International Journal of Artificial Intelligence in Education (2024) 34:136–143
https://doi.org/10.1007/s40593-023-00350-5

COMMENTARY

Navigating Ethical Benefits and Risks as AIED Comes of Age

Ken Koedinger[1]

Published online: 18 October 2023

Wayne

Our initial provocation begins with two observations, one that AIED's successes are increasingly commercialised, and the second that AIED is being criticised for various reasons. If we can take those two separately: first, has AIED become more commercialised recently, and if so what's the impact?

Ken

There's more EdTech in schools than there ever has been, and there certainly is EdTech that advertises itself as AI. Some such systems do not have AI inside, depending, of course, on how one defines AI. Some, but not all, may nevertheless deserve the AIED label if they are scientifically driven and have been demonstrated to enhance student learning. One quite important and valuable outcome of this commercialization, is that by building sustainable systems we can have more impact on student learning. So, in that sense, commercialization has been quite positive. More funding and more innovation are positive things for education.

However, there are some risks inherent in the profit motive of commercial entities that aren't always aligned with the longer-term goals of education. These longer-term goals are hard for a consumer driven or profit driven company to keep in focus. The buyers and the users aren't the same in this context, as in healthcare, which makes it hard for the invisible hand to operate effectively. Further, the incentive structures in education itself aren't all that well aligned. For example, in the United States, we have these school boards in every little region of the country, and our schools are essentially being run by those school boards – and what expertise do they have in

✉ Ken Koedinger
kk1u@andrew.cmu.edu

[1] Carnegie Mellon University, Pittsburgh, USA

producing educational quality? Very little. So, it's not like the process is set up to be all that great from the start. So, yes, there are reasons to be worried about the growth in commercialization.[1] We in academia and in Ed Tech companies have to work hard to navigate toward the positives while avoiding the negatives of increasing commercial interests in this space.To be sure, the AIED academic community wants research and development to be impactful. However, we must continually push (might I even say market!) – our messages if they are going to have impact. Many recent commercialization efforts have ignored the long history of AIED research and have built technology and school integration processes as though nobody had done anything similar ever before[2]. Of course, many of them are painfully relearning lessons we already knew. But, it would be better if our results and lessons were known and utilized from the start.

Wayne

Let's move on to the second observation: that, because AIED has become more visible in recent years, partly because of that commercialization and its head being above the parapet far more than it was, it's being criticised for various things, such as poor pedagogic practices, datafication, classroom surveillance. What's your take?

Ken

I guess my general theme here is glass half empty, glass half full. I think that with respect to perpetuating poor pedagogical practices, if you look at the recent MOOC boom, what they mostly were doing was putting recorded lectures online, which is not a really effective educational practice. So, they were perpetuating poor pedagogical practices. I find it quite striking how many commercial folks and many educators as well, perhaps more so at the college level than at the K12 level, persist in thinking that the best way to teach is, I know stuff and I'm going to tell it to you. And, if we do that at scale, maybe more people get access to those lectures than they would have before, but it's missing the really powerful opportunity to make effective practices better. It's missing active learning, where students are actually more in charge of the doing and the thinking, which is a much more effective way to support learning than lectures.

Now, there is a more general question about technology replacing human workers. Is AI going to replace teachers? If people are framing AI in Education as striving to replace teachers, then, yes, that is a bad idea. But, a more productive and realistic framing is to think about AI as replacing doing homework without any available

[1] In Koedinger and Aleven (2016), I reflected on how despite the potential opportunities afforded by the 1998 commercialization of Cognitive Tutors in Carnegie Learning, Inc., research and development advances were "slow to be incorporated in fielded Cognitive Tutor products".

[2] Reich (2020) tells a story of how the creator of Khan Academy rediscovered, decades later, a key message of an early AIED paper about how to integrate EdTech in schools and enhance student learning (Koedinger et al., 1997).

support. A key principle of effective learning and expertise development is deliberate practice. Students learn the most when they are given tasks at the edge of their competence. In order to select a task that's at the edge of the student's competence, you have to know where the student's competence is. So, that means assessment. AI can help automate that assessment, select an appropriate next task, and provide a student with feedback and instruction just when they need it.

Another framing I've heard is replacing human tutors, but that isn't really a replacement. It's not like we have one-on-one tutors for every student. But the idea here is, to a large extent, that active learning with support from a guide on the side can really be scaled much more effectively with a technology component added in. So, your homework assignment is more interactive than just a passive piece of paper. Further, the interaction data that is naturally collected can inform caretakers of students, including their teachers and parents, about their progress and their needs. The data can also help improve the system and, more generally, improve our understanding of how best to support student engagement and learning.

Some of the criticism that AIED involves poor pedagogic practices is based more on opinion than science. As human learners, we all have opinions about how learning works and when we see something that is different from our intuitions, we conclude that it's bad. The trouble is, many of our intuitions about how learning works are wrong. Physics students at Harvard think passive lectures are better for their learning but they're not[3]. Active learning is better for learning. Why do they think that? Well, in active learning they get questions wrong, they experience failure, and failure is hard. Failure is psychologically aversive. One might say that's being mean to students, and that's not good pedagogy. But good pedagogy involves both supporting and challenging students. Good pedagogy provides challenges while both emphasising to students that failure is positive and supporting students, with feedback and reactive instruction, to learn from those failures.

When these kinds of active learning interactions are supported by AIED technology, there are many benefits like the use of data for multiple purposes as I mentioned before. There is also the potential, which I find quite exciting, that we can use this data to replace annual high-stakes standardized testing – because the data from this year-long interaction provides more information about where a student stands than a hour-long test. And, further, this assessment data from learning interactions does not take valuable student and teacher time away from learning.

That goes to the next point about datafication. No doubt, it is important to reaffirm that the AIED position is about issues of human rights and ethics and social justice, doing right for kids. And part of doing so is to remember that it isn't the data by itself that should drive our decisions, but it's the interpretation of the data that leads to decisions. This point can get lost in the excitement about machine learning. While I find explainable AI a great idea, much of the current efforts are great big Band-Aids that are not addressing the fundamental issues. We need to build models that are designed for explainability and ease of interpretation from the start, not as an after-the-fact attempt to poke into a complex black-box model after it has been built. Models should be built around a theoretical understanding of the relevant constructs, the psy-

[3] See Deslauriers et al. (2019).

chological domain, the content and motivational constructs, the social need, and the interaction constructs. These constructs should be built in from the start. Otherwise, researchers are forced to do the discovery themselves by trying to figure out (e.g., by labeling example clusters) what is driving how the deep learning model is making its predictions. This movement of labeling and construct development from the beginning to the end does not save time. Further, it opens the door to the deep learning model making the right predictions for the wrong reasons, for example, because it is making biased predictions based on indicator variables that are spuriously correlated with the true causal variables. I think there is huge power in these new big data, deep learning, and large language models. There's a lot of potential there, but let's not forget that it's the interpretation of data that should drive decision making, and that's a very human collaborative thing. It's not about the data per se.

Wayne

That all makes sense but one study that I came across recently shows that clinicians using AI-assisted systems to provide a diagnosis quite quickly stopped interpreting the data and just trusted the output without questioning it.

Ken

Do we know how to navigate in our cars anymore, for example? Just turn on the map, and my car does all that kind of lower-level stuff. I just turn the wheel. I think it's fine when it's highly reliable. And in some of those clinical situations, the offloading that the technology provides hopefully means that the doctor is attending to other issues. In medicine and in education, diagnosis is just step one. But, what's the remedy? And again, I find in both communities sometimes we get too obsessed with the diagnosis and forget the remedy.

The diagnosis is about figuring out which of the possible remedies to pursue and that's where interpretation is involved and some creativity. In both fields, we're learning that it's not always simply about giving you a pill or tweaking something in the math content. The bigger issue might be more social context, behavioral in terms of changing how you approach your health or changing how you approach your learning. There's magic in what teachers do in those situations and what parents do, and AI shouldn't get in the way. Ideally it should allow us to do the things we do best – more easily and effectively. But, it is also a new environment where we can become dependent on our technology. If the airline pilot turns on the autopilot, and in an emergency doesn't remember how to do stuff, that's bad. And so, pilots do a lot of simulated training to make sure that they're prepared for those rare but high-risk situations. Maybe, we need to do more of that in education.

For surveillance, like a lot of these issues, there are two sides of a coin. Through course assessment or more dynamic assessment that occurs naturally, as part of instruction rather than a standardised test, it definitely has a feeling of surveillance – because it's much more continuous. Is it the feeling of being watched all the time,

while you're doing your math, that's problematic? Or is it something about how the data gets used downstream? Maybe a student doesn't get a job because of this data. But let's compare that against the standardised test. If you take your SAT or your GRE test and you don't get admitted to a university, there's the same potential criticism there. In both cases we need to be careful that those measures are as valid and reliable as possible.

Because it's easier to assess reliability, psychometricians obsess about it, but validity is much more important. It's just harder to get at. And this also goes back to the interpretation issue. Anybody who's making a hiring or admissions decision based on one number, it just doesn't make sense. That number should be a part of a review, that includes other numbers like grades and other qualitative things like a personal statement and the letters of recommendation, and maybe some background factors too.

I think that the trickiest thing about this surveillance thing is that psychological sense of which we all feel to some extent when we're on the Internet and we know that unless we turn off the cookies, we get focused ads. Is it the creepiness of that or is it something deeper that's part of the conversation that we need to have? If it is the creepiness, do we put up with it when there are these other potential benefits, or should we help people better understand or deal with the creepiness? I think that's what a lot of this is about. Some deep psychological issues about how we think and how we behave and how and what we react to, which kind of interesting loops back around to education? And I can't resist saying, but I find it super strange that the science of ourselves is a science that's hardly taught at all in K12 education. Which science could be more important than the science of ourselves? Like I care about stars and plate tectonics, they're pretty interesting, but maybe the science of ourselves is not only as interesting but may be more important.

Wayne

Do you think a focus on social justice, human rights and ethics can actually facilitate innovation in AIED?

Ken

I think there's some kind of basic science application debate that has some similarities, but I think it's more nuanced because there are plenty of applications of AI that aren't about ethics, human rights, or social justice. They're more about making money for companies. At the same time though, if they're saying they have some kind of new result in a basic research sense, can they apply it in a way that makes a difference for people, especially the people that need it the most? Now that's an element of the application side that is more oriented toward social justice and human rights and ethics.

Here, at Carnegie Mellon we have a School of Computer Science where I'm in a department called Human Computer Interaction. When we started the School of Computer Science, there was a Computer Science Department, a Robotics Institute,

a Center for Automated Learning and Discovery, which later became the Machine Learning Department, and soon after, a Language Technologies Institute, which has become huge. But what has been interesting, and I think positive is the acceptance and prevalence of human computer interaction, whereas when we started, people said, human computer interaction… I don't know, that's not real computer science. It started to turn around and now almost every department in our School of Computer Science is hiring HCI people. Software engineering is another specific department where a lot of work is being done. How do we build our software engineering systems, so they have an impact on people? And the same question fits machine learning and AI.

There is always going to be a push and pull. Some researchers are still going to say I don't want to deal with that application stuff. And, if they're doing great basic research, that's fine. But at the same time with the notion of use-inspired research, so many innovations in science have been driven by application. We shouldn't just think of it at the end, but it's part of the driver of innovation and of basic science as well. That's what I love about this field: that we can be solving real problems but we can also be inspired to tackle hard issues. Real problems bring up causal inference which none of these machine learning systems do, they just do correlation. There's a lot of work that has been done and some progress has been made on building systems that actually make inferences about cause, but there's still much to be found in that area. I think that's a good example. Pay attention to bias in a deep way and do new research on causal inference. And so you can do good ethics and good science at the same time. And they should be taught to our undergraduates, Masters, and PhD students. All of them can benefit from that pathway.

Wayne

The final part of initial provocation asks whether the AIED community should carry on doing what it does now regardless of the issues we've discussed. If not, what should the AIED community do now?

Ken

I guess there's a nuance on repositioning ourselves versus reaffirming our position. I think it's worthwhile for people to be reminded that in the 40 plus years of AI and education there have been multiple efforts. For example, in the 90s we were working in urban schools, highly diverse schools, and we were working in those schools in the big city (ITS Goes to the Big City), because we cared about social justice issues. We read 'The Algebra Project', you might know that book by someone who was part of the 60s racial justice movements, who then recognized that his daughter wasn't getting very good math preparation and that was risking her future. It was essentially an element of systemic racism, although it wasn't communicated that way in the 1990s. However, we were thinking about those issues and we were working with teachers and schools to try to address social justice. There are lots of people in the community

who have cared about these kinds of issues for a long time, and some of them were members of the panel (at AIED 2022 in Durham). So we should be reminding folks of that. John Self said it well when he said that a key theme of AI in education is to build systems that care. And I think that gets at the kind of broader, deeper, notions of making a difference for kids and doing it to help them. Let's not lose our way. Let's make sure it doesn't get lost in the excitement over deep learning and large language models and big data.

I think we have to be careful to move from a diagnosis phase to a remedy phase, but sometimes we get a little stuck in saying what's wrong when, ultimately, we want to say what's right. I think there's a very important part of scholarship in this space, which is about finding injustices and finding inequities, identifying those and identifying biases. But we can't stop there as academics or as educators. We also have to think about, how do we move forward? From inequities to equities, from injustices to justices, and sometimes there are hard trade-offs. That's what sometimes complicates these things.

As for the community, we should be reaffirming our position that these issues are important, I think there are opportunities and risks. The opportunity is that these issues are more generally of interest in academics across disciplines, that is, the FATE movement in AI broadly, not just AI in education. But, I also think that we have to be aware of the risks of the black box machine learning kind of mentality. Where biases that are inherent in the data sets that are being processed in a very correlational way get reflected in the decisions of those systems – the correlations might be good but correlation isn't causation. That gets back to interpretation and that risk. Opening up the black box to build better causal models provides super opportunities for the AIED research community. We can build causal models and models that move from diagnosis to remedy. I've found that in the excitement about educational data mining and AI and large language models. We write tons of papers that stop at prediction and don't get to these other important issues. And if that is the future of AIED, then we have lost our way and we're not making a real difference for students and improving education. But I think we can, and I think it is a matter of reaffirming that we want to, and making sure our community supports work that takes those next steps.

I think if you look at the history of AIED, in the early days, you could write a paper about how my system can actually do geometry proofs, and teach geometry, and that was a publishable paper. But then at some point people said, well, you say it can, but you haven't proven it right? You haven't given any data that shows that using it makes students better than if they just used paper to do their geometry proofs. So, we got into more causal inference through experimentation and I think, appropriately, put more weight on that. But that's not happening at the moment. I've heard computer scientists in our field saying you can't write a technical paper in AIED anymore. I think we want to have a broad portfolio. It should be that you can write a purely technical innovative paper, if it's truly a first, technically. We're at a stage now where we've had plenty of those purely technical papers and we have to get rebalanced by getting the review process to ask, what evidence do you have of impact on student learning?

One very practical thing we need is to think carefully about the large language models and deep learning systems that can produce this amazing work, that can

answer these questions and are 70 to 80% correct. We have to probe that, because 70 or 80% correct is C or D grade work afterall. You can pull out lots of great examples that make it look super smart, but it doesn't necessarily mean it's going to be very effective. So, in our review process, we have to be saying that papers showing feasibility are fine, but papers can't be just showing feasibility. Papers have to be showing how the model or system can be used for more than prediction. How can the prediction actually be used to make things better for teachers, students, and schools?

Wayne

Thank you, Ken.

Funding Open Access funding provided by Carnegie Mellon University

References

Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences, 116*(39), 19251–19257.

Koedinger, K. R., & Aleven, V. (2016). An interview reflection on "Intelligent tutoring goes to school in the big city". *International Journal of Artificial Intelligence in Education, 26*(1), 13–24.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education, 8*(1), 30–43.

Reich, J. (2020). *Failure to disrupt: Why technology alone can't transform education*. Harvard University Press.