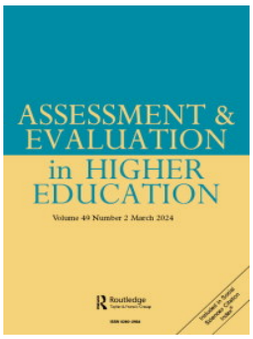

Assessment & Evaluation in Higher Education

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/caeh20

Developing evaluative judgement for a time of generative artificial intelligence

Margaret Bearman, Joanna Tai, Phillip Dawson, David Boud & Rola Ajjawi

To cite this article: Margaret Bearman, Joanna Tai, Phillip Dawson, David Boud & Rola Ajjawi (10 Apr 2024): Developing evaluative judgement for a time of generative artificial intelligence, Assessment & Evaluation in Higher Education, DOI: 10.1080/02602938.2024.2335321

To link to this article: https://doi.org/10.1080/02602938.2024.2335321

Published online: 10 Apr 2024.

ASSESSMENT & EVALUATION IN HIGHER EDUCATION
https://doi.org/10.1080/02602938.2024.2335321

Developing evaluative judgement for a time of generative artificial intelligence

Margaret Bearman[a], Joanna Tai[a], Phillip Dawson[a], David Boud[a,b,c] and Rola Ajjawi[a]

[a]Centre for Research in Assessment and Digital Learning (CRADLE), Deakin University, Melbourne, Australia; [b]Faculty of Arts and Social Sciences, University of Technology Sydney, Sydney, Australia; [c]Work and Learning Research Centre, Middlesex University, London, UK

ABSTRACT

Generative artificial intelligence (AI) has rapidly increased capacity for producing textual, visual and auditory outputs, yet there are ongoing concerns regarding the quality of those outputs. There is an urgent need to develop students' evaluative judgement – the capability to judge the quality of work of self and others – in recognition of this new reality. In this conceptual paper, we describe the intersection between evaluative judgement and generative AI with a view to articulating how assessment practices can help students learn to work productively with generative AI. We propose three foci: (1) developing evaluative judgement of generative AI outputs; (2) developing evaluative judgement of generative AI processes; and (3) generative AI assessment of student evaluative judgements. We argue for developing students' capabilities to identify and calibrate quality of work – uniquely human capabilities at a time of technological acceleration – through existing formative assessment strategies. These approaches circumvent and interrupt students' uncritical usage of generative AI. The relationship between evaluative judgement and generative AI is more than just the application of human judgement to machine outputs. We have a collective responsibility, as educators and learners, to ensure that humans do not relinquish their roles as arbiters of quality.

KEYWORDS

Generative artificial intelligence; evaluative judgement; assessment for learning; higher education

Introduction

As generative artificial intelligence (AI) rapidly integrates into society, it becomes pressingly important for every user to be able to recognise the quality of its outputs. Generative AI is noteworthy for its 'hallucinations' or 'confident responses that seemed faithful and non-sensical when viewed in light of the common knowledge in these areas' (Alkaissi and McFarlane 2023, 3). University graduates should be able to effectively deploy the disciplinary knowledge gained within their degrees to distinguish trustworthy insights from the 'hallucinatory'. Therefore, there is an urgent need to develop students' e*valuative judgement* or 'the capability to judge the quality of work of self and others' (Tai et al. 2018, 472). While evaluative judgement may be critical for working with generative AI, at the same time generative AI provides the opportunity to

CONTACT Margaret Bearman margaret.bearman@deakin.edu.au

better develop evaluative judgement. By exploring the intersections between generative AI and evaluative judgement, we lay out a case for assessment practices that prepare students for a world where judging AI outputs occurs on a daily basis.

Assessment and associated pedagogical strategies are strongly connected to developing students' evaluative judgement. Indeed, recognising the quality of one's own work is one of the key drivers behind sustainable assessment approaches. Boud (2000, 152) writes that, through assessment, students should 'be prepared to undertake assessment of the learning tasks they face throughout their lives… in ways *which identify whether they have met whatever standards are appropriate for the task in hand*' [italics ours]. Boud argues that this will enable students to 'undertake related learning more effectively', including broadly seeking feedback information to avoid being 'dependent on teachers or other formal sources of advice'. This work forms the foundation to the idea that evaluative judgement – or coming to know 'what good looks like' – is a key means to sustain learning from assessment (Tai et al. 2018).

Evaluative judgement takes on particular significance with respect to AI, because of the role that humans play as the arbiters of quality (Bearman and Luckin 2020). In other words, while a generative AI may offer a nuanced and sophisticated response to a prompt, a *person* needs to judge the quality of its output. We offer the example of writing a poem: a human prompts a generative AI to write a sonnet for a colleague's birthday. In this case, the generative AI produces text within long-standing rules about sonnet rhyme and meter but the decision about whether the poem 'works' rests with the prompter. Thus, quality can be considered as socially constituted boundaries that indicate success (Bearman and Ajjawi 2023). In this case, and in many others in the educational context, it is human judgement of quality that ultimately prevails.

While generative AI is newly emerging as a significant presence in higher education, early reports of use suggest that many, although not all, students are using it to supplement their study or in undertaking assessment tasks (Nam 2023; Ziebell and Skeat 2023; Freeman 2024). In March 2023, a survey of 1000 US college students described that a substantial minority were using AI tools like ChatGPT within their studies; however, 17% of this unspecified minority reported generating an assignment text without edits (Welding 2023). By November 2023, a survey of 1000 US college students showed that 56% used AI tools to complete assignments or examinations and 53% reported that their coursework required them to use AI tools (Nam 2023). A small sample of Australian university students provides descriptions of how students are using ChatGPT as a 'study buddy' for activities such as gaining a broad overview, clarifying assessment instructions and generating different options (Ziebell and Skeat 2023). Similarly, about a third of 1250 surveyed UK students noted that they used generative AI as a 'personal tutor' (Freeman 2024). This latter survey also indicates that about a third of respondents noted that in any instance of use, they 'did not know' how often generative AI was producing 'hallucinations'. These are early findings and therefore must be read with caveats about their generalisability, but at the same time they are suggestive of general trends. We also note that the variety of generative AI-backed tools which may be used in higher education is ever increasing, both across various means of communication (e.g. text, images, sound), but also as inherent to a domain, subject or discipline – e.g. writing assistance (Escalante, Pack, and Barrett 2023), qualitative analysis methods (Siiman et al. 2023) or computer code generation/programming (Prather et al. 2023).

Some university educators are actively exploring how to employ generative AI to develop learner capabilities such as critical thinking (Guo and Lee 2023). A key idea is that higher education should focus on developing uniquely human capabilities (Aoun 2017) rather than those that can easily be undertaken by readily available technologies. In other words, if machines can perform a cognitive task, there is a diminished need for graduates to learn the associated know-how, though they do need to know the circumstances in which they can off-load tasks to a given machine (Dawson 2020). This makes sense: computerisation automates routine cognitive tasks and this shifts the tasks people do in workplaces (Autor, Levy, and Murnane 2003). But there is less consensus about what these human capabilities are. Clearly students will require digital literacies, but the

focus is often on even higher-level skills. For example, a capability such as critical thinking does not *necessarily* involve technology but is required for graduates to work with the digital (Bearman, Nieminen, et al. 2023). Markauskaite et al. (2022) describe an array of possible human capabilities necessary for an AI-mediated world – 'creativity', 'deliberate engagement' with technologies, managing 'visual representations' and so on. Bearman and Luckin (2020) argue that evaluative judgement is one of these unique human capabilities, due to its emphasis on identifying quality.

Defining evaluative judgement

Tai et al. (2018) describe evaluative judgement as having two core components. Firstly, a person must hold an internal understanding of what constitutes quality, and secondly they must make a judgement about work – whether it be theirs or someone else's. Thus, evaluative judgement is distinguished from generalised capabilities, such as critical thinking or problem-solving, by its specific focus on complex appraisal of texts, artefacts or performances. Sadler (1989, 130) describes this type of judgement as 'multicriterion'. While such judgements can be analytic – built through stepwise appraisal of certain factors or criteria – Sadler (1989, 132) particularly notes the importance of global judgements of quality, where 'imperfectly differentiated criteria are compounded as a kind of gestalt and projected onto a single scale of quality, not by means of a formal rule but through the integrative power of the assessor's brain'. That is to say, expert evaluative judgement when appraising the quality of a text or artefact or performance is often tacit, holistic and not reducible to individual parts.

We conceptualise the act of evaluative judgement as bringing together a person's internal conception of quality with holistic disciplinary standards (Bearman 2018). It must therefore always be contextualised. As Tai et al. (2018, 472) note: 'Making an evaluative judgement requires the activation of knowledge about quality in relation to a problem space'. This is consonant with what Kuhn, Cheney, and Weinstock (2000) call an 'evaluativist epistemology' where knowing is neither entirely absolute nor entirely relativist but requires people to make alignments between internal perspectives and evidence from the surrounding world. Indeed, we regard making an evaluative judgement as not just a matter of selecting 'yes/no' or 'correct/incorrect' – but weighing up the alternatives in particular situations. For example, a piece of computer code is judged according to how well it meets its requirements, including its efficiency and conceptual elegance. Similarly, a philosophy essay is judged according to how well it demonstrates original argumentation and appropriate refutations. As a novice, these judgements tend to be analytical – in reference to set criteria such as those found in rubrics – but they increasingly become more holistic as a person becomes more expert.

Our view of evaluative judgement emphasises this contextualised capability. There are other perspectives: for example, Luo and Chan (2023) explore the relationship between evaluative judgement and holistic competencies. This casts evaluative judgement as a process, and therefore becomes more similar to process-based capabilities such as Nicol's (2021) view of self-feedback where making comparisons with different kinds of reference material is central to feedback and learning. These approaches are valuable and consonant with our approach but as mentioned our focus is on evaluative judgement as the capability required to appraise a specific instance of work.

Evaluative judgement as a contextualised capability has three distinctive features. Firstly, it is generally always *exercised* as part of complex tasks but learners don't always have the capability to make good appraisals. For example, a student can examine an academic paper and not understand that it has fatal flaws. Secondly, a person has 'good' evaluative judgement when their contextualised claims are defensible in relation to the broad acceptable standards of quality associated with the text, artefact or performance that is being appraised. In our example, a different student might defend their appraisal of the academic paper, relying on general standards of rigour known within the discipline. Finally, people can be oriented towards *developing* good

evaluative judgement. As Tai et al. (2018, 472) write: 'we consider that the process of developing evaluative judgement needs to be deliberate and be deliberated upon'. Thus, drawing on our academic paper example, students can be taught 'critical appraisal' of the literature and thus develop good evaluative judgment with respect to scholarly work within their discipline.

Assessment practices, which generally rely on students making judgements about whether the task is ready to submit, offer the primary vehicle in higher education for developing good evaluative judgement. As we articulate later, this includes assessment-related pedagogical strategies such as self-assessment; peer assessment/review; feedback; rubrics; or exemplars.

Intersections between evaluative judgement and generative AI

There is a deep need when working with generative AI for students to recognise the quality of its outputs as they can often appear plausible and relevant even when they may be unsuitable. This is not just a matter of 'hallucinations' – that is information that is clearly incorrect – but also the overall value of these outputs. For example, they can be too generic or contain bias or draw from outdated assumptions and texts. Therefore, assessment designs should promote the development of students' evaluative judgement capabilities with respect to such outputs (Bearman and Luckin 2020). In addition, the availability of generative AI tools itself offers convenient opportunities for students to deliberately develop their evaluative judgement. We propose three points of intersection between generative AI and the development of good evaluative judgement capability (Figure 1). Firstly, a learner can *assess generative AI outputs*, for example judging the quality of a text produced by ChatGPT. Secondly, a learner can *develop their own generative AI processes*, for example assessing the quality and utility of the prompts that they use to interrogate ChatGPT. Finally, learners can employ *generative AI to assess their own evaluative judgement*. We articulate what we mean in each of these situations before discussing how assessment strategies might deliberately orient learners towards improving their evaluative judgements and issues that they can face in so doing.

Developing evaluative judgement through assessing generative AI outputs

When evaluative judgement is mentioned in relation to generative AI, the most immediate response is that students need to discern the quality of its output. For some, the automatic assumption may be that students are using generative AI to present its outputs as their own. However, early survey work suggest that this is far from how a student uses, for example, ChatGPT (Ziebell and Skeat 2023). When working with generative AI, students can: seek an

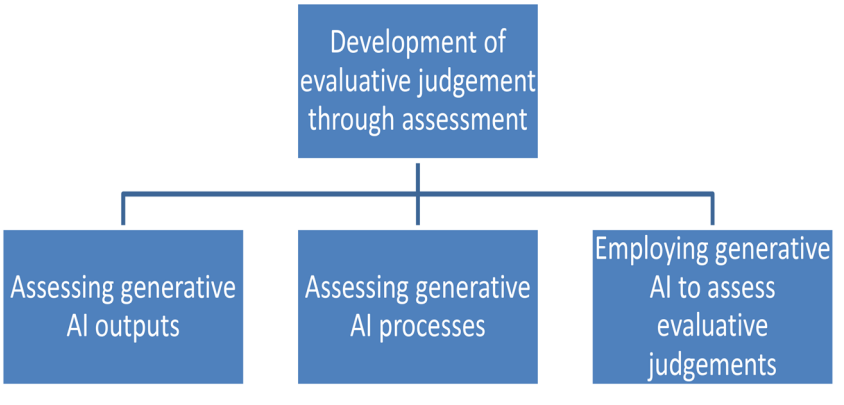

Figure 1. Three assessment strategies for developing evaluative judgement.

exemplar response; or a particular point of view; or a broad scoping of information about a topic of interest; or wish to receive comments from the generative AI on their own work. In all these instances, students will exercise evaluative judgement through appraisals that are essentially no different to that which they might make in relation to information from teachers, peers or the scholarly literature.

The feedback literature suggests that information cannot be disentangled from how the student perceives the source (Telio, Regehr, and Ajjawi 2016; Watling et al. 2012). According to this literature, learners often tacitly consider: what do I know about this source? How much can I trust it? Through the lens of these initial credibility impressions, the learners can begin to assess the generative AI responses to their prompts. They may ask: is the content directly relevant to what I need to know? Does the response address the quality of output I am seeking? Is it in a form I can readily work with for the task I have in hand? If the information from generative AI doesn't meet these basic requirements, students will likely rephrase prompts, to focus more precisely on the issue at hand. Or they can turn to another source. But we should not assume that because the act of using generative AI *exercises* evaluative judgement, students are developing good evaluative judgement.

One of the dangers of generative AI is that those with poor understandings – and hence an inability to make good evaluative judgements – will uncritically rely on inappropriate outputs. Developing evaluative judgement means building the ability to critically reflect on information which is superficially persuasive. Beyond familiarity with the subject, this requires an understanding of how to judge a piece of work within that domain, and how to employ generative AI to help build high quality work. This is where assessment practices can assist.

Educators can employ familiar general assessment pedagogical strategies to develop evaluative judgement but adapted to generative AI, as we outline in Table 1. In many ways, generative AI is just another source of information about their work or understanding. However, a key difference to previous strategies is that they must learn to weigh up the credibility of the generative AI outputs by charting inconsistencies between the texts produced by generative AI and information or standards from known credible sources. In addition, they can directly and iteratively compare multiple generative AI examples with their own work. For example, they can prompt the generative AI to produce introductory arguments, substantive content, and so on, and then compare their own. This is a similar process to comparing one's own work with that of a peer. Judgements are made both about the quality of the example, and the quality of one's own work using appropriate criteria for comparison. Such comparisons feed into developing holistic impressions of quality.

This act of evaluative judgement can lead to either the generation of further prompts to the tool, or to modifications to one's own work. Both are of interest but particularly the latter: evaluative judgement here is not just about outputs but also about how to integrate the outputs into the task the student is working on. However, the most significant focus should be the development of students' own capability to judge the quality of work in the topic at hand. This is how learning is sustained. The particular judgement at the time is less important than how it underpins the development of evaluative judgement capability with respect to similar tasks in the future.

While assessing generative AI outputs forms the most obvious point to develop evaluative judgement, it is not the only point of intersection. We explore two additional ways in which students may need to develop their evaluative judgement with respect to generative AI.

Developing evaluative judgement about generative AI processes

Students may develop their evaluative judgement in relation to the quality of their own use of generative AI tools within assessment tasks. Discerning 'good practices' for working with

generative AI technologies is important – not just from a quality of output/performance sense, but regarding the quality of ethical and moral engagement. There may be circumstances in which it is inappropriate to engage with generative AI processes, and learners should develop their evaluative judgement in relation to when or when it is not acceptable to use generative AI.

In addition to generalised teaching about using generative AI, assessment practices also offer opportunities to navigate some of this complexity. For example, many assessment tasks may preclude certain uses of generative AI and students will need to know how to negotiate this. As outlined in Table 1, there are a range of assessment pedagogic strategies that can be adapted with respect to building such capabilities.

One obvious aspect of using generative AI is the need to develop appropriate and relevant prompts, and to iteratively determine if the prompting (as done by the student) is of sufficient quality to use the tool effectively. Such evaluative judgement about prompt quality is likely to only be developed when students interact with, and use, generative AI tools in context. Whilst as educators we may be able to provide some heuristics and overarching principles of how to 'engineer' a good prompt, it may be that students can only fully grasp how series of particular prompts produce desired outputs through spending time iterating prompts and seeing what happens. This aligns with the role of a 'prompt engineer' – someone whose job is to write prompts for generative AIs. Such a role relies on the capacity to judge quality of these interactions with generative AI, rather than just the output. We do not want students stumbling upon the 'right answer' through erroneous assumptions or methods. Therefore, it is essential to ensure that learners also know something about the quality of processes required to arrive at outputs, answers or solutions.

A less obvious process question is when to stop using an AI tool, from an ethical or moral point of view. The product of the generative AI may be both relevant and of high quality, but from a generative AI process perspective there may be ethical or moral reasons to disengage. This could be because the task requires a person to complete the work without any AI input, but it might also be because the generative AI itself is known to be unethical or the particular generative AI may breach privacy or intellectual property considerations. If students submit their own work to generative AI, do they know where it is stored and what happens to it? These processes may not immediately impact the student's work but form part of the broader context for the usage and we should orient students to assess when and why AI use is not acceptable.

Generative AI assessment of student evaluative judgement

Finally, generative AI can be used to assess students' evaluative judgement: a student can ask a generative AI not just 'have I done a good enough job?', but 'have I accurately appraised the quality of my work?' The most obvious way this can be done is through substitution of external human appraisals with AI appraisals. For example, a student might already assess their own work against a rubric – as mentioned above self-assessment forms a means of developing evaluative judgement – and then a teacher might comment on how their expert judgement differs from the student's. However, a generative AI tool could easily substitute the teacher's appraisal (although it is unlikely to produce the same result). The student must then employ their evaluative judgement of the generative AI output, as described in an earlier section, to discern if the AI judgement of their judgement has merit.

Another possibility is that generative AI can also be used to model evaluative judgements, and the students can then compare their own judgements with these outputs. Consider those generative AI tools which are intended to critique or synthesise scholarly literature. Students (and scholars) can provide such tools with a reference and ask it to identify strengths and weaknesses, or even ask specific methodological questions. This can be helpful in identifying which papers are worth reading, and whether a particular paper has some flaws in it that the student might

not have thought of. Exposure to this sort of scholarly critique may develop evaluative judgement. However, just as with concerns about generative AI diminishing students' ability to write, there is the potential for this use of generative AI to inhibit the development of evaluative judgement, if they do not make judgements of their own but purely rely on cognitively offloading.

Students using generative AI to perform evaluative judgement and calibrate their judgement do not need to regard the generative AI outputs as expert or even accurate. In fact, generative AI may provide the opportunity for students to encounter dubious evaluative judgements that they may wish to argue against, and in doing so, develop their ability to articulate their own evaluative judgement. Unlike teacher judgements which can be seen to carry the weight of authority in an education setting, judgements from generative AI can be ignored or challenged without any degree of interpersonal conflict or considerations of power differentials. While generative AI is capable of dialogic exchanges about its judgements, in some instances it can act eager to please, and may change its outputs to suit the user's preferences or instructions. Thus, all the cautions that accompany the previous two modes of developing evaluative judgement also apply here.

Assessment practices that develop evaluative judgement in a time of generative AI

In the Tai et al. (2018) paper, five assessment-related pedagogical strategies were proposed that could be harnessed to develop evaluative judgement. These were: self-assessment; peer assessment/review; feedback; rubrics; and exemplars. While generally these would be considered parts of an overarching and holistic approach to developing evaluative judgement, in Table 1 we now focus on each individual strategy to assist educators considering how and why they might employ these in a time of generative AI. We aim to shift these strategies beyond how they are currently employed to develop evaluative judgement, which we describe alongside an exemplar from the published literature. Interestingly, we note that employing generative AI in pursuit of evaluative judgement development permits more teacher-like roles to be taken on by students, although this does not mean that evaluative judgement needs to be unguided or untaught. Rather, we suggest that this is 'teacherly work' that students will need to sustain themselves in the future, rather than a substitution of the educator by AI.

Building evaluative judgement systemically over a program of assessment

Taken together, these types of assessment practices interrupt students' uncritical usage and acceptance of generative AI at the same time as building better evaluative judgement capabilities. While we have separated out the three intersections of evaluative judgement and generative AI, these are inter-related as the development of this capability is iterative and processual. For example, as students produce generative AI outputs, they will dynamically appraise the products and processes involved in working with generative AI. Moreover, they can recursively use generative AI itself to assess these early evaluative judgements. For most students, some form of generative AI will be readily available, so they may rely on it at any time during their learning journeys – either as part of the formal design or informally as part of their general learning processes.

Good evaluative judgement is underpinned by a contextualised understanding of quality. To come to this understanding of quality, students need to appraise many different examples. As Sadler (1989, 128) notes: 'Students need, in many educational contexts, to be presented with several exemplars (for a single standard) precisely to learn that there are different ways in which work of a particular quality can find expression'. Therefore, we recognise the need for iterative employment of generative AI for the same task, as well as across tasks and even across subjects,

Table 1. Assessment strategies that develop evaluative judgements in a time of generative AI (categories replicated from Tai et al. (2018)).

Strategy	Current means to develop evaluative judgement (and illustrative reference)	Developing evaluative judgement with respect to:
Self-assessment	Self-assessment allows the opportunity for students to examine their own work against external criteria, which is a useful means to develop evaluative judgement. Ideally, these same criteria, whether in rubrics or other form, will span units, allowing the students to trace their developing understanding of how to identify quality in their own work (e.g. McIver and Murphy 2023).	*Assessing generative AI outputs* Self-assessment practices can be explicitly expanded to consider how to identify quality when working with generative AI. This means developing: a) evaluative judgements about the quality of the outputs; and b) evaluative judgements about how well these outputs have been integrated into the task, possibly employing external criteria such as rubrics. *Assessing generative AI processes* Self-assessment tasks can also include students' evaluative judgements about their processes pertaining to generative AI – e.g. how iterative prompts were utilised and the quality of these strategies. *generative AI assessment of student evaluative judgement* The generative AI can directly assess the student's work and then the student can compare their own judgement with the generative AI's.
Peer assessment or peer review	Peer assessment allows students to review others' work or performance, articulating evaluative judgements on how they address the relevant criteria or standards and how the work could be improved. This may be a reciprocal process where students also receive comments back on their own work, or may be engaged in purely for the practise of making and articulating judgements (e.g. Chen et al. 2022).	*Assessing generative AI outputs* Peers can jointly review each other's evaluative judgements of generative AI outputs. Again, this is both: a) appraisal of the outputs; and b) appraisal of the student's ability to integrate the AI into the task vis-à-vis external criteria. *Assessing generative AI processes* Peers can jointly review each other's generative AI processes that underpin an assessment task. Peers can work together to iteratively develop prompts and judge their quality in relation to the outputs. *generative AI assessment of student evaluative judgement* A peer review role could develop into being a 'human in the loop' who either approves or corrects a range of generative AI judgements on others' work. It still requires students to interact with work and make judgements independently of the generative AI tool, as well as with it.
Feedback	Students solicit and receive feedback inputs from others that focus on the evaluative judgements the learner makes about their own work, rather than the substantive features of that work. They compare their own judgements with those of others (e.g. De Mello Heredia, Henderson, and Phillips 2023).	*Assessing generative AI outputs* Students can iteratively request feedback outputs from generative AI about their work and compare these texts with others' judgements, including their own. Thus, the student's evaluative judgement is built around generative AI's capability to provide useful information about own work, about the student's own feedback processes, and about the quality of the work itself. *Assessing generative AI processes* Students can reflect on the variety of feedback inputs and outputs they have garnered from generative AI and start to form explicit judgements about their own processes in working with generative AI. In addition, students can solicit feedback outputs from generative AI about how they have worked with generative AI and compare these outputs with others' judgements, including their own. *generative AI assessment of student evaluative judgement* With generative AI, feedback and self-assessment strategies become linked: the generative AI can directly assess the student's work, and the student compares their own judgement with the generative AI's.

(*Continued*)

Table 1. Continued.

Strategy	Current means to develop evaluative judgement (and illustrative reference)	Developing evaluative judgement with respect to:
Rubrics	Students and/or teachers construct rubrics as a way to articulate and explore relevant quality standards. Students use rubrics as a scaffold to support their evaluative judgement when undertaking self- and peer-assessment (e.g. Gyamfi, Hanna, and Khosravi 2022).	*Assessing generative AI outputs* When using outputs from generative AI as part of their assessed work, students can use rubrics to evaluate the quality of this supplemented work and identify aspects of the work that require human intervention to improve. Thus, evaluative judgement is built about the generative AI outputs and the work itself. A caution is that analytical rubrics may be limited when it comes to holistic judgements about quality (Sadler 1989). *Assessing generative AI processes* Peers can employ a rubric to jointly review each other's generative AI processes that underpin an assessment task. Where there is a difference in judgement, they engage in a dialogue with generative AI and/or others to decide which information to trust. The rubric itself may focus on assessing generative AI processes, both skills and ethical judgements. *generative AI assessment of student evaluative judgement* The student can ask the generative AI to assess how well the student has used the rubric; the student and the generative AI may also assess the rubric itself and the student can compare comments about its qualities. Students can construct rubrics with generative AI as a means of exploring relevant quality standards.
Exemplars	Students use exemplars from a variety of sources to enable them to identify features of good work of the type they are seeking to produce. They compare their own subsequent work with the exemplars (e.g. Chong 2021).	*Assessing generative AI outputs* Students can use exemplars generated from a variety of generative AI outputs to enable them to identify features of good work of the type they are seeking to produce. They compare their own subsequent work with these exemplars and then seek further exemplars from the AI source. This iteratively develops evaluative judgement through these comparisons. They may need to draw on other exemplars from credible sources for calibration. *Assessing generative AI processes* In the above exemplar process, students can review how well their processes have worked in generating exemplars. Students could also be provided with exemplars of how to work with generative AI within tasks. *generative AI assessment of student evaluative judgement* Students prompt the generative AI to assess their judgements about exemplars and these can be compared to peers, teachers and own assessments.

as students progress in their degrees. This requires external feedback information, in addition to that provided by generative AI, in order to compare students' judgements with credible external insights. This will improve students' evaluative judgement capabilities with respect to generative AI; they can move between judging generative AI processes and outputs in relation to feedback information from peers or teachers.

This type of comparison over time allows students to calibrate their evaluative judgements against AI, peers, teachers and exemplars; they can come to understand they are developing complex higher order understandings of quality relevant to their discipline (such as philosophical argument or clinical reasoning or architectural design). Such calibrations might themselves be tracked using generative AI. For example, as students progress in their degrees, they might use feedback comments from first year as a prompt for conversation with generative AI, triangulating these against what the student has since learned. This is a form of ipsative feedback (Malecka and Boud 2023) or comparing current work against previous to provide insight into development. In this scenario, the student is using generative AI to track the development of their evaluative judgement of relevant higher order learning outcomes, in a way that spans years rather than months.

Finally, none of these changes can take place without teaching students about safe and effective use of generative AI. This requires dialogue about the use of generative AI in classrooms and with teachers, to remove the mystery and any fear of wrongdoing, and to discuss ethical use. It may be that a teacher employs intellectual candour (Molloy and Bearman 2019) about their own limitations to model some of the challenges faced when using the tools by presenting the intersection between generative AI and their own evaluative judgements. In particular, a discussion about credibility – of people and AIs – may be very enlightening for all.

The unknowns ahead of us

AI is often seen in opposition to the human (Bearman, Ryan, et al. 2023) and we cannot deny the impact of this on our students and ourselves. In previous research even when the feedback comments were exactly the same, students judged those from a human as more credible that those that were computer-generated (Lipnevich and Smith 2009a, 2009b). Perhaps over time, if working with generative AI leads to good grades or improved performance it is possible that students may rethink relative machine-human credibility. They may start to think that a generative AI has their best interests at heart, which in human relations leads to not only increased credibility but a strengthened relational bond (Telio, Regehr, and Ajjawi 2016). However, how students will come to regard generative AI in the future is unknown. But whatever it is, we must emphasise that because a technology performs well in one task (writing computer code or elegant prose) does not mean it will necessarily do well in another (presenting a nuanced rationale). Students must come to recognise that in some circumstances generative AI use will be appropriate and ethical and useful, while in other moments it will be none of those things.

We cannot talk about generative AI without discussing the corpus of data that underlies it. Student use of generative AI may depend on them knowing how that data is constituted. At present, the nature of generative AI corpora is obscure but we suggest is likely to reflect a Global North perspective of the world, among other biases. This is highly significant because students' development of evaluative judgement whilst working with generative AI at university will shape their future notions of 'quality'. The statistical tendency of generative AI towards dominant forms of knowledge and ways of seeing the world perpetuates existing inequities (Bender et al. 2021); the 'silenced' will have even less of a voice. Thus, any discussion of evaluative judgement and generative AI must also talk about the body of work it draws from, in order to understand both its limitations and its influence even when these limitations are understood. How generative AI will impact diversity in society remains uncertain.

Finally, we live in a time of accelerated technological change. While generative AI may feel as if it is beyond anything we have seen before, for many years the capabilities of this type of AI have been increasingly understood by many within the AI and education field. As with any new technology, it is less the power of the technology itself that is challenging but how these powers integrate with society. Social media presents an example of a relatively simple AI-powered technology that has had an outsized impact on social functioning and, arguably, upon notions of truth and associated knowledge practices (Barnett and Bengtsen 2017). Where will technologies lead next? In a time of great change, we argue that evaluative judgement– recognising quality of work of and with AI – becomes increasingly important.

Conclusions

AI has widened the gap between our capability to produce work, and our capability to evaluate the quality of that work. It is easier than ever to produce something that has the superficial appearance of quality—but knowing if it is good enough for any given purpose requires expertise. For learners to work productively with generative AI, it is important to work towards narrowing this gap. However, this is not the complete story on the relationship between evaluative judgement and AI. In this paper we have also argued that the relationship between evaluative judgement and generative AI is more than just the application of human judgement to machine outputs. Learners need to develop their ability to judge the quality of the processes they use with generative AI, and generative AI can also be a partner in the development of human evaluative judgement capability. While evaluative judgement might be developable through tasks not specifically designed to do so, in an age of AI not paying attention to evaluative judgement carries the risk that learners may start to adopt the understanding of quality that AI has inferred from its data, programmers or owners. If evaluative judgement is to remain a uniquely human capability (Bearman and Luckin 2020), it follows that the development of evaluative judgement is to remain a responsibility held by humans—both educators and learners.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Margaret Bearman is a Professor with the Centre for Research in Assessment and Digital Learning (CRADLE), Deakin University, Australia. Her current research investigates feedback, assessment, and digital learning in higher and professional education, with particular interests in sociomateriality and clinical contexts.

Joanna Tai is a Senior Research Fellow at the Centre for Research in Assessment and Digital Learning (CRADLE), Deakin University, Australia. She researches student experiences of learning and assessment from university to the workplace, including feedback and assessment literacy, evaluative judgement, and peer learning.

Phillip Dawson is a Professor and Co-Director of the Centre for Research in Assessment and Digital Learning (CRADLE), Deakin University, Australia. He researches higher education assessment, with a focus on feedback and cheating.

David Boud is an Alfred Deakin Professor and Foundation Director of the Centre for Research in Assessment and Digital Learning (CRADLE) at Deakin University, Australia. He is also Emeritus Professor at the University of Technology Sydney and Professor of Work and Learning at Middlesex University. His current work is in the areas of assessment for learning in higher education, academic formation, and workplace learning.

Rola Ajjawi is a Professor of Educational Research at the Centre for Research in Assessment and Digital Learning (CRADLE), Deakin University, Australia, where she has led an extensive programme of research into student failure and success, with particular interest in equity, feedback, and workplace learning cultures.

ORCID

Margaret Bearman http://orcid.org/0000-0002-6862-9871
Joanna Tai http://orcid.org/0000-0002-8984-2671
Phillip Dawson http://orcid.org/0000-0002-4513-8287
David Boud http://orcid.org/0000-0002-6883-2722
Rola Ajjawi http://orcid.org/0000-0003-0651-3870

References

Alkaissi, H., and S. I. McFarlane. 2023. "Artificial Hallucinations in ChatGPT: Implications in Scientific Writing." *Cureus* 15 (2): e35179. doi:10.7759/cureus.35179.

Aoun, J. E. 2017. *Robot-Proof: Higher Education in the Age of Artificial Intelligence*. Cambridge, Massachusetts: The MIT Press.

Autor, D. H., F. Levy, and R. J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *The Quarterly Journal of Economics* 118 (4): 1279–1333. doi:10.1162/003355303322552801.

Barnett, R., and S. Bengtsen. 2017. "Universities and Epistemology: From a Dissolution of Knowledge to the Emergence of a New Thinking." *Education Sciences* 7 (1): 38. doi:10.3390/educsci7010038.

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*, edited by David Boud, Rola Ajjawi, Phillip Dawson and Joanna Tai, 147–155. Abingdon: Routledge.

Bearman, M., and R. Ajjawi. 2023. "Learning to Work with the Black Box: Pedagogy for a World with Artificial Intelligence." *British Journal of Educational Technology* 54 (5): 1160–1173. doi:10.1111/bjet.13337.

Bearman, M., and R. Luckin. 2020. "Preparing University Assessment for a World with AI: Tasks for Human Intelligence." In *Re-Imagining University Assessment in a Digital World*, edited by Margaret Bearman, Phillip Dawson, Rola Ajjawi, Joanna Tai and David Boud, 49–63. Cham: Springer.

Bearman, M., J. H. Nieminen, and R. Ajjawi. 2023. "Designing Assessment in a Digital World: An Organising Framework." *Assessment & Evaluation in Higher Education* 48 (3): 291–304. doi:10.1080/02602938.2022.2069674.

Bearman, M., J. Ryan, and R. Ajjawi. 2023. "Discourses of Artificial Intelligence in Higher Education: A Critical Literature Review." *Higher Education* 86 (2): 369–385. doi:10.1007/s10734-022-00937-2.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Virtual event, Canada.

Boud, D. 2000. "Sustainable Assessment: Rethinking Assessment for the Learning Society." *Studies in Continuing Education* 22 (2): 151–167. doi:10.1080/713695728.

Chen, L., S. Howitt, D. Higgins, and S. Murray. 2022. "Students' Use of Evaluative Judgement in an Online Peer Learning Community." *Assessment & Evaluation in Higher Education* 47 (4): 493–506. doi:10.1080/02602938.2021.1933378.

Chong, S. W. 2021. "University Students' Perceptions towards Using Exemplars Dialogically to Develop Evaluative Judgement: The Case of a High-Stakes Language Test." *Asian-Pacific Journal of Second and Foreign Language Education* 6 (1): 12. doi:10.1186/s40862-021-00115-4.

Dawson, P. 2020. "Cognitive Offloading and Assessment." In *Re-Imagining University Assessment in a Digital World*, edited by Margaret Bearman, Phillip Dawson, Rola Ajjawi, Joanna Tai, and David Boud, 37–48. Cham: springer.

De Mello Heredia, J., M. Henderson, and M. Phillips. 2023. "Using Video Feedback to Support Students." Evaluative Judgement Society for Information Technology & Teacher Education International Conference 2023, New Orleans, LA. https://www.learntechlib.org/p/222090

Escalante, J., A. Pack, and A. Barrett. 2023. "AI-Generated Feedback on Writing: Insights into Efficacy and ENL Student Preference." *International Journal of Educational Technology in Higher Education* 20 (1): 57. doi:10.1186/s41239-023-00425-2.

Freeman, J. 2024. *Provide or Punish? Students' Views on Generative AI in Higher Education*. Oxford: Higher Education Policy Institute. https://www.hepi.ac.uk/2024/02/01/provide-or-punish-students-views-on-generative-ai-in-higher-education/

Guo, Y., and D. Lee. 2023. "Leveraging ChatGPT for Enhancing Critical Thinking Skills." *Journal of Chemical Education* 100 (12): 4876–4883. doi:10.1021/acs.jchemed.3c00505.

Gyamfi, G., B. E. Hanna, and H. Khosravi. 2022. "The Effects of Rubrics on Evaluative Judgement: A Randomised Controlled Experiment." *Assessment & Evaluation in Higher Education* 47 (1): 126–143. doi:10.1080/02602938.2021.1887081.

Kuhn, D., R. Cheney, and M. Weinstock. 2000. "The Development of Epistemological Understanding." *Cognitive Development* 15 (3): 309–328. doi:10.1016/S0885-2014(00)00030-7.

Lipnevich, A. A., and J. K. Smith. 2009a. "Effects of Differential Feedback on Students' Examination Performance." *Journal of Experimental Psychology. Applied* 15 (4): 319–333. doi:10.1037/a0017841.

Lipnevich, A. A., and J. K. Smith. 2009b. "I Really Need Feedback to Learn:" Students' Perspectives on the Effectiveness of the Differential Feedback Messages." *Educational Assessment, Evaluation and Accountability* 21 (4): 347–367. doi:10.1007/s11092-009-9082-2.

Luo, J., and C. K. Chan. 2023. "Conceptualising Evaluative Judgement in the Context of Holistic Competency Development: Results of a Delphi Study." *Assessment & Evaluation in Higher Education* 48 (4): 513–528. doi:10.1080/02602938.2022.2088690.

Malecka, B., and D. Boud. 2023. "Fostering Student Motivation and Engagement with Feedback through Ipsative Processes." *Teaching in Higher Education* 28 (7): 1761–1776. doi:10.1080/13562517.2021.1928061.

Markauskaite, L., R. Marrone, O. Poquet, S. Knight, R. Martinez-Maldonado, S. Howard, J. Tondeur, et al. 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. doi:10.1016/j.caeai.2022.100056.

McIver, S., and B. Murphy. 2023. "Self-Assessment and What Happens over Time: Student and Staff Perspectives, Expectations and Outcomes." *Active Learning in Higher Education* 24 (2): 207–219. doi:10.1177/14697874211054755.

Molloy, E., and M. Bearman. 2019. "Embracing the Tension between Vulnerability and Credibility: 'Intellectual Candour' in Health Professions Education." *Medical Education* 53 (1): 32–41. doi:10.1111/medu.13649.

Nam, J. 2023, November 22). "56% of College Students Have Used AI on Assignments or Exams." *BestColleges.com*. https://www.bestcolleges.com/research/most-college-students-have-used-ai-survey/

Nicol, D. 2021. "The Power of Internal Feedback: Exploiting Natural Comparison Processes." *Assessment & Evaluation in Higher Education* 46 (5): 756–778. doi:10.1080/02602938.2020.1823314.

Prather, J., P. Denny, J. Leinonen, B. A. Becker, I. Albluwi, M. Craig, H. Keuning, N. Kiesler, T. Kohn, and A. Luxton-Reilly. 2023. "The Robots Are Here: Navigating the Generative AI Revolution in Computing Education." In Proceedings of the 2023 Working Group Reports on Innovation and Technology in Computer Science Education, 108–159. Turku, Finland. doi:10.1145/3623762.3633499.

Sadler, D. R. 1989. "Formative Assessment and the Design of Instructional Systems." *Instructional Science* 18 (2): 119–144. doi:10.1007/BF00117714.

Siiman, L. A., M. Rannastu-Avalos, J. Pöysä-Tarhonen, P. Häkkinen, and M. Pedaste. 2023. "Opportunities and Challenges for AI-Assisted Qualitative Data Analysis: An Example from Collaborative Problem-Solving Discourse Data." International Conference on Innovative Technologies and Learning, Porto, Portugal.

Tai, J., R. Ajjawi, D. Boud, P. Dawson, and E. Panadero. 2018. "Developing Evaluative Judgement: Enabling Students to Make Decisions about the Quality of Work." *Higher Education* 76 (3): 467–481. doi:10.1007/s10734-017-0220-3.

Telio, S., G. Regehr, and R. Ajjawi. 2016. "Feedback and the Educational Alliance: Examining Credibility Judgements and Their Consequences." *Medical Education* 50 (9): 933–942. doi:10.1111/medu.13063.

Watling, C., E. Driessen, C. P. van der Vleuten, and L. Lingard. 2012. "Learning from Clinical Work: The Roles of Learning Cues and Credibility Judgements." *Medical Education* 46 (2): 192–200. doi:10.1111/j.1365-2923.2011.04126.x.

Welding, L. 2023, March 23. "Half of College Students Say Using AI on Schoolwork Is Cheating or Plagiarism." *BestColleges.com*.

Ziebell, N., and J. Skeat. 2023. "How Is Generative AI Being Used by University Students and Academics?." Semester 1, 2023.