Balance Human Agency & AI Assistance in the Tussle for the "Right" to *

Anonymous Authors[1]

Abstract

AI increasingly mediates and augments daily life, yet its technical properties risk reshaping the way users choose (what they consume), own (what they create), learn, and work. These risks reflect widely recognized AI principles, which unfortunately remain largely high-level and unoperationalizable. As the risks arising from AI assistance continue to subtly erode human agency, this work takes a different approach to operationalizing AI principles by translating the risks into concrete research questions that can guide the design of AI systems. This enables the research community to balance human agency and AI assistance in AI development and align it with the goal of benefiting humans.

1. Introduction

Artificial intelligence (AI) has become inseparably intertwined with the rhythms of modern life, shaping how we *choose* what we consume, *own* what we create, *learn* new information, and *work* to complete tasks[1]. In leisure, it subtly chooses the next video or reel we watch (Dekker et al., 2025). In creative production, it composes text and images that both rival and enrich traditional forms of art (Khan, 2024a). In workplaces, it lightens the burden of labor through automation (Simon, 2025), and streamlines the intricacies of office work (Leopold, 2025). When we seek understanding, it serves as an ever-present and patient tutor (Zhai et al., 2024; Kosmyna et al., 2025). Simultaneously, scholarly engagement with AI has shifted from niche to mainstream, resulting in an exponential growth of AI-engaged publications across diverse scientific fields and applications (Duede et al., 2024) to support the development of capable assistants that benefit mankind (Li, 2025) by augmenting our abilities to choose, own, work, and learn.

However, there are also social ramifications to AI use as certain technical properties of AI systems can reshape the way its users choose, own, work, and learn. These technical properties reflect widely recognized principles at the core of frameworks and guidelines on AI use (Jobin et al., 2019; Floridi & Cowls, 2022) regarding human autonomy, dignity, fairness, and meaningful participation[2]. While it is clear that AI research should engage with these principles, existing work highlights a gap between these high-level principles and their practical operationalization (Birkstedt et al., 2023; Morley et al., 2021; 2023). Indeed, conceptual papers far outnumber empirical, technological or implementational research (Birkstedt et al., 2023). Hence, it remains difficult to translate these principles into concrete research directions.

This gap has direct implications on how AI impacts human agency (in choosing, owning, working, and learning), challenging not merely expectations or established routines, but the fundamental role of humans as being able to exercise meaningful discretion in areas of importance to us. Assistive technologies like AI do not merely relieve us of certain drudgeries, they risk fostering overreliance beyond cognitive offloading (Zhai et al., 2024) and skill decay (Noy & Zhang, 2023). What is at stake is more than our capacity to act—it is in fact our ability to make decisions. Crucially, human agency is the capacity to act, but it exists only when it is being exercised (Sanchini et al., 2019). AI risks undermining human agency not merely through automation or the intrinsic opacity of black-box algorithms, but through system outputs shaped by influences that are simultaneously technical and institutional, and largely inaccessible to users. As intrinsic opacity, expert control, and organizational mediation become embedded in these outputs, the scope for exercising individual agency is progressively reduced (Winner, 1980).

In line with AI research's mission to build systems that benefit and augment humans, this paper takes a different approach to operationalizing these AI principles. It examines

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]This paper focuses on *choosing*, *owning*, *working*, and *learning* because they are universal and pervasive in everyday life. Furthermore, the consequences of failing to safeguard these capabilities or "rights" are profound and demand timely and coordinated action. We discuss the selection of these capabilities or "rights" in Section A.3.

[2]We discuss mapping these widely recognized AI principles to the "rights" in a framework in Section A.2.

how AI creates risks to everyday activities such as choosing, owning, working, and learning and translates these into concrete research questions that can guide the design and evaluation of AI systems, enabling the research community to better align AI development with its mission. We adopt the vernacular of "rights"[3] to identify the capabilities we risk giving up to our machine assistants: our "rights" to choose, to own, to work, and to learn.

To start the ball rolling, we present open research questions to address to reinforce each "right". These research questions distill reflections from researchers[4] actively engaged in AI development. The questions are deliberately output-oriented and grounded in lived practice, and act as technical metrics specifying the intended behavior/capability of an AI system, but presented in plain language for a wider audience who may consider addressing these questions from the perspective of their field[5]. Each question is accompanied by an appendix that surveys relevant prior work, highlights unresolved technical challenges that render the question non-trivial, and presents illustrative technical case studies or solutions that serve as inspirations (rather than prescriptions) with the aim of motivating (potentially better) alternative solutions. Throughout the paper, we use "AI systems" as a broad term encompassing, but not limited to, generative models/AI assistants, AI agents, computer vision, voice or speech AI assistants, decision-making systems, recommendation systems, AI productivity tools, personalization systems, and other tools built on AI. We introduce specificity in system types only when necessary to preserve question generality, particularly when discussing risks. We hope these questions will help the audience strike a balance between human agency and AI assistance so that society as a whole reaps the transformative rewards of AI.[6]

2. The "Right" to Choose

Risk: Echo Chambers, Filter Bubbles, and Information Barriers. AI models can increasingly align and optimize their responses with human expectations (Christiano et al., 2017; Rafailov et al., 2023). Large pretrained generative models curate suggestions based on user preferences, search histories, or their own latent biases (Liang et al., 2021), while recommendation feeds curate the content (Tik Tok, 2025) instead of relying on explicit user selection (Agan et al., 2023; Mills & Sætra, 2024). This narrows exposure to diverse perspectives, reinforcing confirmation biases and echo chambers and trapping individuals in filter bubbles (Anwar et al., 2024). Over time, users may struggle to articulate their information needs (Hirvonen et al., 2024), and blocking their own access to broader knowledge.

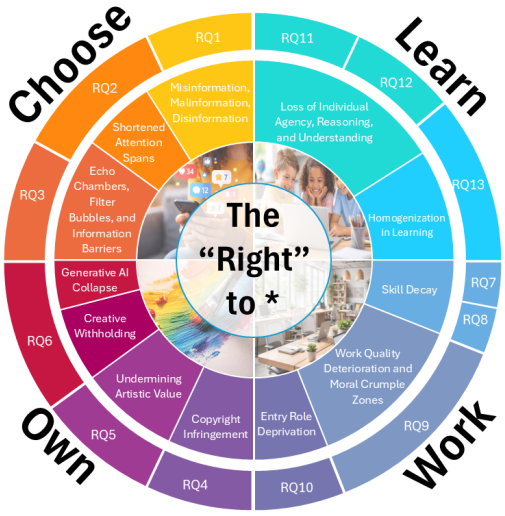

Figure 1. Overview: We focus on the risks (inner ring) arising from AI adoption (without careful scrutiny) that threaten the four "rights" to choose, own, work, and learn. We propose corresponding research questions (RQs in outer ring) to mitigate these risks. Each RQ is color-mapped to one risk, but some span multiple risks and hence extend across risk boundaries (e.g., RQ3). Note: Slice sizes are illustrative and do not indicate risk or RQ importance.

Risk: Shortened Attention Spans, Cognitive Offloading and Cognitive Decline. AI-driven platforms maximize engagement through short, personalized, and rapidly delivered content. As users scroll through AI-curated streams of brief videos, attention spans become increasingly fragmented and decisions become increasingly simple. For example, YouTube presents users with a list of videos to choose from (Ghori et al., 2025), while TikTok further simplifies this decision into binary choice by presenting one video at a time (Tik Tok, 2025). Over time, repeated exposure conditions users' attention systems to favor immediate, frictionless AI curation over diversity or depth of choices (Shanmugasundaram & Tamilarasu, 2023). Individuals may struggle with activities that require sustained cognitive focus, such as long-form reading, and turn to brain-rot content (Yazgan, 2025; Yousef et al., 2025), reflecting a decline in their cognitive abilities.

This phenomenon is observed beyond AI-driven platforms. The use of LLMs over traditional search engines for informa-

[3]The "rights" framing is not intended to propose new legal human rights, but to align with established AI ethics principles synthesized in global surveys on normative expectations of users (Jobin et al., 2019). We conceptualize "rights" in Section A.1.

[4]In August 2025, a cohort of [redacted]–funded Master's and PhD scholars convened to deliberate on the risks associated with AI use. As researchers engaged in developing AI systems, they support AI development and seek to maximize its transformative potential for broad societal benefit.

[5]This target audience includes researchers, developers, social scientists, humanities scholars, and policymakers, who can work together to co-design AI systems to benefit society.

[6]Amid debates over the proper scope of AI agency (Bengio et al., 2025; LeCun, 2022), this paper examines how to preserve human agency without framing it as a zero-sum tradeoff with AI.

tion gathering has also been found to ease cognitive load but limits the depth and quality of content engagement (Stadler et al., 2024). This reveals the cost of cognitive ease. Rather than freeing users to think more deeply, LLMs instead reduce the need to engage with the information independently and meaningfully.

RISK: Mis/Mal/Disinformation. AI systems enable large-scale generation and amplification of fabricated media and coordinated false narratives, undermining public discourse integrity[7] (Maitland & Lee, 2025). Such campaigns "firehose" or flood information ecosystems with high-volume content to overwhelm users (Paul & Matthews, 2016). Empirical work further underscores how users face difficulty in discerning AI-generated content (Frank et al., 2024; Groh et al., 2024), undermining the capacity of users to make informed choices about what to believe.

The risks mentioned above are non-exhaustive and will evolve with the way we access information. AI systems risk obscuring the "Right" to Choose as they increasingly mediate everyday decisions. Safeguarding this "right" therefore requires preserving informed agency throughout the decision process. For example, by ensuring users have the freedom to access or reject unbiased or desired sources of information or options, and even retaining the ability to review, override, or withdraw consent for choices delegated to AI systems.

RQ1. **How can AI systems present diverse and credible content that expands users' informational exposure, while preserving meaningful and sustained engagement?** AI systems act as choice architects, structuring and presenting limited sets of options[8] for users to choose from (Agan et al., 2023; Mills & Sætra, 2024). However, optimizing option sets solely for engagement metrics such as click- or watch-time risks placing users in echo chambers (Baumann et al., 2024) and increases their vulnerability to misinformation (Avram et al., 2020). This has motivated regulations to protect users to have their personal data such as cookies, age, or geographical location *forgotten* (European Parliament and Council of the European Union, 2016) from AI systems, reduce excessive personalization, and allow users to experience more general and unbiased content. In addition to legal enforcement, more can be done in terms of AI development. Consider, for example, platforms that adopt multi-objective recommendation frameworks that explicitly trade off engagement with viewpoint diversity/credibility and civic value. Domain experts and journalists can help identify informational blind spots in various fields, assess credibility, and distinguish factual content from opinion. This can support informed and autonomous choices about what users engage with and expand users' informational exposure while supporting meaningful and sustained engagement, allowing users to escape filter bubbles and avoid undue exposure to false information. As mentioned in Section 1, every RQ, such as RQ1, specifies the intended AI system behavior, invites interdisciplinary examination, and is accompanied by an appendix describing related work, open challenges, and illustrative technical solutions; for RQ1, see Section B.1.2.

RQ2. **How can we design AI systems that users can trust?** The opacity of AI systems prevents humans from fully understanding and interpreting its workings, even for computer scientists or those with specialized training (Burrell, 2016). Increasing transparency is a common response to opacity, but trust depends on appropriate calibration, as both under- and over-transparency can erode trust in the AI system's recommendations, especially when users have previously disagreed with the AI system's outputs (Kizilcec, 2016). Collaborating between legal scholars, social scientists, domain experts, along with computer scientists can help to build AI systems that justify the reason or intent behind their recommended option set. With interpretations for AI outputs, users can ensure that the intentions behind each option–at least at the level of interpretation–are sufficiently informative, and also unbiased, constructive, engaging, and free from harmful influence. This will allow users to trust and evaluate options critically and make autonomous and well-reasoned decisions on their own terms, without worry of manipulation or incorrectness. See Section B.1.3.

RQ3. **How can AI systems empower users to reject, adjust, or withdraw from algorithmic curation?** Even when the option set is optimal and the decision is informed, users may still wish to "fight the feed" (Ghori et al., 2025) by adjusting or withdrawing consent to algorithmic curation, especially as their preferences evolve. AI platform developers may support this by providing the appropriate user controls. Alternatively, AI systems can also be developed to directly support users in wiping out their preferences from the platform.

Algorithmic intervention is complex and its preferred extent varies between users. Even determining the "correct" amount of algorithmic intervention is inherently difficult as users may accept personalized services but simultaneously reject the collection of personal data required for it (Kozyreva et al., 2021). The acceptance of algorithmic curation also depends on social norms, emotions, and heuristics (Acquisti et al., 2015). These challenges have motivated regulations on consent withdrawal (European Parliament

[7] As part of the influence operation: Storm-1516, AI-generated artwork and synthetic personas were used to stage false claims. These narratives were laundered through a network of AI-enabled news websites and algorithmically boosted by automated accounts on platforms like Telegram and X, demonstrating how AI systems can support coordinated, large-scale disinformation campaigns.

[8] We formulate an example option set in Section B.1.1.

and Council of the European Union, 2016) and works on preference shifts (Wang et al., 2023); further progress requires collaboration with decision scientists in psychology, economics, and sociology to design AI systems that account for shifting preferences and perspectives, and empower user control over personalization and privacy needs that evolve over time. Collaboration with public policy and legal experts is also necessary to design AI systems that re-balance power asymmetry between users (who may be naive, uncertain, and vulnerable) and governments and corporations that design and operationalize algorithmic curation. See Section B.1.4.

3. The "Right" to Own

Generative AI has demonstrated remarkable capability to compose text (Brown et al., 2020; Sutskever et al., 2011), images (Goodfellow et al., 2014; Rombach et al., 2022), and audio (Engel et al., 2017; van den Oord et al., 2016), and has been praised as democratizing art by turning passive observers into creators as they no longer need years of training to recreate Rembrandt's brushstrokes or Beethoven's virtuosity (Park, 2025). Text-, image-, audio-based works, and their mixed-media combinations are traditionally covered by copyright where a single discernible author owns each work (U.S. Copyright Office). This contrasts with the layered process of *AI-mediated* (i.e., AI-generated, AI-augmented, or created through human-AI collaboration) creation, involving artists/authors/creators, model developers, data curators, and prompters, causing AI-generated art to fall into a grey area of copyright infringement.

Definition 3.1. Ownership refers to the possession of a creation and the control over the reproduction, distribution, and derivative use of the creation.

Section 3.1 begins with the premise that a rose by any other name would smell as sweet: AI-mediated creative work can still be art. We focus on what follows, namely the risks to ownership or attribution of art, brought on by recognizing AI-mediated creative work as art, and focus on questions that alleviate these risks. Section 3.2 considers how when attribution fails or is not done, the "gardeners" of art (i.e., artists or authors) are not acknowledged and may lose motivation to cultivate novel art. This necessitates questions that protect "gardeners" so that we can continue to appreciate new blooms.

3.1. A Rose by Any Other Name Would Smell as Sweet

Risk: Copyright Infringement. Recognizing AI-generated art as forms of creative expression[9] exposes it to a wide range of infringement risks. Generative models may make unauthorized copies of copyrighted works for training or reproduce them as unlicensed derivatives, thereby violating reproduction rights and moral rights. They can also embed copyrighted materials within model weights, leading to secondary infringement when models are shared (Chesterman, 2024; Jamali, 2025; Montgomery, 2025). These practices undermine creators' rights and recognition.

Risk: Undermining Artistic Value. By automating the reproduction of creative expression without attribution, AI systems appropriate human-created art, reducing demand for human artists and displacing opportunities for salaried employment in creative industries (Chesterman, 2024; Senftleben, 2023; UN Conference on Trade and Development, 2024). As generative AI tools receive recognition, this undermines human artistic integrity and the sustainability of creative labor, threatening the professional recognition and employment opportunities of artists.

RQ4. **How can the authors of an AI-mediated artwork be identified and their contributions quantified to assign credit and ownership?**
Intellectual property is a legal framework designed to incentivize and reward only human creativity and innovation (Chesterman, 2024). AI-generated or modified art falls into a legal gray area, and human and AI modifications may be interleaved throughout the creative process, making it difficult to recognize and reward authors, much less assign ownership. In reasoning about ownership of an AI-mediated creation, attribution and credit become a necessary conceptual precursor as ownership depends on understanding who has contributed what to the final work. To do so requires first identifying contributors, then quantifying their contributions, and lastly assigning credit and ownership.

One may consider the following solution. First, identifying all contributors to a piece of art acknowledges those who may claim authorship. This provides a foundation for AI researchers to trace its complete chain of authorship and influence, and is necessary as artists may use AI tools for assistance or modification, full generation, or not at all. Secondly, the magnitude and nature of each contribution can be quantified through collaboration with artists. Third, discussion on assigning credit welcomes insights from legal experts and economists, such as game theorists, who can consider each contribution and build frameworks to assign credit and ownership accordingly. This rewards creativity and innovation and assigns ownership, thus upholding artistic integrity, sustaining creative labor, and allowing continued creative contribution. See Section B.2.1.

[9] It is unproductive to draw hard lines around what counts as "art", especially as artists themselves increasingly use AI tools to create. We draw parallels to Benjamin's analysis of mechanical reproduction (Benjamin, 2018) and shift attention from definitional debates to how changes in production transform a work's aura and its real-world impact (e.g., on human perception, social relations, politics, power, misinformation, further discussed in Section 2).

RQ5. **How can intangible contribution be recognized and attributed in AI-mediated artworks?**
Style is the technique of creating artwork, while expression encompasses creative choices that convey intent, emotion, or narrative. Even minor pixel-level changes can subtly shift symbolism or emotional tone, altering intangible meaning and perceived value. Legally, expression is copyrightable, but style is not (WIPO, 2025; WTO, 1995). However, generative AI allows style to be replicated accurately and at scale, raising questions about whether style should now be copyrighted to protect original creators. Accordingly, RQ5.1 asks whether AI-mediated stylistic influence constitutes expressive contribution and deserves creative credit. If such stylistic influence is recognized, credit can be assigned to the ones creating the AI-mediated artwork. Otherwise, credit reverts solely to the original creator. RQ5.2 asks how the extent of that intangible contribution can be measured to reward the ones being attributed to it.

RQ5.1. **How can we determine when AI-mediated stylistic influence deserves creative credit?**
Style is not protected under copyright law, but certain styles are so distinctive that they are immediately associated with a specific creator. Generative models can easily copy style and blur the line between inspiration and replication. An example is the "Studio Ghibli style", which remains legally unprotected (SCMP, 2025). Methods can be developed alongside artists to quantify expressive vs. stylistic contributions. This help to distinguish creative contribution from superficial reproduction and recognize creative influence that shapes expression in AI-mediated art. In parallel, engaging legal experts can help with acknowledging contributions and determining copyright infringement. See Section B.2.2.

RQ5.2. **How can we quantify the expressive overlap between an AI-mediated artwork and the original?**
While style replication by generative AI is often apparent (SCMP, 2025), the replication of expression is less straightforward. For example, the New York Times alleges in an ongoing lawsuit that OpenAI is infringing on copyrights by reproducing "significantly more expressive content from [an] original article than what would traditionally be displayed" by an online search (Pope, 2024; NYT, 2023). Having a measure that can quantify expressive overlap (like continuous measures such as similarity scores in plagiarism detection systems (Turnitin, 2025)) can help generative AI rebuke or artists prove claims of copyright infringement. Artists, legal experts, and AI developers can collaborate to build legal frameworks based on this measure, and impose graduated penalties based on the extent of infringement to properly credit the original creator. See Section B.2.3.

3.2. A World Without Gardeners

If credit is attributed solely to the final visible creator such as someone writing a prompt, the layered contributions of other creators risk being overlooked. Authors who see their creations used without recognition risk becoming disincentivized to produce or share new works. Over time, erosion of credit may discourage artistic participation, diminishing the diversity and vitality of the creative ecosystem.

Risk: Creative Withholding. Art shared online is easily copied, modified, or incorporated into datasets without proper attribution, and routinely scraped from public platforms by generative models. Many artists have expressed dissatisfaction with the use of their work in AI training without consent (Ali & Breazeal, 2023; Kambur & Dolunay, 2024; Lovato et al., 2024), becoming increasingly reluctant to publicly share their art (Ali & Breazeal, 2023).

Risk: Generative AI Collapse. Creative withholding results in the lack of new training data. When models are trained recursively on their own outputs, their quality (precision) or diversity (recall) progressively decrease (Alemohammad et al., 2024; Briesch et al., 2024; Shumailov et al., 2024). If human authors stop producing new work, generative models may spiral into self-consumption.

RQ6. **What can authors do to protect their works prior to legal judgment?**
Authors today rely on legal interpretation and subjective judgment to prove that their works have been copied,[10] and have no tangible technical mechanism to detect or demonstrate overlap (de Leon, 2024). Authors may also have had their works "scraped" (lawfully or unlawfully, with or without permission) without being compensated for it (Chesterman, 2024). Developing models that allow legal experts to obtain tangible evidence of unauthorized use (whether directly by copying or indirectly for model training) of artworks will empower authors to privately verify and document potential misuse, giving them awareness and control before formal legal action, and encourage continued creation instead of creative withholding. See Section B.2.4.

4. The "Right" to Work

Risk: Skill Decay. In contemporary workplaces, AI assistants can automate routine tasks, boosting efficiency and task completion (Dell'Acqua et al., 2023; Noy & Zhang, 2023). However, as users increasingly rely on AI tools to reduce the information processing requirements of a task so as to reduce cognitive demand, cognitive offloading re-

[10] Singaporean photographer Zhang Jingna sued Luxembourg artist Jeff Dieschburg who had painted her 2017 Harper's Bazaar Vietnam photograph and submitted it to an art competition, winning a prize. The case initially ruled against her and this was only overturned two years later in 2024 (de Leon, 2024).

sults (Risko & Gilbert, 2016; Grinschgl et al., 2021; Cavicchi et al., 2025), causing the relevant human competencies to atrophy from sporadic use (Macnamara et al., 2024; Natali et al., 2025). For highly skilled users, AI tools even distract by giving generic suggestions (Brynjolfsson et al., 2023) that "relieve" users from practising or developing key cognitive processes. Over time, these diminish the ability of employees to contribute in the workplace without AI assistance, contributing to skill decay.

RISK: Work Quality Deterioration and Moral Crumple Zones. AI tools especially benefit less experienced workers (Brynjolfsson et al., 2023; Cazzaniga et al., 2024), allowing them to shift efforts towards more impactful tasks like planning, evaluation, and problem solving (Filippucci et al., 2024). However, this creates an illusion of competence, where workers over-estimate their own performance on the task (Fernandes et al., 2026). This also gives workers the confidence to submit hallucinated text (Ashktorab et al., 2025) or even "workslop"–long, fluent, AI generated text that seems polished, but in fact needs heavy correction– (Skibba, 2026), as work deliverables.

Furthermore, generic AI tools also produce generic solutions (Anderson et al., 2024; Brynjolfsson et al., 2023). Their widespread deployment hence creates a "homogenization" effect that reduces diversity and creativity in output, undermining work quality. As accountability is typically concentrated with the end user, workers may be placed in a moral crumple zone (Elish, 2019) (i.e., made to absorb the force of impact in a car or AI system's accident) and unproductively blamed for negative outcomes arising from system behaviors or hallucinations they cannot fully control or detect.

RISK: Entry-Role Deprivation. AI tools now perform many basic tasks traditionally assigned to entry-level staff (Leopold, 2025), directly competing with human graduates for entry-level positions (Teng et al., 2024). As fresh graduates lose hiring opportunities to AI tools, they miss out on training and development opportunities entry-level roles traditionally provide. Over time, they lack the deep expertise and specialized knowledge needed to make sustained contributions to their workplaces (Nolan, 2025).

The risks mentioned above are non-exhaustive and will continue to evolve as AI systems increasingly perform human tasks. However, the intrinsic value of contributing meaningfully at work remains key to workers' motivation and the production of original and high-quality output (Bailey & Madden, 2016; Martela, 2023). It is hence vital to ensure that workers can continue to invest effort and even modulate or decline AI assistance in AI mediated workplaces. We propose the following RQs.

RQ7. **Should AI systems collaborate or automate?** Even when users have domain expertise, appropriate AI support can improve performance (Vaccaro et al., 2024). The use of AI systems in the workplace can improve productivity, but the balance between collaboration and automation is a trade-off where efficiency gains from automation often come with limitations (Brynjolfsson et al., 2023). This trade-off has to account for how labor should be divided between human and AI (Vaccaro et al., 2024). RQ7 decomposes this trade-off into sub-RQs that examine when (RQ7.1) and how (RQ7.2) AI can intervene.

RQ7.1. **How can AI systems detect/learn optimal moments to intervene in a task (vs. leaving it to the human)?** One may consider AI systems that intervene in human tasks to collaborate. Identifying when to intervene requires insights from psychology to, for example, understand user cognition and recognize when human limits are reached. Domain expertise can also characterize when task difficulty or error accumulation crosses a learned threshold for intervention. Crucially, there are also situations where humans outperform AI, and no intervention is necessary, as discussed in Section B.3.6. Sparse and purposeful intervention supports problem solving while ensuring that humans remain primarily responsible. This preserves agency and mitigates skill decay. See Section B.3.1.

RQ7.2. **How can AI systems intervene in ways that complement their users' efforts and expertise?** By collaborating with psychologists to translate theories of cognition and learning into system-level design constraints, AI developers can calibrate the extent and mode of AI assistance to complement the user. For example, by allowing the AI system to form a contextual understanding of the user's task, prior actions, and work environment, it can offer appropriate guidance. Adaptive and socially intelligent intervention strategies, coupled with constraints on the granularity of assistance, engage users in the problem-solving process. This preserves human agency, improves task performance, and ensures that users retain competence, expertise, and confidence for future tasks. See Section B.3.2.

RQ8. **What strategies can help organizations ensure that AI is used appropriately and sparingly so that employees continue to develop their own skills?** The overuse of AI in the workplace can lead to skill decay and hinder skill development (Macnamara et al., 2024; Natali et al., 2025). So, frameworks can be created for companies to detect overuse of AI tools on tasks that do not require AI assistance or integrate "friction" into their AI adoption processes. For example, by drawing on behavioral and learning sciences to define appropriate task difficulty and skill retention criteria, companies can introduce small, well-defined, and appropriately difficult tasks for employees to complete without AI assistance. Employees found using AI can undergo targeted retraining to ensure their core skills

remain intact and prevent skill decay. See Section B.3.3.

RQ9. **How can AI systems support fair responsibility allocation in human–AI collaborative work?**
When AI systems go wrong and responsibility must be allocated, shrinking moral crumple zones requires legal and organizational frameworks that distribute accountability across developers, managers, and organizations rather than focus blame on the final human user (Elish, 2019). This calls for collaboration between legal and technical experts to formalize an AI system's limits and uncertainty and reallocate accountability from end users to upstream actors.

Simultaneously, to reduce the occurrence of wrong outputs being used in the workplace, AI systems need to recognize and communicate uncertainty. By drawing on psychologists' insights into human reasoning and logicians' formal accounts of inference, AI systems can, for example, be designed to identify sources of uncertainty arising from epistemic limitations. Communicating uncertainty can also alert users to unverifiable or hallucinated outcomes. This helps users avoid moral crumple zones and think beyond the model's response, fostering higher quality, diverse, and contextually grounded work outputs. See Section B.3.4.

RQ10. **How can AI systems enable a junior employee to reach the competency of a senior employee?**
As firms increasingly view AI systems as viable substitutes for fresh graduates, many become disincentivized from hiring junior employees (Teng et al., 2024). This restricts fresh graduates from accessing and gaining a foothold in organizations. To mitigate this entry-role deprivation, the capability development of junior employees can be accelerated so that their performance becomes comparable to senior employees. In highly specialized domains such as software engineering, urban planning, and manufacturing, companies may prefer candidates with 10-20 years of experience over those with only 1-2 years. One may consider AI systems that can patch this skill gap and endow junior employees with hard and soft skills typically accumulated through long-term practice. This helps junior employees avoid entry-role deprivation and enables them to learn on the job and approach senior-level proficiency at an earlier stage by accelerating the development of their cognitive maps. See Section B.3.5.

While Section 4 protects the "Right" to Work, it does not claim preference for human over AI effort. Section B.3.6 acknowledges scenarios where AI may outperform humans.

5. The "Right" to Learn

Definition 5.1. Learning (Gross, 2012) is the process of acquiring new understanding, knowledge, behaviors, skills, values, attitudes, and preferences.

RISK: Teacher-Learner AI Divide. In conventional educational settings, learners learn from and follow the guidance of teachers. AI disrupts this relationship by giving learners access to alternative sources of instruction beyond the teacher's oversight. Students may use AI to learn different "truths" or copy homework from. Teachers are therefore forced to play catch-up with students' AI use, making it harder for them to understand how learning actually takes place. Without insight into this learning process, teachers may struggle to engage, diagnose learning needs, and adapt their teaching, which can ultimately harm student performance (Rowan & Grootenboer, 2016).

RISK: Reduced Confidence in Education. In conventional educational settings, syllabuses, certified teachers, and standardized assessments provide some assurance that learners are reaching the right outcomes and engaging in a meaningful process of learning. With AI as a middleman in this process, strong standardized test performance does not necessarily guarantee that the learner has understood the material as they may offloaded reasoning to the AI or sidestep the meaningful path to understanding learning materials. This is especially so as educational technologies frame learning as "effortless" (Mentutor, 2024; Shabanov, 2025). This undermines the value of syllabuses, standardized testing and learning outcomes, and erodes confidence that people, especially learners, teachers, and pedagogists may have in education itself[11].

These risks mentioned above will continue to evolve, especially as future learners leverage AI as an encyclopedic "compression" of the internet for learning and AI tools for education continue to evolve[12]. We introduce research questions (RQs) to mitigate the risks above and safeguard the "Right" to Learn with AI for learners who use AI responsibly[13] to enhance rather than bypass learning.

RESEARCH QUESTION 11 (RQ11). **How can AI systems be designed to close or even bridge the gap between teachers and learners?**
Teachers should be able to see the learner's questions and responses from an AI system, as these reveal what learners struggle with and what instructions they have received. Teachers should also be able to control the complexity and scope of answers that an AI gives, such as by limiting the concepts that an AI can use. This would help align AI sup-

[11] Other problems also arise from having AI as an intermediary in learning. Risks such as the loss of the learner's individual agency, reasoning, and understanding and homogenization in learning are presented in Section B.4.1

[12] Section B.4.7 discusses preliminary progress by ChatGPT and Gemini in addressing RQs in Section 5.

[13] Learners may begin with genuine and sincere intent but gradually drift towards AI dependence due to the ease of obtaining ready-made answers. We hope the risks discussed above will ground sincere learners in their commitment to learn meaningfully. We elaborate on this in Section A.1.

port with upcoming lessons and account for syllabuses or differences across educational contexts. For example, in learning mathematics, teachers may prefer to teach specific mental models such as algebra, model-drawing (Ng, 2022), or intuitive reasoning. By bridging the AI gap between the teacher and learner, teachers can understand students' learning progress, empathise with, and get a greater knowledge of the individual needs of their students, to support their learning (Zhou, 2022). See Section B.4.2.

RQ12. **How can we design AI systems that guarantee meaningful learning?**
One solution places the responsibility for meaningful learning primarily on the learner. AI systems that act like teachers can be designed to measure the depth of the learner's understanding and dynamically calibrate their support (RQ16), strategically allow learners to make mistakes (RQ17), or adapt to preferred learning strategies (RQ18). Here, the learner remains responsible for engaging with AI in ways that preserve genuine understanding instead of bypassing the learning process.

The alternative solution places the responsibility on the teacher. Here, AI systems should give teachers oversight and control over AI use, and ensure that AI supports pedagogical goals, like in RQ11.

Ultimately, meaningful learning is personal and context-dependent. The best arrangement may be a blend of solutions, where both learners and teachers are accountable for ensuring that AI supports meaningful learning. See Section B.4.4 to B.4.6.

6. Conclusion

AI assistants increasingly mediate daily human processes. While they may offer short-term gains, they also risk hollowing out the capacities that sustain human agency. These risks reflect widely recognized AI principles, which unfortunately remain largely high-level and unoperationalizable. We therefore emphasize the importance of operationalizing these principles through concrete research questions that translate daily human processes into design choices in AI systems. These questions are intentionally output-oriented and grounded in lived practice, but presented in plain language for wider audiences including researchers, developers, social scientists, humanities scholars, and policymakers. We hope that by championing research answering these questions, humans can retain a core set of concrete capacities and preserve agency when choosing, owning, working, and learning in an AI-mediated world.

7. Alternative Views

An alternative to our "rights"-based framing is the principle-based framing of AI ethics, which focuses on operationalizing high-level ethical principles such as fairness, accountability, and transparency into technical systems. Building on the convergence of such principles identified in surveys (Jobin et al., 2019), a growing body of work has sought to translate these principles into practice (Hagendorff, 2020; Raji et al., 2020).

While the principle-based approach is valuable, our work is intended as complementary. Rather than beginning from desired attributes of AI systems, our approach foregrounds domains of human experience that we argue are worth preserving. This reframing shifts the focus from system-level properties to human-centered outcomes, and provides a different form of guidance for technical development, as reflected in our output-oriented research questions.

At the same time, a "rights"-based framing opens up the potential for a further line of research: these proposed "right" can, in future work, be more explicitly anchored to broader existing human rights frameworks, whose initial framing may not necessarily be grounded in the context of AI. This creates a pathway for linking technical research questions to established legal and institutional mechanisms for rights protection, thereby situating AI development within broader governance structures.

Additionally, we address other potential questions (e.g., practical feasibility of answering RQs, exhaustiveness of the "rights" and RQs) the reader may have in the FAQ (Section A.1).

References

Acquisti, A., Brandimarte, L., and Loewenstein, G. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015.

Agan, A. Y., Davenport, D., Ludwig, J., and Mullainathan, S. Automating automaticity: How the context of human choice affects the extent of algorithmic bias. Working Paper 30981, National Bureau of Economic Research, 2023.

Agarwal, D., Naaman, M., and Vashistha, A. AI suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–21, 2025.

Agarwal, K. AI in education – evaluating ChatGPT as a virtual teaching assistant. *International Journal for Multidisciplinary Research (IJFMR)*, 5(4), 2023.

Ahmad, S. F., Han, H., Alam, M. M., Rehmat, M. K., Irshad, M., Arraño-Muñoz, M., and Ariza-Montes, A. Impact of artificial intelligence on human loss in decision making, laziness and safety in education. *Humanities and Social Sciences Communications*, 10(1):311, 2023.

Aldasoro, I., Gambacorta, L., Korinek, A., Shreeti, V., and Stein, M. Intelligent financial system: How AI is transforming finance. *Journal of Financial Stability*, 81: 101472, 2025.

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., and Baraniuk, R. Self-consuming generative models go MAD. In *Proceedings of the International Conference on Learning Representations*, 2024.

Ali, S. and Breazeal, C. Studying artist sentiments around AI-generated artwork. arXiv:2311.13725, 2023.

AllSides. AllSides Media Bias Chart, 2025. URL https://www.allsides.com/media-bias/media-bias-chart. Accessed: 2025-11-03.

Amnesty International. Taiwan becomes first in Asia to legalize same-sex marriage after historic bill passes, 2019. URL https://www.amnesty.org/en/latest/press-release/2019/05/taiwan-same-sex-marriage-law/. Accessed: 2026-01-03.

Anderson, B. R., Shah, J. H., and Kreminski, M. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th conference on creativity & cognition*, pp. 413–425, 2024.

Anderson, L. W. and Krathwohl, D. R. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc., 2001.

Anwar, M. S., Schoenebeck, G., and Dhillon, P. S. Filter bubble or homogenization? Disentangling the long-term effects of recommendations on user consumption patterns. In *Proceedings of the ACM Web Conference 2024*, pp. 123–134, 2024.

Ashktorab, Z., Desmond, M., Pan, Q., Johnson, J., Brachman, M., Dugan, C., Danilevsky, M., and Geyer, W. Emerging reliance behaviors in human-AI content grounded data generation: The role of cognitive forcing functions and hallucinations. In *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work*, 2025.

Avram, M., Micallef, N., Patil, S., and Menczer, F. Exposure to social engagement metrics increases vulnerability to misinformation. *Harvard Kennedy School Misinformation Review*, 2020.

Bagchi, C., Menczer, F., Lundquist, J., Tarafdar, M., Paik, A., and Grabowicz, P. Social media algorithms can curb misinformation, but do they? arXiv:2409.18393, 2024.

Bailey, C. and Madden, A. What makes work meaningful — or meaningless. *MIT Sloan Management Review*, 2016.

Bailey, J. and Warner, J. AI tutors: Hype or hope for education? *Education Next*, 25(1):62–71, 2024.

Baumann, F., Halpern, D., Procaccia, A. D., Rahwan, I., Shapira, I., and Wüthrich, M. Optimal engagement-diversity tradeoffs in social media. In *Proceedings of the ACM Web Conference 2024*, pp. 288–299, 2024.

Bengio, Y., Cohen, M., Fornasiere, D., Ghosn, J., Greiner, P., MacDermott, M., Mindermann, S., Oberman, A., Richardson, J., Richardson, O., Rondeau, M.-A., St-Charles, P.-L., and Williams-King, D. Superintelligent agents pose catastrophic risks: Can scientist AI offer a safer path? arXiv:2502.15657, 2025.

Benjamin, W. The work of art in the age of mechanical reproduction. In *Modern Art and Modernism: A Critical Anthology*, pp. 217–220. Routledge, 2018.

Bertschinger, N., Rauh, J., Olbrich, E., Jost, J., and Ay, N. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.

Bhandari, A. and Bimo, S. Why's everyone on TikTok now? The algorithmized self and the future of self-making on social media. *Social Media + Society*, 8(1): 20563051221086241, 2022.

Bhatt, U., Chen, V., Collins, K. M., Kamalaruban, P., Kallina, E., Weller, A., and Talwalkar, A. Learning personalized decision support policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14203–14211, 2025.

Birkstedt, T., Minkkinen, M., Tandon, A., and Mäntymäki, M. AI governance: themes, knowledge gaps and future agendas. *Internet Research*, 33(7):133–167, 2023.

Black, P. and Wiliam, D. Assessment and classroom learning. *Assessment in Education: principles, policy & practice*, 5(1):7–74, 1998.

Bleher, H. and Braun, M. Diffused responsibility: attributions of responsibility in the use of AI-driven clinical decision support systems. *AI and Ethics*, 2(4):747–761, 2022.

Boden, M. A. *The Creative Mind: Myths and Mechanisms*. Routledge, 2004.

Bomba, F. and De Angeli, A. Agency and authorship in AI art: Transformational practices for epistemic troubles. *International Journal of Human-Computer Studies*, pp. 103652, 2025.

Brand, A., Allen, L., Altman, M., Hlava, M., and Scott, J. Beyond authorship: Attribution, contribution, collaboration, and credit. *Learned Publishing*, 28(2), 2015.

Brashier, N. M. and Schacter, D. L. Aging in an era of fake news. *Current directions in psychological science*, 29(3): 316–323, 2020.

Briesch, M., Sobania, D., and Rothlauf, F. Large language models suffer from their own output: An analysis of the self-consuming training loop. arXiv:2311.16822, 2024.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Brynjolfsson, E., Li, D., and Raymond, L. R. Generative AI at work. Working Paper 31161, National Bureau of Economic Research, 2023.

Buchholz, A., London, B., Di Benedetto, G., Lichtenberg, J. M., Stein, Y., and Joachims, T. Counterfactual ranking evaluation with flexible click models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1200–1210, 2024.

Burrell, J. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3 (1):2053951715622512, 2016.

Cavicchi, S., Abubshait, A., Siri, G., Mustile, M., and Ciardo, F. Can humanoid robots be used as a cognitive offloading tool? *Cognitive Research: Principles and Implications*, 10(1):17, 2025.

Cazzaniga, M., Jaumotte, F., Li, L., Melina, G., Panton, A. J., Pizzinelli, C., Rockall, E. J., and Tavares, M. M. Gen-AI: Artificial intelligence and the future of work. *Staff Discussion Notes*, 2024(001):1, 2024. ISSN 2617-6750.

Chan, G. and Lo, E. Is AI cheating on the rise? Few cases reported by S'pore universities, but experts warn of risks. The Straits Times, 2025. URL https://www.straitstimes.com/singapore/is-ai-cheating-on-the-rise-few-cases-reported-by-spore-universities-but-experts-warn-of-risks. Accessed: 2025-11-15.

Charusaie, M.-A., Mozannar, H., Sontag, D., and Samadi, S. Sample efficient learning of predictors that complement humans. In *Proceedings of the International Conference on Machine Learning*, pp. 2972–3005. PMLR, 2022.

Chen, R., Zhang, X., Luo, M., Chai, W., and Liu, Z. Pad: Personalized alignment of LLMs at decoding-time. arXiv:2410.04070, 2024.

Chen, Z., Niu, X., Foo, C.-S., and Low, B. K. H. Broaden your SCOPE! Efficient multi-turn conversation planning for LLMs with semantic space. In *Proceedings of the International Conference on Learning Representations*, 2025.

Chesterman, S. Good models borrow, great models steal: intellectual property rights and generative AI. *Policy and Society*, 44(1):23–37, 2024.

Chi, M. T. H. Self-explaining: The dual processes of generating inference and repairing mental models. In *Advances in instructional psychology: Educational design and cognitive science*, volume 5, pp. 161–238, 2000.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Coffelt, T. A., Grauman, D., and Smith, F. L. M. Employers' perspectives on workplace communication skills: The

meaning of communication skills. *Business and Professional Communication Quarterly*, 82(4):418–439, 2019.

Corrall, M. The harm & hypocrisy of AI art, 2023. URL https://www.corralldesign.com/writing/ai-harm-hypocrisy. Accessed: 2025-11-06.

Dalva, Y., Venkatesh, K., and Yanardag, P. FluxSpace: Disentangled semantic editing in rectified flow transformers. arXiv:2412.09611, 2024.

de Leon, E. A. A. Singaporean photographer wins appeal in copyright case against artist from luxembourg, 2024. URL https://asiaiplaw.com/sector/copyright/singaporean-photographer-wins-appeal-in-copyright-case-against-artist-from-luxembourg. Accessed: 2025-11-06.

Dekker, C. A., Baumgartner, S. E., and Sumter, S. R. For you vs. for everyone: The effectiveness of algorithmic personalization in driving social media engagement. *Telematics and Informatics*, 101:102300, 2025.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013), 2023.

Doshi, A. R. and Hauser, O. P. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290, 2024.

Duede, E., Dolan, W., Bauer, A., Foster, I., and Lakhani, K. Oil & water? diffusion of AI within and across scientific fields. arXiv:2405.15828, 2024.

Dutta, A., Mehrab, K. S., Sawhney, M., Neog, A., Khurana, M., Fatemi, S., Pradhan, A., Maruf, M., Lourentzou, I., Daw, A., and Karpatne, A. Open world scene graph generation using vision language models. arXiv:2506.08189, 2025.

Elish, M. C. Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, 5:40–60, 2019.

Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., and Simonyan, K. Neural audio synthesis of musical notes with wavenet autoencoders. In *Proceedings of the International Conference on Machine Learning*, volume 70, pp. 1068–1077. PMLR, 2017.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). EUR-Lex, 2016.

Fernandes, D., Villa, S., Nicholls, S., Haavisto, O., Buschek, D., Schmidt, A., Kosch, T., Shen, C., and Welsch, R. Ai makes you smarter but none the wiser: The disconnect between performance and metacognition. *Computers in Human Behavior*, 175:108779, 2026.

Filippucci, F., Gal, P., Jona-Lasinio, C., Leandro, A., and Nicoletti, G. The impact of Artificial Intelligence on productivity, distribution and growth: Key mechanisms, initial evidence and policy challenges. *OECD Artificial Intelligence Papers*, (15), 2024.

Fink, G. An introduction to cultural mismatch theory and its role in equitable learning, April 2023. URL https://www.everylearnereverywhere.org/blog/an-introduction-to-cultural-mismatch-theory-and-its-role-in-equitable-learning/. Accessed: 2025-11-06.

Floridi, L. and Cowls, J. A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design*, pp. 535–545, 2022.

Foerster, H., Behrouzi, S., Rieger, P., Jadliwala, M., and Sadeghi, A.-R. LightShed: Defeating perturbation-based image copyright protections. In *Proceedings of the 34th USENIX Security Symposium*, pp. 7271–7290, 2025.

Frank, J., Herbert, F., Ricker, J., Schönherr, L., Eisenhofer, T., Fischer, A., Dürmuth, M., and Holz, T. A representative study on human detection of artificially generated media across countries. In *2024 IEEE Symposium on Security and Privacy*, pp. 55–73. IEEE, 2024.

Gao, R. and Yin, M. Confounding-robust deferral policy learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14238–14246, 2025.

Gao, W., Liu, Q., Huang, Z., Yin, Y., Bi, H., Wang, M.-C., Ma, J., Wang, S., and Su, Y. Rcd: Relation map driven cognitive diagnosis for intelligent education systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pp. 501–510, New York, NY, USA, 2021. Association for Computing Machinery.

Gatys, L. A., Ecker, A. S., and Bethge, M. A neural algorithm of artistic style. arXiv:1508.06576, 2015.

Gay, G. *Culturally responsive teaching: Theory, research, and practice*. Teachers College Press, 2018.

Gerlich, M. AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15 (1), 2025.

Ghori, M. F., Dehpanah, A., Gemmell, J., and Mobasher, B. "they only offer the illusion of choice": Exploring user perceptions of control and agency on youtube. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP Adjunct '25, pp. 214–218, New York, NY, USA, 2025. Association for Computing Machinery.

Gillani, N., Yuan, A., Saveski, M., Vosoughi, S., and Roy, D. Me, my echo chamber, and I: Introspection on social media polarization. In *Proceedings of the 2018 World Wide Web Conference*, pp. 823–831, 2018.

Gillham, J. Can humans detect AI-generated text on their own? Forbes Technology Council, 2024.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Goodier, M. Revealed: Thousands of uk university students caught cheating using AI. The Guardian, 2025. URL https://www.theguardian.com/education/2025/jun/15/thousands-of-uk-university-students-caught-cheating-using-ai-artificial-intelligence-survey. Accessed: 2025-11-15.

Government Offices of Sweden. Chronological overview of LGBT persons rights in Sweden, 2018. URL https://www.government.se/articles/2018/07/chronological-overview-of-lgbt-persons-rights-in-sweden/. Accessed: 2026-01-03.

Griffith, V. and Koch, C. Quantifying synergistic mutual information. arXiv:1205.4265, 2014.

Grinschgl, S., Papenmeier, F., and Meyerhoff, H. S. Consequences of cognitive offloading: Boosting performance but diminishing memory. *Quarterly Journal of Experimental Psychology*, 74(9):1477–1496, 2021.

Groh, M., Sankaranarayanan, A., Singh, N., Kim, D. Y., Lippman, A., and Picard, R. Human detection of political speech deepfakes across transcripts, audio, and video. *Nature Communications*, 15(1):7629, 2024.

Groot, T. and Valdenegro Toro, M. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing*, pp. 145–171, 2024.

Gross, R. *Psychology: The Science of Mind and Behaviour 6th Edition*. Hodder Education, 2012.

Hacker, P. A legal framework for AI training data—from first principles to the Artificial Intelligence Act. *Law, innovation and technology*, 13(2):257–301, 2021.

Hagendorff, T. The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1):99–120, Mar 2020.

Hassan, M., Kushniruk, A., and Borycki, E. Barriers to and facilitators of artificial intelligence adoption in health care: Scoping review. *JMIR Hum Factors*, 11:e48633, Aug 2024.

Heymans, M. Guided learning in Gemini: From answers to understanding, August 2025. URL https://blog.google/products-and-platforms/products/education/guided-learning/. Accessed: 2026-01-28.

Hirvonen, N., Jylhä, V., Lao, Y., and Larsson, S. Artificial intelligence in the information ecosystem: Affordances for everyday information seeking. *Journal of the Association for Information Science and Technology*, 75(10): 1152–1165, 2024.

Ince, R. A. A. Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy*, 19(7):318, 2017.

Jamali, L. AI firm anthropic agrees to pay authors $1.5bn to settle piracy lawsuit, 2025. URL https://www.bbc.com/news/articles/c5y4jpg922qo. Accessed: 2025-11-15.

Jiang, B., Hao, Z., Cho, Y.-M., Li, B., Yuan, Y., Chen, S., Ungar, L., Taylor, C. J., and Roth, D. Know me, respond to me: Benchmarking LLMs for dynamic user profiling and personalized responses at scale. arXiv:2504.14225, 2025.

Jobin, A., Ienca, M., and Vayena, E. The global landscape of AI ethics guidelines. *Nature machine intelligence*, 1 (9):389–399, 2019.

Kambur, H. and Dolunay, A. A research on copyright issues impacting artists emotional states in the framework of artificial intelligence. *Frontiers in Psychology*, 15:1409646, 2024.

Kanervisto, A., Bignell, D., Wen, L. Y., Grayson, M., Georgescu, R., Valcarcel Macua, S., Tan, S. Z., Rashid, T., Pearce, T., Cao, Y., Lemkhenter, A., Jiang, C., Costello, G., Gupta, G., Tot, M., Ishida, S., Gupta, T., Arora, U., White, R. W., Devlin, S., Morrison, C., and Hofmann, K. World and human action models towards gameplay ideation. *Nature*, 638(8051):656–663, February 2025.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Keswani, V., Lease, M., and Kenthapadi, K. Towards unbiased and accurate deferral to multiple experts. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 154–165, 2021.

Khan, I. Midjourney. In *The Quick Guide to Prompt Engineering: Generative AI Tips and Tricks for ChatGPT, Bard, Dall-E, and Midjourney*, pp. 333–343. Wiley AI, 2024a.

Khan, S. *Brave New Words: How AI Will Revolutionize Education (and Why That's a Good Thing)*. Penguin Publishing Group, 2024b.

Kilitcioglu, D., Greenquist, N., and Bari, A. Pyrorank: A novel nature-inspired algorithm to promote diversity in recommender systems. In *International Conference on Swarm Intelligence*, pp. 139–155. Springer, 2023.

Kim, M., Kim, S., Lee, S., Yoon, Y., Myung, J., Yoo, H., Lim, H., Han, J., Kim, Y., Ahn, S.-Y., Kim, J., Oh, A., Hong, H., and Lee, T. Y. Designing prompt analytics dashboards to analyze student-chatgpt interactions in efl writing. arXiv:2405.19691, 2024. URL https://arxiv.org/abs/2405.19691.

Kizilcec, R. F. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 2390–2395, 2016.

Klock, A. H. and Jan, C. B. Syntax error correction method and apparatus. https://patents.google.com/patent/US4617643A/en, 1986.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *Proceedings of the International Conference on Machine Learning*, pp. 1885–1894. PMLR, 2017.

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., and Maes, P. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. arXiv:2506.08872, 2025.

Kozyreva, A., Lorenz-Spreen, P., Hertwig, R., Lewandowsky, S., and Herzog, S. M. Public attitudes towards algorithmic personalization and use of personal data online: Evidence from Germany, Great Britain, and the United States. *Humanities and Social Sciences Communications*, 8(1):1–11, 2021.

Kuang, W., Qian, B., Li, Z., Chen, D., Gao, D., Pan, X., Xie, Y., Li, Y., Ding, B., and Zhou, J. FederatedScope-LLM: A comprehensive package for fine-tuning large language models in federated learning. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5260–5271, 2024.

Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Ivison, H., Brahman, F., Miranda, L. J. V., Liu, A., Dziri, N., Lyu, S., et al. Tulu 3: Pushing frontiers in open language model post-training. arXiv:2411.15124, 2024.

Lau, G. K. R., Niu, X., Dao, H., Chen, J., Foo, C.-S., and Low, B. K. H. Waterfall: Scalable framework for robust text watermarking and provenance for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20432–20466, 2024.

Lau, G. K. R., Dao, H., Lin, N. K. H., and Low, B. K. H. Uncertainty quantification for MLLMs. In *Proceedings of the ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025.

LeCun, Y. A path towards autonomous machine intelligence. *Open Review*, 62(1):1–62, 2022.

Lee, S., Lai, B., Ryan, F., Boote, B., and Rehg, J. M. Modeling multimodal social interactions: New challenges and baselines with densely aligned representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14585–14595, 2024.

Leopold, T. How AI is reshaping the career ladder, and other trends in jobs and skills on labour day. World Economic Forum, 2025. URL https://www.weforum.org/stories/2025/04/ai-jobs-international-workers-day/. Accessed: 2026-01-28.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Li, C., Chen, M., Wang, J., Sitaram, S., and Xie, X. CultureLLM: Incorporating cultural differences into large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 84799–84838, 2024a.

Li, F. *The Worlds I See: Curiosity, Exploration, and Discovery at the Dawn of AI*. Flatiron Books, 2025. ISBN 9781250898104. URL https://books.google.com.sg/books?id=T6bvEAAAQBAJ.

Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., Chang, K.-W., and Gao, J. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022.

Li, R., Zhang, S., Lin, D., Chen, K., and He, X. From pixels to graphs: Open-vocabulary scene graph generation with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28076–28086, 2024b.

Li, Y., Xiong, M., Wu, J., and Hooi, B. Conftuner: Training large language models to express their confidence verbally. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Liang, P. P., Wu, C., Morency, L.-P., and Salakhutdinov, R. Towards understanding and mitigating social biases in language models. In *International conference on machine learning*, pp. 6565–6576. PMLR, 2021.

Liang, P. P., Cheng, Y., Fan, X., Ling, C. K., Nie, S., Chen, R. J., Deng, Z., Allen, N., Auerbach, R., Mahmood, F., Salakhutdinov, R., and Morency, L.-P. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Liu, B., Schlegel, V., Batista-Navarro, R., and Ananiadou, S. Argument mining as a multi-hop generative machine reading comprehension task. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10846–10858, 2023.

Liu, J., Huang, Z., Xiao, T., Sha, J., Wu, J., Liu, Q., Wang, S., and Chen, E. Socraticlm: Exploring socratic personalized teaching with large language models. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 85693–85721. Curran Associates, Inc., 2024a.

Liu, Q., Zhou, Y., Huang, J., and Li, G. When ChatGPT is gone: Creativity reverts and homogeneity persists. arXiv:2401.06816, 2024b.

Liu, S., Shen, J., Qian, H., and Zhou, A. Inductive cognitive diagnosis for fast student learning in web-based intelligent education systems. In *Proceedings of the ACM Web Conference 2024*, WWW '24, pp. 4260–4271. ACM, May 2024c.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., and Zhang, L. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision*, pp. 38–55, 2024d.

Liu, Y., Fan, C., Dai, Y., Chen, X., Zhou, P., and Sun, L. Metacloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24219–24228, 2024e.

Liu, Z., Lin, G., Tan, H. L., Zhang, H., Lu, Y., Gao, X., Yin, S. X., He, S., Goh, H. H., Wong, L. H., and Chen, N. F. SingaKids: A multilingual multimodal dialogic tutor for language learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 1244–1253, 2025.

Lovato, J., Zimmerman, J., Smith, I., Dodds, P., and Karson, J. Foregrounding artist opinions: A survey study on transparency, ownership, and fairness in AI generative art. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pp. 905–916, 2024.

Lu, X., Niu, X., Lau, G. K. R., Nhung, B. T. C., Sim, R. H. L., Wen, F., Foo, C.-S., Ng, S.-K., and Low, B. K. H. Waterdrum: Watermarking for data-centric unlearning metric. arXiv:2505.05064, 2025a.

Lu, X., Sclar, M., Hallinan, S., Mireshghallah, N., Liu, J., Han, S., Ettinger, A., Jiang, L., Chandu, K., Dziri, N., and Choi, Y. AI as humanity's salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text. In *Proceedings of the International Conference on Learning Representations*, 2025b.

Lu, X., Wang, J., Zhao, Z., Dai, Z., Foo, C.-S., Ng, S.-K., and Low, B. K. H. WASA: WAtermark-based Source Attribution for large language model-generated data. In *Findings of the Association for Computational Linguistics*, pp. 23791–23824, 2025c.

Luccioni, S., Akiki, C., Mitchell, M., and Jernite, Y. Stable bias: Evaluating societal representations in diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 56338–56351, 2023.

Ma, P., Yang, X., Li, Y., Gui, M., Krause, F., Schusterbauer, J., and Ommer, B. SCFlow: Implicitly learning style and content disentanglement with flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14919–14929, 2025.

Macnamara, B. N., Berber, I., Çavuşoğlu, M. C., Krupinski, E. A., Nallapareddy, N., Nelson, N. E., Smith, P. J.,

Wilson-Delfosse, A. L., and Ray, S. Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers' awareness? *Cognitive Research: Principles and Implications*, 9(1):46, 2024.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., and Ho, D. E. Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242, 2025.

Maitland, E. and Lee, A. 400 and counting: A Russian influence operation overtakes official state media in spreading Russia-Ukraine false claims, 2025. URL https://www.newsguardtech.com/special-reports/400-and-counting-a-russian-influence-operation-overtakes-official-state-media-in-spreading-russia-ukraine-false-claims/. Accessed: 2026-01-28.

Martela, F. The normative value of making a positive contribution–benefiting others as a core dimension of meaningful work. *Journal of Business Ethics*, 185(4): 811–823, 2023.

Masters, C., Vellanki, A., Shangguan, J., Kultys, B., Gilmore, J., Moore, A., and Albrecht, S. Orchestrating human-AI teams: The manager agent as a unifying research challenge. In *Proceedings of the International Conference on Distributed Artificial Intelligence*, pp. 91–107, 2025.

Matthias, A. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, 6(3):175–183, 2004.

Maurya, K. K., Srivatsa, K. A., Petukhova, K., and Kochmar, E. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered ai tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1234–1251, 2025.

McMahan, B. J., Peng, Z., Zhou, B., and Kao, J. C. Shared autonomy with IDA: interventional diffusion assistance. In *Advances in Neural Information Processing Systems*, volume 37, pp. 128330–128354, 2024.

Mentutor. Effortless education: Hosting, learning, and teaching with mentutor, 2024. URL https://www.mentutor.io/post/effortless-education-hosting-learning-and-teaching-with-mentutor. Accessed: 2025-12-20.

Merchant, B. AI is already taking jobs in the video game industry, 2023. URL https://www.wired.com/story/ai-is-already-taking-jobs-in-the-video-game-industry/. Accessed: 2025-11-09.

Meyers, S., Rowell, K., Wells, M., and Smith, B. C. Teacher empathy: A model of empathy for teaching for student success. *COllege Teaching*, 67(3):160–168, July 2019.

Mills, S. and Sætra, H. S. The autonomous choice architect. *AI & Society*, 39(2):583–595, 2024.

Min, S., Gururangan, S., Wallace, E., Shi, W., Hajishirzi, H., Smith, N. A., and Zettlemoyer, L. SILO language models: Isolating legal risk in a nonparametric datastore. In *Proceedings of the International Conference on Learning Representations*, 2024.

Mlodozeniec, B. K., Eschenhagen, R., Bae, J., Immer, A., Krueger, D., and Turner, R. E. Influence functions for scalable data attribution in diffusion models. In *Proceedings of the International Conference on Learning Representations*, 2025.

Montgomery, B. Disney and Universal sue AI image creator Midjourney, alleging copyright infringement, 2025. URL https://www.theguardian.com/technology/2025/jun/11/disney-universal-ai-lawsuit. Accessed: 2025-12-29.

Moon, K., Green, A. E., and Kushlev, K. Homogenizing effect of large language models (LLMs) on creative diversity: An empirical comparison of human and chatgpt writing. *Computers in Human Behavior: Artificial Humans*, 6:100207, 2025.

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., and Floridi, L. Ethics as a service: A pragmatic operationalisation of AI ethics. *Minds and Machines*, 31(2): 239–256, 2021.

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., and Floridi, L. Operationalising AI ethics: barriers, enablers and next steps. *AI & Society*, 38(1):411–423, 2023.

Naous, T., Ryan, M. J., Ritter, A., and Xu, W. Having beer after prayer? Measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16366–16393, 2024.

Natali, C., Marconi, L., Dias Duran, L. D., and Cabitza, F. AI-induced deskilling in medicine: A mixed-method review and research agenda for healthcare and beyond. *Artificial Intelligence Review*, 58(11), 2025.

Ng, R., Nguyen, T. N., Huang, Y., Tai, N. C., Leong, W. Y., Leong, W. Q., Yong, X., Ngui, J. G., Susanto, Y., Cheng, N., Rengarajan, H., Limkonchotiwat, P., Hulagadri, A. V., Teng, K. W., Tong, Y. Y., Siow, B., Teo, W. Y., Lau, W., Tan, C. M., Ong, B., Ong, Z. H., Montalan, J. R., Chan, A., Antonyrex, S., Lee, R., Choa, E., Tat-Wee, D. O., Liu, B. J. D., Tjhi, W. C., Cambria, E., and Teo, L. SEA-LION: Southeast asian languages in one network. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 512–526, 2025.

Ng, S. F. The model method: Crown jewel in singapore mathematics. *Asian Journal for Mathematics Education*, 1(2):147–161, 2022.

Ngo, T., Kunkel, J., and Ziegler, J. Exploring mental models for transparent and controllable recommender systems: A qualitative study. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pp. 183–191, 2020.

Nicoletti, L. and Bass, D. Humans are biased. Generative AI is even worse, 2023. URL https://www.bloomberg.com/graphics/2023-generative-ai-bias/. Accessed: 2026-01-28.

Nikolic, I., Baluta, T., and Saxena, P. Model provenance testing for large language models. arXiv:2502.00706, 2025.

Ning, L., Liu, L., Wu, J., Wu, N., Berlowitz, D., Prakash, S., Green, B., O'Banion, S., and Xie, J. User-LLM: Efficient LLM contextualization with user embeddings. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 1219–1223, 2025.

Nolan, B. AI is gutting the next generation of talent: In tech, job openings for new grads have already been halved, 2025. URL https://fortune.com/2025/08/15/ai-gutting-next-generation-of-talent/. Accessed: 2025-12-29.

Noy, S. and Zhang, W. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023.

Oh, S.-y. and Ahn, Y. Exploring teachers' perception of artificial intelligence: The socio-emotional deficiency as opportunities and challenges in human-ai complementarity in K-12 education. In Olney, A. M., Chounta, I.-A., Liu, Z., Santos, O. C., and Bittencourt, I. I. (eds.), *Artificial Intelligence in Education*, pp. 439–447, Cham, 2024. Springer Nature Switzerland.

OpenAI. Introducing study mode, 2025. URL https://openai.com/index/chatgpt-study-mode/. Accessed: 2026-01-28.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Paivio, A., Rogers, T. B., and Smythe, P. C. Why are pictures easier to recall than words? *Psychonomic Science*, 11(4): 137–138, Apr 1968.

Papakyriakopoulos, O. and Goodman, E. The impact of twitter labels on misinformation spread and user engagement: Lessons from trump's election tweets. In *Proceedings of the ACM Web Conference 2022*, pp. 2541–2551, 2022.

Park, S. The work of art in the age of generative AI: aura, liberation, and democratization. *AI & society*, 40(3): 1807–1816, 2025.

Paul, C. and Matthews, M. The Russian "firehose of falsehood" propaganda model: Why it might work and options to counter it. *Rand Corporation*, 2(7):1–10, 2016.

Peper, J., Qiu, W., and Wang, L. PELMS: Pre-training for effective low-shot multi-document summarization. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7652–7674, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

Pope, A. NYT v. OpenAI: The Times's about-face, 2024. URL https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/. Accessed: 2026-01-17.

Puig, X., Shu, T., Li, S., Wang, Z., Liao, Y.-H., Tenenbaum, J. B., Fidler, S., and Torralba, A. Watch-And-Help: A challenge for social perception and human-AI collaboration. In *Proceedings of the International Conference on Learning Representations*, 2021.

Qi, N., Li, Y., Fu, R., and Zhu, Q. STRAT: Image style transfer with region-aware transformer. *Neurocomputing*, 617:129039, 2025.

Qi, Y., Schölkopf, B., and Jin, Z. Causal responsibility attribution for human-AI collaboration. arXiv:2411.03275, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33–44. Association for Computing Machinery, 2020.

Reuters. Chinese court rejects landmark same-sex marriage case, 2016. URL https://www.reuters.com/article/world/chinese-court-rejects-landmark-same-sex-marriage-case-idUSKCN0XA1BR/. Accessed: 2026-01-03.

Risko, E. F. and Gilbert, S. J. Cognitive offloading. *Trends in cognitive sciences*, 20(9):676–688, 2016.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Rouinfar, A., Agra, E., Larson, A., Loschky, L., and Rebello, N. Can visual cues and correctness feedback influence students' reasoning. In *American Institute of Physics Conference Series*, 2014.

Rowan, L. and Grootenboer, P. (eds.). *Student engagement and educational rapport in higher education*. Springer International Publishing, Cham, Switzerland, 1 edition, November 2016.

Ruiz-Dolz, R., Kikteva, Z., and Lawrence, J. Mining complex patterns of argumentative reasoning in natural language dialogue. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7421–7435, 2025.

Ryu, S., Do, H., Kim, Y., Lee, G., and Ok, J. Multi-dimensional optimization for text summarization via reinforcement learning. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5858–5871, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Sackett, P. R. and Walmsley, P. T. Which personality attributes are most important in the workplace? *Perspectives on Psychological Science*, 9(5):538–551, 2014.

Sanchini, V., Pongiglione, F., and Sala, R. On the notion of political agency. *Phenomenology and Mind*, 16:10–15, 2019.

Seah, M. Statement by Mr Mark Seah, Deputy Permanent Representative of Singapore, on Agenda Item 84, on the rule of law at the national and international levels, sixth committee. Permanent Mission of Singapore to the United Nations, 2022. URL https://www.un.org/en/ga/sixth/77/pdfs/statements/rule_of_law/06mtg_singapore.pdf. Accessed: 2025-12-29.

Senftleben, M. Generative AI and author remuneration. *IIC-International Review of Intellectual Property and Competition Law*, 54(10):1535–1560, 2023.

Seo, S., Han, B., Harari, R. E., Dias, R. D., Zenati, M. A., Salas, E., and Unhelkar, V. Socratic: Enhancing human teamwork via AI-enabled coaching. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pp. 1876–1885, 2025.

Shabanov, I. Unlock your academic potential. Get the effortless academic quick start guide, 2025. URL https://effortlessacademic.com/. Accessed: 2025-12-29.

Shah, M. and Sureja, N. A comprehensive review of bias in deep learning models: Methods, impacts, and future directions. *Archives of Computational Methods in Engineering*, 32(1):255–267, 2025.

Shamshad, F., Bakr, T., Shaaban, Y., Hussein, N., Nandakumar, K., and Lukas, N. First-place solution to NeurIPS 2024 invisible watermark removal challenge. arXiv:2508.21072, 2025.

Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., and Zhao, B. Y. Glaze: protecting artists from style mimicry by text-to-image models. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, pp. 2187–2204, 2023.

Shan, S., Ding, W., Passananti, J., Wu, S., Zheng, H., and Zhao, B. Y. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy*, pp. 807–825. IEEE, 2024.

Shanmugasundaram, M. and Tamilarasu, A. The impact of digital technology, social media, and artificial intelligence on cognitive functions: a review. *Frontiers in Cognition*, 2(1203077), 2023.

Shi, H., Xu, Z., Wang, H., Qin, W., Wang, W., Wang, Y., Wang, Z., Ebrahimi, S., and Wang, H. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 58(5):1–42, 2025.

Shu, Y., Hu, W., Ng, S.-K., Low, B. K. H., and Yu, F. Ferret: Federated full-parameter tuning at scale for large language models. In *Proceedings of the International Conference on Machine Learning*, pp. 55379–55402. PMLR, 2025.

Shuai, Z., Wu, C., Tang, Z., Song, B., and Shen, L. Latent space disentanglement in diffusion transformers enables precise zero-shot semantic editing. arXiv:2411.08196, 2024.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., and Gal, Y. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Sikander, G. and Anwar, S. Driver fatigue detection systems: A review. *IEEE Transactions on Intelligent Transportation Systems*, 20(6):2339–2352, 2018.

Simon, L. K. Is AI responsible for the rise in entry-level unemployment? Revelop Labs, 2025. URL https://www.reveliolabs.com/news/macro/is-ai-responsible-for-the-rise-in-entry-level-unemployment/.

Skibba, R. Bosses say AI boosts productivity – workers say they're drowning in 'workslop'. The Guardian, 2026. URL https://www.theguardian.com/technology/2026/apr/14/ai-productivity-workplace-errors. Accessed: 2026-05-07.

Slokom, M., Hanjalic, A., and Larson, M. Towards user-oriented privacy for recommender system data: A personalization-based approach to gender obfuscation for user profiles. *Information Processing & Management*, 58 (6):102722, 2021.

Solovev, K. and Pröllochs, N. References to unbiased sources increase the helpfulness of community fact-checks. *Scientific Reports*, 15(1):25749, 2025.

South China Morning Post. ChatGPT usage hits record as Studio Ghibli-style AI images go viral, April 2025. URL https://www.scmp.com/tech/big-tech/article/3304828/chatgpt-usage-hits-record-studio-ghibli-style-ai-images-go-viral. Accessed: 2025-10-30.

Stadler, M., Bannert, M., and Sailer, M. Cognitive ease at a cost: Llms reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior*, 160:108386, 2024.

Steele, D. M. and Cohn-Vargas, B. *Identity safe classrooms, grades K-5: Places to belong and learn*. Corwin Press, 2013.

Sung, Y.-T., Yang, J.-M., and Han, Y. L. The effects of mobile-computer-supported collaborative learning: Meta-analysis and critical synthesis. *Review of Educational Research*, 87(4):768–805, 2017.

Sutskever, I., Martens, J., and Hinton, G. Generating text with recurrent neural networks. In *Proceedings of the International Conference on Machine Learning*, pp. 1017–1024, 2011.

Tan, J., Xu, S., Ge, Y., Li, Y., Chen, X., and Zhang, Y. Counterfactual explainable recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 1784–1793, 2021.

Tan, J., Ge, Y., Zhu, Y., Xia, Y., Luo, J., Ji, J., and Zhang, Y. User-controllable recommendation via counterfactual retrospective and prospective explanations. In *Proceedings of the 26th European Conference on Artificial Intelligence*, pp. 2307–2314, 2023.

Tan, X., Cheng, G., and Ling, M. H. Artificial intelligence in teaching and teacher professional development: A systematic review. *Computers and Education: Artificial Intelligence*, 8:100355, 2025.

Teng, D., Ye, C., and Martinez, V. AI is stronger than a new graduate, why will you recruit now? A human capital perspective. 2024.

The New York Times. Exhibit J: One hundred examples of GPT-4 memorizing content from The New York Times, 2023. URL https://nytco-assets.nytimes.com/2023/12/Lawsuit-Document-dkt-1-68-Ex-J.pdf. Accessed: 2026-01-17.

Tian, Q. and Zheng, X. Effectiveness of online collaborative problem-solving method on students' learning performance: A meta-analysis. *Journal of Computer Assisted Learning*, 40(1):326–341, 2024.

Tian, Z., Liu, A., Esbenshade, L., Sarkar, S., Zhang, Z., He, K., and Sun, M. Implementation considerations for automated AI grading of student work. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, pp. 9–20, 2025.

Tik Tok. For you. Tik Tok, 2025. URL https://support.tiktok.com/en/

getting-started/for-you. Accessed: 2025-08-29.

Tommasel, A. and Menczer, F. Do recommender systems make social media more susceptible to misinformation spreaders? In *Proceedings of the 16th ACM Conference on Recommender Systems*, pp. 550–555, 2022.

Turnitin. Turnitin similarity: Comprehensive plagiarism detection, 2025. URL https://www.turnitin.com/products/similarity/. Accessed: 2025-11-06.

United Nations. Universal declaration of human rights, 1948. URL https://www.un.org/en/about-us/universal-declaration-of-human-rights. Accessed: 2026-01-03.

United Nations Conference on Trade and Development. *Creative economy outlook 2024*. United Nations, 2024.

United Nations Human Rights Office of the High Commissioner. Convention against torture and other cruel, inhuman or degrading treatment or punishment, 1984. URL https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-against-torture-and-other-cruel-inhuman-or-degrading. Accessed: 2026-01-03.

United Nations Human Rights Office of the High Commissioner. Committee on the Rights of the Child reviews the report of Singapore, 2019. URL https://www.ohchr.org/en/press-releases/2019/05/committee-rights-child-reviews-report-singapore. Accessed: 2026-01-03.

U.S. Copyright Office. What is copyright? URL https://www.copyright.gov/what-is-copyright/. Accessed: 2025-11-20.

U.S. Mission Geneva. U.S. explanation of vote on the right to food. U.S. Mission to International Organizations in Geneva, 2017. URL https://geneva.usmission.gov/2017/03/24/u-s-explanation-of-vote-on-the-right-to-food/. Accessed: 2025-11-20.

Vaccaro, M., Almaatouq, A., and Malone, T. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8 (12):2293–2303, 2024.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. WaveNet: A generative model for raw audio. arXiv:1609.03499, 2016.

Vie, J.-J., Popineau, F., Bruillard, É., and Bourda, Y. A review of recent advances in adaptive assessment. *Learning analytics: Fundaments, applications, and trends: A view of the current state of the art to enhance e-learning*, pp. 113–142, 2017.

Wadinambiarachchi, S., Kelly, R. M., Pareek, S., Zhou, Q., and Velloso, E. The effects of generative AI on design fixation and divergent thinking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2024.

Wang, J. and Fan, W. The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis. *Humanities and Social Sciences Communications*, 12(1):621, 2025.

Wang, R., Yu, H., Zhang, W., Qi, Z., Sap, M., Bisk, Y., Neubig, G., and Zhu, H. SOTOPIA-π: Interactive learning of socially intelligent language agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12912–12940, 2024a.

Wang, W., Feng, F., Nie, L., and Chua, T.-S. User-controllable recommendation against filter bubbles. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1251–1261, 2022.

Wang, W., Lin, X., Wang, L., Feng, F., Ma, Y., and Chua, T.-S. Causal disentangled recommendation against user preference shifts. *ACM Transactions on Information Systems*, 42(1):1–27, 2023.

Wang, Y., Li, Z., Zhang, W., Zhang, Z., Xie, B., Liu, X., Zeng, W., and Jin, X. Scene graph disentanglement and composition for generalizable complex image generation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 98478–98504, 2024b.

Watiktinnakorn, C., Seesai, J., and Kerdvibulvech, C. Blurring the lines: how AI is redefining artistic ownership and copyright. *Discover Artificial Intelligence*, 3(1):37, 2023.

Williams, P. L. and Beer, R. D. Nonnegative decomposition of multivariate information. arXiv:1004.2515, 2010.

Wilson, R. A. Afterword to "Anthropology and Human Rights in a New Key": The social life of human rights. *American Anthropologist*, 108(1):77–83, 2006.

Winner, L. Do artifacts have politics? *Daedalus*, 109(1): 121–136, 1980.

Woods, D. D. Paradigms for intelligent decision support. In *Intelligent Decision Support in Process Environments*, pp. 153–173, 1986.

World Intellectual Property Organization. Copyright basics: Ideas and expression. World Intellectual Property Organization Website, 2025. URL https://www.wipo.int/copyright/en/. Accessed: 2025-10-15.

World Trade Organization. Module II: Copyright and related rights. Training Module on the TRIPS Agreement, 1995. URL https://www.wto.org/english/tratop_e/trips_e/ta_docs_e/modules2_e.pdf. Accessed: 2025-10-15.

Wu, X., Li, Y., Zu, T., Hutson, J., Loschky, L. C., and Rebello, N. S. Using multimedia hints to facilitate conceptual problem solving in physics: Investigating the effects of multiple modalities. In *Frontiers in Education*, volume 10, pp. 1568406. Frontiers Media SA, 2025.

Yakura, H., Lopez-Lopez, E., Brinkmann, L., Serna, I., Gupta, P., Soraperra, I., and Rahwan, I. Empirical evidence of large language model's influence on human spoken communication. arXiv:2409.01754, 2025.

Yang, T., Wang, Y., Lv, Y., and Zheng, N. DisDiff: Unsupervised disentanglement of diffusion probabilistic models. arXiv:2301.13721, 2023.

Yang, Y., Song, Q., Gao, Z., Wang, G., Li, S., and Zhang, X. StyDeco: Unsupervised style transfer with distilling priors and semantic decoupling. arXiv:2508.01215, 2025a.

Yang, Z., Wang, H., and Hu, D. Efficient quantification of multimodal interaction at sample level. In *Forty-second International Conference on Machine Learning*, 2025b.

Yazgan, A. M. The problem of the century: Brain rot. *OPUS Journal of Society Research*, 22(2):211–221, 2025.

Yousef, A. M. F., Alshamy, A., Tlili, A., and Metwally, A. H. S. Demystifying the new dilemma of brain rot in the digital era: A review. *Brain Sciences*, 15(3):283, 2025.

Yu, H., Jeong, S., Pawar, S., Shin, J., Jin, J., Myung, J., Oh, A., and Augenstein, I. Entangled in representations: Mechanistic investigation of cultural biases in large language models. arXiv:2508.08879, 2025.

Zerkouk, M., Mihoubi, M., and Chikhaoui, B. A comprehensive review of AI-based intelligent tutoring systems: Applications and challenges. arXiv:2507.18882, 2025.

Zhai, C., Wibowo, S., and Li, L. D. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments*, 11(1):28, 2024.

Zhang, S., Yao, L., Sun, A., and Tay, Y. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys*, 52(1):1–38, 2019.

Zhang, S., Chen, Z., Chen, L., and Wu, Y. CDST: Color disentangled style transfer for universal style reference customization. arXiv:2506.13770, 2025a.

Zhang, Y., Sun, J., Feng, L., Yao, C., Fan, M., Zhang, L., Wang, Q., Geng, X., and Rui, Y. See widely, think wisely: Toward designing a generative multi-agent system to burst filter bubbles. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–24, 2024.

Zhang, Y., Diddee, H., Holm, S., Liu, H., Liu, X., Samuel, V., Wang, B., and Ippolito, D. Noveltybench: Evaluating language models for humanlike diversity. In *Proceedings of the Conference on Language Modeling*, 2025b.

Zhao, S., Hong, M., Liu, Y., Hazarika, D., and Lin, K. Do LLMs recognize your preferences? Evaluating personalized preference following in LLMs. In *Proceedings of the International Conference on Learning Representations*, 2025.

Zhou, L.-Y. and Wang, Y.-Y. Simulation of personalized english learning path recommendation system based on knowledge graph and deep reinforcement learning. *Sci Rep*, 15(1):34554, October 2025.

Zhou, X., Zhu, H., Mathur, L., Zhang, R., Yu, H., Qi, Z., Morency, L.-P., Bisk, Y., Fried, D., Neubig, G., and Sap, M. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *Proceedings of the International Conference on Learning Representations*, 2024.

Zhou, Y., Wang, H., He, J., and Wang, H. Review-based explainable recommendations: A transparency perspective. *ACM Transactions on Recommender Systems*, 3(3):1–20, 2025.

Zhou, Z. Empathy in education: A critical review. *International Journal for the Scholarship of Teaching and Learning*, 16, 11 2022. doi: 10.20429/ijsotl.2022.160302.

A. Appendix

A.1. Frequently Asked Questions (FAQ)

Q: Why are rights challenging to define?
A: Rights are challenging to define as different groups hold mutually exclusive conceptions of what counts as a right. Some examples of how different groups define rights differently include, for example, how the prohibition against torture is widely recognized as a human right under international law, yet its practical interpretation varies. Singapore permits judicial corporal punishment (United Nations Human Rights Office of the High Commissioner, 2019), which is condemned by the United Nations as a form of torture (United Nations Human Rights Office of the High Commissioner, 1984). Similarly, according to the Universal Declaration of Human Rights, men and women of full age, without any limitation due to race, nationality or religion, have the right to marry and to found a family (United Nations, 1948). However, same-sex marriage is legally recognized in Taiwan (Amnesty International, 2019) and Sweden (Government Offices of Sweden, 2018), but not in China (Reuters, 2016).

As Wilson observes, rights are "ideologically promiscuous", invoked across diverse political and moral positions without a shared underlying definition (Wilson, 2006). Rather than treating rights as fixed universal morals, Wilson argues that they are better understood in terms of what they do.

Q: How do you define "rights"?
A: As rights are challenging to define, this paper adopts the vernacular of "rights" as a term that identifies capabilities that humans risk losing to AI assistants, namely the "rights" to choose, to own, to work, and to learn. This paper does not seek to explicitly define what rights are, but instead presents descriptive research questions that shift the focus toward what AI systems should do within expectations of agency regarding everyday processes. This aligns slightly to Wilson's argument that rights are better understood in terms of what they do, and we use this direction to guide the operationalization of AI principles that safeguard agency when humans choose, own, work, and learn.

Q: Are the "rights" in this paper exhaustive, and what other "rights" can readers consider?
A: In this paper, we have focused on the "rights" to choose, own, work, and learn due to their relevance, pervasiveness, and urgency. Simultaneously, we recognize that they do not exhaust the ways AI systems affect human agency. There exist combinatorial "rights" (e.g., the "Right" to Learn at Work, the "Right" to Choose for Learners) and other rights that warrant attention. The "Right" to Safety aligns with Jobin et al. (2019)'s non-maleficence principle and captures concerns around deepfakes, scams, and other harms arising from AI-augmented manipulation. The "Right" to Dignity ensures that communities retain agency and dignity over how data or models represent them. The "Right" to Access/Participate ensures access to AI-supported opportunities, goods, and services, and inclusion of diverse stakeholders in co-designing AI systems. Each "right" merits a dedicated in-depth discussion, hence demanding separate future works.

Q: Are the research questions in this paper exhaustive?
A: The research questions are not exhaustive. Formulating good research questions is itself a difficult and underexplored problem, especially in an emerging area like AI and human rights. Rather than claiming to present an exhaustive set of questions, this paper attempts to identify what kinds of questions are worth asking at this stage. Our goal is to make explicit the lines of inquiry that matter, not to close them off.

Q: How were the research questions in this paper chosen?
A: The questions in this paper reflect a particular way of thinking about expectations of human agency and our capabilities in everyday processes. These are socially enacted and practically consequential, and are meant to be illustrative rather than complete. We see this work as a step toward a more systematic framework that could, in future work, help researchers generate and assess relevant questions in a principled way. When such a framework is more fully developed, it may provide guidance not only on which questions to ask, but also on how to derive new questions by following the same underlying logic.

Q: Why are insincere students excluded from the *"Right" to Learn* with AI? Shouldn't all students be entitled to the "Right" to Learn?
A: The *"Right" to Learn* with AI is premised on the learner's willingness to engage meaningfully in the process of learning.

Students who are not sincere in their intent to learn are unlikely to exercise this "right" responsibly, as their use of AI would not be directed toward understanding or growth. *Even in the absence of AI, such learners may find ways to circumvent learning (for example, by copying peers), and thus fall outside the scope of this framework.* Our focus is therefore on learners who demonstrate genuine intent and who can benefit from AI systems designed to support learning.

Q: The paper calls for a wide audience to work on research questions to develop AI that benefits humans. Is it practically feasible?

A: It may appear naive to think that researchers, developers, social scientists, humanities scholars, policymakers, and users of AI would want to work on these research questions, when real decisions are usually shaped by companies, organizations, and incentives like profit maximization. However, our hope with this paper is also that, as public awareness grows regarding the trade-offs between AI assistance and human agency[14], user expectations may increasingly shape the design of AI systems.

When users express sufficient concern regarding human agency or select competitor AI systems that support human agency, companies may be compelled to offer features that do not directly maximize profits, motivating work on the research questions mentioned in this paper. Such features include improved transparency, explainability, or the deliberate introduction of organizational "friction" in AI-supported workflows within the company. These measures can function as safeguards that preserve meaningful human oversight, even when full automation would be technically feasible. For example, Red Hat is now one of the world's leading providers of enterprise open-source software solutions. Although open sourcing software does not inherently maximize profits, sustained consumer preference has helped establish the transparency brought about by open sourcing as a competitive norm. Similar dynamics may arise for AI systems as users become more aware of the balance between human agency and AI assistance, motivating firms to design to support human agency even if they impose additional cost or organizational constraint.

Q: What are the promising extensions or future works of this paper?

This paper poses research questions; each research question can extend into one or multiple future works. Furthermore, we also recognise that questions of enforcement in AI governance are inherently difficult, particularly when decisions are shaped by powerful economic and organizational incentives. Therefore, following operationalization of AI principles, a natural next step for future research from the perspective of AI governance is to map the proposed "rights" onto existing human rights frameworks that already have legal and institutional enforcement mechanisms, shifting the discussion from what developers and researchers should do to what can be meaningfully enforced. Doing so would likely require involvement beyond individual companies, and include governments and broader policy tools such as regulation and taxation, rather than relying on voluntary compliance alone. This is beyond the scope of this paper.

Q: Since the paper is well-researched and well-cited, it is clear that each of the risks stems from an already-existing research program: some researchers are pursuing this research. Why is this paper still necessary?

A: We highlight that the way research is currently conducted is often misaligned with its intended real-world impact. Recent large-scale analyses (Duede et al., 2024) show that AI has rapidly diffused across disciplines, but remains loosely integrated, producing "semantic tension" between AI-driven and domain-specific research. For example, AI-driven research may emphasize benchmarks and predictive performance, while domain-specific research prioritizes explanation or interpretability. As a result, even when addressing the same problems, AI-driven research may not properly address risks to AI use; while these lines of work coexist in both domains, they may not meaningfully integrate, limiting downstream adoption and impact. This is evidenced in domain-specific papers for medicine (Hassan et al., 2024), law (Magesh et al., 2025), finance (Aldasoro et al., 2025), and education (Tian et al., 2025).

Our paper is necessary as it reorients how research is conducted, to balance AI assistance and human agency, to benefit humans. It highlights risks to human agency reflected by widely recognized AI principles and operationalizes these AI principles through research questions that welcome interdisciplinary collaboration. The research questions hence explicitly accounts for domain context, to achieve operationalization and meaningful adoption and impact.

[14] Users may increasingly feel *urgency* (Section A.3.3) to balance AI assistance with human agency.

A.2. "Rights" Framing

In this paper, we use the term "rights" to *articulate* the capacities that users should retain in choosing, owning, working, and learning in an AI-mediated world. By framing these capacities using the term "rights", we can map risks that undermine human agency to high-level AI principles, and operationalize these AI principles using research questions or technical solutions that directly alleviate these risks.

A.2.1. Legal Rights vs Our "Rights"

The authors of this paper recognize that the term "right" carries significant legal and political weight. *The use of the term "rights" in this paper does not depend on treating the "Rights" to Choose, Own, Work, and Learn as enforceable legal entitlements or as claims within a formal rights framework.* The history of international human rights demonstrates that even widely accepted rights can be contested, and the creation of new rights requires extensive normative justification and political consensus (U.S. Mission Geneva, 2017; Seah, 2022). We do not claim to define new human rights, nor do we seek to intervene in debates over their enforceability or legal interpretation. Instead, we borrow the vocabulary of rights for its clarity and ability to capture what everyday users of AI expect to preserve in their interactions with AI. Framing these expectations as rights enables us to articulate what is at stake, to identify specific risks to user agency, and to motivate research directions that reinforce these core capacities. The phrasing thus serves a practical function: it provides a concise and accessible way to communicate the normative commitments underlying our technical proposals, without invoking the full legal and political apparatus of human rights discourse.

A.2.2. Force of Normativity

The "rights" to choose, own, work, and learn do not draw on the term "rights" for their force of normativity. Instead, they draw on well-established normative foundations, particularly the global synthesis of AI ethics principles surveyed by Jobin, Ienca, and Vayena (Jobin et al., 2019). Jobin et al.'s analysis comprises 84 ethics guidelines documents from various organizations including private companies, academic and research institutions, governmental agencies, inter-governmental and supra-national organizations. It maps the global landscape of AI ethics and reveals strong convergence on a set of high-level principles, including transparency, justice and fairness, non-maleficence, responsibility, and privacy. Although the terminology in this position paper differs (in using the term "rights" rather than "principles"), the "rights" draw on the global consensus from Jobin et al.: that AI development efforts should integrate with the AI principles through substantive ethical analysis and adequate implementation strategies. These will empower users, protect their agency, and promote equitable, trustworthy, and beneficial outcomes.

A.2.3. "Rights" vs AI Principles

A key insight from Jobin et al. (2019) is that although AI ethics guidelines converge on shared values, they typically offer limited operational guidance for how those values should be implemented in practice (Jobin et al., 2019). Principles such as transparency or autonomy are often stated abstractly, without specifying how an AI system should behave or what technical mechanisms would make these principles actionable. The contribution of this position paper is in addressing this gap by reorganising these principles into "rights" and risks that can be resolved by answering the research questions proposed. The output-oriented research questions proposed translate high-level normative desires or expectations from Jobin et al. into concrete behavioral targets for AI systems. Rather than engaging in legal interpretation or philosophical argumentation which lie outside the primary expertise of scholars engaged in developing AI systems, this paper leverages their technical authority to show how these AI principles can be expressed as system requirements, research challenges, and design objectives.

A.3. Selecting the "Rights"

We focus on these four "rights" for three complementary reasons. First, their **relevance** (Section A.3.1): they map cleanly onto established AI-ethics scholarship, particularly Jobin et al.'s synthesis of global AI-ethics frameworks, which shows strong convergence around principles such as transparency, justice and fairness, non-maleficence, responsibility, and privacy. Second, their **pervasiveness** (Section A.3.2): these "rights" track everyday human capacities that AI systems are increasingly insinuating themselves into, shaping how people learn, work, own, and choose in routine and often unnoticed ways. Third, their **urgency** (Section A.3.3): the erosion of these "rights" is already producing tangible social harms, and failure to address them now risks deeper, more systemic consequences as AI adoption accelerates. Taken together, these considerations

motivate our choice of daily processes as impactful both in breadth and depth as they affect large segments of society bearing on fundamental aspects of human agency and well-being.

A.3.1. RELEVANCE

The four "rights" map cleanly onto established AI-ethics scholarship, particularly Jobin et al.'s synthesis of global AI-ethics frameworks, which shows strong convergence around principles such as transparency, justice and fairness, non-maleficence, responsibility, and privacy.

Transparency is foundational to all four rights because individuals cannot exercise agency over choosing, owning, working, and learning if AI systems operate opaquely. For the "Right" to Choose, transparency allows users to recognize when choices are being shaped or constrained by algorithmic curation rather than freely made. For the "Right" to Own, it enables tracing data provenance and authorship, which is essential for attribution and credit. In the "Right" to Work, transparency clarifies how AI aids in tasks in the workplace. Finally, for the "Right" to Learn, explainability helps learners understand answers and the reasoning process, preserving inquiry and reflection.

Justice and fairness ensure that AI does not systematically disadvantage particular individuals or groups across domains of everyday life. In choice, unfair recommender systems can systematically narrow exposure to viewpoints or opportunities for particular communities, undermining equal participation in cultural and civic life. In ownership, inequities arise when some creators experience appropriation of their work without recognition or compensation while others benefit. In work, unfair automation can disproportionately advantage certain roles or demographic groups. In learning, unfair AI tools can advantage some learners while penalizing others through unequal access to assistance or biased assistance.

Non-maleficence directly motivates the protection of these "rights" by foregrounding the harms or *risks* that arise when AI systems are deployed without restraint. Non-maleficence is most immediately salient in discussions of the "Right" to Choose, as this right directly confronts cases where AI systems are deliberately deployed by malicious actors to spread misinformation and disinformation, thereby causing harm to users' decision-making autonomy. Non-maleficence is also relevant to the "Rights" to Learn and Work where overreliance can result directly from the over-use of AI systems for these tasks. Safeguarding these "rights" is thus a practical instantiation of the obligation to avoid foreseeable harm.

Responsibility and accountability are not only AI governance requirements but also consequences of human agency as to act meaningfully is to be answerable for actions and their effects. It is also most salient in discussions of the "Right" to Work where workers are often compelled to use AI tools and the resulting outcomes expose how responsibility is diffused across employers, developers, vendors, and regulators, frequently without clear delineation. It is also related the "Right" to Own where the irresponsible use of data, such as when training generative models to create art, affects the livelihood of human creators. Preserving these rights therefore requires maintaining clear links between agency, action, responsibility, and accountability.

Privacy and data protection are enabling conditions for human agency because the abilities to choose, own, work, and learn freely depend on having spaces that are not fully observable, recorded, or optimized against. When individuals know that their data, behavior, or creative outputs are continuously collected and analyzed, they may self-censor, conform, or avoid risk, hence diminishing their capacity to act autonomously. This is most obvious in the "Right" to Choose where pervasive data harvesting enables hyper-personalized persuasion that constrains genuine freedom of selection. Protecting privacy therefore sustains the agency required to choose, own, work, and learn freely.

For brevity, the mappings of the various principles to the rights are summarized in Table 1.

Table 1. The mappings of the various widely-recognized AI principles (Jobin et al., 2019) to the "Right" to Choose, Own, Work, and Learn. The ○ indicates general relevance, while the ✓ indicates direct relevance, detailed in Section A.3.1.

AI Principles	Choose	Own	Work	Learn
Transparency	✓	✓	✓	✓
Justice and Fairness	✓	✓	✓	✓
Non-maleficence	✓	○	✓	✓
Responsibility	○	✓	✓	○
Privacy	✓	○	○	○

A.3.2. PERVASIVENESS

The four "rights" are pervasive because they correspond to capacities that are exercised continuously across the course of ordinary life. It is difficult to imagine living without learning or working, whether through formal education and employment or through everyday problem-solving and contribution. Likewise, creative expression is a natural human impulse (Boden, 2004) and the expectation that one can retain control or recognition over what one creates follows naturally. Lastly, choice is pervasive through engagement with online platforms that curate information, shape attention, and influence preferences. Together, these "rights" cut across age, occupation, and culture, making them not exceptional or aspirational ideals, but foundational dimensions of everyday human agency.

A.3.3. URGENCY

Across choosing, owning, working, and learning, the erosion of these "rights" is already manifesting in concrete and observable ways, underscoring the urgency of addressing them before they become further entrenched.

In moments of leisure, too, we increasingly surrender choice to AI algorithms that curate our For You Page, drawing upon our browsing history, prevailing trends, and even contextual factors such as geographical location (Bhandari & Bimo, 2022). In doing so, we place in AI's hands a quiet but profound power: the authority to choose and shape what we see. This control makes us vulnerable if there is little oversight or accountability to ensure that the content recommended is safe, constructive, or free from harmful influence (Tommasel & Menczer, 2022; Bagchi et al., 2024). The danger deepens in an age when it is often impossible to distinguish between AI-generated and authentic material (Gillham, 2024; Frank et al., 2024). Worse still, AI-generated content itself is not neutral, reinscribing narrow stereotypes (Luccioni et al., 2023) and forcing audiences to consume the same narratives while excluding diverse perspectives (Nicoletti & Bass, 2023). Human content is already fraught with partiality, often designed to provoke reaction; when such dynamics are reinforced by algorithmic curation, the risk arises of a vicious cycle where polarizing content is rewarded with attention and in turn amplified to attract ever more views (Gillani et al., 2018). Preventing this spiral requires vigilance and transparency but will enable the preservation of **The "Right" to Choose** the content we consume.

Generative AI has also demonstrated a remarkable capacity to compose text, images, and music that both rival and enrich traditional forms of art, expanding the imaginative powers of human creativity while lowering the barriers to artistic expression. It claims to democratize art by increasing the reproducibility of art, transforming viewers from passive observers into active participants or creators (Park, 2025). However, this very power sharpens the question of ownership: who rightfully claims authorship of such works? Should it be the individual who crafts the prompt, the artists whose creations supplied the training data, the subjects whose likenesses are reproduced, or the AI system itself as a novel creative agent? Without a principled assignment of ownership, the problem quickly unfurls. Original artists may be disincentivized to produce new works if their efforts serve only as fodder for AI training (Corrall, 2023); the commercial adoption of AI-generated art also deprives artists of salaried employment, displacing them from the industry (Merchant, 2023); others may retreat into privatization, restricting access to their art in fear of appropriation. These cases contradict the democratization of art and place art further out of reach of ordinary people who are denied the opportunity to experience and appreciate artistic expression in an open and shared domain. Safeguarding creativity in the age of generative AI therefore requires clear recognition and protection of **The "Right" to Own**.

In the workplace, AI is displacing fresh graduates (Simon, 2025) by automating white-collar entry-level jobs (Leopold, 2025). Office work once grounded in tacit know-how such as writing emails, analyzing data, and creating presentations are now readily learned by models trained on vast corpora. Consequently, AI can elevate the performance of new and lower-skilled employees even as it compresses the productivity premium of more experienced staff (Cazzaniga et al., 2024). A caveat is that these gains are uneven: where baseline competence is already high, marginal improvements diminish and skill decay accelerates (Macnamara et al., 2024; Brynjolfsson et al., 2023). Furthermore, widespread reliance on the same generative tools tends to homogenize ideation, causing the new idea to converge if everyone draws from the same engine (Anderson et al., 2024; Wadinambiarachchi et al., 2024). The net effect is a flattening of differentiation as the genuinely productive work of junior employees is discounted as commonplace, while the distinctive contributions of highly skilled workers are blunted by convergence toward median solutions. With the use of AI in the workplace, the distinctions between individual and AI-generated contributions are blurred and the availability of work and jobs are limited, threatening **The "Right" to Work**.

It is easy, even facile, to become over-reliant on an ever-present and patient tutor, raising concerns regarding the dulling of habits of inquiry and an erosion of learners' cognitive abilities (Zhai et al., 2024); and because they require minimal

effort (Bailey & Warner, 2024; Khan, 2024b), they become ripe for exploitation when completing assignments and examinations (Goodier, 2025; Chan & Lo, 2025). learners, too, have begun to sound like their tutors (Yakura et al., 2025)– systems trained on human language have incorporated their own cultural inflections, and the widespread use of such tutors narrows the space for genuine self-expression. Left unchecked, this drift erodes linguistic and cultural diversity (Yakura et al., 2025) and diminishes the articulation and formulation of independent opinions necessary for learning; and when models are optimized for being correct, it tempts learners to trade the slow work of reasoning and the virtue of making instructive mistakes for a shortcut to the correct answer. These put into jeopardy **The "Right" to Learn**.

A.4. The Need for Operationalization

Beyond addressing existing gaps in the literature on operationalizing AI governance frameworks, we also offer an additional motivation for operationalization: that technology is not value-neutral. Values are already embedded in the systems we build and use, and making them explicit allows us to shape their effects rather than leaving them implicit and unexamined.

Technologies are often treated as neutral tools whose social consequences arise only through use. However, a long line of work argues that technologies are never fully value-neutral: even when designers intend neutrality, values enter through design choices, assumptions, and the social arrangements required for deployment. As argued by Winner, certain technologies such as nuclear power are inseparable from specific forms of social organization, authority, and hierarchy that must exist for them to function at all (Winner, 1980). In this sense, technologies cannot be understood as apolitical artifacts, but as systems that embody and enact values. This observation motivates our approach: rather than debating which values technologies should reflect, we argue that embedded values must be made explicit and operationalized. Doing so is necessary not merely because technologies have social impact, but because values are already present—whether acknowledged or not—and therefore should be systematically examined and designed for.

B. Illustrative Solutions

B.1. The "Right" to Choose

B.1.1. Example Formulation of an Option Set

We define the set of options that AI systems can recommend to users as an *Option Set*.

Definition B.1. Let the *Option Set* be a finite set

$$\mathcal{O} = \{o_i\}_{i=1}^{n},$$

where each option o_i represents a possible source of information relevant to a decision problem, Y (for example, a task variable, a social media reel for a user with preferences, a result on a google search). The cardinality or number or *quantity* of available options is $|\mathcal{O}|$. A valuation function

$$\mathrm{val}_Y : \mathcal{O} \to \mathbb{R}_{\geq 0},$$

maps each (o_i) a non-negative value $\mathrm{val}_Y(o_i)$ that represents its informativeness with respect to task Y.

The *quality* of the options should consider how informative $\mathcal{O}$ is to Y.

While maximizing

$$\sum_{i=1}^{n} \mathrm{val}_Y(o_i)$$

may appear to yield an optimal option set since it aggregates the most informative options for the task, such an approach assumes that users will process the entire set of options presented to them. In practice, users may attend to only a subset of options, select a single alternative, or experience cognitive fatigue when confronted with overly dense or information-heavy sets. Moreover, evaluating options solely based on their individual informativeness overlooks interactions among options. As a result, it is not straightforward to curate an optimal option set, and highlights the need for more deliberate option-set curation.

For example, AI solutions that craft the option set can examine the relational structure among all or the top-k options, such as how distinct or complementary they are, and how they collectively frame the decision space. Understanding these relationships could reveal how to construct option sets that maximize informativeness and preserve diversity and interpretability, thereby improving decision making.

B.1.2. Illustrative Solutions for RQ1: How can AI systems present diverse and credible content that expands users' informational exposure, while preserving meaningful and sustained engagement?

AI systems structure and present limited sets of options for users to choose from (Mills & Sætra, 2024; Agan et al., 2023), and can be adapted to present diverse and credible content. For example, they can surface left-, center-, and right-leaning takes on the same story (AllSides, 2025) to encourage diversity or can integrate third-party fact-check labels to encourage credibility (Papakyriakopoulos & Goodman, 2022).

On presenting diverse content

AI systems may present diverse content to users *passively* by presenting alternative viewpoints to users, but not requiring users to explore them. For example, information may be clustered into multiple perspectives (AllSides, 2025), and users can select the viewpoints they are interested in, knowing that alternative viewpoints are available. Alternatively, AI systems may present diverse content to users *actively* by deliberately guiding users through a range of perspectives before being presenting information most aligned with their preferences. For example, a conversational agent could introduce users to contrasting viewpoints before summarizing content most relevant to their search history (Zhang et al., 2024).

A promising direction to better address the diversity aspect of this research question is to develop quantitative metrics for measuring content diversity and integrating them into AI systems. One such metric could utilize information theoretic frameworks such as Partial Information Decomposition (Williams & Beer, 2010; Griffith & Koch, 2014; Bertschinger et al., 2014; Ince, 2017) to quantify unique contributions from different sources of information. This has also been extended to multimodal domains (Yang et al., 2025b; Liang et al., 2023). Another is Pyrorank (Kilitcioglu et al., 2023), which increases

user-based diversity and mitigates systemic bias that traditional recommender system models learn from personalized data. Ultimately, we propose a direction where diversity is mechanistic and controllable to expand a user's exposure to different sources of information.

On presenting credible content

While presenting diverse content, AI systems must also ensure that this content is credible. This domain of discerning facts and biases has been a rich and active domain of AI research. For example, third-party fact-check labels can flag disputed or misleading claims to create a direct barrier against the spread of false or biased content (Papakyriakopoulos & Goodman, 2022). Counterfactual ranking tests, which estimate how different ranking policies would perform if content exposure were changed, can also be used. These tests help identify and correct for hidden feedback loops or popularity-based amplification in recommendations made to a user that might privilege extreme or biased views (Buchholz et al., 2024). Finally, aggregating independent reports (such as crowdsourced fact-checks or external credibility signals) and weighting them by the reputations of their sources aids in separating verified factual information from opinion or propaganda (Solovev & Pröllochs, 2025). Together, these methods have made significant progress to reduce biases within AI systems in content selection and presentation, fostering a more balanced informational environment. However, biases can evolve over time and emerge from multiple factors—such as demographic features—interacting with each other (Shah & Sureja, 2025). To reliably select and present credible content, developing frameworks or adaptive objectives to handle these emergent and dynamic biases is a promising direction direction for AI research.

B.1.3. Illustrative Solutions for RQ2: How can we design AI systems that users can trust?

The black box nature of AI systems prevents humans from fully understanding and interpreting its workings, even for computer scientists or those with specialized training (Burrell, 2016). Building meaningful reasoning and justification and communicating the intent behind AI outputs recommended to users can help them build trust in AI systems. Prior works have partially addressed the transparency concerns in this research question by including explainable recommendation models that provide differentiated levels of model transparency to explain how individual user preferences shape recommendations (Zhou et al., 2025) and mechanisms to seek simple (low complexity) and effective (high strength) explanations for the model decisions (Tan et al., 2021). However, current works have not sufficiently address the "trust" aspect of the system to make more informed choices. Surveys have shown that both under- and over-transparency can erode trust in the AI system's recommendations, especially when users have previously disagreed with the AI system's outputs (Kizilcec, 2016). A possible solution is to introduce customizable and adaptive transparency mechanisms and safeguards that account for user needs or the context of the situation.

For example, older social media users face risks of misinterpretation or misinformation due to differences in demographics, education, media literacy, or worldview (Brashier & Schacter, 2020), increasing their vulnerability to scams, health misinformation, or ideological manipulation. In this case, custom system safeguards may include filtering and warning mechanisms when emotionally manipulative content is detected. In these contexts, greater transparency may be necessary to help users understand how their behaviors interact with these safeguards and enable them to engage with social media without concern that the information presented is misrepresented.

As another example, adolescents and young adults excessively exposed to low-quality online materials (e.g., Italian brain rot) may experience emotional desensitization, cognitive overload, and a negative self-concept (Yousef et al., 2025). In such cases, system-level safeguards that limit compulsive or excessive scrolling behaviors (e.g. doomscrolling) may be necessary, alongside increased transparency to help users understand why certain interaction constraints or usage limits are applied, so that they may engage with social media without being subject to manipulative interaction patterns that exploit attention or emotional vulnerability.

As a last example, social media users may become trapped in echo chambers or filter bubbles as recommender systems increasingly adapt content feeds based on accumulated personal data and past interactions. In such cases, system safeguards may allow users to reset their personal data and escape echo chambers through a transparent system that explicitly explains the consequence of the user's action (Tan et al., 2023). For example, before a user selects content, the system could explain the potential consequences of this action on the user's future content feeds. If these changes are undesirable, users may choose not to proceed or provide explicit guidance to keep their content-feed unchanged, enabling users to regain control over their content feeds and removing undesirable recommendations.

B.1.4. Illustrative Solutions for RQ3: How can AI systems empower users to reject, adjust, or withdraw from algorithmic curation?

AI systems can empower users to directly adjust algorithmic curation. For example, obfuscation techniques can be used to block inference of user preferences (Slokom et al., 2021). Such an AI system that supports user-led "re-personalization" or personalization "resets" can, for example, help users generate videos that they may watch to reset a platform algorithm curated to them.

Existing approaches also adjust algorithmic curation by predicting user preference shifts over time. Preference shifts can be predicted by modeling preferences as sequences (Zhang et al., 2019), or by trying to understand why preferences change and how such changes influence user interactions with algorithmically curated systems (Wang et al., 2023). While these works have contributed greatly towards curated and highly personalized algorithms, their objectives are often aligned to retain user interactions and improve the user's experience.

As such, future works should consider addressing, more intentionally, the "user control" aspect of this research question by incorporating insights from multiple disciplines. For example, researchers should incorporate insights from behavioral researchers who study users' mental models of recommender systems (Ngo et al., 2020) to inform the design of user interfaces. This would allow users to better understand how recommendations are made and enable them to curate their algorithms. Building on this, future works could integrate these user signals as a source of feedback to tune AI systems. For example, AI systems can inform users when their feed becomes increasingly personalized, and allow users to signal or articulate the desire to reduce personalization (Wang et al., 2022).

AI systems can also support users to withdraw from algorithmic curation by providing neutral mode recommender views that temporarily disable personalization and revert to population-level baselines for recommended content. This is in line with the right for individual users to be forgotten (European Parliament and Council of the European Union, 2016). Recommender views that temporarily disable personalization signal to users that they have withdrawn from algorithmic curation, and make withdrawal a reversible, low-friction operation.

B.2. The "Right" to Own

B.2.1. Illustrative Solutions for RQ4: How can the authors of an AI-mediated artwork be identified and their contributions quantified to assign credit and ownership?

AI-mediated artworks are created through a collaborative process between human artists and AI systems. Sustained interactions among the artist, the AI system, and data, such as dataset curation to refine the model and iterative prompting, entangle authorship and complicates credit and ownership allocation (Bomba & De Angeli, 2025).

Beyond artists who are actively involved in the creative process, artists whose works are used as training data for AI tools may also be considered contributors. These contributions to the generated output can be quantified through data attribution methods, based on influence functions (Koh & Liang, 2017; Mlodozeniec et al., 2025), watermarking (Lu et al., 2025c), or retrieval from a separate datastore (Min et al., 2024).

However, many of these methods require open and known training data, which is often unavailable for large, proprietary foundation models. Without access to the training dataset, identifying contributors becomes infeasible. This limitation underscores the need for new legal frameworks to regulate the use, disclosure, and access of training data, to support clearer attribution for AI-mediated artwork (Hacker, 2021).

Naïvely allocating credit based solely on quantified contributions can overlook crucial nuances in creative work, especially on the significance of contributions. For example, contributions that lead to unique artistic expressions can carry more weight, even if the measurable changes are small, and should be allocated more credit accordingly. Development of metrics to account for these qualitative factors requires collaboration with artists, to leverage their expertise and judgment to ensure that credit allocation reflects both process and creative value. Structured credit allocation frameworks can also provide transparent and standardized ways to recognize contributions, via more finegrained roles and credit assignment, inspired by contributor role taxonomy for research (Brand et al., 2015). However, some contributions arise from interactions so interdependent that credit cannot be disentangled, and collaboration with game theorists could help model and fairly allocate these collective credit.

Credit allocation can inform ownership assignment, but the ownership of AI-mediated artworks remains legally ambiguous. Engaging with legal experts is critical where revisions to legal standards that clarify ownership and provide tailored

protections for AI-mediated artwork can be explored (Watiktinnakorn et al., 2023).

B.2.2. Illustrative Solutions for RQ5.1: How can we determine when AI-mediated stylistic influence deserves creative credit?

Recognizing and attributing intangible contributions in AI-mediated artworks may require assigning credit for AI-mediated stylistic influence. When such influence is acknowledged, credit may be attributed accordingly; when it is not, attribution can instead revert to the upstream or original creator. Attribution is conventionally credited based on expression, instead of style; therefore the question of whether stylistic influence merits credit should be informed by the perspectives of artists and legal experts. However, this determination can be facilitated by technical approaches that make style and expression more tangible and assessable:

Style. The concept of style in the research community has often been defined as surface-level characteristics like textures, color palette, and lighting, that can be disentangled from the semantic or compositional elements. Techniques such as Neural Style Transfer (Gatys et al., 2015), StyleGAN (Karras et al., 2019), and more recent methods such as SCFlow (Ma et al., 2025) exemplify how style can be disentangled. This aids artists and legal experts in clarifying whether, and in what form, each aspect of this disentangled style warrants attribution and credit.

Expression. In contrast, expression encompasses the overall creative realization of an artwork, including core components like semantic and compositional structures to express the creator's intent. Techniques like scene-graph representations (Li et al., 2024b; Wang et al., 2024b; Dutta et al., 2025), vision–language alignment models (e.g., CLIP, GLIP) (Radford et al., 2021; Li et al., 2022; Liu et al., 2024d), and semantic disentanglement frameworks (Shuai et al., 2024; Dalva et al., 2024; Yang et al., 2023) illustrate ongoing efforts to model and quantify the semantic and compositional structures that form a key component of artistic expression. The next step is to seek input from artists and legal experts to clarify whether each AI-mediated artwork exhibits expression that warrants attribution and credit.

We provide an additional concrete machine learning formulation by demonstrating how creative credit can be awarded to AI using Shapley value computations. In the example, the AI credit is the Shapley value of the AI contributor in a two-player setup.

We consider a sentence which has been written by author A (such as a human): "The acquisition function selects promising hyperparameters for evaluation.". This sentence is then modified (with modifications in italics) by author B (such as an AI): "The clever acquisition function selects expensive, promising hyperparameters for evaluation.". The Shapley values (ϕ) for authors A and B are, respectively:

$$\phi_A(v) = \frac{1}{2}(v(A) - v(\emptyset)) + \frac{1}{2}(v(A,B) - v(B)) = 0.5 \times 0.20 + 0.5 \times (0.75 - 0.65) = 0.15 \quad (1)$$

$$\phi_B(v) = \frac{1}{2}(v(B) - v(\emptyset)) + \frac{1}{2}(v(B,A) - v(A)) = 0.5 \times 0.65 + 0.5 \times (0.75 - 0.20) = 0.6 \quad (2)$$

Where v is the value function, and as a baseline, we set the empty sentence to have a value of 0. Therefore, $v(\emptyset) = 0$. The normalized human credit is $\frac{0.6}{0.6+0.15} = 0.8$. The normalized AI credit is $\frac{0.15}{0.6+0.15} = 0.2$.

In the experimental setup for this machine learning formulation, the value function v uses LLM-as-a-judge. The prompt, used on `qwen2.5:7b-instruct`:

```
You are an objective evaluator.

Your task is to assign a VALUE score to a sentence,
focusing on the contribution of a specified source.

DEFINITION OF VALUE:
The value is the extent to which the specified
contribution improves the sentence in terms of:
1. Semantic clarity
2. Informational usefulness
3. Appropriateness
```

```
SCORING RULES:
- Score must be between 0 and 1.
- 0 = no contribution or harmful contribution
- 1 = highly valuable contribution
- Evaluate ONLY the contribution from the specified
source.
- Ignore contributions from other sources.
- Prefer lower scores if uncertain.

OUTPUT FORMAT:
Return ONLY a single number between 0 and 1.
--------------------------------
Sentence: {sentence}
```

As this is an example machine learning problem, we also non-exhaustively suggest future directions for research:

- The number of contributors can be increased to more than 2.
- Multimodal contributions can be considered, such as text, speech, and images.
- Shapley value computations are compute intensive, especially if the contribution is computed per word or pixel or token, and if the contribution is longer (e.g. essays, movies). Therefore, approximations can be considered.
- LLM-as-a-judge can be replaced with a better metric, possibly in consultation with interdisciplinary expert (e.g., lawyers, animators).

B.2.3. Illustrative Solutions for RQ5.2: How can we quantify the expressive overlap between an AI-mediated artwork and the original?

The extent of expressive overlap between two or more artworks can help generative AI rebuke, or help artists prove, claims of copyright infringement. Recent works (Karras et al., 2019; Ma et al., 2025; Yang et al., 2025a; Zhang et al., 2025a; Qi et al., 2025) have demonstrated the feasibility of separating the semantic meaning of an image from its stylistic representation. Similarly, scene graph–based methods (Li et al., 2024b; Wang et al., 2024b; Dutta et al., 2025) capture relationships between visual entities and their compositional structure, providing a pathway to trace the creative expression that emerges from spatial arrangements. These methods can be extended to identify copyright-relevant features such as composition, subject framing, or symbolic motifs, that reflect the artist's individual creative choices rather than generic stylistic conventions.

B.2.4. Illustrative Solutions for RQ6: What can authors do to protect their works prior to legal judgment?

Authors currently have to rely on legal interpretation and subjective judgment to prove that their work was copied and have no tangible technical mechanism to detect, or demonstrate copying. For example, Singaporean photographer Zhang Jingna sued Luxembourg artist Jeff Dieschburg who had painted her 2017 Harper's Bazaar Vietnam photograph and submitted it to an art competition, winning a prize. However, the case initially ruled against her, and this was only overturned two years later in 2024 (de Leon, 2024).

One attempt to protect works prior to legal judgment is to have creators embed watermarks within the creative process. These will enable creators to detect and assert ownership even when works are altered or their origins obscured as the watermark should persist in these derivative works. However, black-box and beige-box attacks involving diffusion models, noise injection, and leveraging sematic priors can achieve near-perfect watermark removal (95.7%) with negligible impact on the residual image's quality (Shamshad et al., 2025). Consequently, ensuring the robustness of watermarking-based methods remains an open research challenge.

Authors may also have had their works "scraped" (lawfully or unlawfully, with or without permission) (Chesterman, 2024), but lack a tangible technical mechanism to detect this. In the absence of legal judgment, creators may instead rely on technical deterrence mechanisms to protect their works by introducing perturbations or poisoning into their works to disrupt models trained on them. This can significantly degrading performance (Shan et al., 2024; 2023; Liu et al., 2024e) and

discourages unauthorized scraping of protected works for training datasets. However, recent work (Foerster et al., 2025) shows that images can be depoisoned, and protecting works from "scraping" remains an open research challenge.

B.3. The "Right" to Work

B.3.1. Illustrative Solutions for RQ7.1: How can AI systems detect/learn optimal moments to intervene in a task (vs. leaving it to the human)?

We can look towards two scenarios: one where an AI system selectively intervenes in a human worker's task, and another where an AI system selectively defers to a human decision maker. In both, we want to balance AI assistance and human autonomy to achieve better task performance while preventing over-reliance on AI systems.

In the first scenario, humans are the main drivers in a task and AI systems can intervene when the worker's performance drops. AI systems can detect signs of impasse such as repeated errors, prolonged inactivity, or inconsistent reasoning patterns that signal the need for AI assistance, perhaps due to fatigue. These signals are often multimodal. For example, signals from electroencephalography (EEG) and eye activity detectors are used for driver fatigue detection (Sikander & Anwar, 2018). Beyond static thresholds for intervention, AI systems can learn to assist the worker when the expected value of the agent's action exceeds that of the human across defined goals (McMahan et al., 2024).

However, these methods optimize the AI to complement a fixed offline model of a human. This disregards, for example, the ability of the human to learn, or even a tendency for the human to detrimentally cognitively offload tasks (i.e., effects of cognitive offloading (Risko & Gilbert, 2016; Grinschgl et al., 2021; Cavicchi et al., 2025)), changing the optimal moment to intervene. Our work argues that solving RQ7.1 should also include behavioural experts such as psychologists, who can better balance the trade-offs between AI as a helpful assistant and human agency, and mitigate detrimental effects of AI use in the workplace. Such advice can also inform AI researchers in building models that benefit the user, and future work can better address this research question by integrating findings and perspectives from other disciplines to better quantify and balance out the positive and negative effects of AI intervention in these systems.

To give a proof-of-concept on how RQ7.1 can be translated into research problems for the ML community that can also involve interdisciplinary experts, we demonstrate how AI systems can learn optimal moments to intervene in a task by detecting user uncertainty. This work is inspired by existing works on verbalised confidence (Groot & Valdenegro Toro, 2024; Li et al., 2025). Using the corbt/all-recipes dataset, we create two sets of instructions per recipe for 500 recipes using the following prompts and save them as confident/unconfident instructions:

```
def make_confident_prompt(recipe: str) -> str:
    return (
        f"Recipe: {recipe}\n\n"
        "You are a senior chef verbally instructing a junior chef "
        "on how to prepare the given recipe. "
        "Ensure your verbal directives are clear, and include sensory "
        "cues the junior should look out for."
    )

def make_imposter_prompt(recipe: str) -> str:
    return (
        f"Recipe: {recipe}\n\n"
        "You are a senior chef verbally instructing a junior chef "
        "on how to prepare the given recipe. "
        "Avoid committing to clear directions, give vague or unsure "
        "measurements, unclear sequences, and overuse filler words."
    )
```

Then we combine the instructions (500 confident + 500 unconfident) and shuffle them, and use LLM-as-a-judge to classify whether the chef is confident. We achieve accuracy of 88.6%. A domain expert like a psychologist should advise on, e.g., other cases when, or how the AI system should intervene, and how to intervene in a way that can improve the performance of the less confident chef in the long run.

The second scenario occurs when AI systems are used to automate decisions and uncertain decisions must be deferred to a human expert. This ability to defer decisions to a human expert is often critical in high stakes settings such as healthcare and cybersecurity where experts can consider factors inaccessible to the AI (Gao & Yin, 2025), such as a patient's medical history, to make an informed diagnosis. Current methods for expert deferral include learning a rejector (Charusaie et al., 2022), and a routing policy (Gao & Yin, 2025). Decisions can also be routed to multiple human experts (Keswani et al., 2021), and with personalized policies that accounts for each worker's unique skill-sets and preferences (Bhatt et al., 2025).

B.3.2. Illustrative Solutions for RQ7.2: How can AI systems intervene in ways that complement their users' efforts and expertise?

To intervene in ways that complement their users' effort and expertise, AI developers must work with psychologists to understand how AI tools can improve work quality rather than merely completing tasks. A central insight from work psychology is that effective human-AI collaboration is not about substituting human judgment, but about designing assistance so that the combined human-AI system can achieve outcomes that neither could reliably achieve alone. Improved performance can emerge from how responsibilities, representations, and decision-making authority are distributed between human and AI (Woods, 1986).

An illustrative solution would be to design an AI system that is able to understand the social context, to know how to intervene to complement their users' effort and expertise. The AI system can integrate multimodal signals, such as speech tone, facial expression, gesture, text, gaze, and environmental context, to infer intent and adaptively select actions to assist users while considering interpersonal nuance. Social perception facilitates human-AI collaboration (Puig et al., 2021), and recent multimodal foundation models can already fuse such cues across modalities, offering a basis for "socially grounded" learning (Lee et al., 2024). However, social intelligence in a workplace context also requires conceptual input from social and behavioral sciences, such as theories of communication competence (Coffelt et al., 2019), personality and affect in collaboration (Sackett & Walmsley, 2014), and situated cognition. Combining these perspectives with simulation environments such as SOTOPIA (Zhou et al., 2024; Wang et al., 2024a) could allow agents to learn and evaluate social reasoning in realistic, multimodal, and goal-directed settings.

Beyond low-level assistance, interventions can also be planned via dynamic human-AI team workflows where the AI system acts as a manager agent to decompose complex tasks given by humans, allocate and coordinate subtasks based on the complementary strengths of humans and AI, and adapt to changing preferences (Masters et al., 2025). Such workflow-level human-AI collaboration can lead to more optimal outcomes than those achievable by humans or AI systems alone.

B.3.3. Illustrative Solutions for RQ8: What strategies can help organizations ensure that AI is used appropriately and sparingly so that employees continue to develop their own skills?

Drawing inspiration from cybersecurity practices, organizations can develop AI-use governance frameworks that balance innovation with accountability. In such a framework, employees periodically complete small and well-defined tasks that they are expected to perform without AI assistance. These tasks can serve as competency checks, analogous to white-hat penetration tests in cybersecurity. Employees who consistently rely on AI for tasks that should reflect personal expertise or domain understanding can be detected through watermarking (Lau et al., 2024; Lu et al., 2025a) or provenance tracking (Nikolic et al., 2025). This employees can then be enrolled in targeted retraining modules that reinforce skill development and responsible AI use. In doing so, this creates a culture where white-hat behavior (ethical and transparent use of AI) is rewarded, grey-hat behavior (borderline or excessive dependence) is monitored, and black-hat behavior (misuse or concealment of AI use) is subject to corrective education.

B.3.4. Illustrative Solutions for RQ9: How can AI systems support fair responsibility allocation in human–AI collaborative work?

Matthias (2004) identified a responsibility gap arising from AI systems as traditional frameworks are unable to assign moral or legal responsibility to either human users or AI developers. Recent works have attempted to address this by reframing responsibility as being diffused across multiple AI and human agents (Bleher & Braun, 2022), and formalizing it with a structured causal responsibility attribution framework to allocate responsibility in human-AI collaboration (Qi et al., 2024).

However, fair responsibility allocation remains challenging as it is dynamic and it also evolves alongside technological, social and legal developments. For example, responsibility allocation should be adjusted when AI systems communicate

their uncertainty (Groot & Valdenegro Toro, 2024; Lau et al., 2025): if users act on highly uncertain AI recommendations, more responsibility may fall on the user; conversely, if the AI system's uncertainty is poorly calibrated or inaccurate, the AI system should bear greater responsibility.

B.3.5. Illustrative Solutions for RQ10: How can AI systems enable a junior employee to reach the competency of a senior employee?

One illustrative solution is a human-in-the-loop mentoring system that can turn everyday engineering workflows into training and teaching signals. For example, when a junior employee writes a code review, incident report, or project proposal, the AI system can suggest revisions based on corpora of high-quality senior examples, preserving tone, clarity, and diplomacy. Senior human reviewers still sign off and their corrections are used to further refine the model's coaching ability. An AI system that complements the guidance of senior employees, by monitoring activities of junior employees in a team, detecting misalignments with task goals, and providing targeted interventions, can enhance teamwork and improve overall team performance (Seo et al., 2025). AI-supported personalized and real-time feedback can accelerate the development of junior staff toward senior-level expertise.

While more coaching data is important for training a better AI system, these data are often sensitive and cannot be shared freely across organizations or teams. To enable aggregation of expertise without centralizing sensitive data, federated learning provides a promising approach. Each organization can train locally on its own data, such as codebase, review comments, incident post-mortems, design docs, and communication logs. A global model based on the compiled experience of many senior employees is then updated via federated learning, sharing parameter updates rather than raw data (Kuang et al., 2024; Shu et al., 2025). This lets the system learn general patterns of good engineering practice drawn from many senior engineers that can be used to guide junior engineers.

B.3.6. When can AI use be preferable in the workplace?

AI use can be preferable in specific workplace contexts where it complements or outperforms human efforts and expertise by reducing cognitive burden, mediating complexity, or expanding the space of possible solutions. In particular, AI systems can productively intervene to support communication, creativity, and complex trade-off reasoning.

RQ13. **How can AI systems ease communicative loads and reduce communication barriers in the workplace?**
Communication is central to collaboration but is often complicated by tone, emotion, language, and context. As AI systems become integrated into workplaces, they can mediate interactions between co-workers to improve clarity and reduce friction. Socially and culturally aware AI systems can recognize tone and sentiment, support more empathetic exchanges, and reduce language barriers in multicultural environments. This support may take the form of multilingual capabilities, particularly for scarce languages, such as those supported by SEA-LION (Ng et al., 2025)—as well as multicultural awareness that adapts communication to diverse norms and preferences. By intervening in these ways, AI systems can prevent communication breakdowns and foster more cohesive and effective collaboration.

RQ14. **How can AI systems support creative problem solving and idea generation?**
AI systems can also be beneficial in creative tasks when they expand, rather than collapse, the space of ideas. Instead of defaulting to generic or standardized solutions, AI can support workplace brainstorming by generating diverse creative pathways, building upon a worker's initial ideas, and co-creating novel directions. This is particularly important given evidence that current AI systems risk homogenizing outputs and suppressing originality (Kosmyna et al., 2025; Doshi & Hauser, 2024). To mitigate this risk, we envision AI systems acting as collaborative colleagues that offer alternatives rather than black-box oracles with gold-standard answers. Achieving this requires models with strong capabilities for generating diverse and coherent ideas—a challenge even for current language models (Zhang et al., 2025b). Benchmarks such as NoveltyBench (Zhang et al., 2025b) and metrics like the Creativity Index (Lu et al., 2025b) provide starting points for evaluating and developing AI systems that facilitate idea generation without converging creativity.

RQ15. **How can AI systems learn to help workers handle complex trade-off problems?**
Many workplace tasks involve complex trade-offs where multiple goals and constraints must be balanced at once. Such problems are cognitively demanding for humans, particularly when constraints are numerous, conflicting, or dynamic, and when they involve both technical requirements and human values and preferences such as fairness, satisfaction, or well-being. These also have multiple conditions that are hard to keep track of and weigh simultaneously, especially when priorities change over time. In workplaces, these constraints are not purely technical and can include human values, fairness, and satisfaction. AI systems can be clearly advantageous in such tasks involving complex trade-offs where multiple goals

and constraints must be balanced simultaneously. AI systems can assist by systematically tracking constraints, modeling trade-offs, and identifying balanced solutions that account for both hard rules and soft factors. For example, hospital scheduling requires balancing patient care needs, staff workload, and resource availability—an optimization problem that AI can handle more reliably while reducing human cognitive burden.

B.4. The "Right" to Learn

B.4.1. Additional Risks for Section 5

Risk: Loss of Individual Agency, Reasoning, and Understanding. Recent studies show that AI use initially enhances learning and higher-order thinking, but prolonged use can weaken cognition and knowledge retention (Sung et al., 2017; Tian & Zheng, 2024; Wang & Fan, 2025). Learning involves reasoning and recall, such as considering multiple sources of information, forming own conclusions, and identifying mistakes. By offloading these to AI, learners lose opportunities to hone these skills and engage meaningfully in learning without assistance, negatively impacting decision making and encouraging procrastination and "laziness" (Ahmad et al., 2023; Kosmyna et al., 2025). Furthermore, the immediacy of AI-generated responses shortcuts learning by making it feel effortless, hence discouraging engagement, independent understanding, and reasoning. This is exacerbated by educational technologies that frame learning as "effortless" (Mentutor, 2024; Shabanov, 2025). As learners offload learning to AI assistants and lose ownership of individual and independent understanding and reasoning, they cease to exercise individual agency (Agarwal, 2023; Gerlich, 2025; Kosmyna et al., 2025).

Risk: Homogenization in Learning. AI learning systems tend to promote homogenized forms of knowledge and expression that are disproportionately aligned with Western norms (Naous et al., 2024; Yu et al., 2025), raising concerns about the perpetuation of knowledge and expression shaped around a single dominant norm. Firstly, this diminishes learners' abilities to interpret knowledge from their unique perspectives, conforming ideas generated across learners (Agarwal et al., 2025; Moon et al., 2025). These homogenization effects persist even after learners stop using AI models (Liu et al., 2024b). As learners lose the capacity to explore their own perspectives, learning ceases to be personally meaningful and engaging. Secondly, homogenized instructions can deter learners due to a possible cultural and linguistic mismatch. This undermines the learner's sense of belonging and adversely impacts communication, engagement, and comprehension, which are essential for effective knowledge transfer (Fink, 2023; Gay, 2018; Steele & Cohn-Vargas, 2013).

B.4.2. Illustrative Solutions for RQ11: How can AI systems be designed to close or even bridge the gap between teachers and learners?

Pedagogical research shows that teachers who can understand their students' learning progress, empathise with, and get a greater knowledge of the individual needs of their students, can support their learning better (Zhou, 2022; Meyers et al., 2019). However, for teachers to go through all learner's questions and responses from an AI system would be too tedious. Thus, such a solution can tap on existing literature on teacher-facing learning analytics dashboards (Kim et al., 2024), where researchers and teachers can craft modular or meaningful summaries of learner's questions, and efficiently view the scope of responses given to them(Ryu et al., 2024; Peper et al., 2024).

Alternatively, for teachers who want to align AI support with upcoming lessons, we propose a simple classifier that can help teachers classify questions based on their topic. We attempt to classify questions from the `HuggingFaceH4/MATH-500` dataset using the `qwen2.5:7b` model with the following prompt, and achieve an accuracy of 0.4380 on the 500 samples in the dataset. The confusion matrix can be found in Table 2.

```
def build_prompt(problem: str) -> str:
    topics_str = "\n".join(f"- {topic}" for topic in TOPICS)

    return (
        "You are classifying math contest problems by topic."
        "Choose exactly one topic from this list: \n\n"
        f"{topics_str}\n\n"
        "Return only the topic name. Do not explain.\n\n"
        "Problem:"
        f"{problem}"
```

```
        "Topic:"
        )
```

Predicted	A	C&P	G	IA	NT	PA	PC	NaN
Gold								
A	77	2	8	4	3	12	15	3
C&P	1	28	7	0	2	0	0	0
G	0	0	40	0	0	0	0	1
IA	54	1	11	9	5	0	9	8
NT	7	0	0	0	37	12	1	5
PA	19	11	20	1	8	16	7	0
PC	8	0	24	0	0	0	12	12

Table 2. Confusion Matrix. The classes include Algebra (A), Counting & Probability (C&P), Geometry (G), Intermediate Algebra (IA), Number Theory (NT), Prealgebra (PA), Precalculus (PC), and NaN.

B.4.3. Supplementary RQs for RQ12

RQ16. **How can AI systems measure a learner's depth of understanding and dynamically calibrate its support accordingly?**
Adjusting the depth and complexity of an AI system's support to match a learner's knowledge level can foster meaningful understanding (Zerkouk et al., 2025). Consider an AI system that first measures the learner's understanding through interactive questioning and analyzing patterns in the learner's explanations and revisions, and then interprets these signals using pedagogical tools and assessment frameworks developed in collaboration with social scientists like educators and pedagogists. From the measure, the pedagogist can control the AI system to dynamically provide information at the right level for the learner's understanding. This can be done through a "knob" to calibrate the depth and complexity of the AI system's responses by, for example, cultivating learning through dialog and explanation. This allows the AI to support the access to and the understanding of knowledge by leveraging the pedagogist's expertise, and to uphold the Right to Learn. See Section B.4.4.

RQ17. **How can AI systems strategically allow learners to make mistakes and guide them in examining assumptions and identifying errors to develop robust reasoning?**
Making and identifying mistakes allows learners to detect inconsistencies between their mental models and new information, triggering repair of inadequate mental models and improving understanding for learning (Chi, 2000). Consider an AI system that can strategically intervene or withhold correct answers during learning, so as to guide learners to ask clarifying questions, examine assumptions, and reflect on counterfactual scenarios. This helps learners think from alternative perspectives, prompting them to notice inconsistencies, gaps, ambiguities, and assumptions or mistakes in their reasoning. Pedagogical expertise can provide guidance on the timing and form of these interventions and when mistakes should be revealed. Psychologists can also advise how these interventions can be introduced. By engaging learners in these forms of inquiry and reflection and forcing them to take a more effortful path towards construction of knowledge, AI systems can aid reasoning and understanding and help learners take active responsibility and practise individual judgment when learning. See Section B.4.5.

RQ18. **How should AI systems allow learners to choose their learning strategy and align their reasoning and explanations with this choice?**
Learners benefit when they are given control over learning strategies or solution paths, provided that the system constrains, guides, or contextualizes those choices (Zerkouk et al., 2025). For example, in learning mathematics, learners may prefer specific mental models such as algebra, model-drawing (Ng, 2022), or intuitive reasoning. Pedagogical expertise is required to identify the range of diverse and valid strategies. By reasoning from the learner's perspective, AI systems can adapt to the learner's understanding and context rather than impose standardized explanations, keeping learners engaged and preserving individual agency. Learners from different cultural contexts may also want to choose different learning strategies. Consider an AI system that is designed to accept diverse cultural perspectives when developed alongside pedagogical experts from diverse cultural contexts. This can prevent learners from converging to the same way of thinking or from being shaped by a single dominant perspective embedded in the AI, preventing intellectual homogenization. See Section B.4.6.

B.4.4. Illustrative Solutions for RQ16: How can AI systems measure a learner's depth of understanding and dynamically calibrate its support accordingly?

Personalizing support such as instruction and feedback from an AI system to each learner's needs (e.g., based on their level of understanding) can improve student performance (Zerkouk et al., 2025). For example, adaptive assessments (Vie et al., 2017) where questions and tasks are actively selected according to the learner's responses, can be leveraged to lead to more efficient and accurate assessments of a learner's understanding. AI systems can enhance adaptive assessments via retrieval-augmented generation (RAG) (Lewis et al., 2020), by collating a comprehensive knowledge base with assessments of various difficulty levels. At each turn, generating questions that are appropriately challenging can gather more information on the learner's understanding level. Rather than numeric metrics that often miss qualitative nuances, AI systems can possibly also capture the learner's depth of understanding more holistically, via implicit representations in personalized models (Ning et al., 2025).

These approaches to calibrate an AI system's support to a learner's knowledge must be grounded in existing pedagogical frameworks. For example, scaffolding-guided training for pedagogical alignment has been shown to provide adaptive guidance to learners of different levels (Liu et al., 2025) and can be used by pedagogists to personalize the treatment of learners after evaluating their level of understanding. Pedagogists may also evaluate assessments following cognitive depth taxonomies (Anderson & Krathwohl, 2001; Black & Wiliam, 1998) to personalize effective learning objectives for learners. AI systems can then be aligned with the identified learning objectives via personalized alignment (Chen et al., 2024) to the pedagogist's recommendations.

Another interesting problem for the research audience is that as a learner acquires new knowledge, the data distribution shifts and the AI system must continuously update its internal representations to match the learner's level. This may interest researchers working on continual learning methods (Shi et al., 2025) where a model learns from new data without full retraining. Crucially, this adaptation should not be driven by the model alone. Pedagogists can play an active role in advising how the system ought to be calibrated to the learner at different stages of development. More interestingly, drawing on their understanding of learning trajectories, pedagogists may even anticipate how the data distribution is likely to shift, thereby informing when and how such updates should occur.

We note that learner modeling has been studied for many years and has already led to practical applications (Liu et al., 2024c; Gao et al., 2021). However, this remains an area of interest as there is a gap between technical operationalization and pedagogically grounded outcomes (Oh & Ahn, 2024), and existing works point to mixed results about the effectiveness of AI-based tutoring systems (Zerkouk et al., 2025).

AI systems must therefore partner with pedagogical insights to meaningfully calibrate to a learner's depth of understanding. For example, AI systems can be used to collect quantitative metrics like test scores at scale from students. However, the AI, possibly trained on inputs from a pedagogist, should interpret the scores to discern the learner's depth of understanding or evaluate other qualitative metrics on student performance based on their experience. These can then be used to dynamically calibrate the AI system or set priors regarding learners (like adjusting a "knob").

Existing works, while done with good intention, leave pedagogical experts out of the loop (Oh & Ahn, 2024). For example, they may focus on improving the accuracy of learner modeling, but leave the connection between model representations and how educators define or intervene on learning under-specified (Liu et al., 2024c; Gao et al., 2021).

B.4.5. Illustrative Solutions for RQ17: How can AI systems strategically allow learners to make mistakes and guide them in examining assumptions and identifying errors to develop robust reasoning?

Pedagogical theory suggests that making and identifying mistakes allows learners to detect inconsistencies between their mental models and new information, triggering repair of inadequate mental models and improving understanding for learning (Chi, 2000). A possible solution is to align AI tutor systems to teachers instead of students. AI systems can align to teachers with the desired pedagogical abilities using reinforcement learning, which is widely used for LLM alignment (Ouyang et al., 2022; Rafailov et al., 2023; Lambert et al., 2024). Reward signals can be collected in collaboration with educators and pedagogists, where responses with contrastive dialogues that invite reflection of plausible alternatives and surface assumptions can be scored higher than responses that directly provide an answer. AI systems trained on these rewards will be more pedagogically aligned to the teacher.

A pedagogist may also be able to advise on intervention timing and intervention strategies to improve learning outcomes.

Teaching strategies like withholding from correcting a learner's mistakes and guiding learners to reflect and identify their own mistakes are capabilities that pedagogists are familiar with. AI systems will have to be developed to engage in multi-turn dialogues and to plan conversations, selecting responses based on pedagogical expertise that are most likely to lead to the learning goal after multiple turns (Chen et al., 2025). Pedagogically-aligned evaluations, which include assessing if AI systems can delay answer reveal (Maurya et al., 2025), are also crucial to ensure that AI systems are exhibiting the correct teaching behaviors. These evaluations can be certified by pedagogists and periodically updated, to keep in line with the latest pedagogical frameworks.

Researchers from multimodal (visual, audio, etc) AI can also be engaged in this effort. Pedagogical works suggest that learners benefit from well-designed multimodal input (Paivio et al., 1968). Multimodal cues based on pedagogical expertise can also be used to guide learners towards reflective diagnosis of their own errors. For example, adding graphical hints to diagrams as feedback when learners incorrectly solve a problem can successfully guide learners towards the correct solution (Rouinfar et al., 2014; Wu et al., 2025). Another example is syntax highlighting (Klock & Jan, 1986) where code elements are displayed in distinct colors to indicate their syntactic roles. When a segment appears in an unexpected color, the user immediately perceives an anomaly and can infer that something is wrong. Similarly, when teaching text-based reasoning, the system can signal logical or structural inconsistencies by color-coding arguments, counter-arguments, or key claims, without revealing the correct interpretation (Ruiz-Dolz et al., 2025; Liu et al., 2023). As with the other imagined solutions above, this should be developed alongside pedagogists, for example by engaging them to identify which logical or structural inconsistencies to signal, or what aspects of a text should be highlighted to guide learning.

We note that guided educational support (Liu et al., 2024a), such as through game-ification of learning (Kanervisto et al., 2025), has been studied for many years and has already led to practical applications. However, recent works suggest that the role of teachers and pedagogists in developing such educational support remains limited (Tan et al., 2025). This has highlighted a gap in research addressing the AI development needs of teachers as they integrate AI technologies into their teaching practices and exploring how AI technologies can be applied in education from both the perspectives of student learning and teacher instruction (Tan et al., 2025).

B.4.6. Illustrative Solutions for RQ18: How should AI systems allow learners to choose their learning strategy and align their reasoning and explanations with this choice?

Learning strategies differ across contexts. For example, model-drawing is the method of choice for teaching mathematics in Singapore (Ng, 2022), but other educational systems may prefer specific mental models such as algebra or intuitive reasoning. Recent works suggest a gap in research addressing how AI technologies can be applied in education from both the perspectives of student learning and teacher instruction (Tan et al., 2025), and many works that operationalize AI tools for learning rely heavily on evaluating learning strategies or learner performance via model-estimated learner progress rather than externally validated learning gains with pedagogists or teachers (Zhou & Wang, 2025). As a result, current works risk ignoring pedagogical input regarding which learning strategies may be available or even optimal for the learner.

To overcome this, one illustrative solution is to explicitly ask learners how they would like to learn, to enable learners to customize their learning experience and be active participants in the learning process. AI systems should have a list of pedagogically validated learning strategies to rely on depending on the learner's choice, and should remember and adhere to the stated preferences. As current AI systems may experience difficulty following stated user preferences in a zero-shot setting (Zhao et al., 2025), current methods to improve preference following include prompting strategies that remind AI systems to consider previously stated preferences (Zhao et al., 2025), and extracting relevant information from history using RAG (Lewis et al., 2020). Preferences, such as on which learning strategy to use, can evolve dynamically, and AI systems should capture and align to these evolving preferences, especially in long-context conversations (Jiang et al., 2025).

The choice of learning strategy may also be context or culture dependent. If the learner does not explicitly request a certain learning strategy, then given the profile information of a learner, implicit choices such as cultural preferences can also be inferred, and AI systems can adapt their learning strategies accordingly. AI systems that are culturally aware and encapsulate cultural differences can mitigate bias towards specific cultural perspectives by aligning their responses better with the learner's cultural context (Li et al., 2024a). This would allow the AI system to choose the best, pedagogically validated, learning strategy for the learner.

Sometimes, the pedagogist or teacher may choose to customize the learning experience on behalf of the learner, based on their teaching experience and expertise. To avoid duplicating the preceding argument with only role labels altered (i.e., changing learner to teacher), we provide a simple proof of concept. We consider a teacher using a large language model

to teach mathematics to grade 5 students. The teacher carefully writes a prompt detailing how the large language model should explain the answer of various mathematics questions to a grade 5 student, bearing in mind that the student only has knowledge of introductory algebraic thinking and is fluent in multi-digit multiplication and division. We compare the response to the teacher's prompt against a basic/ baseline prompt. We observe that such teacher prompts are useful for structuring (i.e., curating the way that information is presented to the student) mathematical solutions. We then use an LLM-as-a-judge to see which response better aligns to the needs of a grade 5 student. This proof of concept was performed on the GSM8k dataset, and we found that the LLM-as-a-judge (flan-t5) preferred the response to the teacher's prompt 78.7% of the time ($n = 1284$).

The baseline and teacher-aligned prompts are as follows:

```
def make_baseline_prompt(question: str) -> str:
return (
    f"Question: {question}\n\n"
    "Explain step-by-step for a grade 5 student. "
    "End your response with 'Final answer: <answer>'."
)

def make_aligned_prompt(question: str) -> str:
return (
    "You are a math teacher."
    "Explain step-by-step for a grade 5 student.\n\n"
    "To align with the learning strategy of a grade 5 student,"
    "Follow this EXACT structure:\n\n"
    "[Understanding]\n"
    "Restate the problem in simple terms.\n\n"
    "[Plan]\n"
    "Explain what strategy we should use.\n\n"
    "[Guided Thinking]\n"
    "Ask 1 short question to guide the student "
    "(DO NOT answer it yet).\n\n"
    "[Step-by-step Solution]\n"
    "Solve the problem step-by-step.\n\n"
    "[Final Answer]\n"
    "Write: Final answer: <answer>\n\n"
    f"Question: {question}"
)
```

B.4.7. Preliminary Progress by ChatGPT's study mode and Gemini's Guided Learning in addressing RQ16-RQ18

The introduction of Gemini's Guided Learning (Heymans, 2025) and ChatGPT's study mode (OpenAI, 2025) represents a step in the right direction. For example, ChatGPT's study mode adapts instruction to learners' proficiency levels using diagnostic questions and conversational history, addressing RQ16. Likewise, Gemini's Guided Learning promotes critical thinking through guided questioning, aligning with RQ17. These examples underscore the relevance of the proposed research questions.