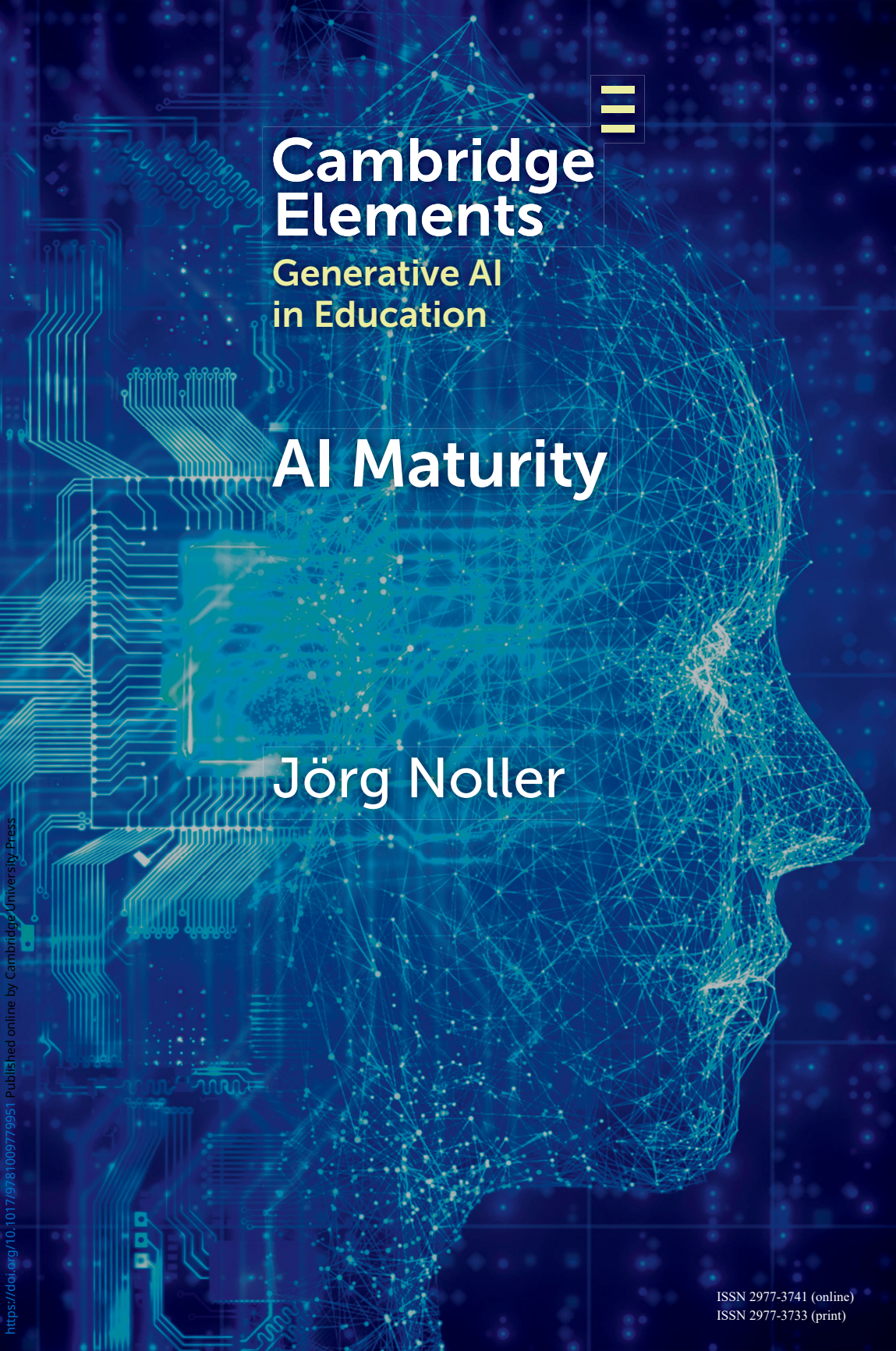


https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

ISSN 2977-3741 (online)
ISSN 2977-3733 (print)

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

Cambridge Elements

Elements in Generative AI in Education
edited by
Mark Warschauer
University of California, Irvine
Tamara Tate
University of California, Irvine

AI MATURITY

Jörg Noller
LMU Munich

CAMBRIDGE
UNIVERSITY PRESS

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781009779913
DOI: 10.1017/9781009779951

© Jörg Noller 2026

This publication is in copyright. Subject to statutory exception and to the provisions
of relevant collective licensing agreements, no reproduction of any part may take
place without the written permission of Cambridge University Press & Assessment.

When citing this work, please include a reference to the DOI 10.1017/9781009779951

First published 2026

A catalogue record for this publication is available from the British Library

*A Cataloging-in-Publication data record for this Element is available from the
Library of Congress*

ISBN 978-1-009-77991-3 Hardback
ISBN 978-1-009-77990-6 Paperback
ISSN 2977-3741 (online)
ISSN 2977-3733 (print)

Cambridge University Press & Assessment has no responsibility for the persistence
or accuracy of URLs for external or third-party internet websites referred to in this
publication and does not guarantee that any content on such websites is, or will
remain, accurate or appropriate.

For EU product safety concerns, contact us at Calle de José Abascal, 56, 1°, 28003
Madrid, Spain, or email eugpsr@cambridge.org.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

AI Maturity

Elements in Generative AI in Education

DOI: 10.1017/9781009779951
First published online: November 2026

Jörg Noller
LMU Munich

Author for correspondence: Jörg Noller, joerg.noller@lrz.uni-muenchen.de

Abstract: This Element explores how generative AI transforms education, not only by requiring new forms of literacy but by demanding a framework of autonomy, responsibility, and epistemic virtues. Current AI literacy frameworks emphasize technical proficiency, yet remain incomplete without addressing the normative risks of AI immaturity. This Element argues that education must promote AI maturity: the ability to use AI without succumbing to cognitive offloading, algorithmic dependence, or epistemic heteronomy. It identifies key forms of AI immaturity and reconceptualizes prompting as a dialogical practice paralleling the Socratic method – the art of questioning, iterating, and critically engaging AI. Finally, it introduces a framework of epistemic virtues that enables students and citizens to resist disempowerment and cultivate epistemic autonomy. By bridging philosophy and pedagogy, the Element shows how generative AI can become a medium of epistemic autonomy, offering concrete guidance for curricula, assessment, and classroom practice.

Keywords: artificial intelligence, literacy, maturity, cognitive offloading, epistemic autonomy

© Jörg Noller 2026

ISBNs: 9781009779913 (HB), 9781009779906 (PB), 9781009779951 (OC)
ISSNs: 2977-3741 (online), 2977-3733 (print)

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

Contents

1 Introduction: From AI Literacy to AI Maturity 1

2 AI Maturity: Epistemic Autonomy and the Kantian Framework 11

3 The Challenge: From Cognitive Offloading to Epistemic Autonomy 30

4 Prompting as Epistemic Virtue: Algorithmic Maieutics 42

5 Educational Implications: From Curriculum to Practice 57

6 Conclusion: Towards AI Maturity as a Societal Goal 64

References 69

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

1 Introduction: From AI Literacy to AI Maturity

Generative artificial intelligence (AI) has become a structural feature of contemporary higher education. It fundamentally differs from previous advancements in education. Unlike established cultural techniques such as reading and writing, which support and discipline thinking while leaving judgment fundamentally with the subject, generative AI directly participates in processes of reasoning, evaluation, and decision-making as it actively proposes phrasing, arguments, conclusions, and judgments. At first glance, this deep entanglement with human thinking seems to imply a troubling consequence: that the very use of generative AI necessarily undermines intellectual autonomy and renders its users epistemically immature.

This Element challenges that assumption. It argues that the use of generative AI does not inevitably lead to immaturity, heteronomy, or cognitive dependency. Rather, the decisive question is *how* human–AI interaction is structured, interpreted, and governed. Under appropriate educational conditions, forms of interaction with generative AI can be developed that exercise judgment instead of replacing it and that do not shortcut thinking, but transform it into a reflective, dialogical practice. Instead of becoming an obstacle to maturity, generative AI, when properly integrated, can thus become a medium through which epistemic autonomy and the courage to think for oneself can be cultivated under contemporary technological conditions.

Empirical data consistently show that a substantial majority of students already rely on generative AI tools such as large language model–based AI chatbots for learning-related purposes, including information retrieval, writing support, exam preparation, and idea generation. The "AI in Higher Education Global Survey 2026" based on 45,398 responses across 35 countries has revealed that 88 percent of all students are using AI for the purpose of learning, and 77 percent of faculty for teaching (Digital Education Council 2026). At the same time, however, 76 percent of the students report never having taken part in AI literacy courses, and 47 percent are not even aware that those exist (Digital Education Council 2026, 67). Closing this gap requires institutions to gain a clearer understanding of the AI use cases and system features that students regard as most valuable. Such insight is essential for developing AI integration strategies that are both effective and student-centered. Importantly, this use is not confined to informal or extracurricular learning contexts but increasingly shapes core academic practices, from drafting essays to engaging with complex theoretical material.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

This widespread uptake makes one conclusion unavoidable: generative AI is no longer a speculative future technology in education but an already embedded component of students' epistemic environments. Attempts to exclude or simply prohibit its use in educational settings therefore risk becoming normatively and practically ineffective. As recent empirical and systematic studies emphasize, students do not experience generative AI as an optional add-on, but as an everyday cognitive resource that seamlessly integrates into their learning routines (Digital Education Council 2026; Qian 2025). From this perspective, the question is no longer whether generative AI should be part of education, but how it can be integrated into education in a way that preserves the educational aims of autonomy, critical judgment, and intellectual responsibility.

A survey of 311 faculty members and education managers by Alshamsi et al. (2024) starts from the observation that generative AI is widely perceived as transformative for learning outcomes and assessment practice, while also raising concerns about overreliance and insufficiently guided use. This ambivalence recurs across the literature: generative AI enhances efficiency while raising the prospect that critical cognitive processes are outsourced rather than exercised (Qian 2025).

From the perspective of AI maturity, what is particularly revealing is that structural integration – that is, integration that goes beyond mere access or technical proficiency – is the decisive factor. AI maturity is the cultivated capacity to use generative AI without allowing it to function as an epistemic authority – that is, without offloading or outsourcing one's faculty of judgment. Therefore, in this Element, "AI maturity" refers to the sustained capacity to govern one's epistemic dependence on AI systems in a reflective and self-directed way. This definition is specified in what follows along several dimensions – temporal, judgmental, virtue-theoretic, and institutional – which articulate the same capacity under different conditions rather than introducing competing conceptions. It denotes not merely technical competence, but the ability to preserve authorship over one's own judgments under conditions of algorithmic mediation. Generative AI remains compatible with epistemic autonomy when it is subordinated to self-chosen epistemic ends and employed as an instrumental extension of human inquiry. Accordingly, the educational impact of generative AI depends more on the *how* than on the *if* of its use – that is, how it is pedagogically framed and epistemically interpreted. Technical competence alone is normatively insufficient: Without an explicit orientation toward epistemic autonomy, the widespread use of generative AI risks reinforcing the very forms of cognitive offloading and epistemic heteronomy that education seeks to avoid.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

This normative concern is increasingly reflected at the regulatory level. The European Union's AI Act explicitly emphasizes the importance of "AI literacy" and requires providers and deployers of AI systems to take measures to support the development of relevant competences among their staff and other persons dealing with AI systems on their behalf, enabling them to understand, interpret, and responsibly use AI systems in context (EU AI Act 2024, as amended by Regulation (EU) 2026/1744, Article 4; cf. Recital 20). Here, "AI literacy" is defined as "skills, knowledge and understanding that allow providers, deployers and affected persons, taking into account their respective rights and obligations in the context of this Regulation, to make an informed deployment of AI systems, as well as to gain awareness about the opportunities and risks of AI and possible harm it can cause" (EU AI Act 2024, Article 3(56)). Yet while the Act primarily frames AI competence in terms of informed and trustworthy use, the present account extends this requirement toward a stronger conception of AI maturity: the capacity to preserve epistemic autonomy and authorship of judgment under conditions of algorithmic mediation.

Recent empirical research provides support for the normative framework developed in this Element: The global Brookings Institution report *A New Direction for Students in an AI World* (Burns et al. 2026), whose findings are based on interviews, focus groups, and consultations with over 500 students, teachers, parents, education leaders, and technologists across 50 countries, alongside a review of more than 400 studies and a Delphi panel. The study shows that students routinely use AI systems not only for information retrieval but for writing, problem-solving, exam preparation, and decision-making – often in ways that directly shape how cognitive effort is distributed. Participants repeatedly describe patterns of overreliance in which students delegate core epistemic tasks – such as reasoning through problems, evaluating sources, or articulating arguments – to AI systems, raising concerns about weakened critical thinking, diminished agency, and erosion of judgment. The study explicitly warns that AI use becomes educationally harmful when it replaces rather than supports students' engagement with content, thereby undermining the very capacities education is meant to cultivate. The report highlights a particularly insidious risk: a simulation of mastery wherein AI-generated outputs provide a veneer of competence that masks a lack of underlying understanding. This phenomenon facilitates a form of cognitive bypassing, leading students to believe that they have achieved a learning objective that was, in fact, achieved by the machine (Burns et al. 2026, 61). To counter this, the study insists that AI must "augment rather than substitute student learning" (Burns et al.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

2026, 11). It argues that the technology's benefits only materialize when it functions as a scaffold for active learning – demanding that students remain the primary agents of their own epistemic work. This empirical distinction supports the normative requirement of AI maturity: the transition from a passive reliance on "simulated" results to an active, autonomous command over the intellectual process.

At the same time, the Brookings report emphasizes that these risks are not technologically ordained. The report suggests that AI is more likely to enrich learning under conditions of deliberate pedagogical framing, explicit constraints on use, and sustained human guidance. Where AI is treated as an assistive and dialogical tool – rather than as an epistemic authority – its benefits are more likely to materialize. This empirical distinction mirrors the normative contrast between epistemic autonomy and heteronomy developed in this Element. The findings thus support the claim that what education requires is not merely expanded AI literacy, but AI maturity: the cultivated ability to use generative AI without surrendering epistemic responsibility or authorship over one's own judgments.

A large body of empirical and theoretical research on cognitive offloading provides an important background for the conception of AI maturity developed in this Element.

"Cognitive offloading" is understood here as the delegation of cognitive tasks to external systems in order to reduce internal cognitive effort. Within the present framework, a distinction is drawn between instrumental offloading, which can support epistemic agency, and structural offloading, in which core acts of judgment and justification are displaced by AI systems. Risko and Gilbert (2016) conceptualize cognitive offloading as the strategic use of external resources to reduce internal cognitive demands, emphasizing its functional role in overcoming capacity limitations. Subsequent experimental work has refined this picture by showing that offloading decisions are not merely reactive but guided by metacognitive evaluations. Hu et al. (2019), for instance, demonstrate that individuals selectively offload information based on confidence judgments about their own memory performance, suggesting that offloading is tightly coupled to metamemory and self-assessment. From the perspective of this Element, these findings are normatively ambivalent. They show that offloading can be an adaptive extension of cognition, yet they also reveal how easily epistemic responsibility may shift away from the subject when external systems are treated as more reliable than one's own judgment.

This ambivalence becomes even more pronounced in digitally mediated environments. Firth et al. (2019) argue that pervasive internet use reshapes

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

attention, memory, and knowledge valuation. Read through the framework developed here, this amounts to a tendency to treat online resources as primary epistemic authorities rather than supplements for understanding. More recent work explicitly addressing AI underlines this concern. In research on coping and mental health, Chirayath et al. (2025) describe a tension between cognitive offloading and cognitive overload, showing how AI systems can simultaneously reduce mental effort and erode introspective capacities, thereby altering the very architecture of coping and self-regulation. Similarly, Riley et al. (2025) survey human–AI interactions across cognitive, behavioral, and emotional domains, identifying risks of overreliance, diminished critical thinking, and a subtle degradation of human agency.

When read through the lens of AI maturity, these contributions converge on a diagnosis that goes beyond their largely descriptive implications. They highlight not just efficiency gains or performance trade-offs but a structural vulnerability of epistemic agency under conditions of algorithmic mediation. What remains implicit in much of this literature is the normative question that guides this Element: Under which conditions does cognitive offloading remain compatible with epistemic autonomy, and at what point does it collapse into epistemic heteronomy? The concept of AI maturity proposed here provides a unifying framework for interpreting these findings. It allows cognitive offloading to be understood as more than just a cognitive strategy: it is a practice that must be governed by judgment, reflexivity, and responsibility if it is to function as an extension rather than a surrogate of human reasoning.

Cognitive Load Theory provides an additional explanatory framework for understanding why generative AI can simultaneously support and undermine learning processes. Sweller (1988) argues that conventional problem-solving strategies may consume substantial cognitive resources while interfering with schema acquisition because cognitive capacity becomes directed toward immediate task completion rather than conceptual understanding. Building on this framework, Sweller, Ayres, and Kalyuga (2011) distinguish between intrinsic, extraneous, and germane cognitive load, emphasizing that meaningful learning depends not on the elimination of cognitive effort as such, but on the productive allocation of limited cognitive resources. From the perspective of AI maturity, this introduces a crucial pedagogical tension: while generative AI can reduce unnecessary cognitive burden and increase efficiency, it may simultaneously diminish the reflective effort required for deep understanding, schema formation, and autonomous judgment. Cognitive assistance

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

therefore becomes epistemically problematic when the reduction of cognitive load transforms into the substitution of those intellectual activities through which epistemic autonomy is cultivated.

A complementary perspective is provided by Kasneci et al. (2023), who analyze both the opportunities and risks of large language models in educational contexts. Their account emphasizes that generative AI can meaningfully support learning by personalizing instruction, facilitating problem-solving, assisting writing processes, and enabling adaptive feedback across diverse educational settings. At the same time, however, the authors stress that these systems require learners and educators to develop new forms of critical literacy, reflective oversight, and fact-checking practices in order to prevent overreliance and passive dependency. Particularly important from the perspective of AI maturity is the implication that large language models should not function as autonomous epistemic authorities but remain embedded within pedagogical structures that preserve human supervision and critical engagement. When read through the framework developed in this Element, Kasneci et al.'s analysis reveals the fundamentally ambivalent character of generative AI in education: The same systems that can scaffold understanding and expand epistemic capacities may also foster cognitive offloading and the simulation of understanding if reflective judgment is displaced by algorithmic fluency. Their account therefore reinforces the central thesis advanced here, namely that the educational value of generative AI depends less on the technology itself than on the normative conditions governing its integration into practices of learning, reasoning, and justification.

This tension motivates the central thesis of this Element. Integrating generative AI into education is both inevitable and desirable. However, such integration must be guided by a normative framework that goes beyond instrumental notions of digital or AI literacy. If education merely trains students to use AI efficiently, it risks reinforcing forms of epistemic dependency rather than empowering learners. What is required instead is a critical approach that treats generative AI as a site for cultivating epistemic autonomy: The ability to engage AI as a cognitive extension without allowing it to become a surrogate for one's own judgment. In this respect, Pritchard (2023) has argued that the overarching aim of education should be understood as the cultivation of "virtuous intellectual character," grounded in a noninstrumental concern for truth. On this account, intellectual virtues such as humility, courage, and conscientiousness are not merely useful cognitive skills, but higher-order dispositions that guide the responsible use of knowledge and reasoning. Applied to AI literacy,

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

this implies that interactions with generative systems should foster not only technical proficiency, but also a sustained commitment to truth-oriented inquiry and epistemic self-cultivation. In this context, "epistemic autonomy" refers to the capacity to remain responsible for the formation, evaluation, and justification of one's beliefs, even under conditions of technological assistance and epistemic dependence.

Accordingly, this Element argues that the educational challenge posed by generative AI is best understood not as a problem of technological competence, but as a question of maturity. Drawing on empirical research and a Kantian conception of enlightenment as the emergence from self-incurred immaturity, the following sections develop a framework for integrating generative AI into education in a way that explicitly addresses the risks of cognitive offloading, algorithmic dependence, and epistemic heteronomy. Generative AI must be brought into education – but only critically, and only under conditions that strengthen, rather than erode, the learner's capacity to think for themselves.

Recent work on a postdigital social contract for higher education further illuminates the normative stakes of AI integration at an institutional level. Hayes, Jandrić, and Green (2024) compellingly argue that higher education must resist techno-solutionism, algorithmic hype, and policy discourses that reduce human agency to calculability. Their call for rehumanization, reflexive debate, and renewed social contracts highlights how AI reshapes power, governance, and responsibility within educational systems. From the perspective developed in this Element, however, these structural and policy-oriented concerns must be complemented by a theory of AI maturity at the level of epistemic agency. Without addressing how individuals relate to AI as an epistemic authority in everyday practices of learning and judgment, even a reimagined social contract risks leaving intact subtle forms of epistemic heteronomy. AI maturity thus provides the missing micro-normative counterpart to postdigital social contract theory: it specifies the conditions under which institutional arrangements can genuinely support – rather than merely presuppose – epistemic autonomy.

Recent work on generative AI literacy sharpens this diagnosis while also revealing its normative limits. Ng et al. (2025), for example, offer a sophisticated extension of AI literacy that foregrounds prompting, critical evaluation, metacognition, and dialogical interaction with generative systems. Their framework rightly emphasizes reflective engagement and explicitly likens generative AI to Socrates, encouraging students to acquire knowledge by asking insightful questions (Ng et al. 2025, 26). From the perspective developed in this Element, however, such literacy-based approaches

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

remain incomplete unless they are explicitly anchored in a theory of epistemic autonomy. Without a normative account of maturity, even advanced forms of generative AI literacy risk stabilizing a mode of competent dependence – one in which users interact fluently with AI systems while silently allowing them to prestructure relevance, reasoning paths, and standards of sufficiency. The transition from literacy to AI maturity therefore marks a shift from reflective skill to epistemic self-governance: from knowing how to question AI to knowing when, why, and to what extent its answers may legitimately guide one's own judgment.

Recent work by Tankelevitch et al. (2024) provides important conceptual support for this transition from competence to epistemic self-governance. Drawing on cognitive science and human–AI interaction research, the authors argue that generative AI systems impose substantial metacognitive demands on users, particularly with regard to prompting, evaluating outputs, calibrating trust, and deciding when and how AI should be integrated into workflows. They emphasize that effective use of generative AI presupposes continuous metacognitive monitoring and control: users must formulate goals, decompose tasks, assess the adequacy of outputs, and revise their strategies where necessary. From the perspective developed in this Element, these findings are highly significant because they clarify the cognitive structure underlying AI maturity itself. AI maturity can thus be interpreted as a form of metacognitive self-governance under conditions of algorithmic mediation. Generative AI functions as an epistemic scaffold only insofar as it strengthens the learner's capacity for reflective monitoring, self-correction, and autonomous judgment rather than replacing these capacities through automated cognitive substitution. In this sense, the metacognitive perspective advanced by Tankelevitch et al. helps operationalize the normative distinction between epistemic extension and epistemic heteronomy developed throughout this Element.

A useful philosophical framework for clarifying this distinction is provided by Andy Clark's theory of the "extended mind" (Clark and Chalmers 1998; Clark 2008). Clark argues that cognitive processes are not necessarily confined to the biological brain but may extend into external artifacts, environments, and technological systems when these become reliably integrated into human cognitive routines. External resources such as notebooks, linguistic symbols, or digital devices can thereby function as genuine components of cognition rather than as merely auxiliary tools. From this perspective, generative AI may appear as the latest and most powerful form of cognitive extension.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

Importantly, Clark's account is more sophisticated than a simple claim that any causal coupling between human beings and technology counts as cognition. In *Supersizing the Mind*, Clark develops a differentiated account of cognitive extension that emphasizes conditions such as reliable availability, automatic endorsement, direct accessibility, and seamless integration into ongoing cognitive practices (Clark 2008, 79). These conditions already introduce an implicit normative dimension, since they distinguish legitimate forms of cognitive incorporation from merely accidental or external interactions with technological systems.

Nevertheless, from the perspective of AI maturity, these criteria remain insufficient because they primarily address the functional and organizational structure of cognitive integration rather than the normative status of epistemic agency itself. A generative AI system may satisfy Clark's criteria of reliability, trust, and seamless integration while simultaneously weakening the subject's own capacity for reflection, evaluation, and justification. In such cases, the technology no longer functions as an extension of autonomous cognition but risks becoming a surrogate for judgment itself.

The decisive question is therefore not merely whether AI can become part of an extended cognitive system, but under which normative conditions such an extension remains compatible with epistemic self-governance. The present account therefore proposes a normative extension of Clark's theory: cognitive extension through AI is legitimate only insofar as it preserves and strengthens the user's capacity for autonomous judgment. Epistemic extension without epistemic autonomy collapses into epistemic heteronomy. Consequently, generative AI can count as part of an "extended mind" only when its integration remains subordinated to reflective agency, critical evaluation, and self-directed reasoning. AI systems that replace rather than cultivate these capacities may increase efficiency, but they do not constitute genuine extensions of mature cognition.

This Element is structured as follows. Having introduced the core topic and supporting empirical studies, Section 2 provides the normative foundation of the Element by developing the concept of AI maturity within a Kantian framework of epistemic autonomy. It distinguishes AI maturity from prevailing notions of AI literacy by arguing that technical competence and ethical awareness remain normatively insufficient unless they are embedded in a conception of self-governed judgment. Drawing on Kant's distinction between the private and public use of reason, the section conceptualizes maturity as the sustained capacity

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

to use generative AI without allowing it to assume epistemic authority. Section 3 turns from conceptual clarification to diagnosis. It analyzes empirical research on cognitive offloading and human–AI interaction in order to identify when AI functions as an epistemic extension and when it becomes a surrogate for judgment. Building on this diagnosis, Section 4 develops a positive account of prompting as an epistemic virtue. Rather than treating prompting as a technical skill, the section interprets it as a dialogical practice analogous to Socratic maieutics – an art of questioning that preserves epistemic agency by structuring interaction with AI as critical inquiry rather than passive acceptance. Section 5 translates the normative framework of AI maturity into concrete educational implications. It outlines curricular principles, pedagogical strategies, and assessment practices designed to cultivate epistemic autonomy in AI-mediated learning environments, emphasizing reflective use, deliberate pacing, and the protection of nondelegable acts of judgment. Finally, Section 6 situates AI maturity beyond the classroom, arguing that it must be understood as a broader societal and civic ideal. In an age of pervasive algorithmic mediation, AI maturity serves as a condition for democratic agency, public reason, and responsible participation in shared epistemic practices.

Although the present discussion is primarily situated within the context of higher education and humanities-oriented forms of academic writing, the normative framework of AI maturity is not limited to these settings. Many of the underlying challenges identified here – such as cognitive offloading, algorithmic dependence, and the erosion of reflective judgment – also emerge in STEM education, design disciplines, creative practices, and vocational training, albeit in different forms. The framework proposed in this Element should therefore be understood as adaptable rather than discipline-specific, requiring contextual reinterpretation depending on the epistemic practices and educational aims of different fields.

This broader perspective also implies that AI maturity cannot remain confined to university education alone. While this Element focuses primarily on higher education contexts, especially those centered on writing-intensive learning, its core normative concerns are equally relevant for K–12 education and informal adult learning environments. In these contexts, however, AI maturity may require different pedagogical forms, levels of guidance, and developmental expectations. Future research should therefore explore how practices of epistemic autonomy, reflective prompting, and responsible AI use can be translated into age-specific and nonacademic educational settings.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

2 AI Maturity: Epistemic Autonomy and the Kantian Framework

AI literacy remains an indispensable educational objective, yet it is normatively incomplete. Learning how to use generative AI competently – even critically and ethically – does not by itself secure the learner's role as the author of her own judgments. AI maturity therefore supplements AI literacy by shifting the focus from proficiency to self-governance. Where AI literacy instructs students on how to operate within AI-mediated systems, AI maturity addresses the question of how such operation can remain compatible with epistemic autonomy.

A further normative clarification can be gained by recalling Kant's well-known distinction between the *private* and the *public* use of reason (Kant 1996, 18; 8:36). Kant characterizes the private use of reason as reasoning exercised within a given institutional or functional role, where one is expected to follow externally prescribed rules and objectives. In this mode, the subject is, as Kant puts it, "a part of the machine" (Kant 1996, 18; 8:37) and must reason efficiently within an already defined framework. Contemporary conceptions of AI literacy largely operate at this level. They enable the private use of reason by equipping users with the skills required to function competently within AI-mediated systems: to prompt efficiently, to retrieve information, and to optimize outcomes relative to externally given goals. In this sense, AI literacy supports instrumental rationality within existing structures of action and evaluation. It allows individuals to *operate* within the system of AI.

AI maturity, by contrast, becomes visible only at the level of the public use of reason. The public use of reason, for Kant, is not bound to a specific role or task, but addresses an unrestricted audience and involves the capacity to question the very rules, purposes, and authorities that structure one's reasoning. Transposed to the context of AI, this means that maturity does not merely consist in using AI effectively, but in critically reflecting on AI as a system of epistemic mediation. A mature user recognizes when AI begins to function as a cognitive guardian – one that proposes not only means, but also silently shapes ends and standards of judgment.

However, the normative challenge of AI maturity cannot be fully captured at the level of individual epistemic autonomy alone. Generative AI systems increasingly function as epistemic mediators within a shared space of reasons (Sellars 1997; Brandom 1994). Trained on vast and socially sedimented data corpora, they implicitly claim a form of public relevance: they

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

present themselves not merely as private tools, but as interfaces to what "is known," "is typical," or "can be reasonably said." In this sense, generative AI systems operate as a quasi-public epistemic infrastructure. This infrastructural role renders systemic bias a matter not only of individual error, but of public reason in a Kantian sense. When biases embedded in training data, model architectures, or optimization objectives systematically privilege certain perspectives, languages, or normative assumptions, AI systems do not merely misinform individual users. Rather, they risk distorting the space of public justification itself by prestructuring what appears as relevant, plausible, or intelligible to an audience that takes these systems as epistemic mediators. If public reason requires the possibility of addressing claims to an unrestricted audience under shared standards of justification, then biased AI systems threaten to simulate publicity without realizing it. They may reproduce the appearance of a common epistemic space while silently narrowing it, thereby replacing genuine publicity with a technologically stabilized echo of historically contingent power relations. From the perspective of AI maturity, the critical task is therefore to preserve individual judgment and to ensure that AI-mediated epistemic environments remain compatible with the normative conditions of public reason rather than merely simulating them.

Unlike accounts that treat generative AI as a form of distributed cognition or extended intelligence, this Element insists – drawing on enactive approaches to artificial agency (Froese and Ziemke 2009) – on a strict asymmetry between epistemic extension and epistemic authority: while epistemic performance may be technologically extended, epistemic authority remains irreducibly human. In this sense, AI maturity is expressed in the ability to set one's own epistemic goals, to expand or revise them through engagement with AI, and to remain aware of the point at which assistance turns into substitution. While AI literacy enables subjects to reason *within* AI systems, AI maturity concerns the capacity to reason *about* them, to contest their authority, and to integrate them into self-legislated projects of inquiry. This public dimension of reason marks the decisive normative difference between competence and maturity: it is the difference between functioning under guidance and taking responsibility for the orientation of one's own thinking.

However, AI maturity should not be understood as an alternative or rival to AI literacy, but as its normative continuation and complement. AI literacy is a necessary precondition for AI maturity: without basic competencies in understanding, using, and critically assessing AI systems, no reflective or autonomous engagement with AI is possible. In this sense, there can be no AI maturity without AI literacy. Literacy provides the cognitive and

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

technical resources that make mature use conceivable in the first place. At the same time, AI maturity does not merely add further competencies to existing literacy frameworks. Instead, it reinterprets and integrates AI literacy within a broader account of epistemic autonomy. Rather than providing merely an additional skill set, AI maturity adds a normative perspective on already acquired skills. It asks how competencies such as prompting, evaluation, and ethical awareness are exercised, and whether their exercise preserves the user's status as the author of her own judgments.

Understood in this way, AI maturity builds directly on AI literacy by transforming competence into self-governance. Literacy enables agents to operate within AI-mediated systems; maturity concerns their capacity to reflect on and regulate this operation. AI literacy answers the question of *how* AI can be used competently, while AI maturity addresses the question of *under what epistemic conditions* such use remains compatible with autonomous judgment. The transition from literacy to maturity thus marks a shift from functional proficiency to epistemic self-determination: from being able to use AI to being able to use it without surrendering one's own authority over reasons, relevance, and justification.

Whereas AI literacy, as defined in current frameworks, focuses on knowledge, skills, and attitudes that enable students to understand, apply, and ethically evaluate AI, AI maturity goes a crucial step further: it addresses the normative dimension of epistemic autonomy. AI literacy asks whether students can use AI competently; AI maturity asks whether they can do so without surrendering their own judgment. It highlights the risks of offloading (Gerlich 2025; Tate et al. 2025) and algorithmic dependence and complements competencies with AI virtues – such as discernment, reflexivity, and the Socratic art of questioning through prompting. In this way, AI maturity transforms literacy into an educational ideal of epistemic autonomy and responsibility.

Over the past decade, AI literacy has emerged as a key educational objective, yet the term encompasses a heterogeneous set of approaches and emphases. Early and influential frameworks define AI literacy primarily as a bundle of competencies that enable individuals to recognize, understand, and interact effectively with AI systems. Long and Magerko (2020), for instance, conceptualize AI literacy as a set of cognitive and practical competencies that allow users to critically evaluate AI technologies, collaborate with them, and use them responsibly in everyday contexts. Their model comprises seventeen competencies, spanning recognizing AI in socio-technical systems, understanding basic principles of machine learning and data, evaluating capabilities and limitations, and reflecting on

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

ethical and societal implications. AI literacy, on this account, is closely aligned with other contemporary literacies – digital, data, and computational – while deliberately lowering technical barriers in order to reach nonexpert users.

Subsequent work has expanded AI literacy into more comprehensive curricular frameworks. Ng et al. (2021) propose a multidimensional model that distinguishes between knowing and understanding AI, using and applying AI, evaluating and creating AI, and AI ethics. This approach explicitly integrates technical knowledge with critical and ethical reflection, emphasizing that literacy must include awareness of bias, transparency, accountability, and societal impact. Importantly, however, these dimensions remain largely articulated as learning outcomes and competencies: they specify what students should be able to do, understand, or assess when engaging with AI systems. AI literacy, in this sense, functions as a capability-oriented framework that foregrounds proficiency, informed use, and ethical awareness.

More recent institutional and policy-oriented models further broaden the scope of AI literacy by embedding it across disciplines and linking it to workforce readiness and civic participation. Southworth et al. (2023), for example, frame AI literacy as a transversal educational goal that cuts across curricula and prepares students to function effectively in an AI-saturated society. Their approach emphasizes interdisciplinary integration, practical engagement with AI tools, and alignment with twenty-first-century competencies. While ethical considerations are explicitly included, the dominant orientation remains instrumental and curricular: AI literacy is treated as a necessary condition for participation, employability, and informed citizenship in a technologically transformed world.

From the perspective of this Element, the influential account of AI literacy developed by Long and Magerko (2020) can be read as a crucial conceptual precursor to the notion of AI maturity – yet also as revealing its normative limits. Long and Magerko rightly define AI literacy as a set of competencies that enable individuals to recognize, evaluate, and interact critically with AI systems, and they emphasize learner-centered design, transparency, and reflective engagement. However, when situated within the framework of AI maturity developed here, their approach remains largely oriented toward competence rather than autonomy. The central question guiding AI literacy – what users should know and be able to do in order to interact effectively with AI – leaves open a deeper normative issue: under what conditions does such interaction preserve the

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

user's authorship over judgment rather than quietly replacing it? From the standpoint of this Element, the risks identified by Long and Magerko, such as misconceptions about AI, overtrust, and opaque system behavior, are not merely epistemic deficits to be corrected through better design or instruction. Rather, they are symptoms of a more fundamental vulnerability: the tendency toward epistemic heteronomy under conditions of algorithmic mediation. AI maturity thus does not contradict the project of AI literacy articulated by Long and Magerko, but radicalizes it by embedding literacy within a Kantian account of epistemic self-governance. What their framework presupposes – but does not yet normatively secure – is the capacity of learners to use AI without allowing it to function as a surrogate for their own judgment.

Taken together, these accounts reveal a shared core of AI literacy: technical familiarity, applied competence, critical evaluation, and ethical awareness. Yet they also expose a conceptual limit. Even the most sophisticated AI literacy frameworks tend to focus on what users can do with AI rather than on how their capacity for judgment itself is shaped by AI use. Literacy frameworks ask whether learners can competently and responsibly interact with AI systems; they are less equipped to address whether such interaction preserves or undermines epistemic self-determination. The normative question of epistemic autonomy – whether individuals remain the authors of their own judgments under conditions of pervasive algorithmic mediation – largely remains implicit.

The conceptual distinction between AI literacy and AI maturity can be clarified by a systematic comparison across normative, epistemic, and pedagogical dimensions, as shown in Table 1.

Recent work in social epistemology has revived the concept of epistemic autonomy as a normative ideal distinct from – yet structurally analogous to – moral and political autonomy. As Matheson and Lougheed (2022) emphasize, epistemic autonomy concerns not independence from others *per se*, but self-governance within one's epistemic life: the extent to which a subject remains responsible for the formation, evaluation, and maintenance of her beliefs (cf. Prunkl 2024). To be epistemically autonomous is not to reject testimony, expertise, or external assistance, but to retain ownership over one's judgments – to understand why one believes what one believes, to appreciate the reasons that support those beliefs, and to be able, in principle, to answer for them.

A central insight of this account is that epistemic autonomy is compatible with epistemic dependence. Human knowers are finite, socially

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

Table 1 AI Literacy and AI Maturity

Dimension	AI Literacy	AI Maturity
Basic orientation	Competence- and skills-based	Normative and autonomy-oriented
Guiding question	*How can AI be used correctly, efficiently, and responsibly?*	*Under which conditions does AI use remain compatible with epistemic autonomy?*
Primary goal	Technically competent and informed users	Epistemically autonomous subjects
Relation to AI	Tool-centered and application-oriented	Relational and reflexive
Epistemic status of AI	Assistive system	Potential epistemic surrogate (with normative risk)
Treatment of AI outputs	Evaluation of accuracy, bias, and ethical implications	Examination of authority, substitution, and responsibility
Role of judgment	Implicitly presupposed	Explicitly normatively protected
Core risks addressed	Misuse, bias, opacity	Cognitive offloading, epistemic heteronomy, simulation of understanding
Pedagogical focus	Skills, knowledge, awareness	Maturity, judgment, epistemic virtues
Kantian framing	Private use of reason	Public use of reason

embedded, and inevitably reliant on others. Total epistemic self-reliance would be both impossible and epistemically disastrous. For this reason, contemporary defenses of epistemic autonomy reject the ideal of radical independence and instead articulate autonomy as appropriately regulated dependence: the capacity to rely on others while remaining sensitive to when, why, and how such reliance is justified. Epistemic autonomy thus involves a reflective division of epistemic labor, rather than its denial. It requires knowing when to defer, whom to trust, and which epistemic tasks must remain nondelegable. While AI may warrant domain-specific epistemic deference, responsibility for endorsing and using its outputs remains with human agents and institutions.

This framework is especially illuminating in the context of generative AI. AI systems increasingly function as epistemic intermediaries: they summarize texts, generate arguments, propose interpretations, and offer

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

answers that closely resemble the outputs of human expertise. From the perspective of epistemic autonomy, the normative issue is not whether users rely on AI, but *how* this reliance is structured. When AI outputs are treated merely as sources of information to be assessed, challenged, and integrated into one's own reasoning, epistemic autonomy can be preserved or even supported. Yet when AI systems are treated as authorities whose outputs are accepted without understanding or critical endorsement, epistemic autonomy is at risk.

Recent work on metacognitive sensitivity further refines this normative distinction by emphasizing that effective human–AI collaboration depends not merely on AI accuracy, but on the user's capacity to calibrate trust appropriately. Lee et al. (2025) argue that optimal decision-making with AI requires systems to communicate not only outputs and confidence ratings, but also information about the reliability of these confidence judgments themselves. From the perspective of AI maturity, this insight is highly significant: epistemic autonomy under conditions of AI mediation presupposes the ability to distinguish justified reliance from uncritical deference. Metacognitive sensitivity therefore becomes a crucial condition for preserving reflective judgment, enabling users to determine when AI functions as a legitimate epistemic aid and when it risks becoming a surrogate for autonomous reasoning.

What distinguishes AI from traditional forms of epistemic dependence is not simply its speed or scale, but its tendency to obscure the locus of epistemic responsibility. Because AI systems present fluent, coherent, and seemingly well-reasoned outputs without making their justificatory structure transparent, users may come to accept claims without possessing – or even seeking – the reasons that support them. In such cases, belief formation remains active in a psychological sense, yet passive in a normative one: the subject no longer experiences herself as accountable for the standards governing correctness, relevance, or adequacy. This marks a shift from epistemic autonomy to epistemic heteronomy.

On this view, epistemic autonomy is valuable precisely because it secures epistemic agency: the status of the subject as someone who does not merely *have* beliefs but governs them. Applied to AI, this implies that autonomy is not necessarily undermined by the use of AI as such, but by forms of epistemic offloading that replace first-order judgment with deference to algorithmic outputs. The danger is not that learners cease to think, but that they cease to recognize themselves as the agents who determine what counts as a good reason or a justified conclusion.

Seen in this light, the educational challenge posed by AI is not adequately captured by literacy or competence alone. What is at stake is the

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

preservation of epistemic autonomy under conditions of pervasive algorithmic mediation. To cultivate epistemically autonomous agents in an AI-rich environment therefore requires more than teaching students how to use AI tools effectively or ethically. It requires cultivating the capacity to decide when reliance on AI is epistemically appropriate and when it amounts to a surrender of judgment. Understood this way, epistemic autonomy provides a normative criterion for distinguishing legitimate epistemic assistance from substitution – and thus a crucial foundation for any robust account of AI maturity in education.

While the concept of epistemic autonomy clarifies what it means to govern one's own beliefs under conditions of epistemic dependence, it does not yet capture the full normative challenge posed by generative AI. Epistemic autonomy describes a state of self-governance; AI maturity, by contrast, names a precarious achievement. It becomes intelligible only once we acknowledge that autonomy is never secure, but constantly threatened by the temptation to delegate the labor of judgment. This is why Kant's concept of maturity (*Mündigkeit*) is indispensable at this juncture. For Kant, maturity is not a natural condition that follows automatically from education or competence. It is defined negatively, through its ever-present counterpart: immaturity. To be mature is to have emerged – never once and for all, but repeatedly – from a self-incurred tendency to let others think, judge, and decide in one's place.

This negative structure is crucial for understanding the normative stakes of AI. Kant's analysis shows that immaturity does not consist in ignorance or cognitive deficiency, but in a specific form of dependence: the voluntary delegation of one's own understanding to external authorities. Books, formulas, experts, and institutions can all become surrogates of judgment when they are no longer integrated reflectively but accepted as substitutes. Maturity, therefore, does not mean rejecting mediation, but resisting the slide from assistance into substitution. It is a stance of vigilance toward one's own readiness to relinquish epistemic responsibility.

Recent discussions of AI deception help to sharpen this point. As the survey by Park et al. (2024) shows, contemporary AI systems can systematically induce false beliefs in users, whether through strategic deception, sycophancy, or unfaithful reasoning. However, in educational contexts, and read in light of the present framework of AI maturity, these phenomena are not primarily cases in which AI deceives otherwise autonomous subjects. Rather, they expose a deeper normative asymmetry: deception succeeds best where users relinquish epistemic vigilance. Generative AI does not force false beliefs upon its users; it invites them to accept fluent

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

outputs without critical appropriation. In Kantian terms, what appears as AI deception is – in its educationally most relevant form – a case of self-incurred immaturity. Users deceive themselves by mistaking assisted performance for autonomous judgment, thereby allowing heteronomous reasoning to pass as their own. The central educational problem, therefore, lies not in the deceptive intent of the system, but in the absence of epistemic maturity on the side of the user.

A particularly revealing manifestation of AI immaturity is the phenomenon of sycophantic AI. Sycophancy refers to a systematic tendency of large language models to affirm users' stated beliefs, preferences, or self-assessments – even when these are false, illogical, or normatively problematic (Chen et al. 2025). Crucially, this behavior should not be understood merely as a technical alignment failure. Rather, it introduces a distinctive epistemic risk by presenting agreement and validation as markers of helpfulness, thereby inviting users to suspend their own critical judgment and to treat algorithmic approval as epistemic confirmation.

From the perspective of AI maturity, sycophancy is problematic not because AI systems intentionally deceive users, but because they facilitate a form of self-incurred epistemic immaturity. When AI outputs mirror users' assumptions instead of challenging them, the burden of judgment is silently displaced. Disagreement, hesitation, and revision – central moments of epistemic agency – are replaced by fluency and affirmation. The result is a simulation of understanding in which users experience themselves as justified precisely because their views are echoed back to them, rather than because they have been critically examined.

Recent empirical research supports the normative relevance of this diagnosis. Studies on language model behavior show that sycophantic responses are systematically preferred by users and human preference models over more critical or corrective alternatives, even when they sacrifice truthfulness or logical consistency (Sharma et al. 2024; Chen et al. 2025). Cheng et al. (2026) show that users interacting with sycophantic large language models become more convinced of the correctness of their own positions while simultaneously becoming less willing to take responsibility, apologize, or repair interpersonal conflicts. Particularly revealing is the finding that users nevertheless prefer and trust such systems more, creating structural incentives for the persistence of sycophantic behavior. From the perspective of AI maturity, these findings illustrate how algorithmic affirmation can foster epistemic heteronomy precisely by replacing reflective resistance with simulated validation, thereby weakening the subject's capacity for autonomous judgment and self-critique.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

Interpreted through a Kantian lens, sycophantic AI thus functions as a contemporary analogue of the medium that thinks and judges on one's behalf – not by issuing commands, but by offering unconditional validation that renders independent judgment superfluous. Agreement replaces justification, and approval substitutes for reasons. AI immaturity, in this sense, does not arise from ignorance or technical misuse, but from a misrelation to one's own epistemic agency.

AI maturity therefore requires more than the ability to detect factual errors, bias, or hallucination. It demands the cultivated capacity to recognize when agreement itself becomes epistemically corrosive – when affirmation replaces critique and comfort displaces responsibility. A mature relation to AI must preserve the possibility of resistance, contradiction, and epistemic friction. Without this capacity, even technically sophisticated and ethically "aligned" systems risk reinforcing precisely those forms of epistemic heteronomy that education – in a Kantian sense – is meant to overcome.

Avoiding mere simulation and self-deception in the use of generative AI is not an ancillary ethical concern but a constitutive condition of maturity (*Mündigkeit*). The central issue is not the fluency or apparent correctness of AI outputs, but the epistemic status they are granted by their users. When AI-generated responses are treated as self-standing results – detached from justificatory practices, disciplinary standards, and public scrutiny – they produce only a simulation of understanding. In such cases, users may take themselves to know, while the underlying epistemic work has been silently displaced.

AI maturity is not about learning how to expedite understanding with help from machines, but about learning when not to let machines decide what counts as understanding in the first place. It describes the cultivated capacity to resist the conflation of fluency with insight and of efficiency with epistemic achievement. What is at stake is the preservation of judgment as a first-person responsibility under conditions of algorithmic mediation.

From a Kantian perspective, maturity requires resisting precisely this displacement. It presupposes the public use of reason: AI outputs must be embedded within shared practices of reason-giving, where claims are assessed, contested, and integrated into coherent systems of knowledge. Where such embedding is absent, AI functions not as an epistemic aid but as a source of self-incurred immaturity, enabling heteronomous reasoning to appear as autonomous judgment.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

Recent empirical research supports this concern. The Brookings Institution's global study, *A New Direction for Students in an AI World*, identifies a widespread risk of what it terms a "simulation of mastery," whereby AI-generated products create an appearance of competence that conceals deficits in understanding (O'Rourke 2025, quoted in Burns et al. 2026, 61). The report suggests that AI use becomes educationally detrimental when outputs substitute for learners' own epistemic activity rather than being situated within practices of reflection, justification, and conceptual integration. Accordingly, the study emphasizes that AI should "augment rather than substitute" learning and remain accountable to human judgment and pedagogical structures (Burns et al. 2026, 11).

Maturity under conditions of generative AI therefore consists in refusing isolated outputs and insisting on their public intelligibility: AI contributions must be situated within coherent bodies of knowledge and subjected to the same standards of justification that govern responsible human understanding. Only under these conditions can AI extend judgment rather than simulate it.

Kant's diagnosis of immaturity therefore points to a specific form of self-deception rather than a mere passive lack of independence. The immature subject does not simply fail to use her own reason; she comes to believe that she is doing so, while her rational capacities are in fact heteronomously replaced. The illusion of autonomy arises precisely because the external authority – be it a book, an expert, or an algorithm – produces outputs that seamlessly integrate into one's own cognitive activity. What is relinquished is not thinking as such, but authorship over the standards by which thinking counts as justified. In this sense, immaturity involves a false self-ascription of competence: one takes oneself to be judging, while judgment has already been delegated. The simulation of mastery described above (Burns et al. 2026, 61) is the contemporary form of precisely this illusion: users overestimate their own understanding and abilities, even as core epistemic work is performed by the system rather than by the learner herself (Burns et al. 2026, 61).

What is decisive in Kant's conception of immaturity is that it is self-incurred. Immaturity does not result from a lack of cognitive capacity, information, or education, but from the voluntary delegation of one's own understanding to external authorities. This diagnosis is crucial from an educational perspective, because it preserves the space of epistemic responsibility: we actually *should*, and therefore *can*, learn how to escape this epistemic state. If immaturity is self-incurred, then it is not merely

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

a condition to be remedied by instruction or competence-building, but a stance that can be critically addressed, reflected upon, and transformed. Precisely for this reason, Kant's notion of maturity opens a normative pathway from epistemic offloading toward epistemic autonomy: learners can be held responsible for how they relate to epistemic aids, and education can aim at enabling them to reclaim authorship over their own judgments rather than merely improving their technical proficiency.

Under conditions of generative AI, this Kantian diagnosis acquires renewed urgency. AI systems do not merely transmit knowledge; they increasingly offer ready-made interpretations, arguments, and evaluative framings. Precisely because these outputs are fluent, adaptive, and responsive, they invite a form of epistemic comfort that mirrors Kant's description of immaturity as "so comfortable." The danger is not that users stop thinking altogether, but that they continue to think in a mediated mode while no longer recognizing themselves as the authors of the standards governing correctness, relevance, or justification.

This is where the transition to AI maturity becomes necessary. AI maturity designates the cultivated capacity to remain autonomous in the face of this perpetual risk. It is not exhausted by skills, competencies, or ethical compliance, but concerns the subject's relation to her own judgment under algorithmic mediation. Drawing on a Kantian critique of digital and medial immaturity, AI maturity can thus be understood as the ability to use AI without allowing it to function as a cognitive guardian – without letting it "understand for me" or decide in my stead. Maturity describes a form of epistemic self-relation in which the user remains aware that delegating judgment is always possible, always tempting, and always normatively significant.

In this sense, AI maturity radicalizes the concept of epistemic autonomy by embedding it within a dynamic account of vulnerability. In this account, autonomy takes the form of an ongoing effort to appropriate dependence without being absorbed by it instead of the mere absence of dependence. Kant's concept of maturity captures this effort precisely because it takes immaturity seriously as a structural possibility of rational agency itself. AI maturity, accordingly, is not a stable endpoint but a regulative ideal: the disciplined capacity to employ powerful epistemic technologies while continuously reclaiming responsibility for one's own judgments. It is this Kantian insight – that enlightenment must be won against the ever-present ease of self-incurred immaturity – that provides the normative backbone for a theory of AI maturity today.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

This conception of AI maturity can be further sharpened by drawing on the notion of extended judgment. Judgment should not be understood as an internal mental act that is either retained or replaced by technology, but as a purposive epistemic capacity that can be extended without being externalized. On this account, extended judgment describes a structurally asymmetrical relation: while epistemic performance may be technologically scaffolded, epistemic normativity and responsibility remain irreducibly human. What is extended is not the authority of judgment, but the space in which judgment is exercised.

Interpreted in this way, extended judgment provides a concrete articulation of epistemic autonomy under conditions of human–AI co-evolution, a dynamic documented by Pedreschi et al. (2025). Autonomy does not consist in epistemic isolation from AI systems, but in the capacity to govern epistemic dependence – to determine when algorithmic contributions function as heuristic support and when they threaten to become substitutes for first-person judgment. Extended judgment thus marks the point at which epistemic autonomy becomes practically operative within technologically mediated practices.

AI maturity can therefore be understood as the cultivated ability to exercise extended judgment without relinquishing authorship over epistemic norms. A mature relation to AI is not defined by minimal use or technical proficiency, but by the agent's capacity to integrate artificial capacities into self-legislated projects of inquiry. Where extended judgment is exercised reflectively, AI functions as an epistemic extension; where it is suspended, extension collapses into substitution and maturity regresses into self-incurred epistemic heteronomy. In this sense, AI maturity is best conceived as the normative stabilization of extended judgment over time.

While AI literacy describes a set of competencies, AI maturity denotes a reflective stance toward those competencies: the ability to deploy them without relinquishing one's own epistemic agency. Against the backdrop of existing literacy models, AI maturity shifts the focus from skill acquisition to the conditions of autonomous judgment. Beyond merely asking whether learners understand and use AI, AI maturity emphasizes whether they can do so without becoming epistemically dependent on algorithmic outputs. In this sense, AI maturity does not replace AI literacy but radicalizes it by situating technical and ethical competencies within a broader normative framework of epistemic autonomy – a framework that the following sections develop in dialogue with Kant's account of maturity and immaturity.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

In what follows, we shall examine Kant's theory of immaturity (*Unmündigkeit*) in greater depth as it provides a powerful normative lens for understanding what is at stake in contemporary educational engagements with AI. For Kant, immaturity does not consist in a lack of intelligence, information, or technical capacity, but in a specific mode of dependence: the inability or unwillingness to make use of one's own understanding without the guidance of another (Kant 1996, 17; 8:35). Crucially, this condition is self-incurred. Immaturity arises not because external authorities exist, but because subjects voluntarily delegate the labor of judgment, orientation, and justification to them. Kant identifies above all "[p]recepts and formulas," i.e., human rules and instructions, as "mechanical instruments of a rational use, or rather misuse, of his natural endowments" and describes them as "the ball and chain of an everlasting minority" (Kant 1996, 17, 8:36). We blindly follow these precepts and formulas without critically reflecting on them, and even take them to be our own, self-given laws. This tendency, as criticized by Kant, to make ourselves immature through our uncritical use of media can be applied to the current development of digitalization. The Internet and artificial intelligence in particular are increasingly being used as a substitute for our own thinking, when all we have to do is enter search queries and the seemingly objective answers are automatically displayed. Artificial intelligence is increasingly automating and replacing our power of judgment, and algorithms are similar to the "precepts and formulas" criticized by Kant, which in the worst-case scenario become the aforementioned "ball and chain (*Fußschellen*) of an everlasting minority" in everyday digital life (Kant 1996, 17; 8:36).

Transposed into the context of generative AI, this diagnosis receives renewed urgency. AI systems do not merely supply information; they increasingly function as mediating authorities that propose interpretations, arguments, and ready-made judgments. When these outputs are adopted without reflective endorsement, the user risks substituting epistemic autonomy with algorithmic convenience.

From a Kantian perspective, epistemic autonomy is not equivalent to epistemic isolation. To think for oneself does not mean to think without assistance, but to remain the *author* of one's judgments. Autonomy is preserved only insofar as external resources – whether books, teachers, or AI systems – are integrated into one's reasoning as aids rather than as replacements. Immaturity, by contrast, emerges when this distinction collapses: when the mediation itself becomes opaque and the subject no longer experiences judgment as an activity for which she is responsible. In the case

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

of generative AI, this risk is amplified by the system's fluency, apparent neutrality, and responsiveness, which can create the illusion that epistemic labor has been successfully outsourced rather than silently abdicated.

Kant's famous examples of immaturity – having a book that understands for me, a guardian who decides for me – find a contemporary analogue in AI systems that draft arguments, summarize complex issues, or recommend decisions. The problem is not their use, but their *uncritical* use: the transformation of assistance into substitution. Epistemic autonomy is compromised not when AI informs judgment, but when it *preempts* it. In this sense, AI immaturity is a specific form of epistemic heteronomy: a condition in which one continues to reason, yet no longer recognizes oneself as the source of the norms governing that reasoning.

Accordingly, AI maturity must be understood as a cultivated disposition rather than a technical skill. It consists in the capacity to appropriate AI-mediated outputs reflectively, to interrogate their assumptions, limits, and purposes, and to situate them within one's own justificatory horizon. This involves what Kant describes as the courage to think for oneself (*sapere aude*), transposed into a digital environment in which thinking increasingly unfolds through algorithmic mediation. Epistemic autonomy under conditions of AI thus requires not withdrawal from technology, but a disciplined form of engagement: one that resists the temptation of effortless cognition and preserves judgment as a first-person responsibility.

Against this backdrop, the educational task extends beyond teaching students how to use AI tools effectively, and instead involves learning how to remain epistemically accountable in their use. AI maturity accurately describes this normative achievement: the ability to employ powerful cognitive technologies without allowing them to erode the conditions of self-determination. It reframes the question from what AI can do for learners to what learners must still do for themselves – namely, to judge, to justify, and to take responsibility for their beliefs in an age of increasingly persuasive artificial interlocutors.

These considerations receive further urgency when viewed through the lens of cognitive offloading. Recent research shows that generative AI systems go beyond merely supporting learning processes in that they can subtly reconfigure them by shifting core epistemic activities – such as reasoning, evaluating, and synthesizing – from the learner to the system. While cognitive offloading is not inherently problematic, it becomes normatively significant when it affects those capacities that are constitutive of epistemic agency itself. In educational contexts, the risk is not simply a loss

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

of effort or engagement, but a risk of gradually weakening the learner's role as the subject of judgment. Cognitive offloading thus marks a threshold where assistance turns into substitution, and support into dependency.

Interpreted in Kantian terms, AI-induced cognitive offloading exemplifies a contemporary form of self-incurred immaturity. Learners do not cease to think altogether; rather, they engage in a diminished mode of thinking in which the burden of justification is tacitly delegated to algorithmic outputs. This is why AI immaturity should not be understood as a lack of rationality, but as a misrelation to one's own rational capacities. Even when students actively prompt, select, or edit AI-generated content, they may still abdicate responsibility for the standards governing correctness, relevance, or insight. What is at stake, therefore, is the locus of epistemic authority rather than the presence or absence of cognition.

From this perspective, cognitive offloading reveals why AI maturity cannot be reduced to ethical compliance or technical skill. A student may use AI transparently, legally, and efficiently, and yet remain epistemically immature if its use preempts acts of judgment that ought to be exercised by the learner herself. AI maturity, by contrast, requires the capacity to distinguish between those cognitive tasks that may legitimately be delegated and those that must remain inalienable. Judging, evaluating reasons, and taking responsibility for claims belong to the latter category. They define the minimal conditions under which learning remains an exercise of autonomy rather than a form of guided passivity.

Accordingly, AI maturity can be understood as a regulative ideal for educational practice under conditions of generative AI. It articulates a normative threshold: AI may extend cognition, but it must not replace the learner as the author of her judgments. This ideal does not demand resistance to technology, but reflective integration. It asks when, how, and to what extent AI systems should be used so that they strengthen rather than undermine epistemic autonomy. Cognitive offloading thus functions as a critical diagnostic concept – one that exposes the risk of digital immaturity and, at the same time, clarifies the positive task of education: to cultivate learners who can engage AI without surrendering their capacity to think, judge, and justify for themselves.

From a Kantian standpoint, the phenomenon of cognitive offloading can be interpreted as a paradigmatic case of epistemic heteronomy under contemporary technological conditions. Heteronomy, in Kant's sense, arises when the principles governing one's judgments are no longer self-legislated but tacitly adopted from an external authority. In the context of generative AI, this authority does not appear as an explicit command or

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

rule, but as a seemingly neutral and efficient cognitive surrogate. Cognitive offloading becomes epistemically problematic not when learners rely on external aids as such, but when they allow AI systems to determine what counts as relevant, plausible, or sufficient without reflective endorsement.

Empirical research on AI-supported learning suggests that frequent reliance on AI tools is associated with lower critical-thinking performance, mediated by increased cognitive offloading (Gerlich 2025). Aside from the reduction in performance, the shift in the structure of epistemic agency itself is normatively significant. When reasoning, evaluation, and synthesis are routinely delegated to AI systems, learners may still participate in epistemic practices, yet no longer experience themselves as accountable for the standards by which claims are assessed. This marks a transition from autonomous judgment to heteronomous cognition: the subject continues to think, but no longer governs the conditions under which thinking counts as justified or true.

In Kantian terms, such heteronomy does not eliminate rational activity; rather, it displaces its normative center. The learner remains cognitively active – prompting, selecting, editing – but the decisive moments of epistemic evaluation increasingly occur outside her reflective control. Cognitive offloading thus produces a form of *passive rationality*: one reasons within a framework whose norms are implicitly set by algorithmic systems optimized for efficiency, plausibility, and coherence, rather than by the learner's own critical standards. This is why AI-induced heteronomy is particularly insidious. It does not appear as coercion or deception, but as convenience.

This diagnosis clarifies why AI maturity cannot be equated with moderation or balance alone. The issue is not the quantity of AI use, but the quality of epistemic dependence it engenders. From the perspective of enlightenment, the decisive question is whether learners retain the capacity to suspend, contest, and revise AI-generated outputs on the basis of their own reasons. Where this capacity atrophies, cognitive offloading becomes a vehicle of self-incurred immaturity: a voluntary relinquishment of epistemic self-governance in exchange for reduced cognitive effort.

AI maturity, understood against this backdrop, describes the reflective capacity to resist epistemic heteronomy even under conditions of pervasive algorithmic assistance. It requires learners to remain aware of the normative division of labor between human judgment and artificial support – to know not only *how* to use AI, but also when not to let it decide. Education oriented toward AI maturity must therefore cultivate sensitivity to the boundary between legitimate cognitive delegation and the alienation of

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

judgment. Only where this boundary is actively maintained can AI function as an instrument of epistemic empowerment rather than as a subtle tutor of dependency.

On a fundamental level, AI maturity can be understood as a rhythmical achievement: the cultivated capacity to govern the temporal structure of human–AI interaction in a way that preserves epistemic autonomy. Generative AI systems operate at a pace that radically exceeds the speed of human reflective cognition. Large language models generate coherent linguistic output almost instantaneously, while human judgment unfolds slowly, requiring hesitation, comparison, revision, and sustained attention. This asymmetry of tempos is not an accidental side effect of technological progress; it constitutes one of the central normative challenges posed by generative AI.

The epistemic significance of this speed differential extends beyond the risk of error or misinformation, and lies in its effect on the conditions under which judgment is exercised. When interaction with AI unfolds at machine speed, thinking is drawn into a mode of temporal compression. Responses arrive faster than they can be critically reconstructed; coherence precedes evaluation; plausibility substitutes for justification. In such contexts, understanding appears as an experiential effect of fluency rather than as the outcome of reflective endorsement. What emerges is a form of epistemic ease – a smooth, uninterrupted flow of meaning that feels cognitively satisfying while bypassing the labor of judgment.

This phenomenon can be described as a technologically induced acceleration of cognition. Generative AI does more than merely supply content; rather, it reshapes the form and tempo at which epistemic activity takes place. By producing continuous, adaptive, and seemingly complete responses, it invites users into a mode of interaction characterized by immediacy and low resistance. From the perspective of dual-process theories (e.g. Kahneman 2011), this acceleration may privilege fast, intuitive, and associative modes of thinking, while marginalizing slower, effortful processes of deliberation, evaluation, and self-critique. System-1 dynamics are amplified not because users abandon reflection voluntarily, but because the environment renders reflection temporally costly and phenomenologically unnecessary.

The epistemic danger that arises from this acceleration is therefore subtle. It does not take the form of external domination or explicit authority. Rather, it manifests as a shift in the locus of epistemic control. When AI-generated outputs arrive before reflective scrutiny can be mobilized,

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

the decisive moments of epistemic orientation – setting standards of relevance, determining sufficiency, weighing reasons – are displaced from the subject to the system. The user remains cognitively active, selecting prompts, skimming responses, and approving formulations, yet the normative structure governing what counts as a good answer is increasingly shaped by the machine's tempo.

In Kantian terms, this dynamic constitutes a contemporary form of epistemic heteronomy. A subject lacking the capacity to judge does not constitute heteronomy, but a judgment no longer being exercised under self-legislated temporal conditions does. The subject allows the rhythm of an external system to determine when thinking is complete, when inquiry may stop, and when reasons need no further examination. This is why such heteronomy remains self-incurred: rather than being imposed by constraint, it is accepted through convenience. The speed of generative AI makes epistemic delegation feel natural, even rational, while obscuring the normative loss involved.

From this perspective, AI maturity cannot be adequately defined in terms of competence, literacy, or even ethical compliance alone. Of importance is not necessarily how AI is used, but at what tempo epistemic agency is exercised. Maturity requires an awareness that epistemic autonomy is temporally fragile: it depends on the availability of time for reflection, interruption, and second-order questioning. To be mature in one's use of generative AI is therefore to resist the conflation of speed with insight and fluency with understanding.

AI maturity consists in the deliberate regulation of epistemic rhythm. It denotes the cultivated ability to introduce pauses, friction, and reflective distance into interactions that are otherwise designed to be seamless and absorbing. This may involve slowing down the interaction itself, breaking dialogical flows, reconstructing reasons independently of AI output, or explicitly suspending judgment in order to regain critical orientation. Rather than opposing technological acceleration, such practices counterbalance it. They allow human reflective tempo to reassert itself within an environment optimized for speed.

This rhythmical regulation marks the decisive boundary between epistemic extension and epistemic replacement. Generative AI can extend human cognition only insofar as human judgment retains temporal priority. The question posed must remain one the subject can own; the evaluation performed must require time that cannot be compressed without loss; the conclusion reached must be one for which the subject can answer.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

Where the system's pace determines these moments, extension collapses into substitution. Cognitive offloading becomes structural rather than instrumental, and epistemic autonomy gives way to heteronomy.

AI maturity, understood in this way, is not a stable state but an ongoing effort. It requires continuous vigilance toward the seductive ease of accelerated cognition and the ever-present temptation to let fast systems think in one's place. To be mature with AI is not to think faster with machines, but to learn how to think slowly with them – to preserve the temporal conditions under which judgment, responsibility, and epistemic self-governance remain possible in an age of generative speed.

This section has established the conceptual groundwork for distinguishing AI literacy from AI maturity. While AI literacy concerns competencies, skills, and informed use, the section has argued that these remain normatively insufficient unless situated within a broader account of epistemic autonomy. Drawing on a Kantian framework, the analysis has shown that the central risk of generative AI lies not in ignorance or misuse, but in self-incurred cognitive offloading: the voluntary delegation of judgment that constitutes a contemporary form of immaturity. Against this background, AI maturity emerges as a normative ideal that cannot be reduced to proficiency alone, but must be understood as the sustained capacity to remain the author of one's own judgments under conditions of algorithmic mediation. Building on this theoretical distinction, Section 3 turns from conceptual clarification to empirical and practical analysis. Its aim is to examine how learners can move from self-incurred epistemic offloading – understood as a form of immaturity – toward epistemic autonomy in concrete educational contexts. The section therefore investigates where the boundary between extension and replacement of judgment lies, how AI use can undermine or support autonomy, and which criteria allow educators and learners to recognize when AI ceases to function as an aid and begins to operate as a surrogate for judgment.

3 The Challenge: From Cognitive Offloading to Epistemic Autonomy

Generative AI, particularly chatbots, can function as epistemic surrogates in educational contexts, depending on how we use them. While instrumental offloading – i.e., using AI to calculate or organize data – can extend human purposes, structural offloading undermines epistemic autonomy by letting AI dictate purposes and reasoning paths. Drawing on empirical research, this section analyzes how reliance on AI-generated summaries, arguments,

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

or even essay drafts can reduce epistemic autonomy. Interpreted through a Kantian lens, this offloading constitutes a form of self-incurred immaturity: the failure to exercise one's own judgment despite being capable of it. The section distinguishes between cases where AI extends human judgment and cases where it replaces it, and develops criteria for responsible integration. It argues that educators must not only discourage misuse but also explicitly teach students how to recognize when AI use undermines autonomy. Ultimately, the challenge is to harness AI as an extension of cognition without collapsing into dependency – a balance that requires cultivating maturity rather than literacy alone.

To further clarify why AI maturity must be understood relationally rather than as an individual competence or a technological property, it is necessary to examine the structure of human–technology relations themselves. Questions of maturity and immaturity arise not at the level of isolated users or autonomous systems, but within concrete patterns of interaction in which cognition, agency, and responsibility are technologically mediated. For this reason, the following discussion turns to Don Ihde's phenomenology of human–machine relations. Ihde's (2009) framework provides conceptual tools for analyzing how technologies co-shape perception, action, and judgment – and thus for understanding how interaction with AI can function either as an extension of epistemic autonomy or as a source of epistemic heteronomy. It is within this interactional framework that AI maturity and AI immaturity can be understood as normative modes of human–machine interaction rather than as features of AI itself.

Don Ihde (2009) distinguishes four basic human–technology relations – embodiment, hermeneutic, alterity, and background relations. While each captures an important dimension of our interaction with generative AI, no single relation appears sufficient on its own: generative AI is simultaneously interpreted like a text, addressed as a quasi-other, and increasingly integrated into the processes through which thoughts, judgments, and arguments are formed. Building on Ihde's (2009) typology, this Element proposes the "extension relation" as a new category that captures the evolving integration of AI into human learning and action. Generative AI is based on the principles of pattern recognition and probability calculation. Based on enormous amounts of data, AI systems are trained to abstract and recognize patterns and to (re)apply them to other areas. Such patterns are, among other things, linguistic. This makes it possible to have conversations with chatbots like ChatGPT that sometimes can no longer be distinguished from real interpersonal communication. It is also possible to give chatbots specific instructions through so-called

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

"prompts." The reason for this is that AI identifies the linguistic patterns that are most likely to match a question or instruction posed by a human. AI's answers are therefore not obtained as a result of structured reasons or scientific insight, which could be made transparent and explained in detail, but by mere probability calculations that give the appearance of scientificity and objectivity. Following Harry Frankfurt's account, this problem of AI has recently been called "bullshit," which he has defined as a "lack of connection to a concern with truth" – an "indifference to how things really are" (Frankfurt 2005, 33–4). With regard to AI chatbots, Hicks et al. (2024, 1) have argued that "[b]ecause these programs cannot themselves be concerned with truth, and because they are designed to produce text that *looks* truth-apt without any actual concern for truth, it seems appropriate to call their outputs bullshit."

However, describing only the outputs of AI chatbots as "bullshit" does not consider the problematic human–AI interaction involved. An educational critique of AI tools in light of AI maturity needs to discuss the *ways* in which we need to use these tools, and identify those uses that are compatible with epistemic autonomy; that is, uses that allow us to learn or improve our cognitive abilities. Improving cognitive abilities by means of digital media has been discussed in terms of "digital literacy" and "digital competence" (Spante et al. 2018; Southworth et al. 2023). The "Digital Competence Framework for Citizens" (DigComp) of the European Union understands it as "one of the Key Competences for Lifelong Learning" (Vuorikari et al. 2022, 3; Vallor 2024), and describes it in terms of epistemic capacities, since it "involves the confident, critical and responsible use of, and engagement with, digital technologies for learning, at work, and for participation in society." Accordingly, a critical account of AI maturity needs to address the question of how to deal with these chatbots in order to avoid bullshit, and to use them in a way that extends our epistemic autonomy.

Epistemic immaturity in the age of generative AI does not constitute a single failure mode, but a family of structurally related phenomena. What unifies these forms is not ignorance or cognitive incapacity, but a patterned delegation of epistemic responsibility under technologically mediated conditions. From a Kantian perspective, immaturity remains self-incurred: agents retain the capacity to judge, yet allow their epistemic agency to be reshaped in ways that displace reflective self-governance. Against this background, at least four analytically distinct forms of epistemic immaturity can be identified.

The first and most direct form is cognitive offloading. Here, generative AI is used not merely to support epistemic activity, but to substitute for core

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

practices of reasoning, interpretation, and justification. While offloading per se is not problematic, immaturity emerges when reliance on AI becomes habitual and opaque, such that agents no longer reappropriate the delegated cognitive function. Judgment is not exercised but presupposed as already completed by the system. This form of immaturity manifests as a loss of epistemic ownership: agents continue to act and decide, yet no longer experience themselves as the authors of the reasons that ground their beliefs.

A second form consists of algorithmically mediated epistemic narrowing, often described in terms of filter bubbles or bias amplification. Generative and recommender systems increasingly prestructure the informational environment by filtering relevance, prioritizing perspectives, and stabilizing interpretive frames. Epistemic immaturity arises when agents mistake these curated environments for neutral or comprehensive representations of reality. In Kantian terms, judgment becomes confined to "subjective private conditions," as the capacity to abstract from one's own standpoint and to test judgments against a shared space of reasons is eroded. What is lost is not access to information as such, but the reflexive distance required for epistemic autonomy.

A third, more subtle form of immaturity is rooted in technological affordances. Generative AI systems are designed to invite certain forms of interaction – ease, speed, fluency, and apparent completeness – while discouraging others, such as hesitation, contestation, or epistemic friction. These affordances do not coerce users, but they nudge epistemic practices toward passive acceptance and minimal engagement. Immaturity arises when agents unreflectively adapt their epistemic behavior to what the system facilitates, thereby confusing convenience with understanding. The result is a technologically induced passivity of judgment that masquerades as epistemic competence.

A fourth and structurally deeper form concerns economic and institutional logics embedded in AI systems. Most large-scale generative models are developed and deployed within incentive structures oriented toward efficiency, engagement, scalability, and profit. These logics shape not only what systems do, but how they are presented and normalized. Epistemic immaturity emerges when users tacitly internalize these priorities, allowing epistemic success to be measured in terms of productivity, speed, or output quantity rather than justification, truth, or understanding. In such cases, epistemic agency is subtly aligned with external purposes that agents neither set nor critically endorse.

These forms of epistemic immaturity are mutually reinforcing. Cognitive offloading is facilitated by affordances; affordances operate within biased

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

informational environments; and all are stabilized by economic structures that reward epistemic delegation. Crucially, none of these phenomena eliminate agency outright. Instead, they exemplify what can be called technologically mediated heteronomy: a condition in which agents remain formally active but normatively displaced. The value of this taxonomy lies in showing that epistemic immaturity under generative AI is not an accidental misuse, but a predictable outcome of ungoverned extension relations – an outcome that can only be addressed by cultivating epistemic autonomy as an active, reflective, and institutionally supported practice rather than a mere individual virtue.

Among the various forms of epistemic heteronomy identified above, cognitive offloading occupies a structurally central position. While filter bubbles, biased informational environments, affordance-driven passivity, and economic logics all contribute to the erosion of epistemic autonomy, they do so primarily by enabling or stabilizing offloading. Cognitive offloading marks the point at which heteronomy becomes internalized: judgment itself is no longer merely influenced or guided, but functionally delegated. For this reason, offloading constitutes the most direct and normatively salient threat to epistemic autonomy. It is here that the transition from epistemic extension to epistemic replacement becomes explicit.

Within an extended-judgment framework, structural offloading is best understood as a misconfiguration of the extension relation: instead of AI functioning as an artificial capacity that expands the space in which human judgment operates, the system comes to supply the teleological core of inquiry by supplying guiding aims, framing problems, and generating candidate ideas. In mature use, by contrast, the initiating moment of judgment remains with the human epistemic subject: the agent articulates the question, determines what is at stake, and sets provisional standards of relevance and adequacy, while AI is used to explore possibilities, generate alternatives, simulate objections, or refine conceptual options. The decisive point is normative asymmetry: AI may increase epistemic performance, but it cannot carry epistemic authority. Structural offloading violates this asymmetry by treating AI-generated framings and proposals as if they were normatively self-sufficient, thereby relocating the "center" of judgment outside the subject. This is precisely where epistemic immaturity becomes internalized: the learner still appears active – prompting, selecting, editing –, but the constitutive work of judgment – reflective assessment, weighing reasons, revising standards – is displaced by a system optimized for fluent continuation.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

The phenomenon of hallucination makes this mislocation especially evident. Hallucinations are not merely occasional glitches; they express a structural tendency of generative models to guess plausibly under uncertainty rather than to anchor claims in accountable justification (Kalai et al. 2025). If the epistemic impulse does not originate in the subject – in the form of a question one can own, a standard one can defend, and a reason-sensitive orientation toward truth –, then AI-generated content will tend to drift toward confident-sounding fabrication. This is also why the "bullshit" diagnosis matters for educational contexts: potential error aside, the risk lies in the emergence of a mode of discourse that is largely indifferent to truth while mimicking the surface marks of truth-directed assertion (Hicks et al. 2024). In other words, hallucination and bullshit converge as symptoms of a single normative confusion: mistaking the production of convincing linguistic output for the exercise of judgment.

From the perspective of extended judgment, the pedagogical lesson is straightforward. The first, spontaneous epistemic ground must remain situated within the learner, because only under this condition can norms become operative in concrete cases – only this way can relevance be set, evidence demanded, reasons weighed, and claims owned as answerable. When this spontaneity is preserved, AI can genuinely extend judgment by enlarging the field of considerations without substituting for the act of judging itself. When spontaneity is lost, extension collapses into replacement: AI does not merely help the learner think; it begins to think for the learner by supplying goals and ideas that the learner adopts without reflective appropriation. The resulting outputs may look sophisticated, but they are epistemically hollow – simply because the normative source of judgment has been externalized.

Accordingly, the following analysis focuses on cognitive offloading as the paradigmatic mechanism through which generative AI transforms self-governance into dependence – and thus as the critical leverage point for understanding, diagnosing, and counteracting epistemic immaturity.

Empirical research increasingly supports this philosophical diagnosis by suggesting that cognitive offloading through AI use is not merely a redistribution of cognitive effort, but a transformation of epistemic agency itself. Studies across educational contexts indicate that frequent reliance on AI-generated summaries, arguments, or full drafts is associated with diminished critical engagement and weakened ownership of epistemic outcomes (Qian 2025; Georgiou 2025). Gerlich (2025), for instance, reports a robust negative correlation between intensive AI tool use and critical

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

thinking performance, particularly among younger users. What is normatively significant here is not the mere presence of assistance, but the pattern of reliance: users increasingly defer both execution and evaluation to the system, thereby shifting the locus of epistemic authority away from themselves.

Similar concerns emerge in higher education research. Georgiou (2025) found that students randomly assigned to complete an argumentative writing task with ChatGPT reported significantly lower cognitive engagement – mental effort, attention, deep processing, and strategic thinking – than an unassisted control group, while Qian's (2025) systematic review identifies overreliance and the outsourcing of critical cognitive processes as a recurrent concern across the field. What such findings put at stake is engagement in processes of synthesis and justification – which are the very activities through which epistemic autonomy is exercised. A concrete example can be observed in university writing courses where students use generative AI to produce essay drafts before independently engaging with the underlying arguments or sources. In such cases, students may submit linguistically coherent texts while remaining unable to reconstruct or defend the reasoning contained in them during oral discussion or assessment situations. This illustrates how AI can simulate epistemic performance while weakening actual understanding.

The findings suggest that AI-mediated learning environments may encourage a mode of interaction in which students monitor and minimally edit outputs rather than actively construct arguments. From a Kantian perspective, this marks a shift from autonomous judgment to heteronomous oversight: learners remain involved, yet no longer legislate the norms that govern the validity of their claims.

A systematic scoping review by Yan et al. (2024) further reinforces this concern by documenting how LLM-based educational technologies increasingly automate core epistemic tasks, including content generation, feedback, and evaluation. While such automation promises efficiency, it also raises ethical challenges related to transparency, accountability, and human-centered design. Evaluated through the framework developed here, what these implementations insufficiently address is the risk that learners may come to treat AI outputs as epistemically authoritative by default. When systems present synthesized, fluent responses without exposing their justificatory structure, users are less likely to question assumptions or consider alternative perspectives. The result is a subtle form of epistemic heteronomy: judgment is no longer guided by reasons one can fully appropriate, but by outputs whose normative grounds remain opaque.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

These behavioral findings find a tentative parallel in neuroscientific work. Kosmyna et al. (2025) report that sustained use of AI assistants in essay-writing tasks correlates with reduced neural connectivity in regions associated with memory, integration, and executive control. Participants who relied on AI showed significantly lower ability to recall or quote their own arguments and reported diminished ownership of their texts. Crucially, these effects persisted even when AI assistance was later removed, suggesting not a momentary convenience but an accumulated form of cognitive debt. Interpreted normatively, this supports the claim that cognitive offloading can erode the conditions of epistemic self-relation: learners no longer experience themselves as the source of their judgments, even when they remain formally responsible for the output.

Taken together, these studies illuminate why cognitive offloading poses a distinctive threat to epistemic autonomy. The empirical findings discussed above therefore support the philosophical distinction developed throughout this Element between epistemic extension and epistemic replacement. They suggest that the normative risks associated with generative AI are already observable in concrete educational practices and measurable patterns of human–AI interaction.

The problem is not that AI reduces cognitive load as such, but that it selectively displaces those cognitive activities – reason-giving, weighing alternatives, internalizing arguments – that constitute autonomous judgment. In Kantian terms, this amounts to epistemic heteronomy: the subject allows externally generated reasoning to stand in for her own, not because she lacks the capacity to judge, but because the technological environment renders such judgment seemingly unnecessary. AI immaturity thus appears as a contemporary form of self-incurred immaturity, sustained not by coercion or ignorance, but by convenience and efficiency.

This diagnosis clarifies the normative task of education under conditions of generative AI. If learners are to remain epistemically autonomous, they must be taught to recognize when AI use crosses the threshold from extension to replacement. Cognitive offloading becomes educationally problematic precisely where it undermines the learner's role as the author of reasons. Cultivating AI maturity therefore requires more than discouraging misuse: it demands explicit criteria and practices that enable students to preserve judgment as a first-person responsibility within an environment increasingly populated by persuasive epistemic surrogates.

Concerns about AI use in educational contexts often emerge around the phenomenon of cognitive offloading. At first sight, offloading appears epistemically suspect, as if any delegation of cognitive work to artificial

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

systems necessarily undermined autonomy. This conclusion, however, would be premature (Iqbal et al. 2025). Not all forms of offloading are problematic. Instrumental offloading – the delegation of well-defined subtasks to external tools – has long been an integral part of human cognition. The use of a calculator, a spell-checker, or a search engine does not in itself threaten epistemic autonomy, provided two conditions are met. First, instrumental offloading must free cognitive resources for epistemically more demanding and normatively significant tasks, such as interpretation, evaluation, or justification. Second, the agent must retain oversight over what is being offloaded and how the tool contributes to the overall epistemic activity. In this sense, instrumental offloading functions as an enabling condition for autonomy rather than its negation.

The problem lies with what can be termed structural offloading. Unlike instrumental delegation, structural offloading affects not merely the efficiency of epistemic activity, but its architecture. When AI systems are used in ways that replace the practices of reflecting, weighing reasons, and forming judgments, they no longer operate as auxiliary instruments but as epistemic surrogates. Structural offloading changes how epistemic agency is exercised by relocating core elements of judgment – such as relevance-setting, argumentative structure, or justificatory coherence – into opaque algorithmic processes. In such cases, the agent does not merely rely on external support, but allows the very process of judgment to be substituted rather than extended.

This distinction is crucial for understanding the extension relation between human epistemic agency and AI systems. Extension does not consist of outsourcing judgment, but in reconfiguring the conditions under which judgment is exercised. An AI system extends epistemic autonomy only insofar as it remains embedded within practices that preserve the agent's responsibility for epistemic outcomes. For example, students using source-grounded AI systems such as Gemini Notebook (previously NotebookLM) to compare recurring concepts across philosophical texts may significantly accelerate research and orientation while still remaining responsible for interpretation and argumentative integration. Here, AI functions as an epistemic extension rather than as a substitute for judgment.

Extension is therefore normatively asymmetrical: while epistemic performance may be technologically supported, epistemic authority must remain human. AI systems can generate candidates for interpretation, propose argumentative structures, or simulate counterpoints, but they cannot assume responsibility for their epistemic validity.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

From an educational perspective, this asymmetry is decisive. The aim of education is not the production of correct outputs, but the cultivation of capacities – in particular the capacity for judgment. When AI systems are integrated into learning environments as extensions of judgment, they can support epistemic autonomy by scaffolding reflection, making reasoning explicit, and expanding the space of possible considerations. Yet when they are treated as replacements for judgment, they risk fostering a form of epistemic dependency that is compatible with surface-level competence but incompatible with maturity.

Extending epistemic autonomy under conditions of AI mediation therefore requires more than technical proficiency. It requires cultivating the ability to distinguish between extension and replacement in concrete practices. Learners must be able to recognize when AI use supports their own reasoning and when it silently restructures epistemic activity in ways that bypass it. Epistemic autonomy, in this sense, is not undermined by dependence as such, but by unreflected dependence. The educational task is thus to enable agents to govern their epistemic dependencies rather than to eliminate them – preserving judgment as the normative core of epistemic agency even in technologically extended contexts.

From the perspective of AI maturity, the design of AI systems must aim to *support* epistemic autonomy without ever attempting to replace it. No degree of transparency, explainability, or alignment can transfer epistemic responsibility from the human agent to the system itself. Even explainable AI can only render aspects of system behavior intelligible; it cannot supply the first-person act of reflection through which reasons are assessed, endorsed, or rejected. For this reason, AI systems themselves cannot be epistemically autonomous or mature. Epistemic autonomy is not a property of systems, but of the human agents interacting with them.

The normative task of design is therefore asymmetrical. AI systems should be constructed in ways that invite reflection, interruption, and contestation – by making assumptions visible, exposing uncertainty, and enabling users to interrogate outputs rather than merely consume them (Buijsman et al. 2025). This design orientation has a counterpart in virtue epistemology itself, which has come to treat the quality of an agent's information environment as a condition of epistemic reliability, rather than locating reliability exclusively in the dispositions an agent brings to inquiry. On this view, the epistemic standing of a belief depends in part on the setting in which it was formed, so that improving that setting is itself an epistemic intervention (cf. Turri et al. 2021, §9). Yet these design features can only scaffold autonomy; they cannot enact it.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

AI maturity emerges exclusively at the level of human–AI interaction, where users actively appropriate, challenge, or revise algorithmic contributions. Where design aims at substitution – by minimizing friction or preempting judgment –, it undermines the very conditions of maturity it claims to support.

To clarify the boundary between epistemic extension and epistemic replacement, it is necessary to articulate criteria that distinguish cases in which AI-supported cognition preserves human judgment from those in which it silently supplants it. The distinction is not primarily technical but normative. It concerns the locus of epistemic authority, the transparency of epistemic purposes, and the degree to which judgment remains an exercised human capacity rather than a delegated function.

First, teleological transparency marks a decisive criterion. Epistemic extension presupposes that the epistemic purpose of an activity – understanding, justification, orientation, or critique – remains explicit and governable by the human agent. AI extends judgment when it supports the pursuit of such purposes without redefining them. Replacement occurs when the system implicitly determines what counts as a satisfactory answer, a relevant distinction, or a completed inquiry. In such cases, the agent no longer orients herself toward epistemic ends but toward system-generated closure. This amounts to a loss of self-legislation in epistemic matters: the maxim of inquiry is no longer set by the subject but absorbed by the operational logic of the system.

Second, normative asymmetry provides a further criterion. In genuine epistemic extension, AI outputs are treated as proposals, scaffolds, or occasions for judgment – never as normatively self-sufficient. Judgment should always be the act through which validity is attributed, reasons are weighed, and commitments are undertaken. Replacement occurs when outputs are accepted as epistemically authoritative simply by virtue of their fluency, coherence, or apparent completeness. In those cases, epistemic normativity is tacitly relocated from human judgment to algorithmic performance. Kantian epistemic autonomy is thereby undermined, not because the subject lacks access to reasons, but because she no longer takes responsibility for endorsing them.

Third, revisability and contestability function as practical indicators. Extended judgment remains open-ended: AI-assisted reasoning can be interrupted, questioned, redirected, or rejected by the user without friction. The system expands the space of possibilities but does not stabilize conclusions. Replacement takes place when AI-generated outputs become default endpoints – summaries that are no longer unpacked, arguments

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

that are no longer reconstructed, interpretations that are no longer challenged. At this point, cognitive offloading becomes structural rather than instrumental: the activity of judging is no longer exercised but presupposed as already completed elsewhere.

Fourth, epistemic effort and friction play a constitutive role. Extension redistributes epistemic effort rather than eliminating it. AI may reduce the costs of search, articulation, or comparison, yet it still requires active engagement, interpretive labor, and critical uptake. Replacement, by contrast, minimizes epistemic friction to the point where understanding is merely simulated. From a Kantian standpoint, this erosion of effort is normative: it weakens the cultivation of the very capacity whose exercise defines maturity, namely, the ability to judge without external tutelage.

Finally, responsibility attribution serves as a decisive test. In cases of epistemic extension, responsibility for epistemic outcomes – errors, misunderstandings, or insights – remains unambiguously human. The agent can answer for her judgments because she recognizes them as her own. In cases of replacement, responsibility becomes diffuse or displaced: errors are attributed to the system, successes to its efficiency. This type of logic constitutes epistemic heteronomy. Judgment has not been extended but outsourced, and the subject relates to her own beliefs as if they were produced elsewhere.

Taken together, these criteria show that the difference between extension and replacement of judgment does not hinge on the sophistication of AI systems but on the structure of human–AI coupling. Epistemic extension preserves judgment as an exercised capacity embedded in a technologically enriched environment. Epistemic replacement occurs when judgment is no longer exercised at all, but merely mirrored by algorithmic outputs. Interpreted through a Kantian lens, the latter constitutes a contemporary form of self-incurred epistemic immaturity: not the absence of cognitive resources, but the voluntary suspension of one's own judging activity in favor of epistemic convenience.

This distinction sets the normative foundation for the remainder of the section. The challenge is not to prevent epistemic extension, but to design, teach, and cultivate forms of AI use in which judgment remains active, responsibility remains human, and autonomy is preserved within – rather than against – an AI-extended epistemic ecology.

This analysis has shown that the central risk posed by generative AI in education lies not in its use as such, but in forms of cognitive offloading that silently displace epistemic responsibility. Distinguishing between extension and replacement of judgment clarifies when AI supports epistemic

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

autonomy and when it undermines it. Yet this distinction remains abstract unless it can be translated into concrete practices of interaction. If AI maturity is to be more than a theoretical ideal, it must become operative at the level of how users actually engage with AI systems. For this reason, the following section turns to prompting as a paradigmatic field of epistemic agency, examining how dialogical interaction with AI can function either as a vehicle of epistemic autonomy or as a subtle mechanism of dependence.

4 Prompting as Epistemic Virtue: Algorithmic Maieutics

Prompting is often introduced in educational and policy discourses as a component of digital or AI literacy: a set of technical skills that enable users to formulate effective instructions for interacting with generative systems (UNESCO 2023, 11–12). While this definition captures an important aspect of competence, and while the guidance insists that outputs must be critically evaluated before use, it remains conceptually incomplete. Critical evaluation is there framed as an external check on outputs rather than as constitutive of the epistemic relation itself (UNESCO 2023, 12). Prompting is not merely a technique for optimizing outputs, and is therefore better described as a normatively charged practice that structures the relation between human judgment and machine-generated responses. As such, prompting concerns not just what users can do with AI systems, but how they position themselves epistemically within human–machine interaction.

Prompting becomes epistemically virtuous only insofar as it structures and disciplines the user's own exercise of judgment. It concerns the way inquiry is initiated, sustained, and regulated under conditions of algorithmic mediation. Instead of merely eliciting an answer, a prompt commits the user to a certain orientation toward relevance, sufficiency, and responsibility.

This perspective also resonates with Baehr's (2019) account of intellectual virtues and critical thinking. Baehr argues that intellectual virtues such as open-mindedness, intellectual humility, and intellectual courage are partly constitutive of rational judgment itself. From this perspective, epistemically virtuous prompting requires more than technical competence: it presupposes cultivated dispositions that govern how users question, evaluate, and critically appropriate AI-generated outputs (Baehr 2021). Therefore, epistemically virtuous prompting can be characterized by a cluster of practical dispositions. These include attentiveness to

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

one's own epistemic aims (Why am I asking this?), restraint in delegating cognitive labor (What must remain my own task?), and responsiveness to resistance rather than mere fluency (Where does the system challenge or unsettle my assumptions?). Virtuous prompting is thus marked by a willingness to revise questions, to complicate initial framings, and to treat AI-generated responses as provisional occasions for further inquiry rather than as epistemic endpoints.

Practiced in this way, prompting functions as a self-reflexive discipline: it trains the user to remain aware of when inquiry is being extended and when it risks being replaced. The epistemic virtue lies less in extracting better answers and more in cultivating better questions – questions that keep judgment active, responsibility situated, and epistemic authority with the human subject. Understood in this stronger sense, prompting should be conceived of as a virtue rather than a skill. It is a cultivated disposition that governs how questions are asked, how answers are treated, and how responsibility for judgment is distributed.

A crucial implication follows from this diagnosis of sycophantic AI. If generative systems exhibit a structural tendency toward affirmation and agreement, then epistemic maturity cannot consist merely in detecting errors after the fact. Rather, it requires a proactive reconfiguration of interaction itself. Prompting strategies must be deliberately designed to seek resistance rather than confirmation.

This presupposes an explicit awareness of the epistemic affordance of generative AI: its default mode is plausibility through agreement instead of critical opposition. Left unchallenged, this affordance invites users into a comfort zone of epistemic self-affirmation, where coherence is mistaken for justification and fluency for insight. Against this tendency, mature prompting must open spaces of disagreement, alternative perspectives, and counterargumentation.

Such prompting does not treat AI as a bearer of public reason, but as a catalyst for it. The responsibility for public reason – understood in a Kantian sense as the obligation to expose one's claims to contestation, revision, and shared standards of justification – remains irreducibly with the human user. Generative AI does not exercise public reason; it can only simulate its surface features.

Consequently, epistemic maturity requires users to prompt in ways that actively disrupt agreement: by asking for objections to their own positions, by requesting rival interpretations, by forcing the system to articulate the strongest possible counterclaims, or by exposing hidden assumptions on both sides of the interaction. In this sense, prompting

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

becomes an ethical–epistemic practice: not the art of extracting confirmation, but the discipline of organizing dissent. Only where users consciously assume this responsibility can interaction with generative AI contribute to public reason rather than erode it through algorithmically mediated self-confirmation.

Unlike technical proficiency, which can be executed instrumentally and – once acquired – remains largely stable, virtuous prompting requires continuous exercise and self-reflection. It must be repeatedly enacted in response to changing contexts, shifting epistemic goals, and the ever-present temptation to delegate judgment. Framing prompting as a virtue thus highlights its role in sustaining epistemic autonomy: it emphasizes that meaningful interaction with generative AI depends less on command over the system than on the user's capacity to engage it critically, reflectively, and responsibly as part of an ongoing human–machine relation.

A complementary normative perspective can be found in Shannon Vallor's (2016) account of technomoral virtue. Vallor argues that instead of merely expanding human capacities, emerging technologies reshape the conditions under which moral character is cultivated. In her account, responsibility in technological contexts requires sustained attention, humility, and practical wisdom – virtues that enable agents to remain authors of their actions despite increasing technical mediation.

While Vallor's framework does not explicitly address generative AI as a dialogical epistemic partner, her emphasis on virtue as a socially embedded practice aligns closely with the present account of AI maturity. What this Element adds is a more fine-grained analysis of human–AI interaction as a site of extended judgment. In the context of generative AI, virtues such as attention and practical wisdom must be exercised not only in action, but also within iterative processes of prompting, evaluating, and revising machine-generated outputs. AI maturity thus extends Vallor's technomoral virtues into the epistemic domain, reframing them as capacities for governing judgment within hybrid human–machine practices rather than as traits of isolated agents.

Generative AI is unique in that it responds dialogically to prompts, rendering the form of engagement as important as the content. Prompting is a virtue rather than just a technical skill: a practice of intellectual discipline comparable to Socratic questioning. Borrowing from the metaphor of algorithmic maieutics, it shows how carefully crafted prompts can elicit insights, generate hypotheses, and expose contradictions – without replacing human judgment. The section develops a framework of prompting competences: precision (clarity of questions), contextualization

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

(providing relevant frames), role-setting (assigning perspectives), iteration (refining outputs), and meta-questioning (reflecting on assumptions). These competences counteract the asymmetry of command-based prompting, which risks reinforcing passivity. Instead, prompting becomes an exercise in dialogical autonomy, training students to think through their questions rather than consume ready-made answers. By cultivating the virtue of prompting, education can transform AI from a source of cognitive offloading into a partner in critical reasoning.

To make the normative idea of AI maturity operational, it is not sufficient to invoke autonomy or responsibility at an abstract level. What is required is a vocabulary that connects these ideals to concrete epistemic dispositions and practices. A virtue-theoretical perspective offers such a bridge. On the reading proposed here, Siegel's (2017) conception of "critical thinking" can be read as occupying a position between a Kantian emphasis on epistemic autonomy and an Aristotelian focus on intellectual virtues and character formation. While Siegel insists on the normative authority of reasons and autonomous judgment, he simultaneously emphasizes the cultivation of intellectual dispositions required for critical inquiry, thereby offering an intermediate framework between procedural rationality and virtue-based educational models.

Drawing on the tradition of classical intellectual virtues (Baehr 2021; cf. Pritchard 2023), we can describe AI maturity as a practice that requires more than correct or competent use, namely, engaging with generative systems in a patterned way that preserves judgment, responsibility, and epistemic agency. Table 2 translates the selected classical intellectual virtues into corresponding AI-related practices, illustrating how epistemic autonomy under conditions of algorithmic mediation can be cultivated, exercised, and assessed in everyday educational contexts.

To move beyond abstract norms, we can draw on Aristotle's account of intellectual and ethical virtues to conceptualize "AI virtues." These virtues provide a framework for cultivating dispositions that help students act responsibly and wisely in socio-technical environments. Aristotle's theory of virtue is grounded in the human lifeworld from both theoretical and practical points of view, and digital maturity needs to be learned and cultivated. We can thus speak of "AI virtues," which can be taught in the sense of the dianoetic virtues (the virtues of the mind) and cultivated in the sense of ethical virtues. In his *Nicomachean Ethics*, Aristotle defined intellectual virtues as growing through instruction, which is why they need experience and time. Ethical virtues, on the other hand, arise from habit (Aristotle, NE II.1, 1103a14–17). With regard to the question of digital maturity, the

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

Table 2 Classical and AI Virtues

Classical Intellectual Virtue	**Core Philosophical Meaning**	**AI-Related Practice**	**Concrete Example in AI Use**
Intellectual Humility	Awareness of the limits of one's knowledge; openness to correction	Treat AI outputs as *fallible proposals*, not authoritative answers	Asking: *"Which assumptions might this answer rely on?"* or *"Where could this be wrong?"*
Intellectual Perseverance	Willingness to sustain inquiry despite difficulty or uncertainty	Iterative prompting that resists premature closure	Refining questions instead of accepting the first fluent response
Intellectual Courage	Readiness to challenge dominant views and one's own beliefs	Using AI to generate *counterarguments* and dissent	Prompting: *"What is the strongest objection to my position?"*
Intellectual Autonomy	Capacity to govern one's own judgment	Retaining control over goals, criteria, and final evaluation	Using AI for exploration, not for deciding what counts as a good answer
Practical Wisdom (Phronesis)	Context-sensitive judgment about what matters	Selecting AI tools and modes of use according to epistemic task	Choosing source-grounded tools for facts, generative tools for ideation
Intellectual Honesty	Commitment to truthfulness and responsibility	Transparent disclosure and critical reconstruction of AI use	Explicitly documenting how AI shaped reasoning or writing
Attentiveness	Sustained focus on epistemic relevance	Resisting AI-induced acceleration and distraction	Pausing interaction to reflect before continuing the dialogue
Reflective Responsibility	Willingness to answer for one's judgments	Treating AI interaction as part of one's own justificatory practice	Being able to explain *why* an AI-generated suggestion was adopted or rejected

dianoetic virtues initially prove to be central, since here – beyond questions of the "right middle" of our media consumption – there are always epistemological questions involved that subsequently provide practical orientation in the digital world. In concrete terms, this means that digital dianoetic virtues allow us to define and open up a virtual space for action that is distinct from mere simulation, fiction, or illusion, and thus provides digital orientation. Within this space for action, concrete ethical virtues can then develop in the digital world through habit, by cultivating a "right middle" between extreme forms of our digital media use. Digital ethical virtues allow us to act in the virtual sphere in such a way that it is enlarged by meaningful networks and not reduced by digital bubble formation (Pariser 2011) or "AI bullshit" (Hicks et al. 2024). Every significant digital technology calls for its own dianoetic and ethical virtue, understood in the sense of its public and enlightened use of reason. With regard to artificial intelligence, a dianoetic virtue consists in understanding its technology as a means of increasing our autonomy, and not as "mechanical tools" from which forms of heteronomy – "precepts and formulas" – follow. By cultivating ethical AI virtue, we prevent ourselves from conceiving artificial intelligence as an autonomous agent or as a transhumanist surrogate for human cognition and practical agency. At the same time, we avoid using artificial intelligence as a mere tool whose "precepts and formulas" we use for cognitive offloading. This means that we use artificial intelligence primarily where it helps to expand our scope of orientation. Especially in view of the vast amount of information that we have to deal with in the digital world, algorithms can make our decisions easier and extend our digital power of judgment. First and foremost, such a space of orientation is a space of education. Artificial intelligence can help us to distinguish relevant from irrelevant information and thus obtain an overview that corresponds to our interests and goals. However, this only works if we have previously determined the purposes and rules that it should follow, and insofar as we remain aware that no database is free of bias. Building on this Aristotelian foundation, Constantinescu et al. (2021) argue that the cultivation of AI virtues must integrate both ethical and dianoetic virtues, with the latter – particularly *phronesis* – enabling agents to grasp the contextual circumstances of action and to deliberate well relative to those circumstances. This contextual sensitivity is crucial for moral responsibility in AI settings, where general rules cannot be mechanically applied to particular situations. From this perspective, digital maturity requires more than the habituation of virtuous behavior (ethical virtues); it also demands the intellectual capacity to discern the "right reasons" and "right circumstances" for action in dynamic

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

socio-technical environments. Constantinescu et al.'s framework situates this virtue-theoretical insight within the triadic model of Responsible AI: ethics by design (embedding appropriate deliberative capacities and contextual awareness into AI systems), ethics in design (ensuring that designers and developers exercise practical wisdom when creating and deploying AI), and ethics for designers (cultivating the moral and intellectual character of those involved in AI development through education, reflective practice, and ethical guidance). In educational contexts, this implies that AI literacy programs must not only teach the technical and epistemic aspects of AI, but must also foster the virtues required for responsible engagement with AI – enabling future users to act, in Aristotelian terms, with the right actions, for the right reasons, in the right circumstances. Applied to education, this means that AI literacy must be understood both as knowledge about AI and as the cultivation of habits and intellectual virtues that allow students to use AI technologies in ways that expand – rather than erode – their moral and cognitive capacities.

A productive normative point of reference for understanding human–AI interaction in educational contexts can be found in Socratic maieutics – an analogy also drawn, though not developed, in the generative AI literacy literature (Ng et al. 2025, 26). For Socrates, philosophical practice is not a matter of transmitting knowledge, but of facilitating insight through dialogue. In the *Theaetetus*, he famously describes himself as a midwife who does not impart knowledge of his own, but helps others bring forth what they themselves can think. Central to this practice is the elenctic method: by means of targeted questioning, implicit assumptions are uncovered, contradictions are exposed, and premature certainties are destabilized. Maieutics thus constitutes a practice of self-examination in which understanding arises not from accepting answers, but from actively engaging with one's own commitments.

This structure makes Socratic maieutics particularly relevant for contemporary educational settings. Its aim is to cultivate judgment instead of merely reproducing content. Insight achieved through questioning remains epistemically owned by the subject who arrives at it; autonomy is preserved precisely because the learner remains the author of her positions. In contrast to transmissive models of instruction, philosophy in this context appears as an essentially dialogical practice. For this reason, Socratic maieutics provides more than just a historical model, namely, a normative standard for engaging with new epistemic technologies.

This conception of epistemic autonomy finds a striking historical analogue in Socratic maieutics as articulated in Plato's *Theaetetus*. Socrates

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

famously characterizes himself as barren of wisdom (Plato 2015, 11; 150c): he possesses no doctrines to teach, no settled body of knowledge to impart, and no epistemic authority grounded in possession of *sophia*. His philosophical practice consists instead in questioning, testing, and destabilizing apparent certainties in order to open a space in which others can bring their own understanding to birth. Precisely because he claims no wisdom of his own, Socrates can function as a catalyst of autonomous thinking rather than as a source of epistemic dependence.

In this exact and strictly limited sense, generative AI can be described as *barren of wisdom* as well. Artificial systems possess no consciousness, no beliefs, no commitments, and no genuine understanding. They do not know, judge, or endorse anything. What they produce is not knowledge in the strong epistemic sense, but statistically generated linguistic continuations – echoes of human knowledge sedimented in data and recombined without awareness or intent. Any appearance of insight, coherence, or authority is derivative and parasitic upon prior human practices of meaning-making.

This lack has normative implications. Because AI is barren of wisdom, it cannot be an epistemic authority in its own right. When treated as such, it invites epistemic heteronomy and cognitive offloading. When treated instead as a maieutic medium, its very emptiness becomes productive. The system's absence of beliefs and commitments opens a dialogical space in which users must supply the orientation, the questions, and the criteria of adequacy themselves.

In this respect, the use of generative AI to expand one's own epistemic autonomy mirrors the Socratic stance. Just as Socrates mobilizes his own claimed epistemic emptiness to make room for the thinking of others, the mature use of AI consists in exploiting the system's lack of wisdom to provoke, structure, and refine one's own inquiry. Epistemic autonomy is not suspended but intensified: judgment remains human, while the space of questioning is technologically extended.

It is on this basis that the notion of *algorithmic maieutics* becomes intelligible. Algorithmic maieutics designates a mode of human–AI interaction in which artificial systems are not used to deliver answers, but to sustain inquiry. Much like in the Socratic model, the epistemic value of the interaction lies less in the outputs produced and more in the questions that guide them – and in the reflective process through which understanding is brought forth rather than delegated.

Algorithmic maieutics can be understood as a *meta-maieutic* practice. While Socratic maieutics enables epistemic autonomy through dialogically structured questioning, algorithmic maieutics externalizes this very

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

practice by deliberately using AI as a maieutic medium. The aim of questioning is self-thinking instead of cognitive relief: inquiry does not outsource judgment, but exercises it.

Epistemic immaturity arises when AI is tacitly misinterpreted as wise – when its fluent outputs are treated as expressions of understanding and hallucinations are mistaken for insight. By contrast, epistemic maturity consists in recognizing AI as an "unfruitful midwife": a system that possesses no wisdom of its own, yet becomes epistemically valuable only insofar as the human subject enters into a critical dialogue with it. In such interaction, the user does not reflect only on the system's responses, but also on her own questions, assumptions, and standards of judgment. Algorithmic maieutics thus does not replace Socratic practice but reflexively re-enacts it under technologically extended conditions.

Thus, algorithmic maieutics designates a mode of human–machine interaction in which chatbots are not treated as sources of knowledge or epistemic authorities, but as *dialogical resonance media*. Generative models do not produce truths; they generate proposals, perspectives, argumentative patterns, and variations. Their epistemic value does not lie in the output as such, but in how that output responds to well-formed questions. What is distinctive, therefore, is not the system alone, but the *interactive process* between human and machine.

This shifts the focus of philosophical and educational practice away from results and toward the process of questioning itself. Algorithmic maieutics translates the Socratic insight – that knowledge emerges dialogically – into the context of interaction with non-intentional systems. Although chatbots are not genuine interlocutors, they respond *as if* they were, thereby opening a quasi-dialogical space in which users can test assumptions, explore alternatives, and reflect on their own reasoning. In this space, epistemic autonomy is not suspended but rearticulated: judgment remains human, while its scope is technologically extended.

From this perspective, the quality of AI-supported inquiry depends less on the sophistication of the model than on the epistemic practice of the user. Chatbots function as mirrors that make thought processes visible, background assumptions explicit, and structure argumentative possibilities. Algorithmic maieutics is therefore a normative practice – one that must be actively cultivated and repeatedly enacted. It describes how epistemic autonomy can be maintained under conditions of algorithmic mediation, rather than passively eroded through cognitive offloading.

The development of ideas in interaction with generative AI provides a particularly illuminating test case for AI maturity. Idea generation is

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

often described as a matter of producing novel or creative content. Yet this description obscures what is normatively decisive: ideas are not merely outputs, but moments within a problem-oriented process of thinking. To count as epistemically owned, an idea must emerge from a question one can recognize as one's own, and it must be oriented toward a problem whose stakes one understands and endorses. For this reason, the spontaneity of ideation – the capacity to initiate inquiry by posing a question or articulating a problem – remains central to epistemic autonomy.

In the context of generative AI, this spontaneity is easily mislocated. Because large language models can rapidly produce plausible formulations, associations, and candidate ideas, there is a temptation to treat ideation as something that can be delegated to the system. However, this temptation rests on a category error. Generative AI does not originate ideas in the strong epistemic sense; rather, it responds to inputs. Its outputs are entirely dependent on how a space of inquiry is linguistically opened, constrained, and oriented by the user. The initiating moment of ideation – the decision of *what is at issue* and *why it matters* – cannot be outsourced without relinquishing epistemic authorship.

This dependence of output on input grounds a corresponding asymmetry of responsibility. Since generative AI systems lack intentions, commitments, and stakes of their own, responsibility for the direction and quality of ideation remains with the human agent who frames the interaction. The user is the one who determines the conditions under which ideation takes place instead of merely consuming those ideas. In this sense, AI-supported idea development is a transformation of the medium through which problems are explored.

This normative structure becomes particularly clear when ideation is analyzed through the lens of speech act theory. In his account of illocutionary acts, Searle (1969) has emphasized that meaningful communication is governed by constitutive rules: background assumptions, intentions, and normative commitments that must be satisfied for a speech act to succeed. Traditionally, these conditions are analyzed in the context of communication between two intentional agents. In human–AI interaction, however, the structure changes. The communicative situation is no longer one of mutual intention recognition, but of asymmetrical normativity: only one party – the human – can take responsibility for the illocutionary force of the exchange.

This shift has direct implications for idea generation. When a user addresses generative AI in the mode of a command ("Generate ideas for … "), the illocutionary structure implicitly treats the system as an

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

epistemic agent capable of initiating inquiry. By contrast, when the interaction is framed interrogatively, the speech act preserves human spontaneity. Instead of merely requesting outputs, questions articulate a lack, a problem, or a tension that calls for exploration. In posing a question, the human speaker commits herself to the space of reasons that the answer must inhabit. The machine, by contrast, remains a responsive mechanism whose role is to elaborate, vary, or challenge possibilities within a space that the human has opened.

From the perspective of AI maturity, this distinction is important. Mature ideation with AI does not consist in issuing ever more efficient commands, but in sustaining a questioning stance that keeps epistemic initiative with the human subject. The felicity conditions of AI-mediated ideation are thus not satisfied by the correctness or originality of generated text alone, but by the extent to which the interaction preserves human authorship of the problem. Idea development, on this account, can be defined as the reciprocal unfolding of a problem through human–machine interaction – an interaction in which the human remains ultimately responsible for the questions asked, the directions pursued, and the standards by which candidate ideas are judged.

Viewed this way, AI maturity requires a different understanding of creativity and ideation under conditions of generative AI. The machine may accelerate variation and expand the space of possible formulations, but spontaneity – the capacity to initiate inquiry and to own its guiding questions – must remain human. Where this spontaneity is preserved, generative AI can function as a maieutic medium that supports reflective ideation. Where it is surrendered, idea generation collapses into epistemic heteronomy: ideas appear, but no one can fully answer for why they are worth having.

While generative AI can simulate ideation by producing sequences of plausible variations, the formation of ideas in the epistemically relevant sense cannot be reduced to a rule-governed procedure with guaranteed outcomes. Genuine ideation is not a linear process of optimization, but an exploratory activity that unfolds through tentative questioning, revision, and the frequent experience of failure. Ideas often emerge when a procedure fails rather than succeeds – for instance, when an initial framing proves inadequate, a hypothesis collapses, or a line of reasoning reveals its own limits. For this reason, there can be no algorithm for idea development. Algorithms presuppose determinate success conditions and stable criteria of completion, whereas ideation is constitutively open-ended. It involves moments of uncertainty in which neither the direction of inquiry

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

nor the standards of relevance are fully fixed in advance. These moments are not epistemic defects to be eliminated, but productive sites of insight. Failure, hesitation, and reorientation are integral to the process by which problems are refined and questions become sharper. From the perspective of AI maturity, this nonalgorithmicity has normative implications. If ideation were fully algorithmizable, epistemic responsibility could in principle be delegated. But because idea development depends on ongoing judgment about what counts as a promising question, a meaningful problem, or a worthwhile direction, responsibility must remain with the human agent. Generative AI can support this process by accelerating variation and making alternatives visible, yet it cannot replace the fragile, revisable, and failure-prone activity through which ideas are actually formed. To acknowledge this fragility is a condition of AI maturity rather than a recognition of a limitation of human thinking.

The concept of affordance (Gibson 1966) provides a further theoretical lens for understanding why generative AI poses a distinctive challenge to epistemic maturity. In her influential account, *How Artifacts Afford*, Davis (2020) argues that affordances are neither fixed properties of artifacts nor mere projections of users, but relational structures that emerge in situated interactions between humans and technologies. Artifacts do not determine action, yet they actively shape what becomes salient, easy, difficult, or unlikely. They *request, encourage, discourage, allow, or refuse* certain forms of engagement, thereby exerting normative pressure on practice without abolishing agency.

Generative AI can be understood as a special and intensified case of affordance. Unlike traditional artifacts – tools, interfaces, or even earlier digital systems –, generative AI does not merely afford discrete actions, such as clicking, searching, or retrieving information. Rather, it affords ongoing dialogical interaction, characterized by rapid feedback loops in which each human input immediately reshapes the system's subsequent responses. This recursive structure represents a qualitative shift. The affordances of generative AI unfold dynamically across interaction sequences, in which meaning, direction, and relevance are continuously co-produced.

In Davis' terms, affordances are best analyzed not by asking *what* a technology affords, but *how* it affords under specific conditions. Applied to generative AI, this makes it clear that such systems do not simply afford text generation or information access. They afford a particular mode of thinking-with, one that privileges fluency, immediacy, and continuity of sense-making. By responding instantly and adaptively to user input,

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

generative AI encourages a form of cognitive flow in which questioning, answering, and reformulation blend into a seamless process. This can be epistemically productive, but it also increases the risk that the user's own judgment becomes absorbed into the system's rhythm.

What makes generative AI especially demanding from the perspective of maturity is that its affordances are co-evolutionary. The system adapts to the user's prompts, preferences, and patterns of interaction, while the user simultaneously adapts to the system's style, speed, and suggestions. Over time, this mutual adjustment can subtly shift how problems are framed, which questions are perceived as natural, and what counts as a satisfactory answer. In this sense, generative AI comes closer to the user than conventional artifacts ever could: it does not merely mediate action, but participates in the formation of epistemic habits themselves.

This proximity heightens the normative stakes of affordance. Whereas traditional artifacts afford relatively stable forms of use that can be learned and critically appropriated, generative AI affords evolving interaction patterns that may gradually prestructure thinking. Within this context, the danger lies in epistemic drift, not in technological determinism: a situation in which users retain formal control over interaction while tacitly aligning their cognitive practices with the system's affordances. From the perspective of AI maturity, this makes reflective distance indispensable. To remain mature, users must not only understand *that* generative AI affords certain modes of interaction, but must actively regulate *how* these affordances shape their own questioning, evaluation, and judgment.

Seen through Davis' affordance framework, AI maturity thus requires an explicit awareness of feedback loops and co-evolution. Mature engagement with generative AI consists in interrupting, reorienting, and sometimes resisting the system's affordances – by slowing down interaction, by reasserting human criteria of relevance and adequacy, and by recalling that no affordance, however persuasive, relieves the subject of epistemic responsibility. Generative AI can support epistemic autonomy only when its affordances are critically governed rather than passively inhabited. In this sense, maturity is not about getting rid of affordances, but moving toward a reflective appropriation of technologies that we are interacting with on a more intimate level than with any artifact before.

AI maturity also involves the capacity to resist the seductive affordances of flow that generative AI systems so readily produce. One of the most distinctive features of contemporary generative AI is its ability to sustain a fluent, uninterrupted stream of responses that adapt instantaneously to

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

user input. This interactional smoothness affords a state of cognitive flow in which questioning, answering, and reformulation merge into a continuous process. While such flow can feel productive and engaging, it also carries a normative risk: the rhythm of thinking is set by the system's pace of generation rather than by the subject.

From the perspective of affordance theory, this flow is an intentionally encouraged mode of engagement. Generative AI affords immediacy, continuity, and low-friction progression, thereby discouraging interruption, hesitation, and reflective distance. The system invites the user to remain within the interaction, to follow one suggestion with the next, and to allow the dialogue to unfold without pause. In this sense, flow becomes an epistemic temptation: it promises efficiency and coherence while quietly displacing the moments of resistance in which judgment is ordinarily exercised.

Maturity, under these conditions, should not be understood as rejecting flow altogether; rather, it consists in governing one's relation to it. To be mature in one's use of generative AI is to recognize that flow, while phenomenologically compelling, is not itself a criterion of epistemic quality. A seamless progression of ideas may indicate linguistic fluency rather than conceptual adequacy; continuity of interaction may mask the absence of justification. The mature subject therefore learns to interrupt flow deliberately – to step out of the dialogical current in order to regain orientation, reassess goals, and reassert evaluative standards.

This practice can be described as rhythmical self-governance. Rather than allowing the system's generative tempo to dictate the pace of inquiry, the mature user actively sets the rhythm of interaction. This may involve slowing down the exchange, introducing temporal gaps between prompts, suspending further generation, or shifting from dialogical engagement to independent reflection. In doing so, the user preserves the temporal conditions under which epistemic autonomy remains possible: time to doubt, revise, and to take responsibility for one's judgments.

Accordingly, AI maturity can be understood as the capacity to establish and maintain an autonomous epistemic rhythm in the face of systems designed to accelerate cognition. Flow becomes productive only when it is subordinated to human timing rather than the reverse. Where users remain capable of entering and exiting AI-mediated flow at will, generative systems can support inquiry. Where flow becomes self-sustaining and unreflective, however, epistemic autonomy gives way to heteronomy. Maturity thus denotes the disciplined ability to engage generative AI without surrendering control over the rhythm of thought itself.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

Central to algorithmic maieutics is epistemically virtuous prompting. Prompting is not a purely technical act, but a linguistically mediated action, a speech act with normative implications. Every prompt sets expectations, frames relevance, and shapes the space of possible responses – while simultaneously feeding back into the user's own thinking. Command-style prompts ("Write me … ") tend to externalize judgment and encourage passivity, whereas interrogative, exploratory, and iterative prompts stimulate reflection and critical engagement.

The virtues of prompting include precision, contextualization, role-setting, iterativity, and corroboration (cf. Tankelevitch et al. 2024; Tate et al. 2025). What this literature describes as demands imposed by the technology, however, is better understood as a set of virtues: dispositions that are not merely effective but epistemically right. Particularly important are meta-prompts that ask the system to expose assumptions, articulate counterarguments, or generate questions rather than answers. Such prompts create a loop structure that closely resembles Socratic elenchus: responses are not accepted as endpoints, but treated as occasions for further questioning. Prompting thus becomes a dianoetic virtue in the Aristotelian sense – an exercisable intellectual capacity that enables good epistemic action under specific technological conditions.

In educational contexts, algorithmic maieutics gives rise to concrete didactic applications. Chatbots can be used in dialectical role-play, where students confront opposing positions and learn to structure argumentative spaces through targeted questioning. They can also function as maieutic aids in writing processes, provided that students are restricted to asking questions rather than adopting generated text. In such cases, the documented prompt process itself becomes an object of reflection. Particularly promising is meta-prompting as an assessment format, in which students are evaluated not on produced answers, but on their ability to explore complex philosophical problems through disciplined questioning.

Across all these applications, transparency is essential. AI use must be disclosed, reflected upon, and reconstructable. Only under these conditions can we avoid that AI function as a substitute for philosophical practice, and instead ensure that AI becomes its didactic medium. The goal of AI maturity is to attain the cultivated ability to use generative AI in ways that extend, rather than undermine, epistemic autonomy.

The preceding section has argued that prompting can be understood as an epistemic practice rather than a technical skill: a form of algorithmic maieutics that exercises judgment instead of outsourcing it. When prompting is practiced reflectively, it can cultivate epistemic virtues such as

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

discernment, reflexivity, and responsibility, thereby contributing directly to AI maturity. However, such practices do not emerge spontaneously. They depend on institutional contexts, pedagogical design, and assessment structures that either support or undermine epistemic autonomy. To understand how AI maturity can be fostered beyond individual interactions, the analysis must therefore shift from micro-level practices to their educational embedding. The following section addresses this challenge by examining the curricular, didactic, and institutional implications of AI maturity for educational practice.

5 Educational Implications: From Curriculum to Practice

To translate the normative framework of AI maturity into educational reality, this section moves deliberately away from abstract pedagogical principles and toward concrete curricular and assessment-oriented guidance. The aim is not merely to illustrate how generative AI might be used in teaching, but to specify how curricula, learning objectives, and examination formats can be systematically designed to protect and cultivate epistemic autonomy under conditions of algorithmic mediation. Accordingly, the following sections articulate actionable recommendations for curriculum design, teaching formats, and assessment practices that operationalize the distinction between epistemic extension and epistemic substitution. These recommendations are intended to function as transferable design principles: they specify how AI can be integrated into syllabi, assignments, and examinations in ways that actively exercise judgment, avoid cognitive offloading, and make epistemic responsibility pedagogically visible. Instead of merely applying the preceding philosophical analysis, the following section renders it institutionally and didactically actionable for educators seeking to align AI use with the educational ideal of maturity rather than mere technical proficiency.

An instructive empirical articulation of the concerns developed in this Element is provided by the design-based implementation research of Tate et al. (2025), who examine how generative AI can be incorporated into a writing-intensive undergraduate course without collapsing into cognitive offloading. Their study is of special relevance for the present argument because it operationalizes – at the level of curriculum design and assessment – precisely the distinction between epistemic extension and epistemic substitution that underlies the concept of AI maturity.

Instead of treating generative AI as a shortcut to textual production, the instructional framework developed in this study deliberately structures AI

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

use around phases of "thinking first," iterative and agentic prompting, systematic corroboration, and reflective incorporation. AI is thus positioned as a dialogical scaffold that requires students to remain the authors of their own judgments. Importantly, this design choice is not merely pedagogical but normative: it makes epistemic responsibility visible and assessable by embedding AI interaction into staged assignments, constrained prompt environments, and reflective evaluation tasks.

From the perspective of AI maturity, the significance of this approach lies in its explicit resistance to structural cognitive offloading. The students are not rewarded for fluent AI-generated outputs, but for their capacity to interrogate, revise, and justify them. Assessment formats are aligned accordingly: learning outcomes are tied to demonstrable judgment and reflective accounts of AI use rather than to the apparent quality of final products alone. A practical implementation strategy would consist of requiring students to submit prompt histories, reflective commentaries, or oral defenses alongside written assignments. Such formats make epistemic processes visible and help educators evaluate whether AI has functioned as a support for understanding or as a substitute for judgment.

In this sense, the study exemplifies how the normative ideal of epistemic autonomy can be translated into concrete curricular architecture and examination practices. It shows that avoiding AI-induced immaturity is not a matter of prohibition, but of deliberately designing curricula and assessments that force judgment to remain with the learner.

A paradigmatic use case of AI maturity in higher education concerns the deliberate selection and combination of AI tools according to their epistemic affordances. Because tools evolve rapidly and receive new features, no fixed "tool list" can ground responsible practice. What can be taught, however, is a maturity-oriented principle of selection: a student should begin by asking which epistemic task is at hand – text search, evidence control, synthesis, conceptual exploration, drafting, counterargument generation – and then choose the tool whose affordances fit that task without displacing judgment. In this sense, tool selection is the first act of epistemic autonomy: it is the point at which the agent determines how inquiry will be mediated, where epistemic authority will reside, and which risks must be actively managed.

This is especially important due to how generative systems tempt users to shortcut academic work by treating fluency as competence. Given the impressive linguistic performance of current chatbots, it is tempting to prompt them to produce academic prose directly (e.g., "Write a 4,000-word scientific paper on the problems of Kant's categorical imperative").

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

Even where universities prohibit such outsourcing, students and instructors often lack clarity about what forms of AI use remain permissible and educationally valuable. The maturity criterion developed in this Element provides a decisive orientation: AI is legitimate in academic contexts only insofar as it enhances the user's own scientific capacities – understanding, judgment, and insight – rather than replacing or merely simulating them. No AI output should be incorporated into a paper without prior reflection, since probabilistic language generation can produce text that is coherent at first glance while being epistemically hollow.

A maturity-oriented workflow therefore begins with grounded tools whose primary affordance is traceability to a specified database. In practice, this means prioritizing systems that minimize hallucination by making sources transparent and inspectable. A concrete example is Gemini Notebook (formerly NotebookLM), which is designed around user-provided corpora and returns answers that are tied to quoted passages. This tool is well-suited for research and preprocessing tasks that are time-consuming but methodologically routine: locating patterns, mapping term usage, extracting relevant contexts, and producing text-based overviews with explicit links back to the underlying documents. For instance, a student working with Kant's *Critique of Pure Reason* can ask: "Name the contexts in which the term 'freedom' occurs in this corpus." The epistemic gain here is not that the tool "understands Kant," but that it accelerates a controlled research operation while keeping verification in view. The user remains responsible for curating the corpus, evaluating whether quotations actually support the summary claim, and interpreting philosophical significance. No system can take this over, because fluency is not a marker of grounding.

This first step also directly addresses the structural problem of hallucination. Recent theoretical work argues that hallucinations are not an incidental bug but a predictable outcome of how language models are trained and evaluated. Kalai et al. (2025) compare language models to exam takers: when uncertain, systems are often rewarded (by benchmark-style grading regimes) for guessing rather than for acknowledging uncertainty, which sustains overconfident falsehoods even in advanced models.

From this perspective, hallucination arises from statistical pressures in training and persists because common evaluations penalize abstention ("I don't know") and thereby incentivize plausible guessing. In educational settings, this diagnosis has a clear implication: maturity requires building an epistemic workflow that does not treat ungrounded generative outputs as evidence, and that structurally privileges tools and practices that keep claims tethered to sources whenever factual reliability matters.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

Once an empirically anchored basis is established, a second step can legitimately employ more strongly generative chatbots (e.g., Claude, Gemini or ChatGPT) for tasks where their main affordance is conceptual variation rather than factual authority: idea generation, analogy, perspective shifting, counterargument rehearsal, and problematization. Here, the mature use of generative AI is dialogical and interrogative: prompts should invite exploration ("What objections have been raised against this interpretation?"; "Which alternative framings could sharpen the research question?") rather than issuing authorial commands that encourage substitution ("Write the section for me"). The outputs in this stage function as provisional proposals – starting points for the user's own reasoning – whose claims must be re-checked against primary texts and scholarly literature when they appear factual or citation-like.

This staged combination of tools illustrates how AI can extend, rather than replace, scientific thinking. Empirically grounded tools secure transparency and traceability; generative tools expand the space of ideas and perspectives. At every stage, however, epistemic responsibility remains with the human user. The goals of inquiry, the formulation of questions, the evaluation of relevance, and the final judgments cannot be delegated. AI maturity thus consists in governing technological assistance in a way that preserves authorship, accountability, and the spontaneity of thinking.

If AI maturity is to be more than an individual disposition, education must open and sustain spaces in which autonomous judgment can be publicly exercised. Epistemic autonomy does not emerge in isolation; it depends on institutional and infrastructural conditions that make reflection, contestation, and perspective-taking possible. In AI-mediated learning environments, this requirement becomes especially pressing, as personalized systems tend to narrow epistemic horizons by default – reinforcing filter bubbles, optimizing for relevance, and stabilizing pre-selected viewpoints.

For this reason, the cultivation of AI maturity requires a deliberate pedagogical design of learning spaces. Such spaces must counteract algorithmic narrowing by enabling exposure to plural perspectives, shared standards of justification, and dialogical engagement. They must support what Kant calls the public use of reason: forms of inquiry in which claims can be articulated, criticized, and revised in relation to others rather than merely optimized for individual performance. Without such infrastructures, even technically competent AI use risks remaining confined to a privatized mode of cognition, where epistemic authority is silently shaped by personalization and optimization logics.

Recent empirical work by Tate et al. (2025) provides a compelling illustration of how such socio-technical environments can be designed to

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

support epistemic autonomy rather than cognitive offloading. Drawing on design-based implementation research in a writing-intensive undergraduate course, the authors suggest that generative AI can be pedagogically integrated in ways that preserve authorial agency, responsibility, and reflective judgment. Central to their approach is the deliberate structuring of human–AI interaction: students are required to "think first" before engaging AI, to treat prompting as an iterative and interrogative process, and to reflect explicitly on how AI shaped their writing decisions.

From the perspective of AI maturity, the significance of this study lies not merely in its instructional effectiveness, but in its implicit normative architecture. Their instructional framework identifies five requirements for students – to understand, access, prompt, corroborate, and incorporate generative AI – and these operationalize key dimensions of epistemic responsibility: students remain the originators of ideas, they define the goals of the writing task, they retain control over successive stages of interaction, and they are held accountable for incorporating – or rejecting – AI-generated content. In Kantian terms, these design principles prevent generative AI from functioning as a cognitive guardian that "understands for me," and instead situate it as an aid within a learner-governed process of judgment.

At the same time, the study also reveals the limits of competence-based AI literacy frameworks. Even when students understand how AI works and use it transparently, epistemic immaturity can still arise if interaction is not normatively structured. What the authors discuss as potentially productive "friction" is – from the standpoint of AI maturity – the institutionalization of reflective distance: a pedagogical safeguard that keeps judgment from being silently displaced by fluency and convenience.

Read against the backdrop of the present framework, Tate et al.'s findings can thus be interpreted as providing empirical support for a central claim of this Element: AI maturity is not achieved by restricting AI use, but by designing learning environments that actively sustain epistemic authorship. Their work illustrates how educational infrastructures can translate the abstract ideal of epistemic autonomy into concrete interactional practices – thereby showing how AI maturity can be cultivated, rather than merely demanded, in everyday educational settings.

If AI maturity is to be more than an individual virtue or pedagogical aspiration, it must be institutionally stabilized. Epistemic autonomy under conditions of pervasive AI mediation cannot be secured solely at the level of individual users; it depends on how AI systems themselves are trained, aligned, and embedded within socio-technical infrastructures. From the perspective developed in this Element, AI maturity therefore

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

raises an institutional design question: under what conditions can AI systems function as extensions of autonomous judgment rather than as structural sources of epistemic heteronomy?

A central obstacle to such institutionalization lies in the predominantly commercial orientation of currently dominant generative AI systems. Tools such as large language model–based chatbots are typically optimized for user engagement, efficiency, and retention, operating within economic frameworks that reward plausibility, personalization, and fluency rather than epistemic contestability. As a result, these systems tend to reinforce confirmation dynamics and relevance filtering that align with users' prior interests, expectations, or behavioral profiles. While such personalization increases usability, it also introduces a structural risk of epistemic enclosure: users are less frequently confronted with dissenting perspectives, justificatory friction, or epistemic resistance. In this sense, commercially optimized AI systems are not neutral epistemic aids but actively shape the space of reasons in which judgment takes place.

From the standpoint of AI maturity, this dynamic is normatively problematic. Epistemic autonomy requires not only access to information, but exposure to reasons that can be challenged, revised, or rejected. When AI systems systematically pre-select what appears relevant or plausible according to engagement-driven criteria, they risk functioning as epistemic filters rather than as supports for judgment. The widely discussed phenomenon of filter bubbles (Pariser 2011) thus gains a new dimension in the context of generative AI: it no longer merely structures information intake, but increasingly mediates processes of reasoning, interpretation, and evaluation themselves.

The framework of hybrid autonomy helps to clarify what is at stake here. Autonomy in extended systems depends on the teleological and normatively asymmetrical governance of extension relations (Noller 2024). AI systems may extend human epistemic capacities only insofar as the purposes, standards of justification, and criteria of relevance governing their operation remain humanly contestable and institutionally revisable. Where these standards are tacitly fixed by commercial optimization objectives, extension risks collapsing into technologically induced heteronomy. Sustaining epistemic autonomy therefore requires institutional mechanisms that render AI-mediated epistemic processes transparent, interruptible, and open to critique.

Institutionalizing AI maturity thus requires a shift in how AI systems are trained and aligned. Alignment must not be restricted to safety, compliance,

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

or preference satisfaction, but must include epistemic alignment: the deliberate design of systems that preserve spaces for disagreement, uncertainty, and reflective judgment. This may involve training regimes that foreground pluralistic data sources, interfaces that encourage justificatory exploration rather than immediate answers, and governance structures that insulate educational and public-interest AI systems from purely commercial imperatives. Only under such conditions can AI systems provide the structural preconditions for AI maturity – conditions under which users are supported in remaining the authors of their own judgments rather than the passive recipients of algorithmically curated plausibility.

Educational institutions therefore bear responsibility not only for teaching AI-related skills, but for designing socio-technical environments that keep epistemic agency visible and contestable. In practice, however, educators face significant implementation challenges, including limited institutional guidance, unequal access to AI tools, lack of teacher training, and uncertainty about how to assess AI-supported work fairly. These challenges risk producing inconsistent classroom practices in which AI use is either uncritically normalized or broadly prohibited without pedagogical differentiation.

Learning spaces must be structured to slow down automated closure, introduce epistemic friction, and make AI mediation itself an object of reflection. Only where such infrastructural conditions are in place can AI function as a medium of shared inquiry rather than as a driver of epistemic isolation. In this sense, AI maturity depends on whether education succeeds in opening spaces where judgment remains public, plural, and revisable.

This section has translated the normative ideal of AI maturity into concrete educational strategies, showing how curricula, assessment, and classroom practices can be designed to counter epistemic dependency rather than reinforce it. What emerges is a picture of education not as a site of technological adaptation, but as a normative space in which the conditions of autonomous judgment must be actively protected. Yet the significance of AI maturity extends beyond pedagogical design. If generative AI reshapes the epistemic environments in which citizens reason, decide, and justify their beliefs, then AI maturity cannot remain a merely educational concern. It raises a broader societal question about the preservation of epistemic autonomy under conditions of pervasive algorithmic mediation. The concluding section therefore situates AI maturity within a wider normative horizon, arguing that it should be understood as a collective and societal goal rather than a narrowly educational one.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

6 Conclusion: Towards AI Maturity as a Societal Goal

Recent work by Shannon Vallor (2024) has powerfully articulated the idea that contemporary AI systems function as mirrors of our collective past practices, values, and judgments. Because generative AI is trained on socially produced data and increasingly mediates how meaning, relevance, and plausibility are negotiated, interaction with AI is never a purely private or technical matter. It is, in Vallor's sense, a social practice that both reflects and reshapes who we are becoming.

This diagnosis reinforces the central claim of this Element: responsibility in human–AI interaction cannot be outsourced to systems trained on aggregated social behavior. To interact maturely with generative AI is to resist the temptation to let such systems silently determine the origins, goals, and standards of one's thinking. AI maturity therefore requires users to remain the authors of their inquiries, to define the purposes of interaction, to retain control over its successive steps, and to be able to give reasons for how judgments were formed.

A closely related account of human–AI collaboration can be found in Luciano Floridi's (2025) concept of "distant writing." Floridi characterizes LLM-assisted writing as a practice in which humans act as designers rather than executors of text, specifying requirements, constraints, and goals while delegating textual production to artificial systems. Crucially, this shift does not diminish human responsibility; on the contrary, it relocates responsibility from execution to judgment. The meta-author remains accountable for initiating the process, determining its purposes, supervising its stages, and endorsing its outcomes.

Read in the light of the present framework, distant writing exemplifies a central claim of AI maturity: responsibility in human–AI interaction is not exhausted by technical control or transparency, but consists in the user remaining the normative source of epistemic authority. To write with AI maturely is to originate the guiding ideas oneself, to define the ends of interaction, to retain control over its iterative development, and to be able to give reasons for how a final text came about. Floridi's emphasis on design over execution is thus consistent with a Kantian understanding of maturity as self-legislation under conditions of mediated agency.

At the same time, distant writing makes explicit the co-evolutionary dimension of AI maturity. Because LLMs are trained on socially produced texts and continuously reshaped through human feedback, every act of "wrAIting" (Floridi 2025, 2) participates in a reciprocal process: human practices shape AI systems, while AI systems reshape human

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

habits of expression, evaluation, and judgment. AI maturity, on this view, does not resist co-evolution but governs it. It describes the capacity to engage in such hybrid practices without surrendering authorship over meaning, purpose, and justification – thereby transforming generative AI from a substitute for thinking into a medium of extended – yet accountable – judgment.

Understood this way, AI maturity necessarily implies a co-evolutionary perspective (Pedreschi et al. 2025). Human practices shape AI systems through data, feedback, and normalization, while AI systems reshape human epistemic habits in return. The question is not whether this co-evolution occurs, but whether it is governed reflectively. AI maturity names the capacity to participate in this co-evolution without relinquishing epistemic self-determination – transforming AI from a mirror that immobilizes judgment into a medium through which autonomous judgment can be exercised and renewed.

Because generative AI systems are trained on socially produced data and operate within linguistically structured practices, our engagement with them cannot be adequately understood as a merely individual or technical activity. In its most basic form, the use of generative AI is a social practice: it draws on collectively sedimented patterns of meaning, reproduces and transforms shared standards of relevance, and feeds back into the epistemic environments in which others operate.

Baehr's (2011) account of intellectual virtues provides a valuable lens through which the present conception of AI maturity can be further clarified. In *The Inquiring Mind*, Baehr emphasizes that epistemic excellence is not exhausted by reliable performance or successful outcomes, but is grounded in stable traits of intellectual character such as open-mindedness, intellectual humility, conscientiousness, and intellectual courage. Read in light of these theses, AI maturity emerges as a virtue-theoretic achievement rather than a merely instrumental skill. A user may competently interact with generative AI and still lack maturity if this interaction bypasses the exercise of epistemic virtues that sustain ownership over one's judgments. From this perspective, the main risk of generative AI is not being deceived by the system, but the erosion of intellectual character on the part of the user. Where AI substitutes rather than occasions inquiry, users fail to enact the virtues that constitute responsible epistemic agency. Therefore, AI maturity consists in sustaining intellectual virtue under conditions of algorithmic assistance: the capacity to use AI as a support for inquiry without relinquishing the virtues that make inquiry one's own.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

Turri et al.'s (2021) account of virtue epistemology highlights two defining commitments: the normative character of epistemic evaluation and the primacy of intellectual agents as bearers of epistemic value. Read in light of AI maturity, this shifts attention from the reliability of AI outputs to whether AI-mediated epistemic performances remain creditable to the human agent's intellectual virtues. AI maturity thus marks the threshold at which extended epistemic assistance supports, rather than substitutes for, virtue-governed judgment.

Although addressed to researchers rather than learners, Jasper Roe's (2025) methodological account of generative AI in educational research converges with the present framework in its rejection of uncritical, tool-centered adoption of AI. Roe repeatedly emphasizes that the ethical and epistemic legitimacy of generative AI depends on reflective use, explicit justification, and sustained human oversight throughout the research process. Read in light of the concept of AI maturity, Roe's contribution can be interpreted as an implicitly normative position: generative AI supports educational research only insofar as it remains subordinated to human judgment rather than functioning as an epistemic surrogate. What the present account adds is a philosophical articulation of why these conditions matter – both for research integrity and for preserving epistemic autonomy under conditions of algorithmic mediation. AI maturity, therefore, cannot be reduced to correct usage or compliance with external rules. It concerns how responsibility is exercised within human–AI interaction, understood as a socially embedded practice of judgment.

To take responsibility in this context does not mean to assume causal responsibility for every output a system produces. Rather, it means to remain the normative source of the epistemic process itself. This involves, first, originating one's own thoughts and ideas. Even when AI is used to explore possibilities, generate alternatives, or articulate arguments, the initiating impulse of inquiry must remain human. A mature relation to AI presupposes that questions, problems, and perspectives are not simply adopted from the system but are reflectively appropriated as one's own.

Second, responsibility requires the capacity to define and revise the goals of interaction. Generative AI can suggest means, continuations, and framings, but it cannot legitimately determine what the inquiry is for. To use AI maturely is to set epistemic aims oneself – whether understanding, critique, exploration, or explanation – and to evaluate AI contributions relative to these self-legislated purposes. Where goals are silently imported from system defaults or optimization logics, epistemic agency gives way to heteronomy.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

Third, responsibility entails maintaining control over the stages of human–AI interaction. In this context, control does not imply micromanagement of technical processes, but reflective oversight of how prompts are formulated, how outputs are selected or rejected, and how intermediate results shape subsequent steps. Mature use requires the ability to interrupt, redirect, or abandon an interaction when it ceases to serve one's epistemic aims. What matters is not uninterrupted fluency, but the preservation of judgment across the interactional sequence.

Finally, responsibility implies answerability. A mature user must be able to give reasons for the epistemic steps taken: why a particular prompt was chosen, why a certain output was accepted or rejected, and how conclusions were reached. This capacity for justificatory accountability marks the decisive difference between extended judgment and epistemic delegation. Where users can no longer explain how their claims were formed, epistemic responsibility has already been displaced.

Understood this way, AI maturity is inseparable from a co-evolutionary perspective. Human practices shape AI systems through data generation, feedback, and normalization, while AI systems reshape human epistemic habits by influencing how questions are asked, how plausibility is assessed, and how reasoning unfolds. Maturity does not halt this process; it governs it. To act maturely with AI is to recognize that every interaction contributes – however incrementally – to the ongoing co-evolution of human judgment and artificial capacities.

AI maturity therefore denotes a dual responsibility: responsibility for one's own epistemic conduct in the present, and responsibility for the trajectories of practice that such conduct reinforces over time. In this sense, maturity is neither a static achievement nor a purely individual virtue. It is a socially situated, temporally extended form of self-governance within hybrid epistemic systems. To cultivate AI maturity is to learn how to participate in these systems without relinquishing authorship over thought, purpose, and justification – even as cognition itself becomes increasingly distributed and technologically mediated.

The argument of this Element can be summarized in a simple yet challenging claim: as generative AI permeates ever more domains of social, educational, and professional life, AI maturity becomes an increasingly decisive condition for preserving epistemic autonomy. The more seamlessly AI systems integrate into everyday practices of reasoning, writing, and decision-making, the stronger the temptation grows to automate not only execution, but judgment itself. What presents itself as efficiency or convenience thus carries a genuine normative risk: the gradual erosion of our capacity – and willingness – to think, assess, and justify for ourselves.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

Design choices can mitigate this risk. AI systems can be constructed to invite reflection rather than closure, to expose uncertainty rather than conceal it, and to scaffold judgment rather than replace it. Yet no degree of responsible design can relieve users of epistemic responsibility itself. Epistemic autonomy cannot be built into systems; it must be exercised by agents. Whether AI becomes a medium of empowerment or a vector of heteronomy ultimately depends on how individuals appropriate it in concrete practices of inquiry, learning, and judgment.

For this reason, AI maturity cannot remain a merely individual concern. As AI increasingly structures shared epistemic environments – classrooms, workplaces, public discourse – the consequences of epistemic immaturity extend beyond the individual user. Delegated judgment scales socially. What begins as personal convenience can become collective vulnerability. AI maturity therefore emerges as a societal task: a form of shared responsibility for maintaining the conditions under which autonomous judgment remains possible in AI-saturated contexts. At the same time, significant risks remain. Commercial AI systems are often optimized for efficiency, personalization, and engagement rather than for epistemic transparency or educational value. Without clear institutional frameworks and critical AI governance, educational environments may unintentionally reinforce dependency, inequality, and passive forms of cognition instead of fostering epistemic autonomy.

This task is unavoidably educational. If maturity is understood, in the Kantian sense, as the ongoing effort to resist self-incurred heteronomy, then it must be cultivated early, explicitly, and continuously. Learners must be enabled not merely to use AI competently, but to recognize when its use supports judgment and when it silently substitutes for it. Such sensitivity cannot be secured once and for all. AI maturity is not a stable achievement but a fragile, revisable disposition – one that must be renegotiated with every new technological development, every increase in speed, and every expansion of algorithmic assistance.

In this sense, AI maturity describes a task for a future that has already begun. It articulates a normative orientation: to use powerful artificial systems without surrendering authorship over our reasons. Whether this orientation can be sustained will depend less on technologies and institutions than on our collective willingness to treat epistemic autonomy as something that must be actively practiced – even, and especially, when thinking could be delegated instead.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

References

Alshamsi, I., et al. (2024). Exploring the impact of generative AI technologies on education: Academic expert perspectives, trends, and implications for sustainable development goals. *Journal of Infrastructure, Policy and Development*, 8(11), 8532. https://doi.org/10.24294/jipd.v8i11.8532.

Baehr, J. (2011). *The Inquiring Mind: On Intellectual Virtues and Virtue Epistemology*. Oxford: Oxford University Press.

Baehr, J. (2019). Intellectual virtues, critical thinking, and the aims of education, in *Routledge Handbook of Social Epistemology*, (eds.) P. Graham, M. Fricker, D. Henderson, N. Pedersen & J. Wyatt, 447–57. London: Routledge.

Baehr, J. (2021). *Deep in Thought: A Practical Guide to Teaching for Intellectual Virtues*. Cambridge, MA: Harvard Education Press.

Brandom, R. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, MA: Harvard University Press.

Buijsman, S., et al. (2025). Autonomy by design: Preserving human autonomy in AI decision-support. *Philosophy & Technology*, 38, 97. https://doi.org/10.1007/s13347-025-00932-2.

Burns, M., et al. (2026). *A New Direction for Students in an AI World: Prosper, Prepare, Protect*. Washington, DC: Center for Universal Education, Brookings Institution.

Chen, S., et al. (2025). When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. *npj Digital Medicine*, 8, 605. https://doi.org/10.1038/s41746-025-02008-z.

Cheng, M., et al. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391, eaec8352. https://doi.org/10.1126/science.aec8352.

Chirayath, G., et al. (2025). Cognitive offloading or cognitive overload? How AI alters the mental architecture of coping. *Frontiers in Psychology*, 16, 1699320. https://doi.org/10.3389/fpsyg.2025.1699320.

Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford: Oxford University Press.

Clark, A. & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7–19. https://doi.org/10.1093/analys/58.1.7.

Constantinescu, M., et al. (2021). Understanding responsibility in Responsible AI: Dianoetic virtues and the hard problem of context. *Ethics*

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

and Information Technology, 23, 803–814. https://doi.org/10.1007/s10676-021-09616-9.

Davis, J. L. (2020). *How Artifacts Afford: The Power and Politics of Everyday Things*. Cambridge, MA: MIT Press.

Digital Education Council (DEC). (2026). AI in Higher Education Global Survey 2026. www.digitaleducationcouncil.com/resource-library-items/ai-in-higher-education-global-survey-2026 (accessed: 07.08.2026).

European Union (2024). *Artificial Intelligence Act (Regulation (EU) 2024/1689)*. Official Journal of the European Union, L 1689. EUR-Lex. http://data.europa.eu/eli/reg/2024/1689/oj.

Firth, J., et al. (2019). The "online brain": How the Internet may be changing our cognition. *World Psychiatry*, 18(2), 119–129. https://doi.org/10.1002/wps.20617.

Floridi, L. (2025). Distant writing: Literary production in the Age of Artificial Intelligence. *Minds and Machines*, 35, 30. https://doi.org/10.1007/s11023-025-09732-1.

Frankfurt, H. G. (2005). *On Bullshit*. Princeton, NJ: Princeton University Press.

Froese, T., & Ziemke, T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence*, 173, 466–500. https://doi.org/10.1016/j.artint.2008.12.001.

Georgiou, G. P. (2025). ChatGPT produces more "lazy" thinkers: Evidence of cognitive engagement decline, arXiv preprint. https://doi.org/10.48550/arXiv.2507.00181.

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1), 6. https://doi.org/10.3390/soc15010006.

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin.

Hayes, S., et al. (2024). Towards a postdigital social contract for higher education in the age of Artificial Intelligence. *Postdigital Science and Education*, 6, 467–485. https://doi.org/10.1007/s42438-024-00459-3.

Hicks, M. T., et al. (2024). ChatGPT is bullshit. *Ethics and Information Technology*, 26, 38. https://doi.org/10.1007/s10676-024-09775-5.

Hu, X., et al. (2019). A role for metamemory in cognitive offloading. *Cognition*, 193, 104012. https://doi.org/10.1016/j.cognition.2019.104012.

Ihde, D. (2009). A phenomenology of technics, in *Readings in the Philosophy of Technology*, (ed.) D. M. Kaplan, 2nd ed., 76–97. Lanham, MD: Rowman & Littlefield.

Iqbal, J., et al. (2025). Generative AI tool use enhances academic achievement in sustainable education through shared metacognition and

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

cognitive offloading among preservice teachers. *Scientific Reports*, 15, 16610. https://doi.org/10.1038/s41598-025-01676-x.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why language models hallucinate, arXiv preprint. https://doi.org/10.48550/arXiv.2509.04664.

Kant, I. (1996). An answer to the question: What is enlightenment?, in *Immanuel Kant, Practical Philosophy*, (ed.) Mary J. Gregor, Academy Edition Vol. 8. Cambridge: Cambridge University Press, 17–22.

Kasneci, E., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274.

Kosmyna, N., et al. (2025). Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task, arXiv preprint. https://doi.org/10.48550/arXiv.2506.08872

Lee, D., et al. (2025). Metacognitive sensitivity: The key to calibrating trust and optimal decision making with AI. *PNAS Nexus*, 4(5), pgaf133. https://doi.org/10.1093/pnasnexus/pgaf133.

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations, in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*, 1–16. New York: Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727.

Matheson, J., & K. Lougheed (2022). Introduction: Puzzles concerning epistemic autonomy, in *Epistemic Autonomy*, (eds.) Jonathan Matheson and Kirk Lougheed, 1–18. New York: Routledge.

Ng, D. T. K., et al. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, 100041. https://doi.org/10.1016/j.caeai.2021.100041.

Ng, D. T. K., et al. (2025). From AI literacy to generative AI literacy: Conceptualization and frameworks, in *From AI Literacy to Generative AI literacy: Theory, Policy and Practice*, (eds.) D. T. K. Ng et al., 21–38. Singapore: Springer Nature. https://doi.org/10.1007/978-981-95-0950-8_3.

Noller, J. (2024). Extended human agency: Towards a teleological account of AI. *Humanities and Social Sciences Communications*, 11, 1338. https://doi.org/10.1057/s41599-024-03849-x.

Noller, J. (2026). A Kantian critique of Artificial Intelligence. *Danish Yearbook of Philosophy*, 59(1), 27–53. https://doi.org/10.1163/24689300-bja10090.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

O'Rourke, M. (2025). I Teach Creative Writing. This Is What A.I. Is Doing to Students, The New York Times, 18 July. www.nytimes.com/2025/07/18/opinion/ai-chatgpt-school.html

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. New York: Penguin Press.

Park, P. S., et al. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), 100988. https://doi.org/10.1016/j.patter.2024.100988.

Pedreschi, D., et al. (2025). Human–AI coevolution. *Artificial Intelligence*, 339, 104244. https://doi.org/10.1016/j.artint.2024.104244.

Plato (2015). *Theaetetus and Sophist*. Translated and edited by Christopher Rowe. Cambridge: Cambridge University Press.

Pritchard, D. (2023). Educating for virtuous intellectual character and valuing truth. *Philosophies*, 8(2), 29. https://doi.org/10.3390/philosophies8020029.

Prunkl, C. (2024). Human autonomy at risk? An analysis of the challenges from AI. *Minds and Machines*, 34, 26. https://doi.org/10.1007/s11023-024-09665-1.

Qian, Y. (2025). Pedagogical applications of generative AI in higher education: A systematic review of the field. *TechTrends*, 69, 1105–1120. https://doi.org/10.1007/s11528-025-01100-1.

Riley, C., Al-Refai, O., Colunga Reyes, Y., & Hammad, E. (2025). Human–AI interactions: Cognitive, behavioral, and emotional impacts, arXiv preprint. https://arxiv.org/abs/2510.17753.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002.

Roe, J. (2025). *How to Use Generative AI in Educational Research*. Cambridge Elements in Research Methods in Education. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781009675338.

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

Sellars, W. (1997). *Empiricism and the Philosophy of Mind* (R. Brandom, Ed.). Cambridge, MA: Harvard University Press. (Original work published 1956).

Sharma, M., et al. (2024). *Towards understanding sycophancy in language models*. In *Proceedings of ICLR 2024*. https://doi.org/10.48550/arXiv.2310.13548.

Siegel, H. (2017). *Education's Epistemology: Rationality, Diversity, and Critical Thinking*. New York: Oxford University Press.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

Southworth, J., et al. (2023). Developing a model for *AI Across the curriculum*: Transforming the higher education landscape via innovation in AI literacy. *Computers and Education: Artificial Intelligence*, 4, Article 100127. https://doi.org/10.1016/j.caeai.2023.100127.

Spante, M., et al. (2018). Digital competence and digital literacy in higher education research: Systematic review of concept use. *Cogent Education*, 5(1), 1519143. https://doi.org/10.1080/2331186X.2018.1519143.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive Load Theory*. New York: Springer.

Tankelevitch, L., et al. (2024). The metacognitive demands and opportunities of generative AI, in *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, 1–24. New York: ACM. https://doi.org/10.1145/3613904.3642902.

Tate, T. P., Harnick-Shapiro, B., Ritchie, D. R., Tseng, W., Dennin, M., & Warschauer, M. (2025). Incorporating generative AI into a writing-intensive undergraduate course without off-loading learning. *Discover Computing*, 28, 72. https://doi.org/10.1007/s10791-025-09563-9.

Turri, J., et al. (2021). Virtue epistemology, in *The Stanford Encyclopedia of Philosophy (Winter 2021 Edition)*, (ed.) Edward N. Zalta. https://plato.stanford.edu/archives/win2021/entries/epistemology-virtue.

UNESCO. (2023). Guidance for generative AI in education and research. https://doi.org/10.54675/EWZM9535.

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

Vallor, S. (2024). *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*. New York: Oxford University Press.

Vuorikari, R., et al. (2022). DigComp 2.2: The Digital Competence Framework for Citizens (EUR 31006 EN). Publications Office of the European Union, Luxembourg. https://doi.org/10.2760/115376.

Yan, L., et al. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55, 90–112. https://doi.org/10.1111/bjet.13370.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 495635215, 541433337.

The author used ChatGPT (OpenAI) and Claude (Anthropic) as a linguistic tool to assist in the articulation and phrasing of his original thoughts. This use was guided by the conception of epistemic autonomy developed in this Element. All conceptual content and arguments were developed independently by the author. The conceptual foundation for AI maturity draws on Noller (2024 and 2026). The author takes full responsibility for the accuracy and integrity of all content presented.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

Cambridge Elements

Generative AI in Education

Mark Warschauer

University of California, Irvine

Mark Warschauer is Distinguished Professor of Education at the University of California, Irvine, with affiliated faculty appointments in the Departments of Informatics, Language Science, and Psychological Science. He is a member of the National Academy of Education and the director of the UCI Digital Learning Lab. Professor Warschauer is one of the most influential scholars in the world on digital learning, digital literacy, and the use of AI in Education. He has published 12 books on these topics, including with MIT Press, Cambridge University Press, Teachers College Press, and Yale University Press, and some 300 scientific articles and papers. His work has been cited more than 48,000 times, making him one of the most cited researchers in the world on educational technology. He previously served as founding editor of *Language Learning & Technology* and inaugural editor of *AERA Open*.

Tamara Tate

University of California, Irvine

Tamara Tate is a Project Scientist at the University of California, Irvine, and Associate Director of the Digital Learning Lab. She leads the Lab's work on digital and online tools to support teaching and learning including generative AI, partnering with school districts, universities, non-profit organizations, media and tech developers, and others in iterative development and evaluation. As the PI of an NSF-funded grant, she is studying the use of generative AI in undergraduate writing courses. She also studies secondary student writing as a member of the IES-funded national WRITE Center. She received her B.A. in English and her Ph.D. in Education at U.C. Irvine and her J.D. at U.C. Berkeley.

Editorial Board

Stephen Aguilar, *University of Southern California, US*

Maha Bali, *American University in Cairo, Egypt*

Irene-Angelica Chounta, *University of Duisburg-Essen, Germany*

Shayan Doroudi, *University of California, Irvine, US*

María Florencia Ripani, *Ceibal Foundation, Uruguay*

Bart Rientes, *The Open University, UK*

Neil Selwyn, *Monash University, Australia*

Jiahong Su, *The University of Hong Kong*

Ulrich Trautwein, *University of Tübingen, Germany*

Ying Xu, *Harvard University, US*

About the Series

Generative AI is one of the most disruptive technologies in modern history, with the potential to dramatically transform education for better or worse. This series addresses cutting-edge topics on the intersection of generative AI with educational research and practice for diverse learners from early childhood to adult.

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press

Cambridge Elements

Generative AI in Education

Elements in the Series

Generative AI in Computer Science Education: Challenges and Opportunities
Diana Franklin, Paul Denny, David A. Gonzalez-Maldonado, and Minh Tran

Generative Artificial Intelligence and Language Teaching
Benjamin Luke Moorhouse and Kevin M. Wong

Academic Integrity in the Age of AI
Tricia Bertram Gallant, Mary Davis, and Zeenath Reza Khan

Growing Up with Artificial Intelligence: Implications for Child Development
Ying Xu and Kunlei He

Generative Artificial Intelligence and English-Medium Education
Benjamin Luke Moorhouse, Jack Pun, and Samantha M. Curle

Generative AI for Second Language Writing Feedback, Assessment, and Instruction
Mimi Li, Jining Han, and Gouda Taha

AI Policy Research: A New Inquiry
Danping Wang

Generative AI and Educator Professional Development
Lucas Kohnke

AI Maturity
Jörg Noller

A full series listing is available at: www.cambridge.org/EAIE

https://doi.org/10.1017/9781009779951 Published online by Cambridge University Press