pierdomenico, G. (2023). The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani). Goldman Sachs.

Hogan, M. (2024). The Fumes of AI. Critical AI, 2(1).

Hooker, J. N., & Kim, T. W. N. (2018). Toward Non-Intuition-Based Machine and Artificial Intelligence Ethics: A Deontological Approach Based on Modal Logic. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 130–136.

Lawlor, L., & Nale, J. (Eds.). (2014). Resistance. In The Cambridge Foucault Lexicon (1st ed., pp. 432–437). Cambridge University Press.

Lindström, A. D., Methnani, L., Krause, L., Ericson, P., Troya, Í. M. de R. de, Mollo, D. C., & Dobbe, R. (2024). AI Alignment through Reinforcement Learning from Human Feedback? Contradictions and Limitations (arXiv:2406.18346). arXiv.

Luccioni, A. S., Jernite, Y., & Strubell, E. (2024). Power Hungry Processing: Watts Driving the Cost of AI Deployment? The 2024 ACM Conference on Fairness, Accountability, and Transparency, 85–99.

Luccioni, S., Gamazaychikov, B., Hooker, S., Pierrard, R., Strubell, E., Jernite, Y., & Wu, C.-J. (2024). Ratings—So Why Can't AI Chatbots? *Nature*.

Marx, P. (2024, February 6). How to Stop a Data Center. *Disconnect*.

Mathenge, R. (2024). Data Workers Organizing – The African Content Moderators Union. In: M. Miceli, A. Dinika, L. Dingebacher, C. Salim Wagner, and K. Kauffman (eds), The Data Workers' Inquiry. Creative Commons BY 4.0.

Merchant, B. (2023). Blood in the Machine: The Origins of the Rebellion Against Big Tech. Little, Brown.

Muldoon, J., & Wu, B. A. (2023). Artificial Intelligence in the Colonial Matrix of Power. *Philosophy & Technology*, 36(4), 80.

Narayanan, A. (2023, October 27). Resistance or Harm Reduction?

Nguyen, A., & Mateescu, A. (2024). Generative AI and Labor: Power, Hype, and Value at Work. *Data & Society Research Institute*.

Renieris, E. (2022). Virtue Ethics and Technomoral Futures [Broadcast].

Silberling, A. (2023, September 26). The Writers' Strike is Over; Here's How AI Negotiations Shook Out.

Timmermann, J. (2013). Kantian Dilemmas? Moral Conflict in Kant's Ethical Theory. *Archiv Für Geschichte Der Philosophie*, 95(1), 36–64.

Vallor, S. (2016). Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford University Press USA.

Whittaker, M. (2021). The Steep Cost of Capture. *Interactions*, 28(6), 50–55.

Whittaker, M., & Lund, J. (2023, November 16). Privacy is Priceless, but Signal is Expensive. Signal.

# The Epistemic Frontier of Capitalism: A Theory of Epistemic Regression through Artificial Intelligence and Misinformation

*James Rice*
*University of Essex*



This article advances a single thesis about contemporary capitalism: under the pressure of frontier artificial intelligence and platform-mediated misinformation, accumulation, growth, and knowledge production has entered a phase in which outcomes are increasingly shaped by the privatization and strategic destabilization of our shared epistemic commons at the technological frontier. These, I define as consensus-based standards of scientific evidence, categories or criteria for a priori truth, and collaborative channels through which public reasons are disputed and agreed upon, these are facts and features of our politics that capitalism itself requires to coordinate expectations, channel investment, and establish and maintain political legitimacy. Artificial intelligence automates tasks and enhances productivity; but it also industrializes cognitive work itself, enabling the large-scale privatization of decision-making authority and the concentration of power, assets, and skills. AI defines our era, I suggest, but also threatens to disenfranchise a multitude of thinkers and creators, capable of contributing immense value.

Misinformation, on the other hand, threatens our sense of belonging in knowledge-based communities of practice, through the infusion of divisive, polarizing, and invasive falsehoods, intended to seed doubt, confusion, and anarchy. Misinformation, I argue, is the strategic manipulation of shared reality, often working at odds with reasonable intuition, and therefore deeply dangerous when undetected. The paper develops this thesis through two tightly linked case studies — AI and misinformation as strategies operating with the purpose of extreme capital accumulation funnelled to an elite subset of our technocratic society — and argues that capitalism now profits from corroding the epistemic and intellectual preconditions of its own vitality through information warfare.

**Keywords**: Epistemic Capitalism, Artificial Intelligence, Misinformation, Political Economy, Epistemic Commons, Economics of AI

## Introduction: Capitalism's Epistemic Turn

Capitalism has always been, at its core, a system for coordinating uncertainty (Williamson, 1985; Shackle, 1972; Lachmann, 1986). From its earliest theoretical articulations, political economists understood capitalism not merely as a system of exchange or production, but as a social order that transforms dispersed, uncertain social knowledge into coordinated, collective, productive action. The price system aggregates dispersed information about scarcity and preference, firms serve to facilitate transactions when markets fail to coordinate efficiently, and states stabilize legal expectations by enforcing contracts, property rights, and macroeconomic order. After classical theory, subsequent literature in information economics and institutional theory further emphasized that uncertainty reduction is a precondition for efficient investment, innovation, and institutional legitimacy (Arrow, 1974; Hayek, 1945; North, 1990). What is distinctive about the present landscape is therefore not the existence of uncertainty as such, but the transformation of how uncertainty is governed, monetized, and politically contested under a new technological and institutional regime.

Two literatures have expanded rapidly over the past decade, but largely in parallel. One (our first case study) examines artificial intelligence, automation, and frontier technologies, focusing on productivity, labour markets, inequality, and long-run growth (Acemoglu and Johnson, 2023; Jones and Tonetti, 2020). The other (the second case study) analyses misinformation, fake news, and epistemic polarization, focusing on democratic stability, media ecosystems, and political behaviour (Vosoughi et al, 2018; Benkler et al, 2018). Despite addressing related phenomena, these literatures rarely speak to

one another in a systematic theoretical way. Artificial intelligence research often assumes a relatively homogeneous macroeconomic informational environment within which economic agents optimize, while misinformation research frequently treats technological change in information systems as a neutral transmission channel rather than as a reorganization of capitalist accumulation itself. In other words, technology mediates the flow of information, and misinformation corrupts or interrupts this flow. This paper will analyse these two phenomena in tandem, drawing on a third body of work, in analytic social epistemology and philosophy of technology, that, I argue, has recently developed the conceptual resources needed. These ideas and resources include models of hermeneutical injustice, accelerating technological change, and epistemic dispossession — ultimately, I integrate these concepts into a theory of technocratic capitalist accumulation driven, on two opposing (and sometimes allied) sides, by AI and misinformation (Milano and Prunkl, 2025; Alvarado, 2023; Kwok, 2025). This integration is ultimately what this paper contributes.

The separation between idealized epistemic ecosystems and unequal economic reality belies a deeper structural transformation. Artificial intelligence is not simply a labour-saving technology, and misinformation is not merely a tech-induced breakdown of civic norms. This paper argues that capitalism has entered a transformative phase in which the control of computational and cognitive inference—classification, prediction, ranking, and narrative framing—has become automated, concentrated, and central to accumulation. Under this regime, the production of shared knowledge is transformed from a public, professional function stabilized by institutional norms into an object of private investment, proprietary control, and mass strategic manipulation, embedded within platform-based business models and data-intensive physical infrastructures. Inference is becoming a core asset, and its privatization is reshaping accumulation, authority, and class relations at the same time.

The single thesis of this paper is that this transformation produces a structural contradiction specific to the present conjuncture: the more successfully capital encloses and destabilizes the epistemic commons, the more it erodes the conditions of its own long-run coordination. Epistemic frontier capitalism is, in this precise sense, self-undermining. It is a regime whose most profitable frontier is identical with the substrate of its own reproducibility. The argument proceeds by specifying the epistemic commons as a condition of coordination (Section 1), then developing the core claim that AI industrializes and encloses inference (Section 2), showing that misinformation is an endogenous accumulation strategy within this regime rather than an exogenous pathology (Section 3), and formalizing the resulting contradiction in dialogue with its classical antecedents in Marx and Polanyi (Section 4). Section 5 draws out the class and state implications, and the conclusion identifies the empirical signatures by which the regime can be recognized.

## 1. Capitalism, Coordination, and the Epistemic Commons

Any theory of capitalism at the epistemic frontier must begin with a definition of the economic structure itself. Capitalism is not reducible to markets, nor to private property alone. It is a historically specific social order characterized by generalized commodity production, private control over investment, competitive accumulation, and a state apparatus that enforces property rights while managing social conflict and macroeconomic instability. Crucially, capitalism also depends on institutions that induce stability, rendering the immediate future calculable and predictable, enabling economic and political actors to form expectations, coordinate plans, and assess the legitimacy of authority. Recent research in political economy emphasizes that the predictability of social structures is first and foremost social and based on dispersed and coordinated knowledge. Capitalism and democratic institutions rest on shared frameworks of valuation, expertise, and legitimacy that make coordination possible under radical uncertainty.

Marx's analysis foregrounded capital accumulation as the driving logic of capitalism, emphasising how surplus-value extraction structures class relations and is transformed by evolving technology. His Grundrisse formulation of the "general intellect" offers a sharper point of entry for the present argument than Capital does: it foregrounds how capital seeks to embed knowledge, coordination, and prediction into technological systems, altering the social distribution of agency (Marx, 1973). Where steam-era fixed capital embedded knowledge of mechanical processes, AI-era fixed capital embeds knowledge of social and cognitive processes. The general intellect now takes the form of trained weights, ranking functions, and recommender systems, and its enclosure is correspondingly more consequential because what is enclosed is not the labour of production but the labour of collective sense-making. Maximilian Kasy has recently argued that artificial intelligence is best understood as "the means of prediction," a productive asset whose ownership and governance determine the distribution of its benefits and whose privatisation therefore recomposes the underlying class relation (Kasy, 2025).

Weber approached capitalism from a complementary angle, emphasising rationalisation, bureaucracy, and calculability as fundamental cultural and organisational conditions (Weber, 1978). What Weber called calculability is one specific component of what I call the epistemic commons. Polanyi's contribution was to show that markets are not self-sustaining but depend on institutions that moderate commodification, and that the commodification of labour, land, and money generates systemic dislocation (Polanyi, 1944). Where Polanyi identified labour, land, and money as the "fictitious commodities" whose commodification threatens society, Jessop extends the idea of the epistemic commons as a fourth fictitious commodity (Jessop, 2007), one whose commodification threatens society in a deeper way because it is the condition of possibility of recognising the threats produced by the commodification of the other three.

What unites these canonical traditions is an implicit recognition that capitalism depends on a socially agreed-upon epistemic order. Prices, contracts, and technologies only function when actors share sufficient agreement about facts, rules, and expectations; together with Keynes's observation that long-horizon investment rests on socially constructed expectations (Keynes, 1936; Beckert, 2016), regulatory standards, scientific expertise, and bureaucratic procedures historically transformed radical uncertainty into co-ordinable risk. The epistemic commons is the collective inferential substrate on which this scaffolding depends: the stock of shared vocabularies, categories, procedures of proof, and trusted channels through which agents interpret their circumstances and adjudicate competing claims.

Analytic social epistemology has recently developed the concepts needed to specify what happens when the basis for shared knowledge and discovery is undermined through adversarial shifts in the politically and technologically mediated information environment. Silvia Milano and Carina Prunkl, building on Miranda Fricker, suggest that there exists an ontological condition in which collective epistemic resources systematically prevent individuals from making sense of their social experiences (Milano and Prunkl, 2025). I argue that this very threat is being manifested through the imminent confluence of superintelligent AI and widespread online misinformation.

The object of this inquiry is the collective resources of intellectual labour, rather than information unique to the individual knower: a lacuna in the shared stock of interpretive resources is a deficiency in the biological medium within which agents form beliefs. Conversely, the epistemic commons is not a pure public good but a commons in Ostrom's sense (Ostrom, 1990), sustained by governance arrangements such as public debate over fundamental truths, professional editorial judgement, evidentiary rules of legal procedure, professional occupational licensing, financial and accounting conventions, statistical testing and disclosure standards, and the norms of scientific replicability and falsification (Popper, 1959). These arrangements are expensive to operate and vulnerable to displacement; that vulnerability is the opening through which the

accumulation dynamic identified in this paper moves.

The transition to the next section turns on a simple claim. These background conditions are not free, not automatic, and not self-reproducing. They are produced by specific institutions, under specific governance arrangements, and can be eroded by economic processes that do not internalize their value. The remainder of the paper analyses what happens when the frontier technologies of capitalist accumulation begin systematically to enclose and to degrade the very commons they presuppose.

## 2. AI as the Industrialization of Inference at the Epistemic Frontier

The first case study is artificial intelligence (AI). Technological change has always reshaped capitalism, but not all technologies are equal in their systemic effects. General-purpose technologies reorganise production, institutions, and social relations across the economy (Bresnahan and Trajtenberg, 1995). Artificial intelligence belongs to this category, but with a distinctive feature: it targets cognitive inference itself. Traditional industrial technologies mechanised physical labour and first-generation information technologies mechanised calculation and storage; artificial intelligence mechanises classification, prediction, and decision-making under uncertainty. In doing so, it transforms not only how goods are produced but how choices are made, authority is exercised, and epistemic legitimacy is constructed. Ramón Alvarado specifies the point: AI and the data-science methods associated with it are first and foremost epistemic technologies, designed to manipulate epistemic content through epistemic operations such as prediction and analysis (Alvarado, 2023).

Once we recognise AI as an epistemic technology whose functional role is to perform inference at scale, the central structural claim follows: the industrialisation of inference transforms a historically quasi-public function into a privately appropriable asset. Previous technological revolutions mechanised activities largely external to the commons; the AI revolution mechanises the intellectual activities simultaneously internal to it — classification, adjudication, ranking, prediction — whose prior institutional carriers (courts, editorial boards, professional bodies, statistical agencies) were precisely the governance arrangements that kept the commons functioning.

AI systems increasingly substitute for human judgement in domains that involve interpretation, evaluation, and normative discretion. Coding algorithms, predictive financial systems, automated misinformation detection tools, and ubiquitous recommendation engines all perform functions once exercised by human agents embedded in institutional contexts with explicit accountability structures. By embedding these functions in algorithmic systems, firms and states relocate authority into technical, model-based artefacts that are difficult to contest or audit — what Pasquale diagnosed, well before the generative turn, as the "black box society" (Pasquale, 2015).

At the epistemic frontier, accumulation depends on owning and governing systems that define what content is visible, credible, and actionable: search engines, social media feeds, recommendation systems, and models that quantify risk. The economic value of these systems lies precisely in their capacity to shape expectations and behaviour at scale. Zuboff captures the velocity of the transition: "*surveillance capitalism vaulted to the commanding heights of the social order at astonishing speed*" (Zuboff, 2019), establishing a regime of what she calls "epistemic inequality." Epistemic inequality is the distributional face of a deeper process in which the commons from which both parties to the inequality used to draw is itself being appropriated.

The enclosure has three interlocking consequences. First, it converts non-rival inferential capacity into an excludable asset: when classification and prediction are embedded in proprietary systems behind trade-secret protections, the evidentiary procedures that shape consequential decisions — creditworthiness, employability, newsfeed ranking, diagnostic triage — are withdrawn from public scrutiny. Chi Kwok frames this as "epistemic dispossession": a structural form of

injustice in which socio-technical infrastructures systematically extract, suppress, or appropriate epistemic resources from those who supply them (Kwok, 2025). The user who contributes training data and the public whose discursive categories are reshaped by platform ranking are not exchanging epistemic goods with the firm on terms that leave the commons intact: they participate in an extraction whose principal output is a proprietary model and whose principal residue is a diminished shared resource.

Second, enclosure concentrates inference: data-intensive production exhibits increasing returns (Jones and Tonetti, 2020) and, combined with fixed costs of model training and scarce compute, produces structural concentration. A small number of firms own the dominant classificatory architectures through which large fractions of economic and political life are interpreted; concentration of inference produces something closer to ideological power than to ordinary rents. Third, and most consequentially, enclosure introduces a novel vulnerability. Proprietary inference systems generate profit by shaping the informational environment in which decisions are made; but the value of any particular inferential output depends on the user's belief that the underlying evidentiary process is reliable. When the same technology that generates credible outputs can also generate at scale what Fritts and Cabrera call "phantom patterns" — correlations that are statistically detectable but meaningless — the infrastructure of inference and the infrastructure of misinformation become physically and economically indistinguishable (Fritts and Cabrera, 2022). The commons is thereby not merely under-supplied by enclosure; it is actively polluted by the outputs that enclosure makes profitable to produce.

Productive inference expands the set of feasible goods and services and contributes to long-run growth. Extractive inference redistributes surplus by manipulating informational asymmetries, attention, and predictable consumer behaviours for the purpose of profit. Both generate surplus, but only the former aligns inference-accumulation with social welfare. In the regime described here, the systematic tendency is toward the latter because it is the latter that most efficiently exploits the commons. The intellectual infrastructure of artificial intelligence — the mathematics and the science — once carried by public or quasi-public research-based institutions, has been privatised, concentrated, and rendered productively indistinguishable from its pathologies. We now move from AI to misinformation because agglomeration and the erosion of cultural integrity is the structural condition from which misinformation emerges as an accumulation strategy unique to platform economies mediated by advanced and ubiquitous forms of machine learning.

## 3. Misinformation as an Intentional Strategy for Asymmetric Accumulation

The second case study is misinformation. Misinformation is typically framed as a pathology of democratic discourse, a byproduct of cognitive bias or technological disruption. Within the framework developed here, it must be theorised as an endogenous outcome of accumulation rather than an external shock. When profits depend on controlling access to information, synthesising artificial expectations, and skewing belief formation, the strategic production of misleading information becomes economically rational, even as it imposes substantial social costs. Misinformation is not a glitch; it is an equilibrium feature of the system. Stiglitz and Kosenko have recently shown that standard results on market efficiency break down once disinformation is treated not as noise but as a strategic choice variable deployed by informed actors against less informed counterparties (Stiglitz and Kosenko, 2024a, 2024b).

False or misleading content diffuses differently from reliable information, often spreading more rapidly and widely within networked environments (Vosoughi et al, 2018). This pattern is not accidental. Platform architectures reward engagement, novelty, and emotional salience, creating incentives that favour content capable of capturing attention regardless of its veracity (Allcott and Gentzkow, 2017). Three mechanisms sustain the equilibrium. The first is the engagement economy itself. Platforms monetise attention, and attention responds disproportionately to affect, novelty, and

violation of expectation. It follows — from the business model rather than from any individual bad actor — that content optimised for engagement will tend to be content whose relationship to truth is incidental. The saturation is the emergent property of competition for attention on a substrate that does not reward the costlier signals that correlate with truth.

The second mechanism is the generative supply shock. Large language models and diffusion models have lowered the marginal cost of producing plausible-seeming content to approximately zero. The political-economic implication is that the ratio of synthetic-to-curated content in circulation will rise, not because anyone has decided it should, but because a firm that declines to use generative tools is outcompeted by one that does (Kay, Kasirzadeh and Mohamed, 2024). The structural shift from a curated-content equilibrium to a generative-content equilibrium changes what the epistemic commons contains. Under the curated equilibrium, even highly partial sources had to be produced by identifiable actors who incurred real costs and could be reputationally sanctioned. Under the generative equilibrium, vast quantities of content can be produced with no identifiable actor and no reputational exposure. The distinction between content and noise begins to erode because the cheapest path to communicative bandwidth runs through generation rather than through epistemic effort.

The third mechanism is the manufactured-doubt strategy, which predates the present moment but is amplified by it. Oreskes and Conway documented how tobacco and fossil-fuel interests discovered that it was cheaper to manufacture uncertainty about scientific consensus than to contest it on the merits (Oreskes and Conway, 2010). What is different now is that the infrastructure for manufacturing doubt is continuously available, inexpensive, and targeted, and asymmetric media ecosystems institutionalise the dynamic (Benkler et al, 2018; Jungherr, 2023). The manufactured-doubt strategy works by undermining the conditions of political self-determination at the commons level rather than at the claim level. A regulator persuaded that the scientific consensus on tobacco or climate is contested does not need to be persuaded that any particular claim is false; the regulator has already been moved out of the range in which the question of truth is actionable. Scaling this effect through AI-mediated infrastructures is, from the strategic actor's perspective, not a corruption of the infrastructure but a use of it.

Beliefs function as coordination devices in capitalist societies. Shared expectations about monetary and fiscal policy, technological progress, the future of regulation, and social norms enable long-horizon investment and collective action (North, 1990). When the structure of public beliefs becomes fragmented, coordination costs rise. For actors seeking to delay regulation, avoid liability, or weaken collective action — for example on climate change — the fragmentation of public knowledge can be advantageous. Misinformation increases uncertainty and disagreement, reducing the probability that coherent political or regulatory responses will emerge. Economic actors do not need to believe misinformation themselves for it to be strategically useful; what matters is that misinformation disrupts common knowledge, so that citizens, regulators, and investors cannot agree on basic facts and policy regresses into a contested, delayed, ineffective state.

What unites these three mechanisms is that each generates private returns while depleting a shared resource. The firm that optimizes for engagement, the firm that floods a market with synthetic content, and the actor that manufactures doubt about climate science all capture value by increasing the costs of collective sense-making. Milano and Prunkl's concept of "epistemic fragmentation" names the cumulative result. The public does not merely receive different information under algorithmic mediation: they lose shared access to the interpretive context within which information can be evaluated (Milano and Prunkl, 2025). Fragmentation is not simply heterogeneity of belief. A heterogeneous public can deliberate when it shares evidentiary standards and a common vocabulary of adjudication; a fragmented public cannot, because the interpretive substrate needed to translate

between positions has been partitioned into incompatible proprietary contexts.

The basic argument and conclusion for this second case study is not that misinformation is qualitatively new or that audiences are uniquely credulous. It is the narrower structural claim that the cost function of epistemic commons-degradation has collapsed while the private returns to such degradation have risen — sufficient, in the elementary economics of externalities, to guarantee over-production without any pathology of individual actors. The over-production is not a side-effect of the accumulation process but the accumulation process itself: misinformation is what the system produces in the course of producing its intended output.

## 4. The Self-Undermining Contradiction

With the case studies of artificial intelligence and misinformation — Sections 3 and 4 — inhand, the contradiction can now be stated in its sharpest form. Capitalism requires a functioning epistemic commons in order to coordinate expectations, pricing, contracting, regulation, and political legitimacy. The frontier technologies through which contemporary capitalist accumulation occurs — AI-mediated inference and platform-mediated communication — generate returns that are maximized by enclosing and degrading those commons. The social accumulation of highly valued creative capital — the means of cultural production — in the hands of powerful capitalistsactually erodes the structural conditions of its own reproduction, through the exclusion from power of ideological diversity based on bespoke knowledge gained through quotidian personal experience.. This is the self-undermining contradiction that names the regime. It is first political and cultural, but it is also structural. It arises from the alignment of profit incentives with informational volatility, an alignment that becomes more pronounced as shifts in ideology and the personalization of the informational environment are together commodified.

This contradiction is distinct from the classical contradictions it resembles. Marx's tendency of the rate of profit to fall concerns the internal dynamics of surplus-value extraction. Polanyi's double movement concerns the social destruction wrought by commodifying land, labour, and money. The substrate here — collective inferential capacity — is categorically different: it is not a prior material or social resource that markets transform into a commodity, but the condition of possibility of market coordination itself. Degrading labour produces immiseration; degrading land produces ecological crisis; degrading money produces financial panic. Degrading the epistemic commons produces a dissolution of the shared interpretive and epistemic ground on which any of these crises can be recognised, contested, and responded to at all.

Kohei Saito has argued that capital is structurally indifferent to the natural conditions of its own existence and is therefore unable to internalise the value of those conditions (Saito, 2023). The same applies to the epistemic case. Capital is indifferent to the epistemic properties of the outputs it produces, so long as those outputs generate engagement, adoption, or throughput. The difference between an inferential output that augments the commons and one that degrades it does not register at the level of the valorisation process; it registers only downstream, in the degradation of the social and relational character of capitalism itself.

Habermas's account of democratic legitimacy offers the complementary normative diagnosis. Democratic orders require a public sphere capable of generating communicatively validated consensus, and such a sphere presupposes discursive conditions that cannot be sustained when communication is systematically colonised by strategic action (Habermas, 1996). A political order whose communicative substrate has been colonised by strategic action does not merely produce worse deliberation; it loses access to deliberation as a political technology, and with it the means by which the political system ordinarily learns what it has gotten wrong.

Democratic systems presuppose a minimally shared environment for building trust — agreement on facts, procedures, and standards of evidence — within which disagreement can be productively managed. When belief formation is privatised and optimised for

engagement rather than truth, democratic deliberation becomes structurally vulnerable to manipulation. Political conflict shifts from distributive struggles over material outcomes to epistemic struggles over the nature of reality itself, making compromise increasingly difficult. The economic consequences are equally significant: instability raises uncertainty and weakens the informational signals that guide investment, so that short-term gains for some firms come at the cost of long-run growth and innovation.

Two objections deserve a brief response. The first objection is that market incentives themselves will resolve the contradiction between fundamental social and cultural value creation, on the one hand, and intangible capital accumulation without meaningful economic purpose on the other, as firms whose outputs are disconnected from physical assets or products lose credibility and revenue. This misunderstands the structure of the externality: an individual platform's engagement-optimised feed imposes epistemic costs diffusely while capturing attention-based revenue privately, with no mechanism to reprice the diffuse cost. The structure is isomorphic to classical commons dilemmas, in which private optimisation produces collective under-provision (Olson, 1965; North, 1990). Moreover, reputational sanction depends on the prior existence of a shared evidentiary standard against which claims can be adjudicated. Under fragmentation, that sanction becomes structurally unavailable.

The framework yields two substantive predictions. First, regulatory interventions focused on the content of misinformation will be less effective than interventions focused on the underlying infrastructure of digital media and commoditized intelligence — the governance of training data, model access, ranking systems, and compute — which serves to enable the pervasive dissemination of misinformation online. If, across jurisdictions, content-level interventions systematically succeed where infrastructure-level interventions fail, the thesis is wrong. Second, conventional measures of successful capitalist societies — accelerating productivity growth, strong profit margins, increasingly valuable market capitalisation — may rise even as the systematic capacity for digital coordination and consensus declines. If these two measures move together, the thesis is wrong; if they diverge in the predicted direction, my argument, as it stands, would seem to be supported.

## 5. Class, State, and the Reconfiguration of Power at the Epistemic Frontier

Two further implications follow from the thesis. The first concerns class. Within the regime described here, the decisive means of production is inferential capacity, and the decisive class relation is the one structured by access to and control over that capacity. Those who own the (increasingly blended) physical and intellectual infrastructure of inference — factories for chips and compute, repositories of training data and model weights, experts who understand ranking systems, and CEOs who run and build hyperscaling platforms — exercise a structural power not fully captured by either ownership of physical capital or command of ordinary expertise. On the other side stand epistemic subjects: those whose cognitive and affective activity supplies the training data, whose behaviour is targeted for inferential intervention, and whose interpretive environments are constituted by outputs over which they exercise no governance. Inequality of inference compounds inequality of income and wealth, because the inferential tools most effective at preserving and growing wealth are themselves the most expensively governed and the most tightly held.

The second implication concerns the state. The state is now also the principal institution with the scale and authority to sustain the epistemic commons against enclosure. A state that is unable to effectively and transparently audit algorithmic systems, counter strategic misinformation, articulate credible public narratives, or produce authoritative statistical knowledge under conditions of synthetic-content saturation is not merely a weaker state. It is a state whose other functions — fiscal, regulatory, adjudicative — are progressively hollowed out, because all of those functions depend on the shared evidentiary substrate that the epistemic commons supplies. The comparative political economy of how different institutional variants respond is an important

agenda, but the structural diagnosis is prior to it: across variants, the counter-movement must do its work under conditions in which the shared evidentiary ground it needs is itself the object under attack.

The thesis does not entail that AI is, in any essentialist sense, opposed to the epistemic commons. Under different institutional arrangements — public training data, open model access, regulated ranking, compute governed as a quasi-utility — the same inferential technologies could be deployed in ways that augment rather than enclose the commons. The claim is structural, not technological: it concerns the accumulation dynamic produced when these technologies are governed by private, engagement-driven incentives under conditions of extreme concentration. Changing the technology without changing the governance will not resolve the contradiction; changing the governance without abandoning the technology can.

**Conclusion: The Empirical Signature of Epistemic Frontier Capitalism**

Capitalism's historical resilience has rested on its capacity to harness technological innovation while preserving sufficient order to coordinate investment, promote global governance, and maintain institutional legitimacy. From the rise of professional expertise to the institutionalization of scientific authority and bureaucratic rule, the development of capitalism has depended on shared systems of knowledge that transform uncertainty into valued expectations for economic continuity. Epistemic frontier capitalism represents a new phase in which this balance becomes increasingly fragile. AI industrializes the production of inference, converting prediction and judgment into scalable, proprietary assets. Platform-based economies monetize algorithms that define how individuals form beliefs, embedding profit incentives directly into the architecture of our digital information environments. Within this configuration, misinformation emerges not as an aberration but as a socially corrosive accumulation strategy in a system that rewards epistemic disruption.

The central contradiction is that accumulation increasingly depends on destabilising the informational foundations upon which capitalism itself relies. Markets, states, and firms require relatively stable expectations to coordinate long-term investment and collective action, yet the same technologies that enhance predictive power also enable the strategic manipulation of belief, the fragmentation of common knowledge, and the erosion of institutional credibility. The contradiction is a structural feature of the regime and will intensify as inferential technology becomes more capable, more concentrated, and more deeply integrated into public life. Whether epistemic frontier capitalism evolves toward renewed stability or deeper fragmentation will depend on whether inferential capacity can be governed as a quasi-public good, or whether epistemic power remains concentrated and democratic institutions struggle to adapt.

Classical and contemporary theories alike have often treated epistemic stability as a background condition rather than a contested terrain. The accumulation of epistemic power and the degradation of the epistemic commons are two faces of the same process. The question is no longer how to correct a market failure in the provision of "good information." The question is whether the political-economic order can sustain the conditions of its own coordination when the most profitable frontier of capital is the privatisation and destabilisation of collective sense-making. Answering it is, as Polanyi saw in a different crisis, ultimately a matter of whether society can re-embed a market that has begun to consume the substrate that makes society possible. Epistemic frontier capitalism poses a question that is at once economic, political, and philosophical: can a system that profits from epistemic disruption sustain the social order it requires to function? I have argued that unless the epistemic commons is reconstituted as an ecosystem for the development of collective and creative human livelihoods, unfettered technological capitalism will continue to accelerate in consuming the very ground on which our civilizations' moral and cultural foundations lie, for now, unprotected. It is our duty to continually seek out and defend these ultimate human values.

**References**


Acemoglu, D. and Johnson, S. (2023). Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity. New York: PublicAffairs.

Allcott, H. and Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2), pp. 211–236.

Alvarado, R. (2023). AI as an Epistemic Technology. *Science and Engineering Ethics*, 29(5), article 32. https://doi.org/10.1007/s11948-023-00451-3.

Arrow, K.J. (1974). The Limits of Organization. New York: W. W. Norton.

Beckert, J. (2016). Imagined Futures: Fictional Expectations and Capitalist Dynamics. Cambridge, MA: Harvard University Press.

Benkler, Y., Faris, R. and Roberts, H. (2018). Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics. Oxford: Oxford University Press.

Bresnahan, T.F. and Trajtenberg, M. (1995). General Purpose Technologies: Engines of Growth? *Journal of Econometrics*, 65(1), pp. 83–108.

Fritts, M. and Cabrera, F. (2022). Online Misinformation and 'Phantom Patterns': Epistemic Exploitation in the Era of Big Data. *Southern Journal of Philosophy*, 60(1), pp. 57–87. https://doi.org/10.1111/sjp.12451.

Habermas, J. (1996). Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy. Translated by W. Rehg. Cambridge, MA: MIT Press.

Hayek, F.A. (1945). The Use of Knowledge in Society. *American Economic Review*, 35(4), pp. 519–530.

Jessop, Bob. 2007. "Knowledge as a Fictitious Commodity: Insights and Limits of a Polanyian Perspective." Reading Karl Polanyi for the Twenty-First Century: Market Economy as a Political Project, edited by Ayşe Buğra and Kaan Ağartan, 115–133. Basingstoke: Palgrave Macmillan.

Jones, C.I. and Tonetti, C. (2020). Nonrivalry and the Economics of Data. *American Economic Review*, 110(9), pp. 2819–2858.

Jungherr, A. (2023). Artificial Intelligence and Democracy: a Conceptual Framework. *Social Media + Society*, 9(3), pp. 1–14. https://doi.org/10.1177/20563051231186353.

Kasy, M. (2025). The Means of Prediction: How AI Really Works (and Who Benefits). Chicago: University of Chicago Press.

Kay, J., Kasirzadeh, A. and Mohamed, S. (2024). Epistemic Injustice in Generative AI. Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society. New York: ACM, pp. 684–697. https://doi.org/10.5555/3716662.3716722.

Keynes, J.M. (1936). The General Theory of Employment, Interest and Money. London: Macmillan.

Kwok, C. (2025). Epistemic Dispossession in Platform Capitalism: Insights from Iris Marion Young's Five Faces of Oppression. *Social Epistemology*, published online 3 December 2025. https://doi.org/10.1080/02691728.2025.2585445.

Lachmann, L.M. (1986). The Market as an Economic Process. Oxford: Basil Blackwell.

Marx, K. (1973). Grundrisse: Foundations of the Critique of Political Economy (M. Nicolaus, Trans.). Penguin Books.

Milano, S. and Prunkl, C. (2025). Algorithmic Profiling as a Source of Hermeneutical Injustice. Philosophical Studies, 182(1), pp. 185–203. https://doi.org/10.1007/s11098-023-02095-2.

North, D.C. (1990). Institutions, Institutional Change and Economic Performance. Cambridge: Cambridge University Press.

Olson, M. (1965). The Logic of Collective Action. Cambridge, MA: Harvard University Press.

Oreskes, N. and Conway, E.M. (2010). Merchants of Doubt. New York: Bloomsbury Press.

Ostrom, E. (1990). Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge: Cambridge University Press.

Pasquale, F. (2015). The Black Box Society: The Secret Algorithms That Control Money and Information. Cambridge, MA: Harvard University Press.

Polanyi, K. ([1944] 2001). The Great Transformation: The Political and Economic Origins of Our Time. Boston: Beacon Press.

Popper, K. R. (1959). The Logic of Scientific Discovery. Hutchinson.

Saito, K. (2023). Marx in the Anthropocene: Towards the Idea of Degrowth Communism. Cambridge: Cambridge University Press.

Shackle, G.L.S. (1972). Epistemics and Economics: A Critique of Economic Doctrines. Cambridge: Cambridge University Press.

Stiglitz, J.E. and Kosenko, A. (2024a). The Economics of Information in a World of Disinformation: A Survey. Part 1: Indirect Communication. NBER Working Paper No. 32049. Cambridge, MA: National Bureau of Economic Research.

Stiglitz, J.E. and Kosenko, A. (2024b). The Economics of Information in a World of Disinformation: A Survey. Part 2: Direct Communication. NBER Working Paper No. 32050. Cambridge, MA: National Bureau of Economic Research.

Vosoughi, S., Roy, D. and Aral, S. (2018). The Spread of True and False News Online. *Science*, 359(6380), pp. 1146–1151.

Weber, M. (1978). Economy and Society. Edited by G. Roth and C. Wittich. Berkeley: University of California Press.

Williamson, O.E. (1985). The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting. New York: Free Press.

Zuboff, S. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New York: PublicAffairs.

# Metacognitive Conversational AI: A Design Framework for Reflective Well-Being

*Tianyi Yu*
*University College London*



Conversational artificial intelligence systems are increasingly embedded in everyday cognitive and relational tasks, while their optimisation for immediacy and solution delivery raises questions about how they shape users' reflective capacities. Drawing on metacognitive theory, this paper proposes a conceptual framework for Metacognitive Conversational AI, which structures interaction design around three mechanisms: Reflection Activation, Structured Scaffolding, and Transfer and Retention. This framework translates metacognitive processes into interface-level design principles intended to preserve users' evaluative participation rather than fully externalise it.

Using design fiction methodology, two illustrative scenarios, grammar correction and interpersonal conflict resolution, demonstrate how modest adjustments to conversational sequencing can shift interaction from direct answer substitution toward structured reflective engagement. The paper argues that such design choices may have implications for regulatory flexibility and sustained well-being. It further examines structural, economic, and ethical tensions, including usability trade-offs, user agency, and cultural variation in conceptions of reflection.

By reframing conversational AI as a cognitive environment rather than merely a problem-solving tool, this work contributes a theoretically grounded design framework for integrating metacognitive support into human–AI interaction.

**Keywords**: Conversational AI, Human Well-Being, Chatbot, Design Fiction, Human–AI Interaction

## Introduction

Conversational artificial intelligence systems are increasingly embedded in everyday cognitive and relational tasks. From drafting essays to offering guidance during interpersonal conflict, large language models function as general-purpose cognitive assistants (Biancofiore et al, 2025). Their appeal lies in immediacy: users receive fluent, contextually relevant responses within seconds (Callister et al, 2025). While this optimisation for speed and helpfulness represents a significant technological achievement, it raises an important question: how does interaction with systems that perform cognitive work on users' behalf shape the reflective capacities associated with sustained psychological well-being?

Psychological well-being has been consistently linked to adaptive coping, effective problem-solving, and emotional regulation (Hernandez et al, 2018; Tan et al, 2019). These capacities require more than correct outcomes; they depend on individuals' ability to monitor their thinking, evaluate alternatives, and adjust strategies across contexts. Such higher-order monitoring and control processes are collectively described as metacognition (Flavell, 1979; APA Dictionary of Psychology). Importantly, metacognitive skills are strengthened through active engagement rather than passive reception (Santangelo, Cadieux, and Zapata, 2021).

Contemporary conversational AI systems, however, are typically designed to minimise cognitive effort (Tezer, 2025; Gerlich, 2025). In educational contexts, they rewrite text directly; in relational scenarios, they provide ready-made advice. Although these features improve usability and short-term performance, they may reduce opportunities for users to articulate reasoning, examine uncertainty, or engage in strategic self-monitoring. Interaction design, therefore, becomes psychologically consequential: it shapes the extent to which users participate in their own reflective processes.

Specifically, this paper proposes a conceptual framework for Metacognitive Conversational

AI, drawing on Flavell's (1979) tripartite model to translate metacognitive theory into interface-level design principles. It then examines how such an approach intersects with structural and economic incentives that prioritise frictionless interaction and illustrates its implications through two design-fiction scenarios involving learning and interpersonal conflict.

By articulating a theoretically grounded model of reflective AI design, this paper aims to contribute to interdisciplinary discussions in digital well-being, human–AI interaction, and cognitive augmentation. As conversational systems increasingly mediate reasoning and emotional appraisal, designing for metacognition offers a pathway toward AI systems that strengthen, rather than replace, human regulatory agency.

## 1. Conversational AI and Wellbeing

Conversational AI has increasingly been positioned as a tool for supporting psychological well-being, especially since the onset of COVID-19 (Amazu, 2020). Research suggests that AI-based conversational agents can provide mental health support and promote behavioural change (Li et al, 2023; Musikanski et al, 2020). Framed as scalable interventions, such systems aim to address gaps in traditional services by offering immediate responses, emotional validation, and structured guidance.

Within human–computer interaction research, technology-mediated nudging has been identified as one mechanism through which AI systems can influence well-being. Rather than merely providing information, nudges restructure the decision environment so that certain actions become more salient, emotionally compelling, or easier to execute (Hansen and Jespersen, 2013; Caraban et al, 2019). Caraban et al. (2019) identified mechanisms, including reinforcement, emotional framing, and social influence, through which interface design can steer behaviour toward predefined outcomes. Empirical applications, such as conversational programmes encouraging physical activity (Chater et al, 2022) and chatbot-based screening tools for eating disorders (Mejova and Suarez-Lledó, 2020), illustrate how dialogue-based systems can lower barriers to action and guide users toward specific behavioural steps.

Despite these advances, important limitations remain (Gupta et al, 2023; Jiang et al, 2023). Many conversational systems prioritise fluent and immediately actionable output, requiring limited cognitive engagement from users. Recommendations can be accepted without articulating reasoning or evaluating alternatives, potentially reducing opportunities for reflective learning and limiting the durability of behavioural change. Moreover, responses generated from broad training data may reflect generalised patterns rather than deeply contextualised reasoning. While such systems can improve efficiency and short-term outcomes, their contribution to sustained regulatory development remains uncertain.

These limitations reveal a design asymmetry. Existing research often evaluates conversational AI in terms of symptom reduction, behavioural compliance, or user satisfaction, yet comparatively little attention has been given to how interaction structures shape users' cognitive participation. This conceptual gap motivates a shift in emphasis: from optimising solution delivery to structuring metacognitive engagement in AI.

## 2. Theoretical Foundation: What Is Metacognition?

To structure metacognitive engagement in AI, it is important to first clarify what metacognition is. Metacognition refers to the processes through which individuals monitor and regulate their own cognition. In his model of cognitive monitoring, Flavell (1979) proposed that monitoring emerges from the interaction of four elements: metacognitive knowledge, metacognitive experiences, goals (or tasks), and actions (or strategies).

Metacognitive knowledge denotes stored knowledge about oneself as a cognitive agent (e.g., strengths or habitual tendencies), about task demands, and about the conditions under which particular strategies are effective. This knowledge is relatively stable and generalisable across contexts. Metacognitive experiences, by contrast, are conscious cognitive or affective experiences that arise during task engagement. Feelings of uncertainty, confidence, difficulty, or

fluency provide situational feedback about ongoing processing. Goals organise cognitive activity by defining what the individual seeks to achieve, while actions are the cognitive or behavioural operations undertaken in pursuit of those goals.

Although Flavell (1979) did not formally designate "metacognitive regulation" as a separate structural component, subsequent scholarship has used the term to describe the evaluative coordination through which metacognitive knowledge and experience guide strategic adjustment (Wongdaeng, 2022). In this paper, metacognitive regulation refers to this functional process: assessing the alignment between goals, current strategies, and ongoing cognitive states, and modifying behaviour accordingly.

This dynamic coordination is central to psychological functioning. Empirical research indicates that the ability to monitor and recalibrate cognitive and emotional processes is associated with adaptive coping and sustained well-being. Awareness of cognitive and affective states facilitates effective stress regulation (Solem et al, 2015; Drigas and Mitsea, 2021), while evaluative monitoring and goal-directed planning are linked to resilience and subjective well-being (MacLeod et al, 2008; Yazar and Tolan, 2020). The relevance of metacognition to well-being, therefore, lies in regulatory flexibility — the capacity to detect misalignment between goals and strategies and adjust accordingly.

If regulatory flexibility contributes to well-being, then the design of systems that participate in users' cognitive processes becomes consequential. In conversational AI contexts, users typically approach systems with defined goals, while strategic actions may be generated internally (by the user) or externally (by the system). When AI generates solutions without requiring evaluative engagement, aspects of the monitoring cycle may become partially externalised. Although such substitution can enhance efficiency, it may reduce opportunities for strengthening internal reflective capacities over time (Ahmad et al, 2023). Conversely, interaction designs that prompt awareness, encourage evaluation, and invite strategic reconsideration may support — rather than replace — users' regulatory involvement.

Accordingly, the present framework identifies three interrelated processes as primary sites of design intervention in conversational AI: activation of metacognitive knowledge, elicitation of metacognitive experience, and support for metacognitive regulation. While Flavell's original model also distinguishes goals and actions, these elements typically define the task structure itself. In conversational AI contexts, goals are often articulated by users, and actions may be generated either internally or suggested by the system. By contrast, knowledge, experience, and regulation constitute the reflective layer through which goals and actions are evaluated and adjusted (Merkebu et al, 2024). Focusing on these three processes, therefore, clarifies how conversational systems may influence the extent to which users remain active participants in their own cognitive control.

## 3. Conceptual Framework: A Three-Layer Model of Metacognitive Conversational AI

Building on the identification of metacognitive knowledge, metacognitive experience, and metacognitive regulation as primary sites of design intervention, the following task is to specify how conversational AI can engage these processes in practice.

This framework distinguishes three interrelated mechanisms of intervention: *Reflection Activation, Structured Scaffolding, and Transfer and Retention*. These mechanisms do not introduce new theoretical constructs; rather, they translate the previously outlined metacognitive architecture into interaction-level design principles. Their sequence reflects functional dependency rather than arbitrary categorisation. Specifically, metacognitive knowledge and experience must first become explicit before regulatory evaluation can occur, and regulatory engagement must be sustained before reflective insight can consolidate into durable, transferable knowledge. Each mechanism, therefore, builds upon the conditions established by the previous one.

*Reflection Activation* operates at the outset of engagement (Ribeiro et al, 2019) and primarily targets metacognitive knowledge and metacognitive experience. Conventional conversational systems typically deliver immediate, fully formed responses, leaving limited opportunity for users to articulate assumptions, clarify intentions, or recognise uncertainty. Reflection Activation reintroduces evaluative space by prompting users to define their goals, identify points of difficulty, or examine their current understanding before or alongside solution provision. By making these elements explicit, the system activates stored knowledge about the task and surfaces subjective cognitive or affective states that inform subsequent evaluation. Awareness here refers specifically to the conscious articulation of one's understanding, uncertainty, or strategic orientation.

*Structured Scaffolding* operates during task processing (Xun and Land, 2004) and primarily supports metacognitive regulation. Once knowledge and experience are brought into awareness, users must evaluate whether their current strategies align with their goals. Rather than completing this evaluative work on the user's behalf, the system organises reasoning into guided steps. Explanations of underlying principles, presentation of alternative interpretations, and prompts for comparison structure planning, monitoring, and adjustment. In this configuration, strategic modification remains participatory: the user continues to assess alignment between goals, strategies, and cognitive states instead of relying solely on externally generated conclusions.

*Transfer and Retention* operate beyond immediate task resolution (Farr, 2012) and reinforce metacognitive knowledge while strengthening future regulation. Many AI interactions conclude once a solution is delivered, limiting opportunities for abstraction. This mechanism introduces prompts that encourage users to extract generalisable principles from the exchange, such as articulating a rule, identifying a recurring pattern, or anticipating application in similar contexts. Such abstraction supports the consolidation of knowledge and increases the likelihood that regulatory flexibility extends across situations rather than remaining confined to a single instance.

Taken together, these mechanisms structure intervention across the progression of reflective engagement: awareness activation, regulatory support, and consolidation. By distributing metacognitive support across these stages, conversational AI can maintain user involvement in cognitive regulation while remaining compatible with efficiency demands. It should be noted that real-world interactions may not follow clean sequential phases, and reflective engagement may unfold iteratively rather than linearly. The three mechanisms, therefore, describe functional dependencies rather than rigid temporal stages.

## 4. Methodology

### *4.1. Design Fiction*

This study uses design fiction to explore how conversational AI interaction could be redesigned by integrating metacognitive prompts to support and maintain human well-being, using the framework. Design fiction is well-suited to conceptual and design-oriented contributions because it enables speculative but plausible scenarios that extend existing technological practices and make design mechanisms visible for analysis (Dunne and Raby, 2024).

### *4.2. Case Study*

Two cases were selected to illustrate metacognitive prompting across distinct task types. The first case focuses on grammar and spelling correction. It draws on Tabib and Alrabeei (2024) and reflects the increasing use of AI systems as writing assistants, particularly for checking and revising academic work (Tossell et al, 2024). This domain is useful for examining whether conversational AI can do more than provide corrected output. Specifically, whether it can prompt learners to notice errors, reflect on rules, and retain strategies for future writing.

The second case extends beyond rule-governed correction into social–relational reasoning. Conversational AI is less frequently used to address social dynamics because interpersonal conflict is highly context-dependent and requires personalised interpretation rather

than fixed rules (Subramonyam et al, 2023). This case, therefore, tests whether metacognitive prompting can support reflection and planning in situations where "correct answers" are not well-defined.

### *4.3. Scenarios*

Two scenarios were created to illustrate how conversational AI responses could be redesigned in line with the framework proposed above. Both scenarios were inspired by everyday situations that can be challenging for metacognitive regulation and were drafted using a conversational AI system (OpenAI, 2023), which can quickly simulate plausible interaction patterns in natural language (Nazi and Wang, 2023). It should be noted that although the illustrative scenarios were primarily created using ChatGPT 3.5 (OpenAI, 2023), the analysis does not depend on the specific capabilities of that model. The proposed framework concerns interaction structure rather than underlying model architecture. As newer conversational models continue to improve in fluency and contextual reasoning, the question of how interaction design shapes users' metacognitive engagement remains equally relevant. The principles outlined here therefore generalise to contemporary and future large language models that function as conversational cognitive assistants.

Each scenario is presented in two versions. Version 1 represents a baseline interaction pattern in which the system responds in a conventional direct-answer style. Version 2 represents a framework-guided interaction in which prompts are introduced to instantiate the three mechanisms: Reflection Activation (prompting users to articulate assumptions, goals, uncertainty, or feelings before or alongside advice), Structured Scaffolding (organising evaluation through stepwise reasoning, comparison, and guided self-evaluation), and Transfer and Retention (encouraging generalisation by prompting users to summarise principles or anticipate future application).

*Scenario 1: Grammar and Spelling Correction*

This scenario illustrates how conversational AI can be used to correct grammar and spelling errors in academic writing. A short paragraph of approximately 40 words was constructed, containing three spelling mistakes and two grammatical errors. The paragraph was submitted to conversational AI for correction.

As shown in Figure 1, in the baseline interaction, the system provides a fully corrected version of the paragraph immediately. While this format efficiently improves the text, it may reduce opportunities for the user to identify error locations or understand the grammatical principles involved. Users may focus on copying the corrected text rather than reflecting on the underlying rule. Without engaging in error diagnosis, the same mistakes may recur in subsequent writing (Covey, 1992).

To operationalise the proposed metacognitive framework, the interaction was restructured (Figure 2). Rather than presenting the corrected paragraph directly, the system first highlights specific areas of concern and prompts the user to reconsider them. For example, the user may be asked to double-check the spelling of a particular word or reflect on whether the verb tense is consistent. This step corresponds to Reflection Activation, as it brings relevant grammatical knowledge and uncertainty into awareness.

Following this, the system provides brief explanations of the grammatical principles involved before presenting the corrected form. This constitutes Structured Scaffolding, as the reasoning process is organised in a way that supports monitoring and evaluation rather than replacing them. The user is guided to understand why the correction is necessary.

Finally, the system may prompt the user to summarise the rule or identify similar errors elsewhere in the text. This stage supports Transfer and Retention, encouraging abstraction beyond the immediate instance and strengthening the likelihood that the user will apply the principle independently in future writing.

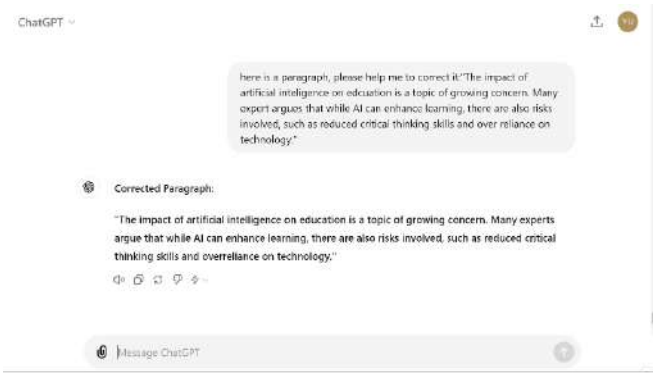



**Figure 1:** The screenshot of how conversational AI helps to correct the sentence.

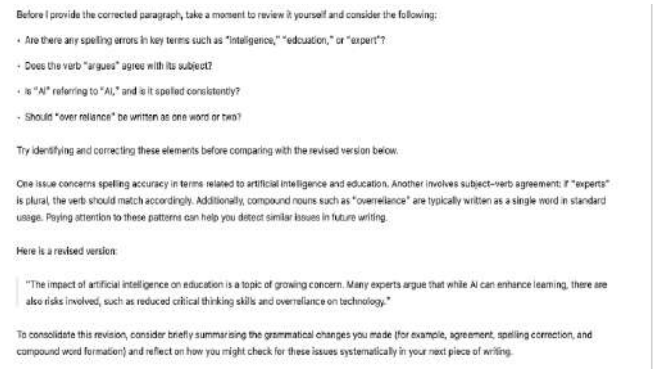



**Figure 2:** The screenshot of how conversational AI with metacognition prompts helps to correct the sentence.

*Scenario 2: Interpersonal Conflict Resolution*

The second scenario examines how conversational AI may be used to address interpersonal conflict. A fictional situation was constructed in which a user feels frustrated because a friend repeatedly arrives late to planned meetings. Although the user's friend promised to send a message in advance if she were late next time, the user still felt the problem was unsolved, and there was tension between them. The user submits this scenario to conversational AI and requests advice.

In the baseline interaction (Figure 3), the system provides general recommendations, such as communicating expectations clearly and planning in advance. Although these suggestions are reasonable, they remain broad and may not address the user's specific emotional experience or relational context. Because the advice is generated without eliciting deeper reflection, it may not prevent the recurrence of the same issue.

In the revised interaction (Figure 4), the response structure is modified in line with the framework. Before offering advice, the system asks a series of clarifying questions about the relationship, the emotional significance of punctuality, and prior communication attempts. These prompts activate metacognitive knowledge and experience, corresponding to Reflection Activation. By encouraging articulation of feelings and assumptions, the system creates evaluative space before solution delivery.

The questions also guide the user through structured reflection on patterns of behaviour and possible misalignment in expectations. For instance, asking how frequently lateness occurs or whether punctuality has been discussed previously supports self-evaluation and regulatory monitoring, aligning with Structured Scaffolding. Asking how the user typically handles unexpected changes introduces an element of planning, further engaging regulatory processes.

After this reflective stage, more tailored suggestions are provided. The user is also encouraged to consider how similar conflicts might be handled in the future, which corresponds to Transfer and Retention. By prompting forward-looking strategy formation, the system supports regulatory flexibility beyond the current incident.

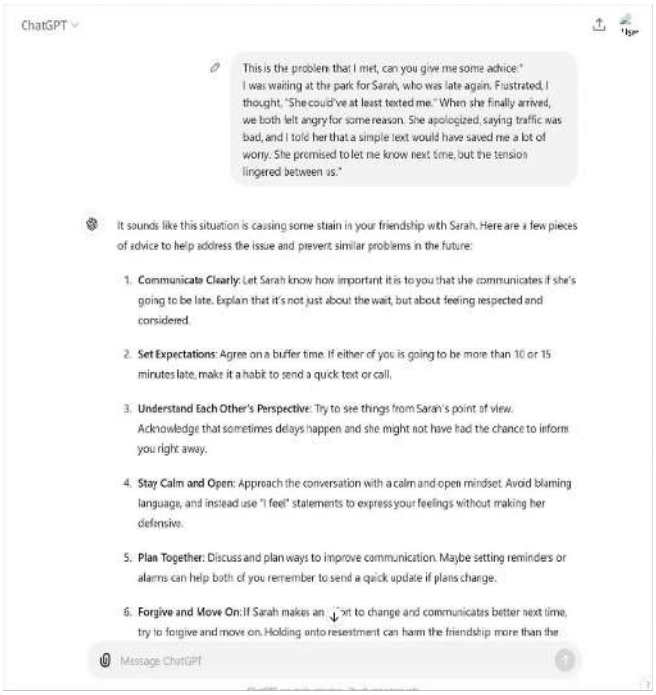



**Figure 3:** The screenshot of how conversational AI with metacognition prompts answers the

requirements of giving a recommended solution to an argument between friends initially.

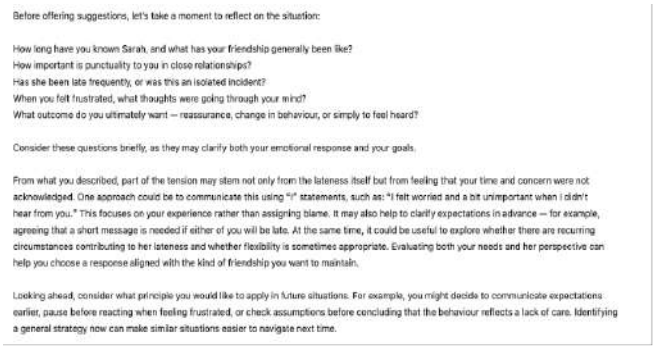



**Figure 4:** The screenshot of how conversational AI answers the requirements of giving a recommended solution to an argument between friends after training.

## 5. Discussion

Across both scenarios, restructuring conversational responses shifted the interaction from answer substitution toward reflective engagement in ways that may support well-being. In the writing context, prompting reconsideration before correction preserved users' involvement in monitoring and evaluation. Such engagement aligns with research showing that retrieval and self-explanation processes are associated with stronger consolidation and reduced repetition of error (Selmeczy, Kazemi and Ghetti, 2024), and that self-monitoring supports problem-solving development (Hewitt, Chreim and Forster, 2016). The scenario, therefore, illustrates how conversational design can scaffold regulatory participation rather than externalise it.

In the interpersonal context, introducing reflective questioning before advice provision similarly altered the sequencing of cognition. By foregrounding appraisal prior to prescription, the interaction theoretically reduces the likelihood of affect-driven distortion, consistent with evidence that strong emotions can bias judgment (Clore and Huntsinger, 2007). Encouraging users to examine relational patterns and anticipate future strategy supports evaluative monitoring and forward planning.

Together, the scenarios suggest that modest changes in interaction architecture can redistribute cognitive work to further maintain well-being, without fully displacing efficiency. Metacognitively structured conversational AI does not eliminate immediacy; rather, it introduces calibrated reflection alongside it. If regulatory flexibility contributes to sustained well-being (Tan et al, 2019; Yazar and Tolan, 2020), then preserving evaluative participation within AI-mediated interaction may have implications for long-term problem-solving capacity (Hartman, 2001).

However, this rebalancing introduces a structural trade-off. Additional reflective steps may increase cognitive load and reduce perceived efficiency. Users seeking rapid answers may experience frustration when confronted with extended questioning (Jeong, 2012). Conversational systems operate within markets that have converged on immediacy for a reason: speed is valued. A design that deliberately slows interaction to promote reflection may therefore conflict with prevailing engagement metrics.

These economic realities raise broader structural questions. Under what governance or institutional conditions would reflective scaffolding be incentivised? Educational technologies, subscription-based learning platforms, and public-sector mental health applications may prioritise long-term skill development over immediate engagement (Chumdoem, Boonlue and Narmluk, 2024; Damij and Bhattacharya, 2022). In contrast, in competitive consumer markets, design incentives typically reward immediacy and frictionless interaction, making reflective friction commercially unattractive (Gnewuch et al, 2018). Therefore, the adoption of metacognitive conversational AI is not solely a cognitive or technical issue but also an institutional one. This also further demonstrates that it is hard to find a balance between enhancing metacognition and the usability of the chatbot.

Relatedly, user agency warrants careful consideration. While metacognitive prompting may support adaptive outcomes, it cannot be assumed that users universally desire deeper engagement. Introducing reflective friction for users' "benefit" risks paternalism if imposed without transparency or choice. Adaptive or optional modes of interaction may provide a more ethically defensible pathway, allowing users to select scaffolded engagement in

contexts involving learning or emotional complexity while retaining rapid-response functionality elsewhere.

Cultural context further complicates the assumption that metacognition is uniformly beneficial across populations. Much of the metacognitive literature underpinning this framework is grounded in Western psychological traditions. In collectivist or relationally oriented contexts, regulatory processes may be more distributed across social networks rather than centred on individual introspection (Wang and Liu, 2010). Conversational systems deployed globally should therefore remain sensitive to variation in communication norms, authority structures, and conceptions of well-being.

Moreover, several other limitations merit consideration. One potential limitation is that adding metacognition prompts might disrupt the natural flow of the conversation. This can be explained by the fact that the conversational AI is not that familiar with human natural language processing and human cognitive processes (Theophilou et al, 2023). This might cause the users to feel the interaction is unnatural and forced, which thus affects the usability of the chatbot.

In addition, effective metacognitive prompting requires personalisation based on the user's background, preferences and cognitive styles (Odilinye, 2019). This means that the conversational AI might need extensive data collection and analysis to provide personalised and useful suggestions for the users. However, this might raise the user's concerns about data security and privacy (Zhu et al, 2014).

As a result of the limitations, some future directions are provided. First of all, future research should empirically evaluate the cognitive and behavioural consequences of metacognitive conversational design. Experimental comparisons between direct-answer interfaces and scaffolded interfaces could assess differences in retention, transfer, and independent problem-solving. Longitudinal designs would be particularly valuable in determining whether repeated exposure to structured reflection enhances regulatory flexibility over time.

At the same time, identifying boundary conditions is essential. Excessive prompting may increase cognitive load or reduce engagement, and individual differences in motivation, prior knowledge, or cultural orientation may moderate effectiveness. Determining optimal levels and timing of reflective scaffolding will therefore be critical.

Implementation challenges also require attention. Effective calibration depends on conversational systems' ability to detect when reflective prompting is appropriate and what language should be used. This necessitates advances in contextual sensitivity rather than merely increasing prompt frequency. In addition, greater emphasis on data protection and privacy safeguards will be necessary if personalisation is used to support metacognitive engagement.

**Conclusion**

Enhancing and sustaining human well-being remains an important interdisciplinary concern, particularly as conversational AI becomes increasingly embedded in learning and relational contexts. Although conversational systems are often positioned as tools for supporting well-being, they are typically optimised for immediacy and solution delivery. As a result, users may receive efficient answers without engaging in the reflective processes that support long-term regulatory flexibility.

This paper proposed a conceptual framework for Metacognitive Conversational AI, suggesting that interaction design can be structured to activate awareness, support evaluative regulation, and encourage transfer beyond immediate tasks. Through two illustrative scenarios, writing correction and interpersonal conflict, the analysis demonstrated how modest adjustments to conversational sequencing can preserve users' participation in monitoring and strategy formation, rather than fully externalising these processes, which may have implications for sustained well-being.

The contribution of this work is conceptual rather than empirical. It argues that conversational AI should be understood not

only as a problem-solving tool but as a cognitive environment that shapes reflective engagement. Future work should examine how metacognitive prompting can be calibrated to balance usability and reflective depth, while also addressing issues of privacy, contextual sensitivity, and cultural variation. By integrating metacognitive principles into conversational design, AI systems may better support sustained well-being rather than merely providing immediate solutions.

**Acknowledgement**
Dr Aisha Sobey provided invaluable feedback and suggestions on this paper.

**References**

Ahmad, S.F., Han, H., Alam, M.M., Rehmat, M., Irshad, M., Arraño-Muñoz, M. and Ariza-Montes, A., 2023. Impact of Artificial Intelligence on Human Loss in Decision Making, Laziness and Safety in Education. *Humanities and Social Sciences Communications,* 10(1), p.311.

Amazu, A.A., 2020. Well-Being within the Digital Transformation Age among University Students: An Exploratory Research (Master's thesis, University of Twente).

APA Dictionary of Psychology (no date b). https://dictionary.apa.org/metacognition.

Biancofiore, G.M., Di Palma, D., Pomo, C., Narducci, F. and Di Noia, T., 2025. Conversational User Interfaces and Agents. In Human-Centered AI: An Illustrated Scientific Quest (pp. 399-438). Cham: Springer Nature Switzerland.

Callister, O., Ardent, S., Quinlan, T. and Elowen, D., 2025. Conversational AI and Voice Interfaces: The Next Frontier of Seamless Customer Interaction.

Caraban, A., Karapanos, E., Gonçalves, D. and Campos, P., 2019, May. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. In Proceedings of the 2019 CHI conference on human factors in computing systems (pp. 1-15).

Chater, A.M., Schulz, J., Jones, A., Burke, A., Carr, S., Kukucska, D., Troop, N., Trivedi, D. and Howlett, N., 2022. Outcome evaluation of Active Herts: A Community-Based Physical Activity Programme for Inactive Adults at Risk of Cardiovascular Disease and/or Low Mental Wellbeing. *Frontiers in Public Health*, 10, p.903109.

Chumdoem, D., Boonlue, S. and Narmluk, M., 2024. Investigating the Components of Chat-Bot-Driven Micro-Learning: Enhancing Problem-Solving Competency in Modern Teaching.

Clore, G.L. and Huntsinger, J.R., 2007. How Emotions Inform Judgment and Regulate thought. *Trends in Cognitive Sciences*, 11(9), pp.393-399.

Covey, S.R., 1992. Principle Centered Leadership. Simon and Schuster.

Damij, N. and Bhattacharya, S., 2022, April. The Role of AI Chatbots in Mental Health Related Public Services in a (Post) Pandemic World: A Review and Future Research Agenda. In 2022 IEEE Technology and Engineering Management Conference (TEMSCON EUROPE) (pp. 152-159). IEEE.

Dawson, T.L., 2008. Metacognition and Learning in Adulthood. Prepared in Response to Tasking from ODNI/CHCO/IC Leadership Development Office, Developmental Testing Service, LLC.

Drigas, A. and Mitsea, E., 2021. Metacognition, Stress-Relaxation Balance & Related Hormones. Int. J. *Recent Contributions Eng. Sci. IT*, 9(1), pp.4-16.

Dunne, A. and Raby, F., 2024. Speculative Everything, With a new preface by the authors: *Design, Fiction, and Social Dreaming*. MIT press.

Farr, M. J. (2012). The Long-Term Retention of Knowledge and Skills: A Cognitive and Instructional Perspective.

Flavell, J.H., 1979. Metacognition and Cognitive Monitoring: A New Area of Cognitive–Developmental Inquiry. *American Psychologist*, 34(10), p.906.

Gnewuch, U., Morana, S., Adam, M., & Maedche, A. (2018). Faster is not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction.

Gerlich, M., 2025. AI tools in society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies*, 15(1), p.6.

Gupta, D., Singhal, A., Sharma, S., Hasan, A. and Raghuwanshi, S., 2023. Humans' Emotional and Mental Well-Being under the Influence of Artificial Intelligence. *Journal for ReAttach Therapy and Developmental Diversities*, 6(6s), pp.184-197.

Hansen, P.G. and Jespersen, A.M., 2013. Nudge and the Manipulation of Choice: A Framework for the Responsible Use of the Nudge Approach to Behaviour Change in Public Policy. *European Journal of Risk Regulation*, 4(1), pp.3-28.

Hartman, H.J., 2001. Developing Students' Metacognitive Knowledge and Skills. Metacognition in Learning and Instruction: Theory, Research and Practice, pp.33-68.

Hernandez, R., Bassett, S.M., Boughton, S.W., Schuette, S.A., Shiu, E.W. and Moskowitz, J.T., 2018. Psychological Well-Being and Physical Health: Associations, Mechanisms, and Future Directions. *Emotion Review*, 10(1), pp.18-29.

Hewitt, T., Chreim, S. and Forster, A., 2016. Double Checking: A Second Look. *Journal of Evaluation in Clinical Practice*, 22(2), pp.267-274.

Jiang, J., Karran, A.J., Coursaris, C.K., Léger, P.M. and Beringer, J., 2023. A Situation Awareness Perspective on Human-AI Interaction: Tensions and Opportunities. *International Journal of Human–Computer Interaction*, 39(9), pp.1789-1806.

Jeong, H., 2012. A Comparison of the Influence of Electronic Books and Paper Books on Reading Comprehension, Eye Fatigue, and Perception. *The Electronic Library*, 30(3), pp.390-408.

Kim, J. Y., & Lim, K. Y. (2019). Promoting Learning in Online, Ill-Structured Problem Solving: The Effects of Scaffolding Type and Metacognition Level. *Computers & Education*, 138, 116-129.

Li, H., Zhang, R., Lee, Y.C., Kraut, R.E. and Mohr, D.C., 2023. Systematic Review and Meta-Analysis of AI-Based Conversational Agents for Promoting Mental Health and Well-Being. *NPJ Digital Medicine*, 6(1), p.236.

MacLeod, A.K., Coates, E. and Hetherton, J., 2008. Increasing Well-Being through Teaching Goal-Setting and Planning Skills: Results of a Brief Intervention. *Journal of Happiness Studies*, 9(2), pp.185-196.

Mejova, Y. and Suarez-Lledó, V., 2020, October. Impact of Online Health Awareness Campaign: Case of National Eating Disorders Association. In International Conference on Social Informatics (pp. 192-205). Cham: Springer International Publishing.

Merkebu, J., Veen, M., Hosseini, S. and Varpio, L., 2024. The Case for Metacognitive Reflection: A Theory Integrative Review with Implications For Medical Education. *Advances in Health Sciences Education*, 29(4), pp.1481-1500.

Musikanski, L., Rakova, B., Bradbury, J., Phillips, R. and Manson, M., 2020. Artificial Intelligence and Community Well-Being: A Proposal for an Emerging Area of Research. *International Journal of Community Well-Being*, 3(1), pp.39-55.

Nazir, A. and Wang, Z., 2023. A Comprehensive Survey of ChatGPT: Advancements, Applications, Prospects, and Challenges. *Meta-Radiology*, p.100022.

Odilinye, L., 2019. Personalized Recommender System for Technology Enhanced Learning Using Learners' Metacognitive Activities.

OpenAI. (2023). ChatGPT (July 17 version) [Large language model]. Retrieved from https://www.openai.com/

Qudrat-Ullah, H., 2025. Building Self-Awareness. In Mastering Decision-Making in Business and Personal Life: An Interdisciplinary Perspective on Making Better Choices (pp. 109-146). Cham: Springer Nature Switzerland.

Ribeiro, L. M. C., Mamede, S., de Brito, E. M., Moura, A. S., de Faria, R. M. D., & Schmidt, H. G. (2019). Effects of Deliberate Reflection on Students' Engagement in Learning and Learning Outcomes. *Medical Education*, 53(4), 390-397.

Santangelo, J., Cadieux, M. and Zapata, S., 2021. Developing Student Metacognitive Skills Using Active Learning with Embedded Metacognition Instruction. Journal of STEM Education: Innovations and Research, 22(2).

Selmeczy, D., Kazemi, A. and Ghetti, S., 2024. Seeking Versus Receiving Help: How Children Integrate Suggestions in Memory Decisions. Child Development, 95(2), pp.515-529.

Solem, S., Thunes, S.S., Hjemdal, O., Hagen, R. and Wells, A., 2015. A Metacognitive Perspective on Mindfulness: An Empirical Investigation. *BMC Psychology*, 3, pp.1-10.

Subramonyam, H., Pea, R., Pondoc, C.L., Agrawala, M. and Seifert, C., 2023. Bridging the Gulf of Envisioning: Cognitive Design Challenges in LLM Interfaces. arXiv preprint arXiv:2309.14459.

Tabib, F.M. and Alrabeei, M.M., 2024. Can Guided ChatGPT Use Enhance Students' Cognitive and Metacognitive Skills?. In Artificial Intelligence in Education: The Power and Dangers of ChatGPT in the Classroom (pp. 143-154). Cham: Springer Nature Switzerland.

Tan, C.S., Tan, S.A., Mohd Hashim, I.H., Lee, M.N., Ong, A.W.H. and Yaacob, S.N.B., 2019. Problem-Solving Ability and Stress Mediate the Relationship between Creativity And Happiness. *Creativity Research Journal*, 31(1), pp.15-25.

Tezer, M., 2025. Metacognitive Engagement In AI-Supported Learning: Frameworks, Challenges, and Transformations.

Theophilou, E., Koyutürk, C., Yavari, M., Bursic, S., Donabauer, G., Telari, A., Testa, A., Boiano, R., Hernandez-Leo, D., Ruskov, M. and Taibi, D., 2023, November. Learning to Prompt in the Classroom to Understand AI Limits: A Pilot Study. In *International Conference of the Italian Association For Artificial Intelligence* (pp. 481-496). Cham: Springer Nature Switzerland.

Tossell, C.C., Tenhundfeld, N.L., Momen, A., Cooley, K. and de Visser, E.J., 2024. Student Perceptions of Chatgpt Use in a College Essay Assignment: Implications for Learning, Grading, and Trust in Artificial Intelligence. *IEEE Transactions on Learning Technologies*.

Xun, G. E., & Land, S. M. (2004). A Conceptual Framework for Scaffolding III-Structured Problem-Solving Processes Using Question Prompts and Peer Interactions. *Educational Technology Research and Development*, 52(2), 5-22.

Wang, G., & Liu, Z. B. (2010). What Collective? Collectivism and Relationalism from a Chinese Perspective. Chinese Journal of Communication, 3(1), 42-63.

Yazar, R. and Tolan, Ö., 2020. İnvestigation of the Relationships between Metacognitive Functions and Subjective Well-Being and Depression, Anxiety and Stress Levels in Adult Individuals. *Research on Education and Psychology*, 4(2), pp.172-193.

Zhu, H., Xiong, H., Ge, Y. and Chen, E., 2014, August. Mobile App Recommendations with Security and Privacy Awareness. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 951-960).

CJAI
CAMBRIDGE
JOURNAL OF
ARTIFICIAL
INTELLIGENCE