Science & Education
https://doi.org/10.1007/s11191-024-00517-z

SI: EPISTEMIC INSIGHT & ARTIFICIAL INTELLIGENCE

The Practical Epistemologies of Design and Artificial Intelligence

William Billingsley[1]

Accepted: 27 March 2024

Abstract
This article explores the epistemological trade-offs that practical and technology design fields make by exploring past philosophical discussions of design, practitioner research, and pragmatism. It argues that as technologists apply Artificial Intelligence (AI) and machine learning (ML) to more domains, the technology brings this same set of epistemological trade-offs with it. The basis of the technology becomes the basis of what it finds. There are correlations between questions that designers face in sampling and gathering data that is rich with context, and those that large-scale machine learning faces in how it approaches the rich context and subjectivity within its training data. AI, however, processes enormous amounts of data and produces models that can be explored. This makes its form of pragmatic inquiry that is amenable to optimisation. Finally, the paper explores implications for education that stem from how we apply AI to pedagogy and explanation, suggesting that the availability of AI-generated explanations and materials may also push pedagogy in directions of pragmatism: the evidence that explanations are effective may precede explorations of why they should be.

1 Reflective Practice and Technology Design

In some fields, "the aim is not to find the truth, but to improve some characteristics of the world in which people live". This characterisation comes from Rittel and Webber's (1973) formulation of wicked problems: the muddy, intertwined, hard-to-define problems that planners and designers face in their practice. Hypotheses become hard to define because every design situation involves interpretation of what the underlying problem is. Solutions become hard to generalise because every situation is imbued with unique context. Experiments become hard to conduct because every implemented solution has consequences: the world after your design has been implemented is, perhaps subtly or perhaps dramatically, different than it was before. The designer, however, "has no right to be wrong", because poorly implemented designs can cause real and lasting harm. This is contrasted with the "tame" problems that science has traditionally tackled, in which problems are definable and separable, hypotheses are falsifiable, and experiments are permitted to fail.

✉ William Billingsley
wbilling@une.edu.au

[1] University of New England, Armidale, Australia

Published online: 18 April 2024

The cost of failure is a common and sometimes wry lament in many practical fields. "When mankind lands on Mars, it will be heralded as a triumph of science. If something goes wrong on the way, it will be blamed on engineering failure" (Billingsley, 2015). When designing for social contexts, not only is the cost of failure present, it may also be infeasible for the designer to model the trajectory of their design before it launches. Designers describe this as the "envisioned world problem" (Hoffman et al., 2010). The designer can only conduct their investigations in the world that exists now, but their design will operate in an envisioned world that does not exist yet: the one in which it has been implemented and may be being used in unexpected ways.

Usually, this results in unexpected user experience problems, but it can also result in uses for a technology that the designer did not originally envisage, with its own consequences. Interviewed in 2015, Evan Williams, co-founder of Twitter (now X) described that "there was this path of discovery … where over time you figure out what it is" (Lapowski, 2013/2019) . First envisaged as a tool for people to share short status updates with a small group of friends, Twitter evolved to have much broader influence on society from global journalism (McGregor & Molyneux, 2020) and politics (Parmelee & Bichard, 2011) to the social dangers faced by children and youth (Bozzola et al., 2022).

Fundamentally, designers are faced with a tension: the situations and participants they study are rich with unique context that limit how safe it is to generalise their findings or predict the impact of any change, but their practice requires them to do so as the products and technology they create are intended to be used more broadly and express a design goal. Design fields have faced this tension down the ages and developed their own ways of knowing and ways of learning (Cross, 1982), which are typically based on reflective practice. In doing so, they gather principles from experience and seek ways to judiciously abstract and generalise them, but without the claims of abstract truth that might be seen in the sciences. Observations stem not from clean experimentation, but from reflecting on situated practice with all its context and complexity, so are more tentatively applicable when seeking to bring them to another context. Schön (1983, 1984, 1987), discussing design pedagogies and how students learn their practice within studios, advocates reflective practice for its ability to leave the high ground of theoretically solvable problems, and enter the "swampy lowlands, where situations are confusing messes … and usually involve problems of greatest human concern" (1983, p42). James (1907), writing much earlier on the philosophy of pragmatism, makes a similar contrast with theoreticians: "The more absolutistic philosophers dwell on so high a level of abstraction that they never even try to come down".

Technology design grapples with these issues in a heightened form. Computer programs are considered more complex for their size than any other human construct (Brooks, 1986). They are also a remarkable mixture of abstract theory and human context. A program's internal workings are described in an abstract manner (a programming language) so precisely that it can be executed by a machine, but most computer programs are not created for their own abstract beauty, but to achieve some real-world goal. So, alongside their engineering complexity, they also carry the rich context and ambiguities of design. Influential AI researcher Herbert A. Simon (1969) described the need for "sciences of the artificial" to explore this, including a chapter proposing a science of design that was later revised and republished separately (Simon, 1988). His work led to some debate as to what attempting to make design more scientific offered to the design community (Schön, 1983, pp. 47–49; Dorst & Dijkhuis, 1995; Meng, 2009). In my own view, the tension between Simon's and Schön's arguments reflects this dual nature of technology design: that it is so precisely described but founded in imprecise human context. This dual nature extends into how practitioners are taught within universities. As well as learning about programming language theory and engineering techniques (computer *science*), students are also often taught their practice using studio methodologies inspired by design studios (Tomayko, 1991; Docherty et al., 2001; Hundhausen et al., 2008).

Modern software is commonly developed using agile methodologies, which make use of a cycle of iterative cycles of development and evaluation. SCRUM, one of the more popular methodologies, describes this aspect in terms of industrial process control (Schwaber, 1997), but action researchers (Lewin, 1946) might also recognise it as a cycle of simultaneous action and reflection, strategists might recognise its similarity to Boyd's (1976) Observe-Orient-Decide-Act (OODA) loop, and designers might recognise agile development's similarity to design thinking cycles (Lindberg et al., 2011).

In technology design, the question of epistemology can seem esoteric. Academic papers submitted to design and technology venues do not typically declare their philosophical stance or ground their methodology in it. (It can be a barrier occasionally and unpleasantly encountered when navigating the criticisms of journal reviewers who come from other fields.) Questions of how to establish design approaches as valid academic methods of inquiry do routinely arise (Zimmerman et al., 2007; Herriott, 2019), along with the question of how to apply rigour to design experiments and investigations (e.g. Zimmerman et al., 2010; Mettler et al., 2014; Frauenberger et al., 2015). Design science (March & Smith, 1995) considered the distinction between research activities and research outputs within these action and reflection loops and Hevner et al. (2004) attempted to describe criteria for evaluating design science research.

However, running through the literature, there is a recognition that the discussion of the epistemology of design is post hoc: reflective practice is ancient, inevitable, and ongoing. James (1907) describes pragmatism as "nothing essentially new", and later simply that "the pragmatist clings to facts and concreteness, observes truth at its work in particular cases, and generalises". Archer (1995), considering the question of design research rigour largely through the lens of action research, concludes with a remark about the artificiality of the question. "All this has been well understood for a long time. What is new, perhaps, today is the introduction of a new quality to be sought for in research: the elusive quality of Research Assessment Exercise-worthiness".

2 Large-Scale AI Is also Pragmatic

In the last 20 years, AI has been applied to many fields of inquiry. In 2005, a report for the United States government (PITAC, 2005) coined the idea of computational science as a "third pillar" of science alongside theory and experimentation, though much of that characterisation centered on simulation and high-performance computing (e.g. Dongarra et al., 2007; Skuse, 2019). Recently, we have witnessed the rise of machine learning based on very large deep learning models: multi-layered neural networks that are developed through unsupervised learning (Bengio et al., 2021). For instance, the natural language model GPT-3 (Brown et al., 2020) contains 175 billion parameters that were trained on 300 billion tokens (approximately, words) of data. This produced a pre-trained model that could then be adapted to new tasks through "few shot" learning: that is, it only requires a few examples of a new task to train the model to be able to perform it. However, specialised models can also be created by more extensive "fine tuning" using large quantities of data from other domains.

Within computer science itself, Codex (Chen et al., 2021) is a language model originally derived from GPT-3 that was fine-tuned by training it with 50 million examples of computer programs from open-source software projects on GitHub. Despite having no formally designed internal model of how a computer works, this trained language model performed well enough at code generation to pass first-year university computer science exams (Finnie-Ansley et al., 2022). Whereas we might instruct students how to use algorithmic thinking in writing a program (Knuth, 1985/2018), a generative language model reflects inductive reasoning: given the prompt of the question, what is the most likely text of a successful answer. The Codex model was made available as a commercial programming assistant under the brand name GitHub Copilot; more recently, this has also begun incorporating GPT-4. Although programming has, perhaps inevitably, been one of the first application areas, researchers are exploring other kinds of data it can generate. Language models have, for instance, proven adept at understanding and generating chemical molecular structures (Flam-Shepherd et al., 2022) and researchers are working on Codex-like tools for chemistry (Ahmad et al., 2022).

The use of large learning models changes the kind of reasoning we are effectively employing in part of our work. When a simulation or traditional computational science technique is applied to a scientific task, it can be argued that the machine is simply a tool of the researcher, enabling them to capture their theory as a model and examine its implications (Diallo et al., 2013). However, a model containing billions of parameters, learned through unsupervised learning on training data, brings in questions about the samples that the model was trained on and their generalisability. For example, deep learning models reflect biases from their training sets (Dale, 2021), and investigations into the reasoning abilities of GPT-3 on vignettes can be confounded by whether it may have encountered the text of similar vignettes in its training set (Binz & Schulz, 2022). For very large models, the training data is typically public data. For example, Codex's use of open-source software repositories published on GitHub is essentially an exceptionally large convenience sample.

These questions about sampling, context, and generalisability are similar questions that design research has faced over the years. Learning from experience and therefore the use of available samples from experience has its basis in pragmatism. In the biases that we see in how AI performs, it is evident that the training data is not neutral and is filled with context that might or might not generalise to new situations. Its use, therefore, represents a post-positivist view that although no sample is objectively generalisable, nonetheless useful insights can be gleaned from it if they are treated judiciously. The use of large convenience samples of public data is a surprisingly close match to a catchphrase of advice given to user experience design practitioners: "recruit loosely but grade on a curve" (Krug, 2006).

Moreover, as a technology artifact, very large deep learning models are inevitably developed using the iterative practices of technology design. It stands to reason, then, that if a model is the product of a pragmatist process, its output will stand on the same philosophical feet.

The way technology is iterated and refined based on how it practically performs can be seen in the literature on its development. For example, the paper "Attention is all you need" (Vaswani et al., 2017/2023) introduced the Transformer architecture that underpins many modern large language models. This replaced recurrence and convolutions (elements of previous machine learning model architectures) with layers using a new technique the authors called "self-attention", because the new technique "allows for significantly more parallelisation and can reach a new state of the art in translation quality after being trained for as little as 12 h on eight P100 GPUs". Large language models have continued to be iterated and developed by testing their practical performance. GPT-4's development is also described in an arXiv preprint (OpenAI: Achiam et al., 2023). Like many large language models, it was pre-trained on a very large corpus of text to be able to predict the next word in a document. However, it was also fine-tuned using reinforcement learning from human feedback (Cristiano et al., 2017), which is an iterative mechanism whereby the model adjusts some of its parameters from human feedback on the perceived quality of its output. Mitigations for some of its harmful responses were also developed iteratively, using a panel of fifty domain experts to test the model by attempting to produce harmful responses, and improving the model in response. Before its release, OpenAI tested GPT-4's performance on a suite of human tasks, including examinations and programming challenges, and compared its performance with GPT3.5.

Once it has been created, technology has a capacity to be unreasonably effective. "Unreasonably effective" is a phrase that was originally coined in mathematics, remarking on the seemingly surprising fact that so many distinct aspects of physics can be modelled using similar simple equations and terms (Wigner, 1960). Why, for example, should circles and the density function of the Gaussian distribution of populations have the value pi in common. The phrase has been taken up in technology, noting the broad applicability of machine learning techniques when given a sufficient amount of data (Halevy et al., 2009; Sun et al., 2017). In this article, however, I wish to use the phrase "unreasonably effective" in a literal sense that there is a triumph of the pragmatic. That is, technology can prove effective at a task before its users understand why it should be.

Among research users of technology, literature can also be found recommending the adoption of technology based on its practical effectiveness. For example, Lemon and Hayes (2020) recommend the use of Leximancer (a well-developed text analysis tool that uses co-occurrence to extract concept maps from text) for triangulation in qualitative research because they found the tool to be effective in practice in helping the researcher refine and add nuance to their conceptual models.

At the time of writing, large language models (such as GPT-4) are proving effective at many different tasks. For academics researching the details, perfectly reasonable explanations for why this is the case emerge, but for most users of the technology the evidence of its effectiveness, rather than a reasoned argument for its soundness, comes first. An epistemology a user is implicitly selecting when they apply a large pre-trained machine learning model, then, is pragmatism.

3 Epistemology in Design Research

Research students are often taught that there are links between metaphysical paradigms and research methods that can be more affiliated with them (Williamson, 2018; Cecez-Kecmanovic & Kennan, 2018; Gallifa, 2018). For example, positivism takes the view that there is an objective reality independent of the observer and that the empiricist's role is to uncover positive truths about this reality. Comte is seen as the founder of modern positivism, although Comte himself claimed "no originality for this conception of knowledge" but that it stems from the scientific practices of Bacon, Descartes, and Galileo (Mill, 1865). Popper (1962) refined this with the notion that a hypothesis typically cannot be positively verified but only supported through attempted falsifications. This view of metaphysics lends itself to experimentation and questions of the reproducibility and external validity of an experiment, as findings are expected to be truths that can be applied outside the experiment's confines. Interpretive paradigms take a more relativist view: that our understanding of social reality is constructed and understood through subjective experience. This inevitably leads to a different view of validity, as it implies a lesser expectation that findings from one investigation will be easily or directly applicable to another. Ethnographies and qualitative methods are often more closely affiliated with this philosophical stance. Post-positivists take the view that there is an objective reality but that observations of it cannot be fully reliable, so the truths and rules that we derive from empirical research are more like instruments. To use Cartwright's (1999) famous example, stating that "aspirin cures headaches" speaks to a capacity that aspirin has in some circumstances, rather than a universal law. From a post-positivist perspective, empirical research is always in a middle ground where the generalisability of its findings is qualified.

Within pragmatism, this tension between subjectivity and objectivity could be argued to flow from philosophical discussions of the mind-body problem: whether we can explain physical things from mental things and what physical things can tell us about experience. James (1907) and Dewey (1908; Leonov, 2018) refer to this question in their work. James particularly was motivated by observing that we never experience either mind or reality in isolation, but only one through the other.

Design has to consider this tension between subjectivity and objectivity because of its nature. It is deeply interested in how its designs will be used by people in context, and therefore in their subjective experience. It is also aware that design situations are rich with context, with many more variables than it would be possible to describe or model (Cash et al., 2022). However, it cannot solely be interpretivist because designs have design goals. Woods (1998) describes designers as experimenters whose "products and prototypes embody hypotheses about what will be useful". In a sense, like James and Dewey, design is less interested in the subjective or the objective in isolation than it is in the combination of the two.

Although it has not always couched the question in philosophical terms, design fields have long grappled with the question of when the cognitive work performed by designers counts as research and when it does not. In 1836, the artist John Constable gave a lecture to the Royal Institution, in which he asked "Why … may not landscape be considered a branch of natural philosophy, of which pictures are but experiments?" A century and a half later, Sir Christopher Frayling (1993) considered this question and this quotation in a seminal paper considering what research in art and design should mean. He introduced a distinction between research *into* design (researching the practice of other artists), research *for* design (the "small r" research that a designer undertakes to design their project), and research *through* design (new insights and knowledge that is discovered through design work and applicable beyond the artifact being designed). Frayling considered research for design to be the "thorny" one of the three, where the thinking is entirely embodied in the artwork and the goal is the art rather than the knowledge or understanding. Frayling also recognised a secondary motivation for not considering this last category research proper: "we feel that we don't want to be in a position where the entire history of art is eligible for a postgraduate research degree". This also expresses a view that Frayling attributes to Picasso, that research must find something.

Frayling's middle category, Research through Design (RtD), has been taken up as a methodology in research, especially in technology design (Zimmerman et al., 2010). The attractiveness of this to technologists can easily be seen from designer Bruce Archer's (1995) exploration of design research and its similarity to action research: "There are circumstances where the best or only way to shed light on a proposition, a principle, a material, a process or a function is to attempt to construct something, or to enact something, calculated to explore, embody or test it".

The growth of this approach has naturally led to questions over where it sits in comparison to scientific research. Herriott (2019) suggests that the action of designing artifacts and observing outcomes is little different from experimentation and that there is only a weak claim to there being special designerly ways of knowing. (I note, however, this bears similarity to William James's comments that pragmatism is nothing new.) Galdon and Hall

(2019, 2021), following observations by Jones (1992, p10), consider the distinction as one of timing: science concerns itself with the present and investigations into what is, while design speaks to the future and is primarily forced to treat as real imagined things that have not been implemented yet. They also propose that design trades off accuracy for “access to areas that are partial and yet-to-be and not-fully formed. Therefore, our output is probabilistic, and research is always preliminary in its nature” (2021). Goldkuhl (2011) draws a similar conclusion about timing—identifying pragmatism as the key epistemology of Research through Design and characterising it through a phrase from Dewey (1931): “An empiricism which is content with repeating facts already past has no place for possibility and liberty”.

The theme that prototypes embody designers’ hypotheses for the future and how it will be used recurs within the literature. As Flores et al. (1988) express it, “technology is not the design of physical things; it is the design of practices and possibilities to be realised through artifacts” .

In education, design has been used as a research method for many years, often as “design experiments” or “design-based research”. As is appropriate, it has often advanced through critical reflection on its uses and shortcomings. “Design experiments” as a term is sometimes credited to Allan Collins (1990), who sought to propose a more systematic methodology for technology innovation in education as “major problems with current design experiments prevent our gaining much information from them”. His proposed improvements included inter-disciplinary teams, including teachers as co-investigators, and a greater focus on design revision and understanding the failures of designs rather than only seeking positive results. Brown (1992) noted a “trade-off between experimental control and richness and reality” when moving between laboratory studies and situated studies in classrooms. Brown wrote from an education and psychology perspective but used a technology analogy for the process of taking an education intervention from development into widespread use, likening it to the “alpha, beta, and gamma phases of software development”. Cobb et al. (2003) described this kind of research as “design-based” in the conclusion of their article describing how to carry out design experiments, and the term “design-based research” appears to have been adopted shortly after (e.g. Barab & Squire, 2004). Design-based research’s position in relation to other methodologies has often been reflected upon. Cobb et al. (2003) saw design experiments as a pragmatic, iterative, and reflective way of generating theory, but “relatively humble” theory about why a design works and how it can be adapted to new contexts, rather than grander but more abstract theories of learning. More recently, Hoadley and Campos (2022) described it as its own tradition, contrasted with qualitative and experimental research. They note that in design-based research, hypotheses take the form of “design conjectures”, as it “attempts to understand the world by seeking to change it”, using iteration and reflection to explore topics where interventions cannot simply be cleanly specified based on theory (in the manner of a scientific experiment), but must be implemented in complex social settings where participants have their own agency.

4 Samples and Context

Questions of research tend to come back to what is found, and questions of rigour tend to come back to questions of reliability and understanding the limits of the findings. As we have seen, design research often finds itself working with one set of participants, situated in all their rich context, as it proposes and tests design conjectures, but then having to make preliminary or probabilistic inferences about what might be useful in future contexts, to future users, or to other similar use cases. When facing questions about reliability or rigour, design research inevitably finds itself dealing with pragmatist and post-positive considerations of what subjective observations drawn from one context can tell us about another.

With the present focus on AI, I would argue that large-scale machine learning encounters a corresponding situation. Each document or item of data the machine learning model was trained with came from a particular context, but the purpose of training the model is so that it can then be applied to different documents, tasks, and contexts. The kinds of models that designers and AI produce are, of course, very different and they consider this probabilistic inference in different ways. (For example, in AI development the error of a model's predictions can be measured and quantified.) However, in both cases, there is a need to draw inferences from one context to apply what is learned to another. This leads each field, in different ways, to adopt a sense of pragmatism and iteration in how samples are chosen as they seek to make this process effective.

In design, a question researchers and practitioners face (for example when conducting usability tests) is how many participants to test with and whom. Cash et al. (2022) consider this question in research, noting that in a design, rich with context, there are typically far more variables than it is possible to enumerate. Citing Lynch (1999) and Wacker (2008), they argue that therefore samples are typically only defined in terms of the factors already known to affect the concepts being studied, what they call the "previous literature convention". In other words, early in the research process, it might be impossible to know what factors a representative sample would need to be representative of. As more is discovered about the problem, more is discovered about what factors are important to consider in sampling. This creates an effective cycle (Cash, 2018) as the sampling that a designer will engage in changes as their theory development moves from discovery to description.

From a practitioner perspective, similar effects can be seen. Designers of new products may make more use of convenience samples as they engage in need-finding and grapple with coarse-level usability issues, whereas refinement of existing products has more capacity to consider sampling variables as there is greater knowledge of the problem being addressed and its stakeholders. Practitioners in interaction design are also taught that the point to stop recruiting new usability test participants is when they feel they have reached "saturation": that is when they are gleaning few additional insights from new participants. In any case, contexts are assumed to be rich (so only probabilistically generalisable—findings must be drawn judiciously), and there is assumed to be a sliding scale of the relevance of a sample to another context. As practitioner Krug (2006) expresses, it "recruit loosely, but grade on a curve".

Design is also unusual in its generation of personas. As a technology has more users than a designer can fully conceptualise and each user has more characteristics than could be enumerated, Cooper (1999, p. 123) proposed the notion of personas: a smaller number of simplified distillations of users and their needs that a designer can use in thought experiments. Each persona becomes an imaginary character, and scenarios can be written about how that character might interact with the technology. These are intended to be grounded in valid research about users and their utility depends on their credibility (Pruitt & Adlin, 2006). However, in practice, the content of personas is not only generated from user research: they may also come from designers' ideas to fill in gaps and possibilities (Chang et al., 2008). What I find interesting here is that it is an example where human design practice has attempted to mitigate problems of context and sampling by artificially generating additional data.

The sliding scale of context, data augmentation, and saturation all have apparent correlations in machine learning. Data in a large model such as GPT-4 is assumed to be highly dimensional: recall that it contains billions of parameters, and it is initially unknown what those parameters will end up corresponding to. Internally, machine learning models use "sparse" representations, pruning connections within the neural network where the connection weights are close to zero. A machine learning model is a highly engineered machine for learning context and what the sliding scale of relevance of one aspect to another should be.

Augmentation and refinement of training data, rather than the traditional empirical approach of seeking a representative sample, are also commonplace. The creators of GPT-3 (Brown et al., 2020) took the CommonCrawl dataset of nearly a trillion words, filtered it for similarity to known high-quality corpora, used fuzzy de-duplication of documents within the dataset, and then supplemented it with other known high-quality training corpora. As described earlier, GPT-4 (OpenAI: Achiam et al., 2023) was fine-tuned after its initial training in order to improve the quality of its responses. The practice of augmenting training data to mitigate biases and known flaws is not limited to text-based machine learning. For example, working with images, Taylor and Nitcshke (2018) report on rotating, cropping, and flipping images within a training dataset to create new examples, so that a neural network will learn to be rotation and position invariant in how it detects objects. Questions of saturation and the necessary size of a training dataset are also common as developers of models consider the accuracy of models against their training time, number of parameters, and size of the dataset, to optimise cost and accuracy trade-offs in the creation of the model.

In each case, the approach to sampling and context is pragmatic and iterative. As insights from one context will only probabilistically transfer to another, they attempt to refine and optimise that inference process as more is discovered about the problem over time.

Design and AI also share the characteristic that they do not usually seek to generalise a single finding but to model the problem as a whole. A trained AI model represents a complex understanding of how each (digested) aspect of its training data relates to each other. The learned model is then applied as a whole when the AI is asked to perform a task. I would argue that this is also true of design: it is comparatively rare for a designer to take just a single conclusion from one study and apply it to another context, in the manner of a scientific finding. Rather, they take the increased understanding of the context and outcomes of each situation that they have gathered in their practice and use that multi-factored understanding when proposing designs.

Compared to human-conducted pragmatic inquiry, however, AI can process much larger quantities of data. The models it produces can also be explored, for example by researchers investigating its biases or the best ways to augment training data to mitigate them. This makes AI-conducted pragmatic inquiry amenable to optimisation, as well as opening an endless array of new possibilities for how each generated model can be used.

5 Implications for Education

AI has had many touch points with education over many decades. Doroudi (2022) describes AI and the learning sciences as having been intertwined since the 1960s, as many of the early AI pioneers were cognitive scientists who were interested in human as well as artificial intelligence. Although, he also describes how those fields have grown further apart in recent years, as AI in education communities gradually moved from "trying to bridge between human and machine learning" to "applying state-of-the-art AI (mostly machine learning) in service of education".

On a practical level, the recent growth in the accessibility and use of AI has significant, sweeping, and open-ended implications for how we should conduct education. Markauskaite et al. (2022) consider the problem that the skills and capabilities that students will need in their careers are likely to be altered—for instance requiring them to be able to understand, interpret, and adapt AI output as it is used in human organisations. The release of ChatGPT and its uptake by students and teachers prompted many academics to rapidly explore the perceived risks and benefits of generative AI to education. For example, Adeshola and Adepoju (2023) conducted topic and sentiment analysis of social media posts to analyse people's views, while Baidoo-Anu and Ansah (2023) based part of their study on asking ChatGPT itself to discuss the issue, as well as exploring academic, news, and social media articles. Uses in adaptive learning, language training, and speeding up the production of content and provision of feedback have been seen as benefits. Biases, inaccuracies, and privacy issues have been identified as risks. In an opinion piece, Sharples (2023) considered generative AI's current limitations—for example that students' conversations with it tend to be short and isolated from each other—and considered new educational roles it could take on if it evolves to become more social. There has also been rapid research on the risks to academic integrity of students misusing generative AI. Hsiao et al. (2023) conducted workshops with academics, to identify ways of redesigning written assignments to be more robust against the misuse of generative AI—a process that also helped some participants recognise where their existing assessment designs "prioritised knowledge telling over promoting knowledge transformation". Gorichanaz (2023) examined social media posts on Reddit to understand the student perspective of being accused of using generative AI to cheat, including the experience of false accusations. His paper also saw assessment redesign as possibly being a more effective approach to handling generative AI than relying on detection.

In this article, however, I wish to focus more narrowly on the direct application of AI in pedagogy. In that setting, the choice of AI model ties in to questions of how it should know what it knows and how students should think about the topic. Particularly, an AI can reflect certain or uncertain reasoning in how students' thinking is modelled and in how explanations and feedback are given.

From as early as researchers were able to build AI models, they started using them in education. Early applications of AI in education often used "classical" AI models that work deterministically, rather than machine learning models. The design of these models considered both the problem and the pedagogy. For example, SOPHIE (Brown et al., 1975) was a reactive learning environment that modelled students' debugging process with electronic circuits—modelling the measurements that students had (virtually) made and the conclusions they could therefore draw from this, to engage in a dialogue leading the student through a problem-solving experience. John Seely Brown pursued these ideas with a view to "cognitive apprenticeships" (Collins et al., 1989), where AI could teach students ways of thinking needed to solve complex problems. Intelligent Tutoring Systems (ITSs) arose as an area of research within AI in education and tended to involve models of the domain alongside models estimating what concepts and processes students do and do not understand, to best know how to advise them. For example, Model Tracing (e.g. Anderson et al., 1995; VanLehn et al., 2005) contained detailed rules about how an ideal student would progress through a problem and estimated students' progress in understanding of this by observing their behaviour. Even where there is no student model (for example, because it is difficult for such a student model to account for learning that takes place away from the computer), the classical AI model of the problem would imply something about how the student should think about the problem. In my own designs, for instance, I observed that AI-generated explanations tended to be large and unwieldy for students to navigate, so I advocated keeping an AI model of the problem alongside a diagram (or other visual model) and pruning the AI explanation to show it on the diagram. The elements shown in the diagram, then, represented the way the teacher wished the student to conceptualise the problem. This is convenient and enables students to explore problems interactively, and there can be a Pygmalion effect where the fact that the interface talks about the problem in a particular way invisibly leads the student to adopt the same conception—Slator et al. (1986) described this effect in text interfaces, but it seems reasonable also to apply it to graphical ones.

However, those classical AI models are deterministic and designed. They convey a defined sense of how the question should be thought about. Where large-scale machine learning models are used, however, we have seen that their insights are probabilistic and their exact internal conceptualisation of the problem is not entirely known. Explainable AI is a highly active area of research (e.g., Hoffman et al., 2018; Goebel et al., 2018), but the 175 billion parameters of GPT-3 (or the reported trillion parameters of GPT-4) are too many for a person to grasp even if we could investigate, measure, and label them all. This puts their use in pedagogy in a different space, where we can investigate and describe what is most important to the model, but not fully describe it in a humanly graspable way.

This is nonetheless useful. For example, Latif et al. (2021) generated an interactive explorable explanation of how historical figures are connected using GPT-3 generated text to augment the visualisations. We can also use AI models generatively to inspect what the AI considered important. For example, if we ask a generative art AI to produce many images in the style of a famous artist, we can explore what it thinks their style is. Or we can also produce data science style visualisations from the model's weights. A common way to explore image recognition models is to produce images highlighted with heatmaps of the where the AI's attention focused, for example when classifying whether the image contained a toy poodle (e.g. Leventi-Peetz & Östreich, 2022).

Direct applications of large language models to AI in pedagogy have proven useful in recent practice. For example, Codex can generate text explanations of code that can help students understand programs (MacNeil et al., 2022 August). It has also been useful for automatic generation of teaching materials and questions (Sarsa et al., 2022; MacNeil et al., 2022, December). What is less clear is what the encoding of the pedagogy within the model is. Explanations and content are generated based on the corpus of examples that the AI has encountered, so are representative of the techniques and concepts that human experts use because that is what it learned from. They are clearly effective (or the teacher would not deploy them), but the decisions for why this is the appropriate explanation are left to the opaque box of the model.

As we expand the scope of what AI can be left to model and generate, this may represent a step towards a radical form of pragmatism in pedagogy: where explanations and examples are chosen because they are unreasonably effective rather than from an a priori known (to the teacher) model of why they should be. GPT-4 was fine-tuned using reinforcement learning techniques, based on human feedback of the usefulness of its responses. At the time of writing, educational applications that use generative AI have not had much time to engage in that kind of iterative refinement process. However, it is reasonable to expect that in the future they will: that educational AI will generate feedback and explanations for students and learn from whether or not its responses were effective. In that situation, the amount of data that the AI would observe over the effectiveness of its feedback is likely to give it an internal model of how to produce good feedback that is successful in practice but difficult to elicit or explore.

At this point in the discussion, it may be useful to distinguish between the way an educational system thinks about the topic and the way it thinks about the student. Recently, intelligent tutoring systems researchers have explored using deep learning for “knowledge tracing” – the process of estimating students’ skill level in the different topics in a domain (Piech et al., 2015). These produce models of student learning that are useful and adaptable, but non-transparent and require explainable AI (xAI) techniques to explore (Lu et al., 2023). It may be that the next stage in evolution in tutoring technology is to forgo the notion of having an explicit and humanly explorable model of the student at all: instead to let reinforcement loops in generative AI learn how best to respond and advise students. Internally, such an AI would learn some representation of student understanding, but it is unlikely to be separable from its model of language, how to generate advice, and other parts of its network.

A potential advantage I see in this is that it would allow a clear distinction between how we ask students to understand a topic, versus how we seek to understand students. We could, for example, supply tasks to students that model problems in a classical and positivist manner—using the Pygmalion-like effect to lead them to see how we would like them to think about the problem. But we can let the AI learn pragmatically how best to help and advise them, from the advice it has seen in its training set (and large language model training sets include very large quantities of educational material) together with reinforcement loops interacting with students.

Despite the criticisms of biases in AI, there may also be equity advantages to this. A recent criticism of student modelling in education has been that it can lead to a “deficit” framing of students’ skills (Ocumpaugh et al., 2024). Where an AI keeps a defined internal model of how students should think about a problem and measures students’ understanding against it, students’ deviations from the model are seen as shortcomings to be fixed, which can lead to a constant emphasis on how students are lacking. A similar effect can be seen in learning analytics, which may focus on classifying which students are “at risk”. There is growing recognition in the learning analytics community that students do not always feel that they are the beneficiaries of the analytics that is performed on their data (Selwyn, 2019). As academics, we naturally wish to understand how our students are learning and why. The value of education to students, however, is in the learning rather than necessarily in being visibly analysed or diagnosed. From a design perspective, then, there may be practical advantages in being clear about what students should know, but pragmatic about what they do know.

6 Conclusion

Design and large-scale AI are fields that work with probabilistic knowledge, rich contexts, and attempt to draw out complex relationships between concepts rather than easily generalised findings. Discussions on their rigour and epistemology inevitably involve questions of sampling, bias, and context. They are also fields that are advanced through iterative modes of practice and reflection on what worked and what can be improved. These features lead them to take pragmatist and post-positive approaches to their subject matter.

AI, being computational, has the additional characteristics of being able to process large amounts of data and inherently modelling its weightings and network (at least internally) numerically. This makes it possible to use these models generatively and probe them for insights. For the teams and organisations that produce them, it is also possible for them to optimise the sampling, data augmentation, and model size in ways that are not practical for human-conducted research.

Past conceptualisations of pragmatism, as far as Dewey and James, have spoken of practitioners acting and reflecting upon experience. The notions of *how much* experience, *what* experience, however, have tended to be qualitative. For example, the user experience designer stops recruiting participants when they feel they have reached saturation. In large-scale machine learning, saturation and biases become explorable properties of the model. This suggests that the kinds of pragmatic inquiry that we see conducted using AI may be more amenable to optimisation than human-conducted pragmatic inquiry.

AI, design, and education have long been intertwined. Many of the early researchers in AI were also interested in human cognition and developed programs and artifacts that were designed to improve teaching. As education technology grew and the design of technology-based interventions in education became a topic of research, design experiments and design-based research methodologies also became intertwined with education research.

More recently, education technology has grown in scope, size and depth. Generative AI in particular has seen very rapid uptake among teachers and students. The questions of how this will impact education are as unknown as design questions always are: the applications of AI are an envisioned world. However, we can speculate that a sense of pragmatism and unreasonable effectiveness may grow in our teaching, as for users, the evidence that the output of the AI is effective comes before the exploration of why this is the case.

In terms of epistemology, I argue that as generative AI takes on some of the functions of student-modelling and feedback generation, we may see an increasing sense of duality and delineation in how educational technology models knowledge. The learning artifacts and interactive tools designed by academics represent how we would like students to think about a problem, so the epistemology they represent will depend on the topic being taught. The way the AI models student understanding, however, is likely to be pragmatic, as it improved by its designers using the pragmatic methods of technology development and learns from experience using the pragmatic methods of machine learning.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Declarations

Conflict of interest No funding was received to assist with the preparation of this manuscript. The author acknowledges a familial personal relationship with a reference cited in the paper and with an editor.

References

Adeshola, I., & Adepoju, A. P. (2023). The opportunities and challenges of ChatGPT in education. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2023.2253858.

Ahmad, W., Simon, E., Chithrananda, S., Grand, G., & Ramsundar, B. (2022). ChemBERTa-2: Towards chemical foundation models. arXiv:2209.01712.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, *4*(2), 167–207. https://doi.org/10.1207/s15327809jls0402_2.

Archer, B. (1995). The nature of research. Co-design, Interdisciplinary Journal of Design, January 1995, 6–13.

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative Artificial Intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI, 7*(1), 52–62. https://doi.org/10.61969/jai.1337500

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *The Journal of the Learning Sciences*, *13*(1), 1–14. https://doi.org/10.1207/s15327809jls1301_1.

Bengio, Y., LeCun, Y., & Hinton, G. (2021). Deep learning for AI. *Communications of the ACM*, *64*, 58–65. https://doi.org/10.1145/3448250.

Billingsley, J. (2015, March 16). An engineer's lament to today's controversial topics. Engineering Outside the Square Series, Engineers Australia, Toowoomba, Australia.

Binz, M., & Schulz, E. (2022). Using cognitive psychology to understand GPT-3. *arXiv Preprint*. arXiv:2206.14576.

Boyd, J. (1976). *Destruction and creation*. US Army Command and General Staff College.

Bozzola, E., Spina, G., Agostiniani, R., Barni, S., Russo, R., Scarpato, E., Di Mauro, A., Di Stefano, A. V., Caruso, C., Corsello, G., & Staiano, A. (2022). The use of social media in children and adolescents: Scoping review on the potential risks. *International Journal of Environmental Research and Public Health, 19*(16), 9960. https://doi.org/10.3390/ijerph19169960

Brooks, F. (1986, September). No silver bullet: Essence and accidents of software engineering. Technical Report 86–020. The University of North Carolina at Chapel Hill. https://www.cs.unc.edu/techreports/86-020.pdf.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of the Learning Sciences, 2*, 141–178. https://doi.org/10.1207/s15327809jls0202_2

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *arXiv Preprint*. arXiv:2005.14165.

Brown, J. S., Burton, R. R., & Bell, A. G. (1975). SOPHIE: A step toward creating a reactive learning environment. *International Journal of Man-Machine Studies*, *7*(5), 675–696. https://doi.org/10.1016/S0020-7373(75)80026-5.

Cartwright, N. (1999). *The dappled world: A study of the boundaries of science*. Cambridge University Press.

Cash, P. (2018). Developing theory-driven design research. Design Studies, 56, pp.84–119. https://doi.org/10.1016/j.destud.2018.03.002.

Cash, P., Isaksson, O., Maier, A., & Summers, J. (2022). Sampling in design research: Eight key considerations. *Design Studies*, *78*, 101077. https://doi.org/10.1016/j.destud.2021.101077.

Cecez-Kecmanovic, D., & Kennan, M. A. (2018). The methodological landscape – information systems and knowledge. In K. Williamson, & G. Johanson (Eds.), *Research methods: Information Systems and contexts* (2 ed., pp. 127–155). Elsevier.

Chang, Y. N., Lim, Y. K., & Stolterman, E. (2008, October). Personas: From theory to practices. In Proceedings of the 5th Nordic conference on human-computer interaction: Building bridges (pp. 439–442). https://doi.org/10.1145/1463160.1463214.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., ... Zaremba, W. (2021). Evaluating large language models trained on code. *arXiv Preprint*. arXiv:2107.03374.

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

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9–13. https://doi.org/10.3102/0013189X032001009

Collins, A. (1990, January). Towards a design science of education. Technical report No. 1. Center for Technology in Education, New York. ERIC Number: ED326179. https://eric.ed.gov/?id=ED326179.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453–491). Laurence Erlbaum Associates.

Cooper, A. (1999). *The inmates are running the asylum*. Sams.

Cross, N. (1982). Designerly ways of knowing. *Design Studies*, *3*(4), 221–227. https://doi.org/10.1016/0142-694X(82)90040-0.

Dale, R. (2021). GPT-3: What's it good for? *Natural Language Engineering*, *27*(1), 113–118. https://doi.org/10.1017/S1351324920000601.

Dewey, J. (1908). What does pragmatism mean by practical? *The Journal of Philosophy and Scientific Methods*, *5*(4), 85–99. https://doi.org/10.2307/2011894.

Dewey, J. (1931). *Philosophy and civilization*. Minton, Balch & Co.

Diallo, S. Y., Padilla, J. J., Bozkurt, I., & Tolk, A. (2013). Modeling and Simulation as a theory building paradigm. In A. Tolk (Ed.), *Ontology, Epistemology, and Teleology for Modeling and Simulation. Intelligent Systems Reference Library.* (Vol. 44). Springer. https://doi.org/10.1007/978-3-642-31140-6_10

Docherty, M., Sutton, P., Brereton, M., & Kaplan, S. (2001). An innovative design and studio-based CS degree. In Proceedings of the thirty-second SIGCSE technical symposium on Computer Science Education, SIGCSE '01 (pp. 233–237). New York, NY, USA: ACM. https://doi.org/10.1145/366413.364591.

Dongarra, J., Gannon, D., Fox, G., & Kennedy, K. (2007). The impact of multicore on computational science software. *CTWatch Quarterly, 3*(1), 1–10.

Doroudi, S. (2022). The intertwined histories of Artificial Intelligence and education. *International Journal of Artificial Intelligence in Education, 33*, 885–928. https://doi.org/10.1007/s40593-022-00313-2

Dorst, K., & Dijkhuis, J. (1995). Comparing paradigms for describing design activity. *Design Studies*, *16*(2), 261–274. https://doi.org/10.1016/0142-694X(94)00012-3.

Finnie-Ansley, J., Denny, P., Becker, B. A., Luxton-Reilly, A., & Prather, J. (2022). The robots are coming: Exploring the implications of OpenAI Codex on introductory programming. In Australasian Computing Education Conference (ACE '22), February 14–18, 2022, Virtual Event, Australia. ACM, New York, NY, USA 10 Pages. https://doi.org/10.1145/3511861.3511863.

Flam-Shepherd, D., Zhu, K., & Aspuru-Guzik, A. (2022). Language models can learn complex molecular distributions. *Nature Communications*, *13*, 3293. https://doi.org/10.1038/s41467-022-30839-x.

Flores, F., Graves, M., Hartfield, B., & Winograd, T. (1988). Computer systems and the design of organizational interaction. ACM Transactions on Information Systems. 6, 2 (April 1988), 153–172. https://doi.org/10.1145/45941.45943.

Frauenberger, C., Good, J., Fitzpatrick, G., & Iversen, O. S. (2015). In pursuit of rigour and accountability in participatory design. *International Journal of Human-Computer Studies*, *74*, 93–106. https://doi.org/10.1016/j.ijhcs.2014.09.004.

Frayling, C. (1993). Research in art and design. Royal College of Art research papers, 1, 1.

Galdon, F., & Hall, A. (2021). (Un)Frayling design research in design education for the 21st century. *The Design Journal, 25*, 6. https://doi.org/10.1080/14606925.2022.2112861

Galdon, F., & Hall., A. (2019). The ontological nature of design: Prospecting new futures through probabilistic knowledge. In Design for change, edited by Rodgers, Paul, 111–128. Lancaster: Lancaster University.

Gallifa, J. (2018). Research traditions in social sciences and their methodological rationales. *Aloma*, *36*(2), 9–20.

Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P., & Holzinger, A. (2018, August). Explainable AI: The new 42? In International cross-domain conference for machine learning and knowledge extraction (pp. 295–303). Springer, Cham. https://doi.org/10.1007/978-3-319-99740-7_21.

Goldkuhl, G. (2011). Design research in search for a paradigm: Pragmatism is the answer. In European Design Science Symposium (pp. 84–95). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33681-2_8.

Gorichanaz, T. (2023). Accused: How students respond to allegations of using ChatGPT on assessments. *Learning: Research and Practice*, *9*, 2, 183–196. https://doi.org/10.1080/23735082.2023.2254787.

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, *24*(2), 8–12. https://doi.org/10.1109/MIS.2009.36.

Herriott, R. (2019). What kind of research is Research through Design? In: International Association of Societies of Design Research Conference (IASDR) 2019.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75–105. https://doi.org/10.2307/25148625

Hoadley, C., & Campos, F. C. (2022). Design-based research: What it is and why it matters to studying online learning. *Educational Psychologist*, *57*(3), 207–220. https://doi.org/10.1080/00461520.2022.2079128. Educational PsychologistHYPERLINK.

Hoffman, R. R., Deal, S. V., Potter, S., & Roth, E. M. (2010). The practitioner's cycles, part 2: Solving envisioned world problems. *IEEE Intelligent Systems*, *25*, 6–11. https://doi.org/10.1109/MIS.2010.89. May-June 2010.

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

Hsiao, Y., Klijn, N., & Chiu, M. (2023). Developing a framework to re-design writing assignment assessment for the era of Large Language Models, Learning: Research and Practice, 9, 2, 148–158. https://doi.org/10.1080/23735082.2023.2257234.

Hundhausen, C. D., Narayanan, N. H., & Crosby, M. E. (2008). Exploring studio-based instructional models for computing education. *ACM SIGCSE Bulletin*, *40*(1), 392. https://doi.org/10.1145/1352135.1352271.

James, W. (1907). Pragmatism: A new name for some old ways of thinking. Project Gutenberg. Retrieved on 12 January 2023 from https://www.gutenberg.org/ebooks/5116.

Jones, J. C. (1992). *Design methods*. Van Nostrand Reinhold.

Knuth, D. E. (2018). Algorithmic thinking and mathematical thinking. *The American Mathematical Monthly*, *92*(3), 170–181. (Original work published 1985).

Krug, S. (2006). *Don't make me think: A common sense approach to web usability*. New Riders Publishing.

Lapowski, I. (2019, April 10). Ev Williams on Twitter's Early Years. Inc. https://web.archive.org/web/20190410151709/https://www.inc.com/issie-lapowsky/ev-williams-twitter-early-years.html (Archived version. Original work published October 4, 2013.).

Latif, S., Agarwal, S., Gottschalk, S., Chrosch, C., Feit, F., Jahn, J., Braun, T., Tchenko, Y. C., Dernidova, E., & Beck, F. (2021, October). Visually connecting historical figures through event knowledge graphs. In 2021 IEEE Visualization Conference (VIS) (pp. 156–160). IEEE. https://doi.org/10.1109/VIS49827.2021.9623313.

Lemon, L. L., & Hayes, J. (2020). Enhancing trustworthiness of qualitative findings: Using Leximancer for qualitative data analysis triangulation. The Qualitative Report, 25, 3, 604–614. Retrieved from https://nsuworks.nova.edu/tqr/vol25/iss3/3.

Leonov, A. (2018). John Dewey and the mind-body problem in the context: The case of «Neutral Monism». Actual problems of Mind. *Philosophy Journal, 19*, 19.

Leventi-Peetz, A. M., & Östreich, T. (2022). Deep learning reproducibility and explainable AI (XAI). arXiv preprint arXiv:2202.11452.

Lewin, K. (1946). Action research and minority problems. *Journal of Social Issues, 2*, 34–46. https://doi.org/10.1111/j.1540-4560.1946.tb02295.x

Lindberg, T., Meinel, C., & Wagner, R. (2011). Design thinking: A fruitful concept for IT development? In C. Meinel, L. Leifer, & H. Plattner (Eds.), *Design Thinking. Understanding Innovation.* Springer. https://doi.org/10.1007/978-3-642-13757-0_1

Lu, Y., Wang, D., Chen, P., Meng, Q., & Yu, S. (2023). Interpreting deep learning models for knowledge tracing. *International Journal of Artificial Intelligence in Education*, *33*, 519–542. https://doi.org/10.1007/s40593-022-00297-z.

Lynch, J. G. (1999). Theory and external validity. *Journal of the Academy of Marketing Science*, *27*(3), 367–376.

MacNeil, S., Tran, A., Leinonen, J., Denny, P., Kim, J., Hellas, A., Bernstein, S., & Sarsa, S. (2022, December). Automatically generating CS learning materials with large language models. arXiv preprint arXiv:2212.05113.

MacNeil, S., Tran, A., Mogil, D., Bernstein, S., Ross, E., & Huang, Z. (2022, August). Generating diverse code explanations using the GPT-3 large language model. In Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 2 (pp. 37–39).

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, *15*(4), 251–266.

Markauskaite, L., Marrone, R., Poquet, O., Knight, S., Martinez-Maldonado, R., Howard, S., Tondeur, J., De Laat, M., Shum, S. B., Gašević, D., & Siemens, G. (2022). *Rethinking the entwinement between artificial intelligence and human learning: What capabilities do learners need for a world with AI?* (Vol. 3, p. 100056). Artificial Intelligence.

McGregor, S. C., & Molyneux, L. (2020). Twitter's influence on news judgment: An experiment among journalists. *Journalism*, *21*(5), 597–613.

Meng, J. C. S. (2009). Donald Schön, Herbert Simon and the sciences of the artificial. *Design Studies*, *30*(1), 60–68.

Mettler, T., Eurich, M., & Winter, R. (2014). On the use of experiments in design science research: A proposition of an evaluation framework. *Communications of the Association for Information Systems,* 34. https://doi.org/10.17705/1CAIS.03410

Mill, J. S. (1865). Auguste Comte and positivism. Project Gutenberg. Accessed on 15 January 2023 from https://www.gutenberg.org/files/16833/16833-h/16833-h.htm.

Ocumpaugh, J., Roscoe, R. D., Baker, R. S., Hutt, S., & Aguilar, S. J. (2024). Toward asset-based instruction and assessment in Artificial Intelligence in education. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-023-00382-x

: OpenAI, Achiam, J. (2023). GPT-4 technical report. arXiv preprint: arXiv:2303.08774.

Parmelee, J. H., & Bichard, S. L. (2011). *Politics and the Twitter revolution: How tweets influence the relationship between political leaders and the public*. Lexington books.

Piech, C., Bassen, J., Jonathan Huang, J., Ganguli, S., Sahami, M., Guibas, L., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. arXiv preprint. arXiv:1506.05908.

Popper, K. R. (1962). *Conjectures and refutations: The growth of scientific knowledge*. NY, Basic Books.

President's Information Technology Advisory Committee (PITAC). (2005). *Computational science: Ensuring America's competitiveness*. National Coordination Office for Information Technology Research and Development.

Pruitt, J., & Adlin, T. (2006). *The persona lifecycle*. Morgan Kaufmann.

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences, 4*, 155–169. https://doi.org/10.1007/BF01405730

Sarsa, S., Denny, P., Hellas, A., & Leinonen, J. (2022). Automatic generation of programming exercises and code explanations using large language models. In Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1 (pp. 27–43). https://doi.org/10.1145/3501385.3543957.

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. Temple Smith.

Schön, D. A. (1984). The architectural studio as an exemplar of education for reflection-in-action. *Journal of Architectural Education*, *38*(1), 2–9. https://doi.org/10.1080/10464883.1984.10758345.

Schön, D. A. (1987). *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions*. Jossey-Bass.

Schwaber, K. (1997). SCRUM development process. *Business object design and implementation* (pp. 117–134). Springer.

Selwyn, N. (2019). What's the problem with learning analytics? *Journal of Learning Analytics*, *6*(3), 11–19. https://doi.org/10.18608/jla.2019.63.3.

Sharples, M. (2023). Towards social generative AI for education: Theory, practices and ethics. *Learning: Research and Practice*, *9*, 2, 159–167. https://doi.org/10.1080/23735082.2023.2261131.

Simon, H. A. (1969). *The sciences of the artificial*. Cambridge.

Simon, H. A. (1988). The science of design: Creating the artificial. *Design Issues*, *4*(1/2), 67–82. https://www.jstor.org/stable/1511391.

Skuse, B. (2019). The third pillar. *Physics World*, *32*(3), 30. https://doi.org/10.1088/2058-7058/32/3/33.

Slator, B. M., Anderson, M. P., & Conley, W. (1986). Pygmalion at the interface. *Communications of the ACM*, *29*, 7, 599–604. https://doi.org/10.1145/6138.6141.

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE international conference on computer vision (pp. 843–852).

Taylor, L., & Nitschke, G. (2018). Improving deep learning with generic data augmentation. In 2018 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 1542–1547). IEEE. https://doi.org/10.1109/SSCI.2018.8628742.

Tomayko, J. E. (1991). Teaching software development in a studio environment. *ACM SIGCSE Bulletin*, *23*(1), 300–303.

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., & Wintersgill, M. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence in Education*, *15*, 3, 147–204.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need. ArXiv preprint: arXiv:1706.03762v7. (version 7. Original version published 2017.).

Wacker, J. G. (2008). A conceptual understanding of requirements for theory-building research: Guidelines for scientific theory building. *Journal of Supply Chain Management*, *44*(3), 5–15. https://doi.org/10.1111/j.1745-493X.2008.00062.x.

Wigner, E. P. (1960). The unreasonable effectiveness of mathematics in the natural sciences. *Communications in pure and Applied Mathematics, 13, No. 1 (February 1960)*. John Wiley & Sons, Inc. https://doi.org/10.1142/9789814503488_0018.

Williamson, K. (2018). Research concepts. In K. Williamson, & G. Johanson (Eds.), *Research methods: Information Systems and contexts* (2 ed., pp. 3–27). Elsevier.

Woods, D. (1998). Designs are hypotheses about how artifact shape cognition and collaboration. *Ergonomics*, *41*(2), 168–173.

Zimmerman, J., Forlizzi, J., & Evenson, S. (2007). Research through design as a method for interaction design research in HCI. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07). Association for Computing Machinery, New York, NY, USA, 493–502. https://doi.org/10.1145/1240624.1240704.

Zimmerman, J., Stolterman, E., & Forlizzi, J. (2010). An analysis and critique of research through design: Towards a formalization of a research approach. In Proceedings of the 8th ACM Conference on Designing Interactive Systems (DIS '10). Association for Computing Machinery, New York, NY, USA, 310–319. https://doi.org/10.1145/1858171.1858228.