Received: 22 November 2022 | Accepted: 29 April 2023

DOI: 10.1111/bjet.13337

ORIGINAL ARTICLE

British Journal of Educational Technology BERA

Learning to work with the black box: Pedagogy for a world with artificial intelligence

Margaret Bearman | Rola Ajjawi

Centre for Research in Assessment and Digital Learning (CRADLE), Deakin University, Docklands, Victoria, Australia

Correspondence
Margaret Bearman, Centre for Research in Assessment and Digital Learning (CRADLE), Deakin University, Tower 2, Level 12, 727 Collins Street, Docklands, VIC 3008, Australia.
Email: margaret.bearman@deakin.edu.au

Abstract

Artificial intelligence (AI) is increasingly integrating into our society. University education needs to maintain its relevance in an AI-mediated world, but the higher education sector is only beginning to engage deeply with the implications of AI within society. We define AI according to a relational epistemology, where, *in the context of a particular interaction, a computational artefact provides a judgement about an optimal course of action and that this judgement cannot be traced*. Therefore, by definition, AI must always act as a 'black box'. Rather than seeking to explain 'black boxes', we argue that a pedagogy for an AI-mediated world involves learning to work with opaque, partial and ambiguous situations, which reflect the entangled relationships between people and technologies. Such a pedagogy asks learners locate AI as socially bounded, where AI is always understood within the contexts of its use. We outline two particular approaches to achieve this: (a) orienting students to quality standards that surround AIs, what might be called the tacit and explicit 'rules of the game'; and (b) providing meaningful interactions with AI systems.

KEYWORDS

artificial intelligence, generative AI, higher education, evaluative judgement, relational epistemology

INTRODUCTION

Artificial intelligence (AI) mediates much of what we do every day. Every Google search, email filter or image manipulator relies on AI in one sense or another, in that the technologies seek to

Practitioner notes

What is already known about this topic

- Artificial intelligence (AI) is conceptualised in many different ways but is rarely defined in the higher education literature.
- Experts have outlined a range of graduate capabilities for working in a world of AI such as teamwork or ethical thinking.
- The higher education literature outlines an imperative need to respond to AI, as underlined by recent commentary on ChatGPT.

What this paper adds

- A definition of an AI that is relational: A particular interaction where a computational artefact provides a judgement about an optimal course of action, which cannot be easily traced.
- Focusing on *working with* AI black boxes rather than trying to *see inside* the technology.
- Describing a pedagogy for an AI-mediated world that promotes working in complex situations with partial and indeterminate information.

Implications for practice and/or policy

- Focusing on quality standards helps learners understand the social regulating boundaries around AI.
- Promoting learner interactions with AI as part of a sociotechnical ensemble helps build evaluative judgement in weighting AI's contribution to work.
- Asking learners to work with AI systems prompts understanding of the evaluative, ethical and practical necessities of working with a black box.

take the ‘best possible action in a situation’ (Russell & Norvig, 2016, p. 30). The rise of generative AI (genAI) underlines the increasing significance of AI to our world. These technologies can offer great advantages. Machines can learn at scale, based on large amounts of data and thus conduct tasks that people are incapable of doing in the same time frame, or even at all. Academic commentators describe a radical shift in job profiles and work practices (Moscardini et al., 2020). At the same time, there are great concerns about the ethics and limitations of AI; it has increased tacit surveillance, promoted racism and enhanced criminal activity (Hayward & Maas, 2020). Many suggest there is therefore an urgent need for ‘explainable AI’, which allows the outputs of a ‘black box’ computational system to be seen as reasonable by stakeholders (Hoffman et al., 2018). In other words, though AI holds great promise, it is less clear how it can productively and ethically integrate into contemporary society. Higher education clearly can play a role in preparing students to live and work in an AI-mediated world.

The literature suggests AI requires imperative response from universities but studies in higher education have not engaged deeply with the implications of an AI-mediated society (Bearman et al., 2022). Aoun's (2017) call to ‘robot-proof’ universities remains influential, proposing that higher education should focus on building skills that are uniquely human, such as teamwork and creativity. While these so-called 21st-century skills are important, they do not necessarily take account of the disciplinary and situated nature of knowledge and practice. Along similar lines, experts in artificial intelligence outline some of the key capabilities required for an AI-mediated world: These are deeply humanistic ideas of agency, ethical thinking and knowledge creation (Markauskaite et al., 2022). These start to address how human expertise can best complement the strengths of artificial intelligence, such as identifying patterns at speed and at scale. However, this literature focuses on what these capabilities are, and not how they might be achieved.

This paper seeks to outline pedagogical approaches fit for an AI-mediated society. This is not, therefore, a call for a curriculum or for specific skills, although these are important, nor is it a singular all-encompassing pedagogy. Rather, we start with the premise, that with such a rapid changing environment, *how* will be more important than *what*. This paper therefore emphasises, not so much *what* we should be teaching, but *how* we invite students to learn and work with indeterminacy. Through such pedagogical approaches, many deeper knowledge practice capabilities can be developed in concert with disciplinary contexts.

Our pedagogical approaches are based around a singular core contention: that AI resembles many other aspects of our complex-socially mediated world in that it can never be fully explainable or transparent. Therefore, rather than explaining or unpacking 'black boxes'—things that we do not understand and cannot explain—our graduates must learn to work and make judgements within complex organisational and social structures, where they need to synthesise information in a sophisticated way as part of their professional practices. After our theoretical defences of this argument, we propose pedagogical approaches for working with 'black boxes' in a complex and ambiguous AI-mediated world.

THEORISING AND DEFINING AI

Theoretical antecedents

We situate our argument in a relational epistemology, where knowing is held between actors. In other words, knowledge is not a singular property of an individual or a machine but is contextualised within particular relationships between people, things and spaces. As with connectivism, we: 'recognize[s] the connection of everything to everything…' (Siemens, 2005). We draw from a sociomaterial perspective, which 'appreciate[es] human/non-human action and knowledge as entangled in systemic webs…' (Fenwick, 2010, p. 111). We see non-humans as active players, drawing from Latour's (2007) emphasis on actors being known by their actions and impact upon others, not by their humanity. For example, Latour (1999) suggests a 'speed bump' is an actor and its agency is expressed through the effects it exerts upon traffic drivers to slow the car. In this way, we consider AIs to be agentic, although not sentient. Thus, we understand knowledge and knowing as producing, and being produced by, the social dynamics bound up with objects and spaces (Foucault, 1963). All of these views are roughly running in the same ontic and epistemic boundaries, and reflect how Johnson and Verdicchio (2017, p. 583) conceptualise *AI systems as*: '*sociotechnical ensembles* … combinations of artefacts, human behaviour, social arrangement and meaning. [italics ours]'.

We also step outside of this frame to note that in higher education, students are asked to demonstrate their knowledge and knowing, from an internal, mostly cognitive perspective. Assessment (and teaching) often assumes learning is an individual and a-contextual activity, yet seeks to prepare students for professional workplaces where people, communities, objects and spaces are bound together in activity. We acknowledge that any pedagogical approaches based on a relational epistemology has to take this tension into account.

Defining AI in a relational epistemology

The famous Turing 'imitation game' proposes that AI is 'reached' when it is impossible to judge whether a conversation was between a human and human or a human and a machine.

While this notion of the 'Turing machine' has been critiqued (Searle, 1980), we feel that Turing's original idea is sound: it is not what the computer is but what *a person thinks the computer is*, that is significant. Inspired by Turing, we conceptualise AI not by what it *is*, but how it is *interacted* with. For example, many people take on trust: the outputs of sophisticated statistical software analyses; a Google search; or a facial recognition algorithm that unlocks our phone. We consider all of these *can be* AI—not because of the underlying computational approach—but because of the contextually bound relationship between the person and the technology. This means, for a small child, a calculator performing simple multiplication may be a form of AI but not for an older child and so on. From this perspective, an AI is determined not just by technology but also through activity.

AI is often defined as having 'weak' forms such as very basic support tools as well as 'strong' forms where machines have some form of thinking consciousness (Bechmann & Bowker, 2019). We deviate from this and define an *AI interaction* as occurring when *in the context of a particular interaction*, a *computational artefact provides a judgement to inform an optimal course of action and that this judgement cannot be traced*. In other words, an AI interaction does not depend on how a technology is constructed, but on the relationship between the humans and the computational artefact, at a particular moment in time. This aligns with our view of knowledge as a sociomaterial production as it emphasises not what things are, but what they *do*, together. From this perspective, an AI interaction must involve information being produced by a 'black box'. Our definition of AI includes those interactions with technologies that suggest optimal courses of actions: search engine algorithms; expert systems that replicate human (often expert) derived rules; and machine learning systems, that can optimise identification of patterns within datasets, like genAI. However, these may not be part of AI interactions, depending on the circumstances of use—a judgement of an expert system used by an expert may be entirely traceable. Our definition excludes technologies that do not present judgements about best courses of action, such as databases and learning management systems (unless they contain predictive elements). For most university students and teachers, engaging with simpler technologies such as spelling and grammar checkers and spreadsheets will not form AI interactions, as outputs can be easily traced.

Why students need to learn to work with 'black boxes'

AI is often referred to as a 'black box' and this is presented as a concern. Zednik (2021, p. 265) writes: 'The Black Box Problem arises when the computing systems being developed in AI are opaque. This metaphorical way of speaking is grounded in the intuition that a system's behavior can be explained by "looking inside." … they might be considered opaque in the sense that it is difficult to know **exactly** how these systems are programmed'. [Bold ours]. From our epistemological perspective, we maintain there is less need to understand *precisely how* AI is constructed. As Latour (1986) describes, a scientific study takes all the human uncertainties, ideas and conversations and condenses it into an inscription (the scientific paper). Indeed, this very text you are reading is a black box of a sort—it is unclear **exactly** how it has come into being, even we as authors could not tell you why or how we were inspired to write certain words in a precise and detailed way. In short, people and things are generally not fully knowable or even explainable.

From an educational perspective, we make the claim that trying to 'see inside' the black box somewhat misses the point. While the 'explainable AI' approach acknowledges that AI may need to be interpretable rather than completely understood, such efforts appear underpinned by a need for transparency (Gilpin et al., 2018). But, the metaphor of transparency suggests that knowledge is fixed, fully known and measurable (Orr, 2007). What if, as we propose, knowledge is relational, partial and indeterminant? We suggest that pedagogical approaches for an AI-mediated world should orient towards what things *do together*, rather than what they *are, separately*.

WHAT DO PEDAGOGIES FOR AN AI-MEDIATED WORLD INVOLVE?

Pedagogies for an AI-mediated world consider AI as part of a *sociotechnical ensemble*, to use Johnson and Verdicchio's (2017) terminology. Therefore, we contend learning to work with AI and other black boxes moves away from analysis of what things *are*, towards an analysis of what things *do*. People encounter AIs, machineries, bureaucracies and people that they do not understand and yet have to engage with them in productive ways. These pedagogies can also prepare students for environments where situations can be ambivalent, unknown and uncertain, but where graduates must continue to work with the information they have, towards outcomes that may be unknown. They must learn when to trust and when to doubt and when doubt turns to mistrust.

We do not claim a singular unifying pedagogy for working with 'black boxes' but to outline different pedagogical approaches that develop students capabilities for navigating a complex and indeterminant AI-mediated world. We identify two particular foci: (1) orienting towards quality standards; and (2) promoting meaningful interactions with 'AI systems'. The first frames interactions with AI as part of a broader negotiation between student and the socially produced 'rules of the game'. The second focuses on learning to work with AI systems in a way that promotes student meaning-making about the nature of their interactions. Both engage students to work with complexity, ambiguity and partial information.

Orienting to quality standards within sociotechnical ensembles

Any sociotechnical system does not exist in a vacuum: it is a social endeavour. When an AI-produced image wins a prize for best artwork (Roose, 2022), it does so within social boundaries that arbitrate: (1) what the quality of beauty is; and (2) what the rules of the competition are. Every current AI, no matter how it is defined, exists within these boundaries. Indeed, examples of AIs that are used to illustrate its power, such as AlphaGo beating grandmaster Lee Sodol (Newcomb, 2016), also illustrate the limitations of what such computational artefacts offer within an AI system. A game such as 'go' has explicit rules and AlphaGo must operate within these socially constructed and clearly defined boundaries. A game also has implicit social rules. Success in playing 'go' might also be about having fun or forming a social bond (Luckin, 2018); AlphaGo cannot 'beat' humans here. This simple example reveals how significant these implicit and explicit social 'rules of the game' are when working with AI systems.

The first overarching pedagogical idea that we offer is that we should orient or attune learners to these social rules for success within their course of study, which we will call *quality standards*. One way of conceptualising this in simple terms is that our students guidance to help them produce good quality work. For example, when a software student 'writes elegant code' or a nursing student 'conducts a sensitive conversation with a patient' or a philosophy student 'develops a coherent and original argument', we have to teach them what constitutes 'elegance' or 'sensitivity' or 'originality' in disciplinary practice (Bearman, 2018). Through engaging with these, students can come to understand 'what good looks like' within their course.

Understanding quality standards, whether tacit or explicit, offers the means for students to critically interrogate the AI systems that they work with: They can offer a useful reference point for students when working with 'black boxes'. Artefacts that represent these *quality standards* are constantly present in higher education, in all disciplines, and are a fundamental part of higher education (Ajjawi et al., 2021). For example (as we describe later), a practice guideline is a disciplinary *quality standard* and a rubric is a local *quality standard* for a given assessment. However, these epistemic artefacts are often taken for granted both by students and educators. We think this is a missed opportunity. Our pedagogical approaches therefore seek to provide opportunities for students to critically engage with quality standards framed within AI systems. We offer specific pedagogical strategies for this: engaging with disciplinary guidelines; designing rubrics for ambiguity and complexity; and discussing the limits of standards, including ethical implications. We describe each in turn.

Engaging with professional or disciplinary guidelines

The most obvious example of explicit quality standards are disciplinary or professional practice guidelines—such as the guidelines for ethical conduct of human research or published critical appraisal tools for judging the quality of papers. Working with such guidelines is a matter of complex and sometimes uncertain professional judgements, not simply following a checklist, and may therefore help individuals develop 'a sophisticated series of beliefs and thoughts about what [disciplinary] knowledge is and how it can be applied', necessary for an AI-mediated world (Bearman & Luckin, 2020, p. 57). These texts can be used to grapple with some of the complexity and ambiguity of working with black boxes within a discipline or profession.

One valuable pedagogical technique is to identify moments when guidelines, AI recommendations and other types of information clash within sociotechnical ensembles. To provide a simple example, physiotherapy students might wish to consider what would happen when a predictive algorithm within an electronic record suggests a hospital patient is at risk of a fall, based on characteristics such as age and admission conditions. They are also asked to consider the *disciplinary guidelines*, which point to the patient being safe for some independent tasks, and patient's wishes to be completely independent. This provides the opportunity for exploration and discussion, as different courses of care plans are investigated. What should the physiotherapist recommend? How do contextual factors play out in the decision-making? What kind of extra information might the physiotherapist need? Who else might the physiotherapist need to work with to ensure a positive outcome? What are the ethical challenges in weighting the patient's wishes, the professional judgement and the algorithmic recommendations? This example illustrates how learning to working with an AI can reveal how many 'black boxes' we are really grappling with all of the time.

Designing rubrics for ambiguity and complexity

If we are asking students to orient to social boundaries through working with disciplinary or professional guidelines, we need also to review the quality standards that we offer to our own students. Most learners pay attention to assessment criteria: the quality standards by which their own performance will be judged. This most often takes the form of a *rubric*. The risk of rubrics is that they convey that knowledge and quality can be neatly contained in simple rows and columns—and if these are simply 'followed' then the students can get the grade/mark that they deserve. By seeking to be transparent and explicit, rubrics can remove ambiguity and complexity and also create endless work through rounds of checking and atomisation to get them 'right' (O'Donovan et al., 2004; Sadler, 2007). But, reductive criteria do not necessarily help students learn; in fact, we would argue that they prevent students grasping complex, indeterminate materials.

An alternative is designing rubrics as invitations to actions (Bearman & Ajjawi, 2019). An invitational metaphor for rubrics avoids the notion of transparency, but rather sees rubrics as inviting students into a 'productive space', where matters of ambiguity and complexity can be embraced. This shifts the emphasis from 'seeing through' (to the black box) to 'working with' (the black box). So how might providing rubrics that promote complex learning help our students with navigating an AI-mediated world? In the first instance, we hold that such rubrics implicitly—and hopefully explicitly—promote the need to work with ambiguity, where 'black boxes' cannot be fully 'known'. Take the predictive algorithm example above. If we were assessing student knowledge of the role of predictive algorithms within electronic health records, where might the rubric focus? A rubric that invited the student to broadly consider decision-making processes within ambiguous situations or to articulate the ethical challenges creates a very different view of quality standards than a rubric that represented success as the ability to categorise predictive algorithms in health. If we are to prepare students for an AI-mediated world, we need to offer standards that reflect the complex indeterminate world they will inhabit. And, we need to emphasise *why* we are not providing reductive check-boxes; we can explicitly acknowledge that standards are framed to reflect the complexity of those social regulating forces that frame an AI-mediated society.

Discussing the limits of standards

A key facet of practice standards is that, like society, they change and what is acceptable one decade is not acceptable the next (Bearman, 2018; Bearman et al., 2020). For instance, no one would wish to return to the accepted gender norms of 50 years ago. Standards can be considered as 'constantly being made and remade, and through this process, [they are] maintained and eroded at the same time' (Ajjawi et al., 2021, p. 735). Following Law (2009), we use the metaphor of rupture—we suggest that standards are like an elastic band that are reused and reused. There are many different ways of stretching the elastic and yet maintaining its fundamental integrity. This elasticity is necessary as it enables stretching of 'rules' to fit specific contexts where appropriate. However, after a certain point, the elastic snaps. A simple example, rules governing chess competition can vary significantly and still be recognisable, but at some point, if the standards are moved too far, the game simply will not be chess anymore. Therefore, we argue that our students need to know what is and is not acceptable in relation to a standard (ie, the social regulating boundaries framing AI systems) rather than focusing on a 'single use elastic band' (ie, a particular and limited occasion of use).

In learning to work with 'black boxes', graduates need to identify straining relationships across quality standards and AI interactions. This allows them to assess whether an AI system exists within the 'rules of the game'—as in our physiotherapy example. However, it is important to know when the rules themselves need to be critically engaged with. We want our students to contribute to the rules rather than simply seeing themselves as upholding and replicating existing norms.

Our graduates need to know under what circumstances, socially regulated boundaries need to be redefined or discarded in an AI-mediated world. A discussion of standards, in context of disciplinary AI practices, allows meaningful discussion of ethical and sociopolitical challenges. In such conversations, we urge educators to raise the particular AI challenges of unacceptable bias (Hayward & Maas, 2020) as well as the general challenge for technological systems that reify out-of-date legacy practices or knowledges, which become very difficult to shift (Selwyn, 2014). Students should be asked to critically engage with quality standards—whether articulated as guidelines, policy, technical specifications or even tacit norms. An AI artefact may embed and promote unacceptable discrimination on the basis of race or gender or disability (Vock, 2022). But if it is doing the latter, it is likely because it is built on a particular knowledge system that holds these biases. It is possible that AI interactions may *help* society understand that standards are insufficient. However, our students cannot attune to the bias of AI if they do not understand what biases there are in society in general. We must prompt our learners to ask questions such as: How do disciplinary and other standards reflect values of good practice? What other forms of quality standards are there, from other knowledge systems? How are the standards adapting? Whose interests are being served by this standard? How do AI interactions stretch or rupture them?

Promoting meaningful interactions with 'AI systems'

If orienting students to quality standards as a pedagogical technique focuses on how AI relates to the broader social context, there is also a need to consider how students must learn to work within the 'AI system'—that sociotechnical entanglement of practices, people and technology (Johnson & Verdicchio, 2017). Others have noted the value of assessing and understanding 'inputs' and 'outputs' as a means of working with AI 'black boxes' (Markauskaite et al., 2022; Zednik, 2021), but this necessarily focuses on the AIs as individual computational artefacts. Within a sociotechnical ensemble, we notice that inputs and outputs are often the same thing; moreover, this is the explicit premise of machine learning—the outputs become inputs of the next iteration. Thus, our pedagogical approach zooms out to consider holistically what might be 'meaningful' interactions and, through this engagement, learn to work with black boxes, where information is partial, opaque and ambiguous.

We use 'meaningful' here to denote that, through interaction, students generate valuable knowledge about themselves and their broader contexts of community. While there are many ways to promote meaningful interactions with sociotechnical ensembles, we propose three particular pedagogic strategies: (1) developing critical digital literacies for an AI-mediated world; (2) tasks that develop evaluative judgement; and (3) acknowledging emotions, and the role of trust and doubt.

Developing critical digital literacies for an AI-mediated world

Freire (1985) described literacy as the ability to *read the world*. We suggest that our students need to be able to work with AI through 'reading' a technologically mediated world, including opacities and ambiguities. We turn to the digital literacies literature. Pangrazio and Sefton-Green (2021) distinguish between digital literacies as a matter of simple skill development or more critical and evaluative approaches. It is the latter that we are interested in here with respect to the 'Black Box Problem'. In particular, learners should learn to seek and evaluate what information is generated within an AI interaction and how can this be used to improve work. For example, professionals such as doctors need to be able to draw from environmental cues to help them monitor their own practice (Bearman et al., 2020) and these cues will be increasingly mediated by AI. This is complex territory for students: They need to know both how to pragmatically engage with the digital but critically assess the effects of the digital upon their work. This is therefore not a matter of 'what' digital skills students need but 'how' they come to critique and work with the digital within an AI system. Some of this will be specific to a discipline, but the overall idea of critique can be understood across courses.

We provide a specific example, which aligns with general academic literacy. To help students understand the role of data within an AI system, we can offer a discussion of the impact of quantification upon how we read, use and produce academic literature. Educators could discuss their interactions with the common academic search engine, Google Scholar. (According to our definition these are often AI encounters, in that academics rely on judgements about 'optimal' papers, in a way that is difficult to trace). Such a discussion could highlight how different literature 'rise' to the top, how citation rates and hits may influence what is seen as quality, and even how academics can be ranked through metrics such as the h-index. This allows a concrete conversation exploring how datafication legitimises quantitative reports and delegitimises other types of texts. At the same time, it acknowledges the reality that these types of data and supporting AIs have become integrated into our professional practices and everyday lives, in often useful ways.

Providing tasks that develop evaluative judgement through AI interactions

Learners do more than 'read the world', they also contribute to it, through their own work. We suggest that in association with building literacies, a pedagogy for an AI-mediated world should offer opportunities to build evaluative judgement. Evaluative judgement is a learner's 'capability to make decisions about the quality of work of self and others' (Tai et al., 2018, p. 471) and encompasses the implicit and explicit cues that learners need to know whether work is meeting quality standards (Bearman et al., 2020). As has been noted elsewhere, promoting evaluative judgement can prepare learners to work in an AI-mediated world as 'it shifts the focus from being successful (yes or no), to coming to understand how success is constituted' (Bearman & Luckin, 2020, p. 57).

Developing evaluative judgement allows learners to grasp how their work meets the broader societal notion of quality, as outlined in the earlier part of this paper, and this capability also allows learners to assess the outputs within an AI interaction. Some might say, but how can the learner know if the AI contributions are good or bad if the AI is not traceable or explainable? We draw from our previous arguments to mount three responses. Firstly, learners can bring knowledge of disciplinary quality standards. Secondly, social actors generally do not explain why they undertake particular actions, but we still judge their contributions. Finally, digital literacies, as outlined earlier, can be brought to bear in developing evaluative judgements within AI interactions.

We believe there is an urgent need to highlight graded and non-graded assessment tasks that ask learners to exercise their evaluative judgement in situations where AI has a strong presence. Students need to be given the opportunity to assess how their AI interactions contribute to a particular task, and to incorporate this understanding into further developing their work. For example, we can explicitly ask first year students to consider the use of everyday computational artefacts that they may be using—such as search engines and genAIs—by asking them to assess both positive and negative impacts upon their work. When and why do they trust its output? How does use intersect with academic integrity and how should they reference their use? What are strengths and deficiencies? How might these AI interactions perpetuate disadvantage? By asking students to examine their assumptions about AI and how it contributes towards their work, we start to build an understanding of what students' relationships with AI produce, within their particular disciplinary contexts.

Evaluative judgement is not just about cognition: emotions form an often powerful part of practice when working with the ambiguous, complex and unknown situations. Learners should understand the role of their emotions in evaluative judgements with AI systems. We expand upon emotions, including trust and doubt, below.

Acknowledging emotions: The role of trust and doubt

There is a sense that in conversations about technology, the affective is often removed to the background as irrational and subjective. From a relational epistemology, emotions aren't internal feelings that we manage through 'reason', rather they are intimately bound with action. As humans we are *affected* by, and *affect*, the world around us. Emotions mobilise how we negotiate our relationships with people, spaces and objects. Feelings of frustration or excitement or loyalty in association with technologies are familiar to most. Learning to work with 'black boxes' also requires acknowledgement of these embodied, tacit and human experiences. We do not see educators' role as helping students 'regulate' emotions but rather to prompt consideration of the role that emotions necessarily play within sociotechnical ensembles. Indeed, partial knowledge, indeterminacy and ambiguity are often associated with feelings of discomfort or even anxiety.

Emotions are particularly pertinent when considering the role of trust within an AI system. Trust has been conceptualised as a form of judgement that has both cognitive and affective elements and involves a 'leap of faith' towards a favourable outcome (Castanelli et al., 2022). If a student 'trusts' the sociotechnical ensemble or a particular AI artefact, then that student will act differently, than if they mistrust a situation or a technology. Mistrust can alert students to ethical concerns such as bias or other forms of injustice, but it can also inhibit necessary actions. Sometimes, on balance, leaps of faith must be taken. How can this be navigated?

Emotional reflexivity is described by Olson et al. (2021, p. 2) as 'a process of drawing on emotions to, potentially, chart a unique path', and through this acknowledgement, understanding how emotions are shaping action and experience. Through considering—and even mobilising—feelings, people can navigate emotionally and relationally through the world. This is a useful framing. Both trust and distrust are powerful affective prompts that should not be overlooked by students—but nor are they sufficient in themselves. Highlighting the need to be emotionally reflexive can prompt students to examine their interactions with AI systems in a critical way.

We also introduce the idea of 'epistemic doubt'. Here, we appropriate Hoeyer and Wadmann's (2020) concept based on the observation that healthcare professionals experience an ambiguous link between data as recorded and real-world objects in clinical practice. We conceptualise epistemic doubt as cognitive and affective: a state of uncertainty and discomfort. Thus, we can ask students to hold AI interactions in 'epistemic doubt'—understanding that information within AI interactions may be partial or biased or possibly incorrect. And, this can lead to them to take information 'on trust', while holding epistemic doubt at the same time. In so doing, they can turn from trust to distrust and possibly back again.

EXAMPLES OF THE PEDAGOGICAL APPROACHES IN USE

To summarise, we propose pedagogical approaches based on a relational epistemology that suggests how students can learn to work with AI systems, by engaging with judgements that cannot be 'traced'. Firstly, we *orient students towards quality standards*. This is key because judgements about what constitutes quality is a social and relational endeavour and underpins how AI systems contribute to our society. These approaches do not require engagement with AI itself, necessarily, but develop students' awareness of the 'rules of the game' that set boundaries for an AI system. We suggest: referencing *disciplinary or professional guidelines*; developing *rubrics* that account for ambiguity and complexity; and prompting students to critical discuss the *limits of quality standards* relevant to their discipline. Secondly, we discuss pedagogical approaches to *promote students' meaningful interactions with AI* *systems*. We describe: developing *critical digital literacies* relevant to AI systems; developing students' *evaluative judgement* with respect to AI systems, particularly in light of the ‘rules of the game’; and finally, we indicate a role for *emotional reflexivity* and *epistemic doubt*.

To illustrate these ideas, we offer two very different task designs, both of which locate AI within a sociotechnical ensemble and work with what AI *does* rather than what it is. In both instances, they focus on genAI and have been chosen to for their relevance across most disciplines.

Task 1: Developing evaluative judgement with genAI

The aim of this simple example task is to develop students' understanding of what ‘good writing’ is in their discipline, with reference to generative AI, such as ChatGPT.

For this task, small groups of students are provided with: (1) exemplars of ‘good writing’ relevant to their discipline; (2) *disciplinary guidelines* (eg, reporting standards for randomised controlled trials or documentation standards for software design or a published description of referencing historical sources); and (3) a teacher-designed *rubric* that is not about ‘rules’ but raises some of the subjectivities and complexities of ‘good writing’. They are asked to assess the exemplars using the disciplinary guidelines and the rubric, to *orient them towards quality standards* for textual communication, whether it be humanities, business or sciences. Next, each student writes a short text on a topic relevant to the discipline and also prompts a genAI (where safe, ethical and legal) to produce a piece of writing on the same topic. They consider their original text plus the generated text and assess both against both the disciplinary guideline and teacher-provided rubric, *developing evaluative judgement*—coming to understand what ‘good writing’ looks like in themselves and others. They then write and submit a piece outlining which text is better and why, and what they *trust, doubt* and feel about the AI interaction, *building critical digital literacies* and *emotional reflexivity*. The task orients students to what the students, the criteria and the AI *do together*.

Task 2: Simulating the challenge of genAI for academic integrity

The next task is located within a Graduate Certificate of Higher Education or similar, the types of preparatory programmes that university educators undertake to develop the quality of their teaching. It is designed to provoke educators (as learners) to engage with the complexity of an AI-mediated world, orienting them to the contextual challenge of ensuring academic integrity in higher education.

In the first part of the task, the learners (ie, educators) are given three assignments: (a) an independently written assignment (with permission); (b) one supplemented with genAI; and (c) one written almost entirely with genAI. They are told that students have been using genAI to produce their work and to consider whether they can identify how. This is based around a conversation about *social standards*: should genAI supplement student writing and, if so, how? It also orients the educators towards considering *trust and doubt*; they must critically reflect on how their own feelings affect their views. By doing so, they are *developing their evaluative judgement* about what constitutes appropriate use of genAI tools.

In the second part of the task, the learners (ie, educators) are asked to design assignments and associated *rubrics* that would (a) assist in preventing inappropriate use of genAI and (b) promote effective use of genAI. In groups, they debate whether these might work or not, and how using genAI might affect students in unexpected ways, building the educators' *critical digital literacies*. The educators would then return to the final part of the task which is to present their own views on the intersection between genAI and academic integrity, particularly highlighting the paradoxes, tensions and '*epistemic doubts*'.

Explaining the pedagogy to students and educators

These example tasks illustrate how students and educators can learn to work with 'black boxes' found within AI interactions. One final key issue to be raised: we must make sure that when we offer students (and educators) these kinds of tasks, they understand *why* we are doing it. This means explaining the reasons that we need these pedagogies. After all, if we are asking students to engage with a complex ambiguous AI-mediated world, it is important that they are persuaded that this is an important part of a university education.

CONCLUSIONS

We present a perspective of AI that is based on a relational epistemology, in line with a sociomaterial ontology. We theorise AI as relational, and focus on its effects rather than its component parts. Framing AIs as 'black boxes' may help university students to shape the future through learning how to deal with indeterminacy, as they learn to work within sociotechnical ensembles. Our pedagogical approaches and strategies propose some means to achieve this. While working with ambiguity has always been part of the world we occupy, the pace and nature of change suggest a pressing need for our graduates to navigate an AI-mediated world.

FUNDING INFORMATION

This research was not funded.

ACKNOWLEDGEMENT

Open access publishing facilitated by Deakin University, as part of the Wiley - Deakin University agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST STATEMENT

We have no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

As this is a conceptual paper, data access does not apply.

ETHICS STATEMENT

No human participants were involved in this research and thus approval for the conduct of ethical research has not been sought.

ORCID

Margaret Bearman https://orcid.org/0000-0002-6862-9871
Rola Ajjawi https://orcid.org/0000-0003-0651-3870

REFERENCES

Ajjawi, R., Bearman, M., & Boud, D. (2021). Performing standards: A critical perspective on the contemporary use of standards in assessment. *Teaching in Higher Education*, *26*(5), 728–741. https://doi.org/10.1080/13562517.2019.1678579

Aoun, J. E. (2017). *Robot-proof: Higher education in the age of artificial intelligence*. MIT Press.

Bearman, M. (2018). Prefigurement, identities and agency: The disciplinary nature of evaluative judgement. In *Developing evaluative judgement in higher education: Assessment for knowing and producing quality work* (pp. 147–155). Routledge.

Bearman, M., & Ajjawi, R. (2019). Can a rubric do more than be transparent? Invitation as a new metaphor for assessment criteria. *Studies in Higher Education*, *1*, 359–368. https://doi.org/10.1080/03075079.2019.1637842

Bearman, M., Brown, J., Kirby, C., & Ajjawi, R. (2020). Feedback that helps trainees learn to practice without supervision. *Academic Medicine*, *96*(2), 205–209.

Bearman, M., & Luckin, R. (2020). Preparing university assessment for a world with AI: Tasks for human intelligence. In *Re-imagining university assessment in a digital world* (pp. 49–63). Springer.

Bearman, M., Ryan, J., & Ajjawi, R. (2022). Discourses of artificial intelligence in higher education: A critical literature review. *Higher Education*, 1–17. https://doi.org/10.1007/s10734-022-00937-2

Bechmann, A., & Bowker, G. C. (2019). Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media. *Big Data & Society*, *6*(1), 2053951718819569.

Castanelli, D. J., Weller, J. M., Molloy, E., & Bearman, M. (2022). How trainees come to trust supervisors in workplace-based assessment: A grounded theory study. *Academic Medicine*, *97*(5), 704–710. https://doi.org/10.1097/acm.0000000000004501

Fenwick, T. (2010). Re-thinking the “thing”. *Journal of Workplace Learning*, *22*(1/2), 104–116. https://doi.org/10.1108/13665621011012898

Foucault, M. (1963). *The birth of the clinic*. Routledge.

Freire, P. (1985). Reading the world and reading the word: An interview with Paulo Freire. *Language Arts*, *62*(1), 15–21.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*.

Hayward, K. J., & Maas, M. M. (2020). Artificial intelligence and crime: A primer for criminologists. *Crime, Media, Culture*, *17*(2), 209–233. https://doi.org/10.1177/1741659020917434

Hoeyer, K., & Wadmann, S. (2020). ‘Meaningless work’: How the datafication of health reconfigures knowledge about work and erodes professional judgement. *Economy and Society*, *49*(3), 433–454.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). *Metrics for explainable AI: Challenges and prospects*. arXiv:1812.04608.

Johnson, D. G., & Verdicchio, M. (2017). Reframing AI discourse. *Minds and Machines*, *27*(4), 575–590. https://doi.org/10.1007/s11023-017-9417-6

Latour, B. (1986). Visualization and cognition. *Knowledge and Society*, *6*(6), 1–40.

Latour, B. (1999). *Pandora's hope: Essays on the reality of science studies*. Harvard University Press.

Latour, B. (2007). *Reassembling the social: An introduction to actor-network-theory*. Oxford University Press.

Law, J. (2009). Actor network theory and material semiotics. *The New Blackwell Companion to Social Theory*, *3*, 141–158.

Luckin, R. (2018). *Machine learning and human intelligence: The future of education for the 21st century*. UCL Press.

Markauskaite, L., Marrone, R., Poquet, O., Knight, S., Martinez-Maldonado, R., Howard, S., Tondeur, J., De Laat, M., Buckingham Shum, S., Gašević, D., & Siemens, G. (2022). Rethinking the entwinement between artificial intelligence and human learning: What capabilities do learners need for a world with AI? *Computers and Education: Artificial Intelligence*, *3*, 100056. https://doi.org/10.1016/j.caeai.2022.100056

Moscardini, A. O., Strachan, R., & Vlasova, T. (2020). The role of universities in modern society. *Studies in Higher Education*, *1*, 812–830. https://doi.org/10.1080/03075079.2020.1807493

Newcomb, A. (2016). Go grandmaster Lee Sedol reflects on losing series to Google's computer. *ABC News*. https://abcnews.go.com/Technology/grandmaster-lee-sedol-reflects-losing-series-googles-computer/story?id=37656509

O'Donovan, B., Price, M., & Rust, C. (2004). Know what I mean? Enhancing student understanding of assessment standards and criteria. *Teaching in Higher Education*, *9*(3), 325–335. https://doi.org/10.1080/1356251042000216642

Olson, R. E., Smith, A., Good, P., Neate, E., Hughes, C., & Hardy, J. (2021). Emotionally reflexive labour in end-of-life communication. *Social Science & Medicine*, *291*, 112928. https://doi.org/10.1016/j.socscimed.2020.112928

Orr, S. (2007). Assessment moderation: Constructing the marks and constructing the students. *Assessment & Evaluation in Higher Education*, *32*(6), 645–656. https://doi.org/10.1080/02602930601117068

Pangrazio, L., & Sefton-Green, J. (2021). Digital rights, digital citizenship and digital literacy: What's the difference? *Journal of New Approaches in Educational Research*, *10*(1), 15–27.

Roose, K. (2022). An A.I.—Generated picture won an art prize. Artists aren't happy. *The New York Times*. https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html

Russell, S., & Norvig, P. (2016). *Artificial intelligence: A modern approach*. Harlow Pearson.

Sadler, D. R. (2007). Perils in the meticulous specification of goals and assessment criteria. *Assessment in Education*, *14*(3), 387–392.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Selwyn, N. (2014). *Digital technology and the contemporary university: Degrees of digitization*. Routledge.

Siemens, G. (2005). Connectivism: A learning theory for the digital age. *International Journal of Instructional Technology and Distance Learning*, *2*(1). http://www.itdl.org/

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education*, *76*(3), 467–481.

Vock, I. (2022). ChatGPT proves that AI still has a racism problem. *The New Statesman*. https://www.newstatesman.com/quickfire/2022/12/chatgpt-shows-ai-racism-problem

Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, *34*(2), 265–288. https://doi.org/10.1007/s13347-019-00382-7

How to cite this article: Bearman, M., & Ajjawi, R. (2023). Learning to work with the black box: Pedagogy for a world with artificial intelligence. *British Journal of Educational Technology*, *00*, 1–14. https://doi.org/10.1111/bjet.13337