

UF's Artificial Intelligence in Education Workshop: A State of the Art Perspective from Academia and Industry

*Seyedahamd Rahimi & Salah Esmaeiligoujar
University of Florida*

Abstract

The University of Florida's AI in Education workshop, conducted in April 2023, brought together three NSF-funded AI institutes and three learning companies together with many other scholars in the fields of Computer Science, Learning Sciences, Instructional Design, and Computational Psychometrics. This online workshop attracted about 750 participants from 42 countries. The workshop had two main parts: (1) presentations: each AI institute and learning company presented their latest work and discussed their findings; (2) panel discussions: in six Zoom rooms with 3 scholars on each panel, discussions around a specific topic in AI in Education were conducted. In this paper, we first briefly describe the six presentations, then we dive into the panel discussions and provide a synthesis of their key takeaways.

Keywords

AI, Artificial Intelligence, AI in Education, Learning, Learning Sciences, Assessment

Workshop's Director: Seyedahmad Rahimi

Presenters & Panels Discussants: Chris Dede, James Lester, Neile Heffernan, Kristen DiCerbo, Peter Foltz, Alina A von Davier, Sidney D'Mello, Valerie Shute, Danielle S. McNamara, Ryan Baker, George Siemens, Diego Zapata-Rivera, Kristy Boyer, Victor Lee, Lin Lin Lipsmeyer, Richard Golden, Russell Almond, and Rene Kizilcec.

Panels' Moderators: Pasha Antonenko, Albert Ritzhoup, Corrine Huggins-Manley, Walter Leite, Anthony Botelho, Jinnie Shin, Robe Moore, and Wanli Xing.

Workshop's Webpage: <https://ialt.education.ufl.edu/ai-in-education-workshop/>

Introduction

Artificial Intelligence (AI) is among the most prominent emerging technologies (Jiang et al., 2022; McCarthy et al., 2022) affecting various aspects of our lives including education. AI, generally, refers to the ability of computer systems to perform tasks that typically require human intelligence, such as learning, reasoning, decision-making, visual perception, speech recognition, and natural language comprehension (Russell & Norvig, 2010; Kaplan, 2016; Kaplan & Haenlein, 2019).

The origins of AI research date back to the 1950s when scientists began exploring the possibility of computers that could think and act like humans (Buchanan, 2005; Frank, 2019). The term “artificial intelligence” itself was coined in 1956 during the Dartmouth Workshop, considered a seminal event in AI history (McCarthy et al., 1955). The proposal submitted by McCarthy, Minsky, and others stated that “every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it” (McCarthy et al., 2006, p. 12). In the early years, research on AI focused on general reasoning tasks using symbolic logic and rules (Russell & Norvig, 2010). In recent decades, advances in computing power and big data have enabled breakthroughs in AI (e.g., neural networks and machine learning approaches) (Saravanan et al., 2017; Kumar, 2017).

AI also refers to a broad *discipline* focused on developing AI-powered computer systems (Russell & Norvig, 2010; Gustafson & Gustafson, 2016). Key focus areas within AI include *machine learning*, which involves algorithms that can learn from data to make predictions or decisions (Michalski et al., 2013; Chen et al., 2020). Machine learning itself includes subfields such as *deep learning* that uses neural networks (inspired by how humans’ neural networks function) trained on large datasets (Lin et al., 2018). AI techniques usually are enabled through supervised (i.e., models learn from human-labeled data), unsupervised (i.e., models learn and find patterns from unlabeled data), semi supervised (i.e., models trained both on labeled and unlabeled data), or reinforcement learning (i.e., models trained based on rewarding desired behavior and punishing undesirable ones) (Berry et al., 2019; Reddy et al., 2018; Zhu & Goldberg, 2022). AI models trained using these techniques can be predictive (e.g., regression and classification models) or generative (e.g., models that can produce new data based on the data they have been trained on).

In the past decade or so, AI has become ubiquitous in modern life through applications such as virtual assistants, recommendation systems, autonomous vehicles, precision agriculture, and financial fraud detection (Hassan et al., 2017; Sarker, 2022). As research continues, AI promises to transform entire industries and further augment human capabilities (McCarthy et al., 1956; Luckin et al., 2016). Education is one of the fields that AI has had and continues to have an enormous impact on and requires further research to understand and even from this impact (Chen et al., 2020; Chiu, 2023; Niemi et al., 2023).

AI in Education

AI in education involves the application of AI technologies such as Intelligent Tutoring Systems (ITSs), chatbots, and machine-learning-based models (both predictive and generative) to support and enhance teaching and learning (Chassignol et al., 2018; Crompton & Burke, 2023). AI in education dates back to the 1950s with the introduction of ITSs - computer programs aimed at providing personalized instruction tailored to individual learners' needs and responses (Shute & Psotka, 1994). ITSs are designed to offer customized and flexible teaching to students (Siemens & Baker, 2012). The early rule-based ITSs would guide students by adaptively branching to new content or activities based on students' answers and performance (Shute & Psotka, 1994; Khosravi et al., 2022).

Advances in AI technology beyond academia (e.g., Chat GPT by Open AI) are rapidly enabling AI innovations in education (Chaudhry & Kazim, 2022; Castelli & Manzoni, 2022). In particular, progress in natural language processing is powering more fluent conversational agents and chatbots that can support student learning through natural dialogue (Vazquez-Cano et al., 2021). Additionally, improvements in computer vision and affective computing now allow AI systems to monitor and respond to learners' cognitive and emotional states in real time (D'Mello et al., 2018). Social robots designed as artificial tutors are also leveraging nonverbal behaviors and social intelligence to naturally connect with and assist students (Kennedy et al., 2016). Furthermore, advances in machine learning techniques are driving more personalized and adaptive educational experiences, such as individualized recommendations and assessment analytics (Holmes et al., 2019). Overall, cutting-edge advances in AI are shaping the future of education by enabling more intuitive, responsive, and customized learning powered by data and algorithms. While AI brings a lot of benefits to education, it also presents challenges (Vincent-Lancrin & van der Vlies, 2020; Akgun & Greenhow, 2021). For instance, magnifying biases, transparency, digital equity, emotional development and privacy risks (Kahn & Winters, 2021), fairness, accountability, and ethics in AI-based educational tools are emerging (DiCerbo, 2021; Memarian & Doleck, 2023). Moreover, AI has been criticized for resembling a black box, where it is difficult to understand how AI systems make decisions (von Eschenbach, 2021). This lack of transparency makes it challenging to fix AI systems when they produce unwanted outcomes (Brožek et al., 2023). A new field called Explainable AI (XAI) is emerging to address the issue of lack of transparency in AI systems (e.g., Arrieta et al., 2020; Minh et al., 2022). XAI refers to processes and techniques that help humans understand and trust the outputs of machine learning models. XAI enables assessing key aspects of AI systems (e. g., accuracy, fairness, transparency, and outcomes) (Minh et al. 2022). By facilitating human understanding of model decisions, XAI promotes trust, auditability, and productive use of AI. Core XAI methods involve explaining model predictions, tracing how outputs are generated, and conveying the rationale behind decisions (Arrieta et al., 2020).

The emergence of generative AI, such as GPT models, presents both opportunities and challenges to the conventional education systems (Chiu, 2023). We need rigorous studies evaluating the efficacy of different AI tools and applications across diverse learning contexts and

populations. It is also essential to investigate best practices for integrating AI in pedagogically sound ways and fostering necessary competencies in both students and teachers to leverage these technologies (Luckin & Cukurova, 2019). Without sufficient research guiding AI implementation, we risk exacerbating educational problems rather than solving them. The growing importance of AI in education (Luckin et al., 2016; Krstić et al., 2022; Chiu et al., 2023) has led organizations such as the National Science Foundation (NSF) to fund multi-site AI research institutes focused on cutting-edge research and development in educational settings. Moreover, major technology companies including OpenAI, Microsoft, and Google have made substantial advances in AI capabilities recently, as evidenced by chatbot systems such as Chat GPT (Thoppil, 2022). The launch of Chat GPT, in particular, has sparked global interest and discussion around AI. It was around the advent of these new large language models that the idea for a workshop started to take shape, seeking to explore the potential impacts of such technologies on education. In this paper, we discuss a workshop focusing on the cutting-edge research in AI in education both from an academic and an industry perspective. Moreover, we discuss the future of AI in education and the challenges we may face.

AI in Education Workshop–April 2023

The University of Florida's AI in Education workshop was conducted in April 2023 bringing together three NSF-funded AI institutes and three learning companies as well as many other AI scholars in the fields of Computer Science/AI, Learning Sciences and Instructional Design, and Computational Psychometrics to discuss the latest developments in the field of artificial intelligence. The NSF-funded AI Research Institutes that were invited included: (1) The AI Institute for Student-AI Teaming (iSAT): an interdisciplinary research community dedicated to transforming classrooms into more effective, engaging and equitable learning environments focusing on collaboration; (2) the ENGAGE AI Institute: an institute that supports and extends the capabilities of teachers and students with AI. This institute's mission is to produce transformative advances in STEM teaching and learning with AI-driven narrative-centered learning environments. And (3) the AI Institute for Adult Learning and Online Education (AI-ALOE) which develops an AI-based transformative model for online adult learning. Each of these institutes received approximately \$20M from US's National Science Foundation. The three well-known learning companies that use AI in their products included: (1) *Duolingo*: a language learning company, (2) *Khan Academy*: a personalized learning company, and (3) *ASSISTments*: a non-profit learning company which creates responsive online technology that is teacher-paced and evidence-based for subject such as mathematics. Finally, the workshop invited 10 other top scientists whose work has been nationally and internationally recognized to discuss important topics around AI-in-Education. This workshop was conducted on April 6th, 2023 online and attracted about 750 participants from 42 countries. The presenters were Chris Dede, James Lester, Neile Heffernan, Kristen DiCerbo, Peter Foltz, Alina A von Davier; the contributors were Sidney D'Mello, Valerie Shute, Danielle S. McNamara, Ryan Baker, George Siemens, Diego

Zapata-Rivera, Kristy Boyer, Victor Lee, Lin Lin Lipsmeyer, Richard Golden, Russell Almond, and Rene Kizilcec.

The workshop had two main parts: (1) presentations: each AI institute and learning company presented their latest work and discussed their findings; (2) panel discussions: in six Zoom rooms with 3 scholars on each panel, discussions around a specific topic in AI in Education were conducted. To elicit the important themes for the panels, we sent a form to the invited scholars before the workshop and asked them to list 10 most important issues they think we need to discuss in the workshop panels. Then, we analyzed the qualitative responses and the following five themes emerged:

- 1) Partnership of Humans and AI / Intelligence Augmented (IA)
- 2) AI Education from K-12 to Adult Learning
- 3) Fundamental Research, Development, and Theory Building
- 4) AI in Assessment and Improving Learning
- 5) Policy and Challenges related to AI in Education

In this paper, we first briefly describe the six presentations, then we dive into the panel discussions and provide a synthesis of their key takeaways.

Presentations: NSF-Funded AI Research Institutes

Institute for Student-AI Teaming (iSAT)

The institute for Student-AI Teaming (iSAT), represented by co-PI Peter Foltz and PI Sidney D'Melloe at the workshop, presented on the topic of "Envisioning Next Generation Human-AI Partnership at iSAT." iSAT brings together researchers, graduate students, and postdocs from 14 research areas, from nine major universities in the United States (e.g., University of Colorado Boulder, Colorado State University, University of California Berkeley). This AI institute also partners with various K-12 and industry network affiliates. The key question that the iSAT team is investigating is "how to promote deep conceptual learning via rich socio-collaborative learning experiences for *all* students?" As a teacher it is often challenging to put students in collaborative learning situations. Addressing such challenges, iSAT researchers investigate several research questions: "what types of communications are going on when students collaborate?" "can we harvest the multimodal-multiparty, naturalistic, social interactions and facilitate those interactions using AI?" "how can we help with the development of the whole student (i.e., students equipped with skills such as collaboration, communication, and critical thinking that are needed for success in life in the 21st-century)?" and "What AI should do in education, rather than what AI can do?" Through these high-level questions, iSAT "views AI as a social, collaborative partner that helps both students and teachers work and learn more effectively, engagingly, and equitably" (see Figure 1).



Figure 1. iSAT vision of AI in classrooms

The iSAT institute, systematically focuses on various areas to provide measurable learning outcomes: (1) domain knowledge acquisition, (2) 21st-century skills (i.e., collaborative problem solving and critical thinking), (3) motivation and affect (i.e., interest, curiosity, and self-efficacy in AI, AI-education and STEM), (4) disciplinary practices (i.e., inquiry learning & academically productive talk), (5) ethics and society (i.e., nature, behavior, and consequences of AI systems in classrooms), and (6) empowerment of students to contribute to an AI-based workforce.

Through a responsible innovation framework (Stilgoe et al., 2022) which encourages participatory research and innovation, iSAT follows three main strands in research from foundational research (i.e., improving AI models and techniques for assessing students' collaboration in real-time) to use-inspired research: (1) understanding and facilitating collaboration using computer science approaches, (2) orchestrating interactions with AI, and (3) broadening participation with co-design. An example of a participatory research conducted by iSAT is about an ideation workshop about how students see future use of AI to enhance collaborative learning. Results of these workshops showed that students expressed interest and desire in interacting with an AI partner, they did not want the AI partner report and poor behavior to their teacher, and they valued agency over the AI partner (e.g., to turn it off if they wanted) among other findings that led to better understanding of how the students like an AI partner in the future. Then, after collecting valuable ideas from these workshops, the team designed their AI systems in many iterations to test and use in classrooms. Through these efforts, the team came up with new metaphors for how they refer to an AI partner: *community builder* (i.e., help students to develop trusting collaborative relationships with each other, their teacher, and the AI partner); *augmenter* (i.e., support for teachers in classroom management using the data from each group); and *Co-Pilot* (i.e., support and facilitate students' work by providing tailored guidance). These AI metaphors have been created in iSAT and tested in various projects for different curriculum.

For example, in a sensor immersion-collaboration curriculum unit, students work in small groups, dyads, and individually to learn about scientific phenomena and use sensors in interesting, creative ways (Southwell et al., 2022). Through a multimodal data collection effort (i.e., audio, videos, eye tracking) for about 900 hours from 500 students and 15 teachers in two school districts, the team has examined the use of AI partners in a real-world classroom situation. The collected data goes through a multifaceted data analysis: (1) annotation and analysis (i.e., looking for evidence for collaborative productive utterances in what the students said during collaboration; code for where the students were looking at during the collaboration); (2) AI-based model building; (3) Model evaluation; and (4) Error analysis. These four steps occur in an iterative process to enhance the accuracy of those models.

After these models are tested and ready, they will be included in AI technologies such as CoBi (Community Builder). CoBi which helps students understand if they show respect, are committed to community, show equity, and show evidence that they are moving forward in their thinking. CoBi provides support by giving the students feedback and showing them examples of what they said related to the desired or undesired behavior and it allows them to agree or disagree with what CoBi said. Students' feedback to the AI system then can be used to retrain the AI models.

Another example of an AI technology being built by iSAT is the iSAT Community Builder dashboard (Figure 2). This dashboard provides information for the schools and other partners showing students' progress toward community building: the extent to which the students commit to the community (e.g., help one another), move their thinking forward (e.g., use evidence, ask follow up questions), and are respectful in their interactions with each other (e.g., give compliments, use each other's names). This dashboard can show this information in a longitudinal way relieving more nuances in students' change and growth when it comes to collaborative learning and community building.

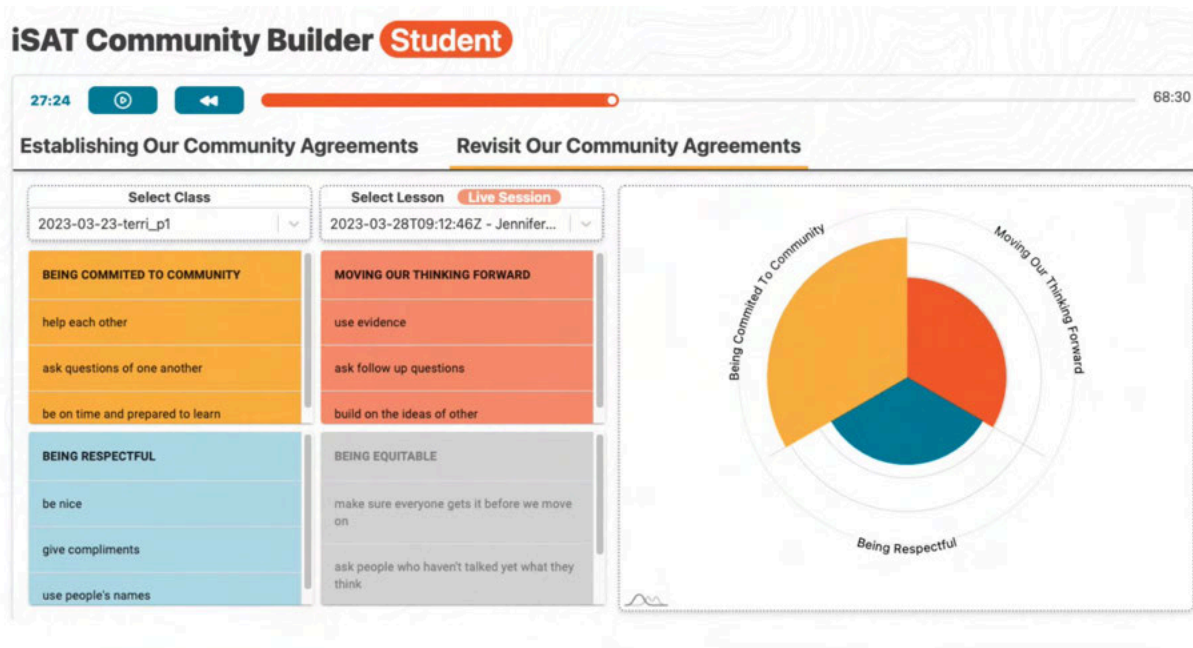


Figure 2. iSAT Community Builder Dashboard

iSAT’s recent studies have demonstrated diverse applications of AI in education, each focusing on unique aspects of learning and collaboration. For instance, Demszyk et al. (2023) explored the use of natural language processing to support educators by automating routine tasks, enhancing professional learning through personalized feedback, and facilitating adaptive lesson planning to meet individual student needs more effectively. In another study, Doherty et al. (2023) investigated the dynamics of human-agent teams, focusing on speech patterns to model “teamness” and providing insights into how AI can improve team interactions in educational and professional settings. Similarly, Cao et al. (2023) analyzed how inaccuracies in automatic speech recognition can affect the understanding and analysis of collaborative discourse in educational settings, highlighting the need for improved accuracy in speech recognition technologies. Lastly, Zhou et al. (2023) applied multilevel vector autoregression to analyze temporal dynamics in collaborative problem-solving, offering a deeper understanding of how successful collaboration is temporally structured and suggesting strategies to enhance the effectiveness of group interactions.

ENGAGE AI Institute

James Lester, PI, presented at the workshop the work he and his colleagues have been doing in Engage AI Institute—a multi-site institute involving North Carolina State University, University of North Carolina, Vanderbilt University, Indiana University, and Digital Promise. Engage AI Institute has two major strands of work: (1) use-inspired AI research which involves narrative-centered or story-based learning, embodied conversational agents, multimodal learning

analytics; and (2) foundational AI research which includes natural language processing, computer vision, and machine learning. Both strands aim to develop engaging learning environments using AI systems that are ethical and effective. Additionally, Engage AI Institute has three major categories of activities: (1) AI-driven narrative-centered learning environments (with the ethical threads of trust and privacy); (2) Core AI methods (with the ethical threads of fairness, accountability, and transparency); and (3) Educational impact in and outside of schools (with the ethical threads of diversity, equity, and inclusion). These three major categories are related to research and strategic outreach and impact. Lester focused on the research aspect of the institute in his presentation.

The AI-driven, narrative-centered, learning environments in Engage AI Institute include narrative-centered learning (i.e., any learning experience that can immerse students in stories), embodied conversational agents that can invite students to participate in the story, and multimodal learning analytics. These environments create learning experiences for students that are effective (targeting science and computer science) and engaging in formal (e.g., classrooms) and informal (e.g., museums) settings.

The core AI team at the Engage AI Institute work in three areas of natural language processing (i.e., natural language understanding and generation), computer vision (i.e., the egocentric vision of a scene from a student's point of view), and machine learning (i.e., addressing problems such as factuality, concerns of private information and privacy). The institute is also concerned about the impact of these technologies on various levels in society. The AI-driven, narrative-centered learning environment prototype has been designed and developed with two interconnected components: (1) participatory design where students are part of the design process, and (2) engineering which can accommodate new narrative planners, embodied conversational agents, and multimodal data integration which can help with use of these environments at scale.

Specifically, through these two components, the Engage AI institute has used reinforcement learning to design and get better at producing new narratives that can learn from narrative generation policies that can reason about the needs of the students (Figure 3). This system also includes generative models that can help produce the narrative-centered stories and environments (i.e., storyworld, plot, and character models). For example, this AI-based system can generate a learning environment based on students' interest in an imaginary storyworld or in a hospital. The institute also used large language models for empowering the embodied conversational agents. Specifically, LLMs are used to support conversation between the students and the agents and allow the agents to take different roles (e.g., pedagogical, companion, or teacher assistant).

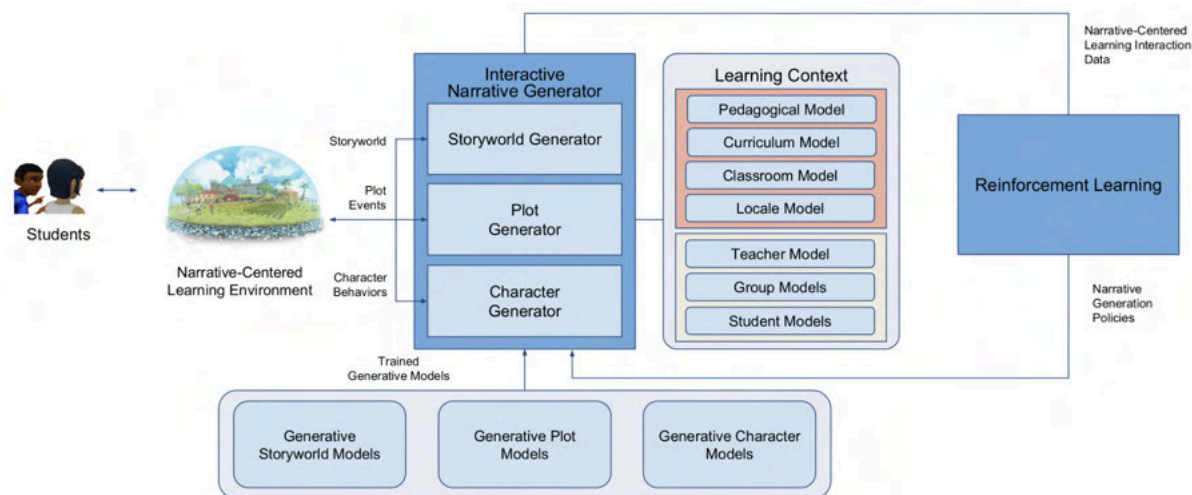


Figure 3. Reinforcement-Learning-based Narrative-Centered Learning in Engage AI Institute

Finally, the multimodal learning analytics infrastructure in Engage AI institute deals with various data streams: video, speech, gaze, and activity logs. This data gets processed through a learning analytics pipeline (Figure 4) which consists of timeline construction (i.e., all the data streams get harmonized using time synching), feature extraction (i.e., the raw data from each stream undergoes an enrichment process; for example, using speech to text systems text gets generated), feature selection (i.e., the data gets enriched in a higher level that feature extraction; for example, the discourse tags are assigned), and behavior classification (i.e., using the enriched data students' behaviors are recognized by the system and can be used for adaptation purposes). The effort in the multimodal learning analytics sector of the institute can be harvested both for research purposes and real-time use in the learning process.

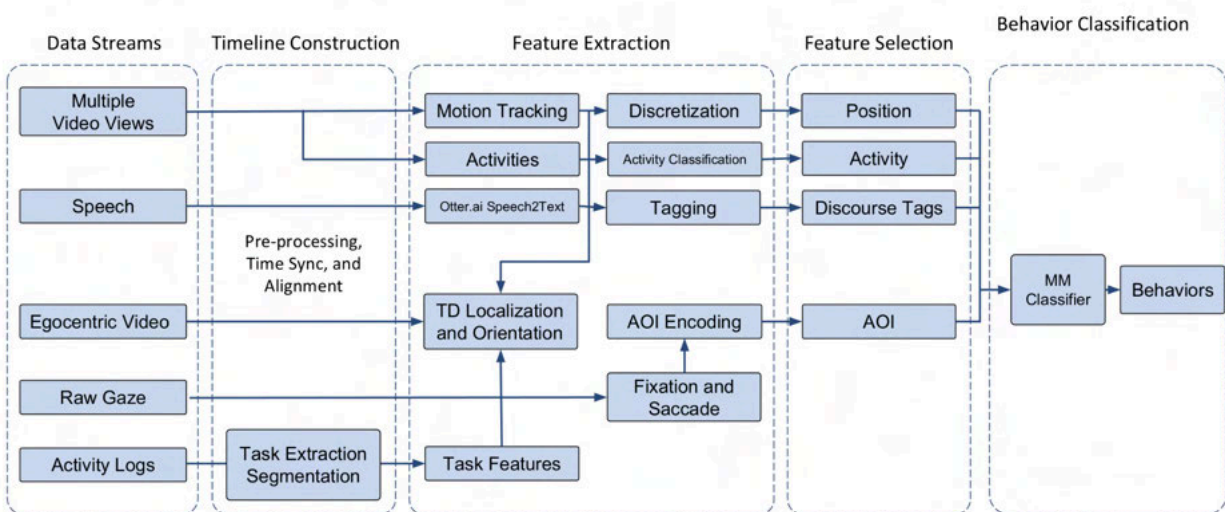


Figure 4. Multimodal Learning Analytics Pipeline in Engage AI.

In recent years, Engage AI Institute has been at the forefront of research in AI-driven education through various impactful studies (e.g., Hmelo-Silver et al., 2022; Vogel et al., 2023). For example, significant study demonstrates how AI supports ambitious learning practices, focusing on equitable and inclusive education, which is important for empowering students and teachers alike (Hmelo-Silver et al., 2022). Engage AI projects include enhancing collaborative problem-solving and knowledge co-construction with AI, where researchers explore how AI can facilitate deeper interaction among learners to build on each other's ideas effectively (Vogel et al., 2023). Together, these studies underscore Engage AI's role in advancing educational practices through AI to foster engaging, supportive, and effective learning environments.

The AI Institute for Adult Learning and Online Education (AI-ALOE)

Chris Dede, Co-PI, shared his insights about the initiatives and progress made by him and his colleagues at the AI-ALOE (Artificial Intelligence for Adult Learning and Online Education) Institute. Dede pointed out that AI-ALOE differs from the other institutes and learning companies (present at the workshop) because of its audience—i.e., adults. The AI-ALOE stands out in its approach to addressing the dynamic needs of adult learners in a rapidly changing world beyond formal education. Dede highlighted the significant demographic shifts in the United States and globally, leading to an increased adult population. This change necessitates a dramatic expansion in adult education, surpassing the resources traditionally allocated for youth education. Dede pointed out that the stability of life in the past has given way to turbulence due to factors such as AI, climate change, and globalization, making reskilling and upskilling essential for adults.

A critical aspect of the AI-ALOE's approach is recognizing that adult learners differ significantly from children. For instance, the diversity in developmental and experiential stages demands a unique pedagogical approach for adult learners, distinct from traditional pre-college and college-level education. There are several conditions that facilitate adult learning: (1) promoting a sense of competence and agency; (2) providing time and space to explore and learn from errors; (3) providing self-paced and spaced practicing opportunities; (4) mitigating age-related cognitive losses and take advantage of existing knowledge (e.g., use of advance organizers, goal reminders); and (5) creating learning episodes within possibility of success and organized by difficulty. Therefore, due to such considerations, changes in developmental and experiential factors the andragogy for adult learners is different from the pedagogy for younger learners in formal education. Analyzing what AI agents we need for adult learners, Dede indicates that we need AI agents that can help with a learning episode, a course, a whole program, and eventually as a career coach in a life-long learning approach for a series of careers one might have in their life.

The AI-ALOE's strategy involves a phased approach in the five-year span of the institute: starting with academic preparation in community colleges and universities (pre-ALOE),

then expanding to non-traditional and marginalized learners (Years 1 and 2), and eventually encompassing workforce-embedded upskilling and reskilling (e.g., corporate training) in Years 3 to 5. This progression is intentionally from easier concepts and skills that adults are familiar with to harder concepts and skills that reflect challenging populations and contexts.

AI-ALOE is designing AI-powered cognitive assistants that are sensitive to age and experience, among other factors, to cater to the diverse needs of adult learners. The theory of action that the AI institute addresses the question of “how to use online education for adult learning effectively?” AI-ALOE hypothesizes that using AI assistants and through personalization of learning for working adult learners can make online learning for adults effective.” This hypothesis is supported by the famous 2-sigma effect of 1-1, personalized tutoring that Bloom talked about (1984). The second hypothesis has to do with the use of AI assistants to address well-known problems in online education which refers to the community of inquiry framework of Garrison et al. (2010). These online problems are generally related to how to engage learners in better ways in online learning? Specifically, how to enhance learners’ cognitive presence, teacher presence, and social presence in online education? Garrison et al. (2010) framework provides various components for how to address these issues; AI-ALOE aims to address these problems using AI-powered assistants.

Dede concludes by emphasizing the importance of both working on present-centered problems in adult education as well as focusing on issues that might present in the future considering the uncertain world we live in. “If we stay on the path of looking at how we look at how to change adult learning in the present, we will miss the target because the target is moving,” Dede asserted. For example, advances in AI at workplaces where the workforce will benefit from AI assistants, the division of labor is going to shift from routine tasks to tasks that require more in-depth human judgment. Moreover, advances in simulations may mean that small-grain-size performance assessments (i.e., formative assessments) can be more suitable rather than proxies such as tests or seat-time (i.e., time on task). Finally, the focus of essential dispositions (e.g., resilience, tenacity, creativity, initiative, and agility) in our complex world seems necessary. Such dispositions help adults successfully maneuver through hard times looking at the challenges as opportunities rather than threats.

Dede closes his remarks by asking an important question: “What business are we in?” We, as a field, should understand that we no longer are in the education business; we are in the learning business. The learning business is different from the education business. We are in the business of continuous capacity building. Keeping this in mind, we can better figure out how to leverage new technology (e.g., AI) to achieve the goals for human learning.

The AI-ALOE initiative has been at the forefront of integrating artificial intelligence with adult learning and online education, striving to transform how educational content is delivered and absorbed on a global scale. Recent research (e.g., Crossley & Holmes, 2022; Dede & Lidwell, 2023; Kim et al., 2023) has shown the potential of AI-enhanced methods to improve

massive digital hybrid learning environments. For instance, Crossley and Holmes (2022) have used advanced NLP techniques to enhance vocabulary assessment, revealing the strengths of semantic embedding approaches over traditional methods in understanding language acquisition among learners (Crossley & Holmes, 2022). Furthermore, Dede and Lidwell (2023) emphasize the incorporation of AI-driven tools and methods to facilitate engagement and personalize learning experiences, ensuring that digital learning resources effectively reach a broad audience (Dede & Lidwell, 2023). Similarly, studies by Kim et al. (2023) have demonstrated the efficacy of AI-enabled knowledge representation systems in advancing understanding of learning progress models, showcasing how AI tools can refine the assessment of educational growth over time (Kim et al., 2023). These studies underscore AI-ALOE's mission to harness AI's capabilities to innovate and expand the reach and effectiveness of online learning platforms.

Next, we discuss the presentations related to the three learning companies who presented their work at the AI in Education workshop.

Presentations: Learning Companies

Duolingo

Alina A. von Davier, Chief of Assessment at Duolingo, presented her work on the use of generative AI for test item development. There are several problems that Duolingo has been addressing using generative AI regarding traditional test development (i.e., time-consuming, expensive, and prone to validity issues and security breaches). Specifically, traditional test item development needed expert-written test items that needed 1000s of pilot administrations, and psychometrics item analyses which, due to the labor intensive nature of this process, would usually lead to limited pool of test items. We need a paradigm in assessment that can provide access to high quality (valid, reliable, and secure tests) to as many people as possible. Digital-first assessments can provide a solution to this problem (Langenfeld et al., 2022).

Digital-first assessments are born, maintained, and modified in a digital environment and are meant to be used in digital form. These assessments are designed with computational models and algorithms in mind. Digital-first assessments are theory-based and evidence-centered accounting for the complexity of the nature of data, and they are valid and reliable. These assessments are integrated through a fluid infrastructure with automatic tools and databases. Most importantly, these assessments drive from and focus on test-takers' experience with technology. For instance, people, these days, are used to the available, on demand, anytime, anywhere nature of technology. The same user experience can be produced by online testing with remote proctoring from the comfort of the test takers using digital-first assessments.

Generative AI can help advance the development of digital-first assessments. But what is generative AI? In short, generative AI involves sensing (through computer vision and audio processing), comprehending (through knowledge representation and natural language

processing), and analyzing various data types (thorough machine learning and statistics techniques). In the context of test development and assessment, the analysis part is called computational psychometrics (von Davier et al., 2022). That is, computational models from machine learning and statistics but put into the psychometrics framework of reliability, validity, and generalizability (Troignon et al., 2022). Generative AI, in general, deals with large computational models encompassing large models dealing with language, text, image, voice, and video. In Duolingo, generative AI is used to develop a large item bank to support a valid and secure digital-first assessment, administered on-demand, at scale. That is, generative AI will replace item writing by experts, pilot administrations, and the limited psychometric analysis steps in the traditional test development process.

This new, AI-based item generation process still involves humans in the loop. The loop includes the following: (1) first, using the expanded evidence-centered design framework (Arieli-Attali et al., 2019) human experts (i.e., experts in TESOL or applied linguistic) design some items; then, using the science of learning and assessment, the features of those items that can provide the optimal evidence for the construct(s) at hand will be identified; (2) then, AI systems will generate new items based on those features extracted from the human-generated items; (3) another group of diverse human experts (trained to identify fairness and bias issues) review and/or edit the AI-generated items (e.g., fact checking happens at this stage as AI systems might provide false information; the items also get reviewed for fairness and bias); and finally (4) the human labels and edits are used to improve the AI systems so they can produce high-quality items in the future. Following this loop, Duolingo's assessment team has used many AI models and technologies (e.g., GPT3, BERT, Chat GPT, Stable Diffusion) to automatically generate various English test item types (e.g., fill-in-the-blank vocabulary questions, multiple choice questions, free response comprehension questions, writing and reading passages).

Alongside the efforts by Duolingo to use generative AI for high quality item development, the team proposed one of the first set of guidelines for responsible AI (Burstein, 2023). Three main areas to think about when using generative AI for item generation responsibly are: (1) responsibility in making sure that the items are fair and bias free; (2) transparency in presenting and explaining the process; and (3) humans should be in the loop for all the important decisions throughout the process. von Davier concluded her presentation by saying that digital-first assessments lend themselves well to automation and AI integration. However, with these powerful AI systems comes responsibility and we, as a field, must put in place suitable guidelines for their use.

Khan Academy

Kristen DiCerbo, the chief learning officer at Khan Academy, presented the work she and her colleagues have been doing regarding use of Open AI's generative system (i.e., Chat GPT4) in Khan Academy's ecosystem from September 2022. Her presentation was entitled "harnessing

large language models for empowered learning at Khan Academy.” Khan Academy has been leveraging Chat GPT through small experimental projects (until the time of the workshop; April 2023) in four cases: (1) a dialogue-based tutor, (2) a student partner in various language activities, (3) a teaching assistant, and (4) an assistant for content creators. DiCerbo mostly focused on the first and the third cases in her presentation in the AI in Education workshop. The aim of the small experimental projects was to gather different types of feedback and enhance the AI-powered Khan Academy.

The first use case DiCerbo discussed was an AI-powered tutor. Through this effort, the Khan Academy team asked questions such as “how to use generative models to do things that we have never been able to do at Khan Academy (e.g., asking students why they have answered a problem that way? Or to ask them to explain their solution, and to get them engaged in a dialogue with the tutor agent)”. The tutor that the Khan Academy team has designed and have been testing is called Khanmigo which can assist students with their math problems (Figure 5).

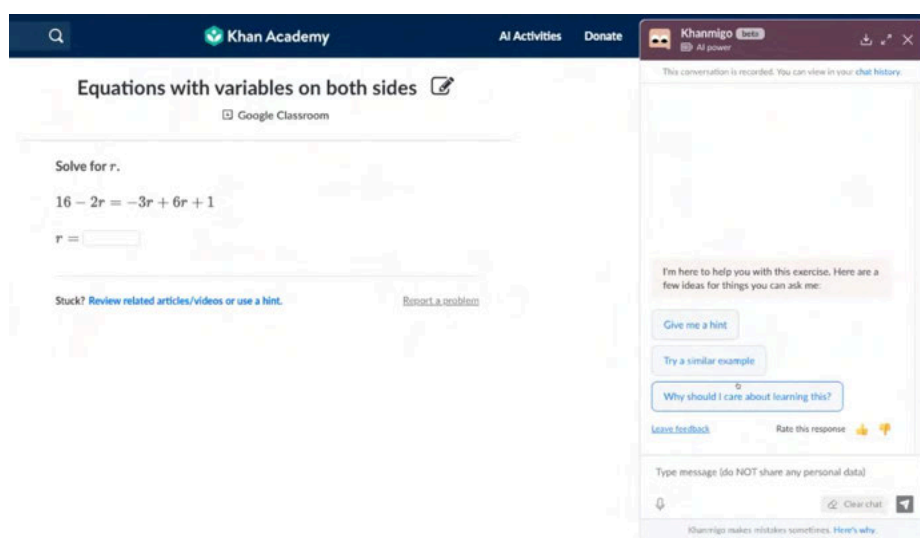


Figure 5. Khanmigo; the AI-powered math tutor at Khan Academy

Khanmigo initially prompts students with three existing commands of “Give me a hint,” “Try a similar example,” and “Why should I care about learning this?” The Khanmigo chat box reminds students not to share their personal data (e.g., their name). This is to protect students’ privacy. Khanmigo sends what students typed, in an anonymized way, in the box along with a long prompt explaining to Open AI’s Chat GPT how to respond (i.e., prompt engineering), so the only way Open AI could know who sent the message is if the students write their personal information in the chat box.

One example of the kind of prompt engineering that has been done to prepare Khanmigo is to tell it not to work on problems other than the current problem at hand (e.g., students’ cannot ask Khanmigo about their homework). Since Chat GPT is not specifically designed for math but is capable of simplifying equations and providing hints to students, another aspect that was

included in Khanmigo's prompt engineering was to help Khanmigo get better at doing math using some reinforcement learning with Open AI. Additionally, Khanmigo was told, through prompt engineering, to be an encouraging, empathetic tutor (i.e., you can tell the AI assistant to take on a specific persona). Sometimes the tutor would overly encourage students (even if the students got the answer wrong). Therefore, through some prompt iterations, the team was able to train Khanmigo to successfully act like a tutor. DiCerbo highlighted the potential benefits of Khanmigo, such as providing real-time support to students without the need for a teacher's presence. Khanmigo is currently being tested in classrooms, with plans for it to be used by students at home as well. DiCerbo also mentioned the creation of various activities and debates that students can engage with.

Additionally, Khan Academy is exploring a prompt playground empowered with AI to help and assist teachers. For example, this playground allows the teachers to interact with the AI system (connected to Chat GPT in the background) and create rubrics (e.g., illustrating different levels of performance when it comes to writing an essay). Such prompt engineered AI assistants can have a lot of potential to augment teachers' performance in the classroom.

DiCerbo also touched on the importance of ensuring equal access to AI and addressing ethical and responsibility considerations, including risk mitigation strategies. All of the efforts at Khan Academy to leverage AI have been through the lenses of ethical AI use (i.e., responsible AI). Adapting the ethical framework for AI in education by The Institute for Ethical AI in Education (n.d.) and the risk mitigation ideas from the National Institute of Standards and Technology (Boyens et al., 2021), Khan Academy developed a set of guidelines related to the risks of using AI (e.g., accuracy, safety) and how to mitigate them (e.g., prompt engineering, transparency, limited access, use of a moderating API). DiCerbo discussed the development and implementation of an AI-based student feedback system that could enhance the accuracy of the AI system. She highlighted the ability of the system to rate responses, provide feedback, and suggest improvements, noting that it has been well-received by students. These guidelines and mitigating actions can be used by other parties in various educational sectors who are interested in using AI—specifically generative AI.

ASSISTments

Niel Heffernan, professor of computer science at Worcester Polytechnic Institution and ASSISTments's director, shared what he and his colleagues have been doing regarding the use of AI in their math tutoring platform. ASSISTments is a platform that teachers use to build their own content (i.e., math problems and feedback) or use what exists in ASSISTments' library to assign to their students. Other than letting teachers curate the problems they want their students to work on, ASSISTments provides learning analytics that can quickly and formatively help teachers address any issues or misconceptions students may have in realtime. Sometimes teachers realize that since a large number of students got a problem wrong, the instructions might

have been confusing or insufficient. Thus, remediating actions will take place to prevent poor performance by students in the future.

ASSISTments' effectiveness has been examined through a large-scale randomized controlled trial (Roschelle et al., 2016). The sample participants in this study were ~2,800, 7th-grade students from 44 schools in Maine. The schools were randomly assigned to a treatment condition (using ASSISTments) and a comparison condition (existing methods for teaching math) for two years. The main findings include: (1) teachers reliably changed their teaching practice if most of the students got the concepts wrong; (2) students reliably learned more using ASSISTments compared to those who learned using traditional methods; and (3) ASSISTments had a greater impact on students with low prior math knowledge (i.e., it closed the achievement gaps).

The ASSISTments team has been paying close attention to how to help teachers with evaluating students' responses to open-ended questions. So far, there have been 45 million times that a student answered an open-ended question, 6 million times a teacher graded those responses, and 1 million times a teacher gave a comment to those responses in ASSISTments. One solution to help teachers in responding to students' responses is similar to Google smart responses (i.e., the teacher can choose a crafted response which is relevant to the student's answer and quickly send it back to the student). Using reinforcement learning and large language models, ASSISTments provides three appropriate suggested messages teachers can choose from after they grade a student's response (Erickson et al., 2020; Baral et al., 2021; Zhang et al., 2022). This is a safe use of LLMs with a human (teacher) in the loop vetting the messages created by LLMs. In a similar effort, the ASSISTments team used GPT-3 to generate explanations based on teachers' explanations and then asked other teachers to rate those AI-generated explanations (Prihar et al., 2023). Results showed the AI-generated explanations were unable to outperform teacher written explanations. These findings could be improved using the more updated GPT models.

Nex, we summarize the rich discussions that happened in the second part of the workshop in each panel.

Panel Discussions

Partnership of AI and Humans: Augmented Intelligence

The first panel included Kristy Boyer from the University of Florida, Kristen DiCerbo from Khan Academy, and Neil Heffernan from ASSISTments and Worcester Polytechnic Institute. The panelists started by stating their definition of AI. The three panelists described AI from a similar perspective—how we use AI defines AI and our definition of AI changes over time depending on how we use AI to augment our capabilities. Heffernan pointed out that "... I don't believe in AI as AI is nothing but [a series] of tricks. For example, Chat GPT tries to predict the best next word in a sentence and this makes it look like it understands natural language."

DiCerbo indicated that “AI acts like what we ascribe to human intelligence. But I think our definition [of AI] changes over time. For example, in the 1970s correcting grammar errors was seen as intelligent. But now those grammar checkers are not seen as intelligent. I think there is a moving goal post there where we are judging whether something is intelligent or not.” Boyer, similar to Heffernan, also mentioned that “... there is no such thing [as AI]... that is, it is not a thing that we can draw boundaries around and say [things outside of these boundaries] are not AI and this *thing* is AI. It is how we use AI to augment our problem solving and our goals that defines what AI is.”

The conversation in the panel, then, revolved around how the panelists’ work relates to human-AI partnership. The three panelists explained their thoughts from AI helping teachers and students as a partner to AI helping content creators and saving them a lot of time. Heffernan described how ASSISTments uses AI as a partner for teachers. He pointed out that “we usually do not call our system *AI system* as it might set high expectations for teachers. We use our system as a partner to the teacher, not as an Intelligent Tutoring System that cares about one student because in education we care about optimizing [teaching and learning] for 25 kids in the room and there is already a human there [that we can support using AI].”

DiCerbo, talked about another human-AI interaction that Khan Academy uses: content creators use of Chat GPT. She points out that “other than helping teachers, we also have a team of content creators that is writing questions and writing articles and different domains found that different [AI] tools (e.g., Chat GPT) are differentially helpful for them. For example, our financial literacy content creator is using [AI tools] the most. She finds [AI] very useful in creating scenarios around financial literacy problems that people would have... [our content creators] reported that using these [AI] tools saves them 30% to 40% time that it would take to create a passage or an article. However, our math and science [content creators] still are not using these tools as much. For example, right now [these tools] are not good at making graphics. If you ask these tools to create a picture of cells and label the parts, they don’t do that very well. But I expect in less time that we think [these tools] will get there and we can use them in these domains as well.”

Boyer talked about two projects in her lab in collaboration with researchers from North Carolina State University (NCSU) and Georgia Tech University (GTU) in how they use AI to help students. In her collaboration with NCSU, she investigates how pedagogical agents can help students (e.g., fifth graders) in how to be better collaborators through use of better dialogues (e.g., asking why questions, taking turns in coding, accepting other team mates’ ideas) in the context of educational games. She pointed out that “like Neil, we also don’t call them AI agents because we don’t want to confound what [AI] means to everybody... Therefore, we are designing these agents that are not anything like what kids see in the rest of their lives [to avoid misinterpretations]. We found encouraging results. Kids do seem to respond to the kinds of modeling that these agents can do (e.g., adopt deeper questions and do a little bit better at turn taking).” Then, Boyer explained the other project in collaboration with researchers from GTU using a platform called EarSketch which is designed for highschool students to write computer

code and remix music. Boyer explained creating a co-creative AI agent in EarSketch “that acts as a partner to highschool learners and we are conducting a lot of usability and HCI investigations on how to create an AI agent as a partner and not as an authority figure and not as less knowledgeable. But similar knowledge and maybe with some complimentary knowledge and abilities to what students have.”

Through several questions by the audience in the room, the panelists reflect on the potential of tools such as Chat GPT and the challenges in front of us using such tools. Boyer indicated that the tool is very promising. However, “it will be very interesting to see how the tool itself evolves and how our paradigms of pedagogy and design evolve with it. What was discussed in this workshop show [how we should bring in the human in the loop] by empowering the learners to give feedback (e.g., by asking what do you want? Did it get it right? If not why?) and avoid some of the missteps.” Boyer continued by pointing out the importance of giving agency to learners rather than taking it away from them even with the best intentions. “Historically, we as a field have taken away learners’ agency and the opportunity to give us feedback about what was wrong and right for them. We specially excluded the considerations of [learners’] culture and knowledge from their own background. That is something I hope we don’t take into the next chapter of using AI to support learners.” This point reminds us that we still need to do a lot of due diligence when using AI tools such as Chat GPT and don’t think that such tools can be used blindly to solve students’ problems. Reinforcement learning methods with a human in the loop can help these systems work more accurately. ASSISTments and Khan Academy have done and continue to do some reinforcement learning using Chat GPT to address this issue.

This issue (i.e., no human in the loop with AI tools) can be very problematic especially in self-directed learning environments (e.g., MOOCs). Khan Academy has been used both in classrooms and in a self-directed learning manner. DiCerbo points out that “[AI] makes stuff up and makes mistakes sometimes. If there is no teacher there to provide some guidance and oversight it can be dangerous sometimes in a self-directed learning environment.” Boyer responds to the same issue by saying that “we should not totally embrace Chat GPT as we should not embrace human teachers as they can get it wrong all the time. However, we can dialogue with the teachers and ask for reasons why they got it wrong. However, the extent to which we can dialogue with AI is still much more limited. And the extent to which things can go in novel scenarios as these models evolve, there is just so many unknowns. I think the toe-in-the-water kind of piloting with a lot of direct research involvement is the responsible way to go.”

Next, the panelists reflected on a question regarding how AI can be used to address the needs of different learners and incorporate individual differences. DiCerbo talked about a dream. She elaborated that “in the future, there will be more to the use of AI than the conversation with GPT and the prompt engineering we do (e.g., telling more to GPT about the context). For example, there will be more connection between the student profile and student model and the AI tool. A type of connection with more depth than typical student models that provide students’ proficiency levels on different skills. For instance, the student model will include students’

reading levels and other dispositions and even their interests included and influence what the interactions with the AI system will look like.” Heffernan also described their recent work on investigating if the interventions using ASSISTments are helping or harming students with special needs. We need to examine our AI tools to see how these tools work for learners from different backgrounds and demographics.

AI Education

The AI Education panel included Lin Lin Lipsmeyer, Victor Lee, and George Siemens. The points of discussion in this panel revolved around several essential topics such as factors that influence people’s trust in AI and how education can help address these factors, transparency in AI, addressing the public’s misconceptions about AI and how it works with AI education, competencies and literacies that we as educators need to cultivate in our students with the dominance of AI, and finally the issue of digital divide that can threaten education.

Addressing the issue of trust in AI, Victor Lee pointed out that “why and how we trust things, in part, comes with some amount of transparency and some amount of working as we predict and expect in service of the goals that we are trying to achieve.” Lee continues with a twist and steps one step back by asking us to before thinking about how we can help people trust AI, to ask ourselves “how we get people who are building AI to build things that are worthy of our trust?” He continued by saying that if people, in general, are educated about what expectations they need to have from an AI system, they can demand the AI tools to continue getting better at trustworthiness.

George Siemens mentioned that “[trust in AI] depends on what we want AI to do for us in educational systems. If I just want, for example, to interact with AI systems such as Chat GPT for idea generation or to build on an essay, I trust it. I know the limitations of it [but I still benefit from it]. But it becomes more problematic if [AI] is part of a system in the education system. For example, as it came up in the iSAT discussion, if AI will be a part of social and emotional dynamics, it is going to be a different trust relationship. I think we should think of trust [in AI] not as a broad system, but as a series of subsystems related to what is being done.” Siemens provides an example of a case where a letter was written by the administrations of a large university in response to a shooting incident using Chat GPT. This letter created a lot of criticism by that community. This example illustrates how AI could be rejected or not trusted by humans when it expresses emotional care. Pointing at what Lee also mentioned, Siemens agrees that the first step is to build trustworthy AI systems by integrating a series of components depending on the task at hand.

Agreeing with what Lee and Siemens pointed out Lin Lipsmeyer indicated that “[trust in AI] is going to evolve; what we trust in now and what we will trust in the future in different contexts. We currently think of AI as a foreigner or an alien. But as we move forward we continue to interact with AI [systems] the way we understand trust in AI will evolve [and we need to be able to be flexible and agile in shaping that trust]. This is an important point to keep in mind when working on AI Education, that what we educate people about AI today may be

obsolete in a couple of years. Therefore, educating people on having the right expectations from AI systems is a key factor. We, as humans, should stay up to date with AI-related developments and expand and adjust how we think of trusting AI moving forward.

The conversation continued around an essential aspect of AI in the future: *transparency*. Siemens pointed out that for us, educators, researchers, students, transparency in using these AI tools will be less and less in our control and more controlled by the big tech companies such as Open AI. These companies are the ones training AI models and then creating AI technologies based on those models. “As those models get more sophisticated we are going to have a less of an important role in the training of those [models]. We are in a small window of agency, where we have some control of [transparency]. The [big tech] companies have an interest in not being transparent and that is a growing concern that we need to confront with networking, collaborating, and structuring with other organizations and demand a reasonable level of transparency,” Siemens mentioned. However, Siemens pointed out that there are some platforms such as Hugging Face that have been improving open access and open source for people. But the big tech AI models and technologies that we are most excited about are not transparent.

Lee looked at transparency from an AI Education perspective. He pointed out that “I think there are going to be different degrees of transparency that are going to be functionally more or less important for different populations.” Lee indicated that the level of transparency that Siemens talked about (e.g., seeing every single line of code) may not be useful for end users (e.g., students) but it is necessary to be there as an option for some technical population. Lee talked about an alternative proposed term instead of AI: SALAMI proposed by Emily Bender and colleagues which stands for Systematic Approaches to Learning Algorithms and Machine Inferences (SALAMI). Such terms dismantle the notion that AI is a smart conscious being and redirects our attention to what actually AI is in a transparent manner. Lee mentioned that “even if [AI] has an icon or avatar associated with it and can do some fun personas on request, it basically is a probabilistic word spitter out.” So, demystifying that can add to everyday functional transparency of what AI is and is not.

Similar to Lee, Lin Lipsmeyer emphasized the importance of context when it comes to transparency. For example, “[transparency in how AI works] is different from childrens’ perspective to adults’ perspective and to corporates’ perspective.” From an educational perspective, “it is great to have these conversations, but on a larger scale and platform, from the general public’s perspective a lot of times things are driven by functions and how [these technologies] can help their daily lives. They don’t care a s much about all the issues [around AI transparency]. So, the transparency issue depends on people’s demands and needs, but for the educational setting, it is important to have these conversations around [transparency in AI].”

With appropriate degree of transparency for various populations, the misconceptions about AI can be demystified and we can prevent new misconceptions about what AI is and how it works. Siemens and Lee pointed out that sense making conversations (such as this event) can help us shape our understanding of this technology. Also, broadening our understanding of the variety of AI technologies (not just Chat GPT, but other technologies such as movie

recommendations on streaming services, facial filters on social media, etc.) can help us avoid misconceptions about AI. Therefore, AI literacy can play an important role in demystifying misconceptions and preventing new misconceptions as these technologies evolve. Lin Lipmeyer added that we need to also think about misinformation and disinformation that can be amplified rapidly using AI technology; another challenge that can be prevented if people have a high level of AI literacy. For example, if one knows that Chat GPT can make things up, they would be cautious in how they accept what it produces.

Another point of discussion in this panel was around the topic of the competencies and literacies we, as educators, need to equip learners with to help them be productive global citizens. Siemens argued that there are two layers of literacies we can focus on: the short term layer which is basically about “you need to know what to do with [AI and how it works]. For example, if you want to download a language model on your computer, do you know what it is and how it works? What are the safety issues? Do you know what it is good for? What are its limitations? What is it bad at? Additionally, literacies about using [AI] in meaningful ways and what happens behind the scenes.” The long term, future focus literacies, as Siemens suggested, are related to “the questions of what is it about the system that needs to be reconceptualized in light of the affordances of these technologies?” That is, “if AI can outperform us in some areas, what is it that we should be doing? This type of literacy can be called “literacy for living in AI which is related to the capacity to think and live in networks, the ability to see systemic relationships and interconnectedness of things, to conceptualize new roles of humans in a world where we have a lot of these sources of knowledge at our fingertips.” Lee added the importance of data literacy and said that “garbage in; garbage out! It is important to know where the data came from, how the data was captured, and how that product of what we get now [in AI technologies].”

Finally, the panelists talked about the digital divide issue that AI technologies might exacerbate. What can we do to address this issue? George Siemens responded that “that is a question that we are all aware of, but it feels very inadequate when we try to answer it. Almost always new technology produces new barriers (e.g., cost). For example, the first version of Chat GPT is free, but for preferred access you need to pay \$20 a month. So, if one wants to use that tool in their curriculum they have to consider that issue.” The economical barriers could be removed at scale by some means (e.g., using business models that social media use with ads and make things free). But the second issue relates to supporting individuals with regard to how to use these tools properly. Siemens ascertained that some people (students) who come across these tools might need support about prompt engineering, for example, and they do not have the social support to go to for their questions. The third issue relates to the bias that exists in these models and how they were trained. Siemens pointed out that “[these models] will favor the perspectives and the world views of certain individuals and even countries. If you don’t see yourself included in the models, even in subtle ways, that can be another form of alienation as well.”

Lee added that “at least from a domestic US perspective, banning those tools in schools would not help with the digital divide issue. We need to strive closer to the ideal of providing

access [to everyone]. Also, this [i.e., the new AI technologies] is a very profit motivated space and all non profit entities need to have a big say in this (e.g., public libraries or government) to make sure that we can keep this broadly available and accessible through many different routes that are open to all. On a world-wide perspective, we need to help and invest in homegrown and local productions of AI [which can help with the issues of bias] and culture.”

Fundamental Research, Development, and Theory Building

James Lester, Peter Foltz, and Diego Zapata-Rivera formed this panel discussing the latest developments around fundamental research, development, and theory building. The discussions in this panel were mainly about the infrastructure needed for AI in education, the opportunities that LLMs have introduced, use of AI in fundamental educational research, and the interplay between AI models and theories of learning.

The panelists first provided their thoughts on the infrastructure needed to promote AI in education. Zapata-Rivera emphasized the importance of continuous evaluation of the models and developing the AI-powered learning environment that can support learners. Foltz added that “to have the right kind of infrastructure is to be able to capture the rich interactions that are happening with students and teachers. We usually instrument the computers and then capture what the students are doing. But we don’t have the full context (e.g., where they are looking, who they are talking to).” Lester broadened the discussion on a national level and pointed out that “I think there is a real need for a national infrastructure for AI in education. That national infrastructure can make it possible for school districts to take advantage of [the new AI technologies]. Part of that infrastructure is hardware [and software], but also the know-how (the people that can mobilize the technology in various places including education) that comes along with that. One of the reasons for the AI institutes program at NSF is that such infrastructure (in universities, non profit organizations, and even industry) has got to be developed. I believe this development is going to be much faster than I thought it would take before.”

With advances in AI technology, especially related to the large language models such as Chat GPT, researchers are more dependent on large companies for infrastructure and computing power. Foltz addressed this comment by saying that “not everyone needs such computing power to create their own large language models. I think these foundational models are going to be there and many people can finetune them in a number of ways [for their purposes]. There may be some dependencies to the large tech companies for providing those foundational models, but there are some [small scale] models that can be run on your laptop. But in the future we need to have people who are both skilled in learning theories, cognition, and prompt engineering to build these [AI-powered] tools.” Zapata-Rivera added that “these LLMs may not know about the current context of the learner. Then, having approaches that bring humans in the loop and provide information about the context for these models is going to be important.” Lester built on this thought by saying that “in some ways we don’t want these models [from the large tech companies] to have deep knowledge of students. We have to think hard about what is the

interplay of LLMs and what we call student modeling and how do you protect the privacy of students in those systems.”

The panelists discussed the role of AI in advancing educational research. Lester responded by saying that “the most regular positive feedback I received has been from my colleagues in education about things that they have been thinking about doing [in their research] and now they think they can do those things. For example, qualitative analyses of any kind of learning interaction [can benefit from AI]. In the past, researchers would take a three-minute video and spend years on it hand coding [the qualitative data]. Now, with the advent of multimodal learning analytics across the field it is possible to do such analyses using NLP. We are now going to be able to answer the old questions, and we are going to be able to ask new questions [and answer them using the new AI technologies].” Zapata-Rivera provided an example in the use of AI in assessment. “We were working on a conversation-based assessment and we used a tutor as the dialogue engine to create the conversations. Then [models] and humans needed to go under training to be able to generate new conversations. Now, with Chat GPT we are able to create conversations much faster.” Foltz agreed with Laster and Zapata-Rivera and added that “I am very excited to use AI to test some of the theories that were not testable before. For example, as we move toward higher-level thinking skills (e.g., collaboration, communication, critical thinking). In the past, someone would read all [conversational, qualitative data] and make some kind of judgment [about those skills]. Now, we have the ability to use AI and go through [the data] and capture something new. We can now analyze students’ discourse and test hypotheses about the effects of the discourses on students’ knowledge gain or even affective response. In theory, if we could have large data sets, we can make some generalizations [and build new theories] or test theories that have been around for a long time but never have been formally tested in this way.”

One of the criticisms of AI models is that those models mostly are not guided by learning theories and are not explainable—i.e., they act as a black box. The panel discussed this issue with regard to the interplay between AI and learning theories. Foltz pointed out that “one of the great things about AI is that you can predict some learning outcome measure using a lot of data independent of any learning theories in a black box manner. You may be able to look at some of the features and tie them back to a [learning theory]. In the field of Computer Science if you can improve your model by 3% you can get published regardless of including a learning theory. The other side is to start by understanding the constructs we want to measure and how they want to be put together and try to understand the theories behind them and how we want to put that in AI models. Explainable AI (XAI) is trying to do this, but it is not there yet.” Zapata-Rivera added that “using hybrid approaches (i.e., top-down [approaches that start from the constructs and get to the data] and bottom-up [approaches that find patterns in the data]) can help improve the AI models. Sometimes learning from data can improve top-down models.” Lester talked about the recent advancements in research related to self-regulated learning and emotion regulation (Lester et al., 2013) and pointed out that “research in SRL and emotion regulation is leveraging the new technology (i.e., AI) to a great extent in a theory-driven way.” Laster also talked about XAI by

saying that “people have been working on [XAI] for about seven years, and I am really not seeing progress. I don’t think weak, tenuous ties of particular features in the data to particular predictions is not satisfying. I feel that in the same way that industry is driving forward advancement in the capabilities of the models, they are also going to drive forward the explainability of the models too and we, in education, are going to benefit from it.”

AI in Assessment and Improving Learning

In this workshop, we had two separate panels on the topic of AI in Assessment and Improving Learning including Valerie Shute, Ryan Baker, and Russell Almond on one panel, and Sidney D’Mello, Alina von Davier, and Richard Golden on another panel. The panelists in both rooms talked about issues around validity, reliability, fairness and equity of AI-driven models in advanced learning environments.

Addressing the issue of validity and reliability, Shute first defined the method that she introduced in her career called *stealth assessment* as “an assessment method that gets embedded in a technology-rich learning environment (e.g., games) and uses evidence-centered design (ECD; Mislevy et al., 2003) that uses various models such as the competency model (what are you interested in measuring?), evidence model (how do you want to measure it?), task model (what tasks and specific features do you want to design?), and assembly model (how do you want to put them all together?). In stealth assessment we always start by establishing the concurrent or convergent validity of a new competency we want to assess (e.g., persistence, physics). Usually, we look at the correlation between the stealth assessment and a validated external measure on various grain sizes. For reliability, we use confirmatory factor analysis, inter class correlations, and Chorbach’s Alpha to figure out the reliability of the actions that the players take in games we use.

Almond drew an analogy between ML and AI approaches and classical statistical methods. He pointed out that “supervised learning is like confirmatory factor analysis and unsupervised learning is like exploratory factor analysis. In supervised learning, you need to start by some expert-driven identification (i.e., expert labeling) of what kinds of performances are indicative of people having the skills that you are trying to assess. In the unsupervised learning (i.e., exploratory approach), you need to also have the experts to look at the exploratory factors that the model uncovers and label them respectively.” Therefore, experts’ opinions can be a means to ensure the validity of an assessment or ML model in both supervised and unsupervised learning. Another consideration that Almond pointed out was calculating weights of evidence. That is, “for each piece of evidence that gets into the system (e.g., each test item, each learner interaction in a game or simulation), how much the conclusion of the system changes (i.e., how much information that new piece of evidence adds). Doing so will give you valuable insights about whether or not you are heavily relying on a piece of evidence when you should not or when you are not considering enough weight for some evidence and you should. This method also helps with the classical problem of AI approaches that is the ability of explanation about the models and how the models came to creation.”

Baker, agreeing with what Shute and Almond pointed out, added that “when we have expert labels, we often want to make sure that the experts agree with each other. Therefore, interrater reliability and having a code book [that can help enhance the interrater reliability] can be valuable and address the validity issue.” Baker also talked about another important issue related to generalizability of the models—algorithmic bias. He mentioned that “[you need to think and ask yourself that] are your models secretly less well for some groups and secretly making the world worse by disadvantaging a group of learners that is already disadvantaged? I think it is important, in this era of validity, to make sure that you have collected the data that is needed for your model and that your model can be validated to work for all the people it is going to be used for.”

von Davier also talked about the issues of validity and reliability related to the high stakes tests at Duolingo. She pointed out that “we use traditional approaches for reliability and validity. We do use those approaches regardless of if the item was developed with the use of AI or by humans. Specific to AI-generated items, we added a new review of the items for fact checking because we learned early on that [AI systems such as Chat GPT] make up some information. Therefore, in addition to our traditional human review of the content (i.e., quality, cultural bias, and fairness), we now have fact checking.” Golden brought up a point about the security of high stakes tests that “sometimes students can memorize the test items and share with others” and suggested that “maybe the future direction could be designing stealth assessments that are in collaborative, immersive learning environments.” von Davier responded specifically to the security issue by saying that “AI is helping us to make the cheating or stealing of the items irrelevant because AI can help us generate a lot of items to the point that cheating by stealing the items becomes irrelevant while maintaining the validity of the tests.”

Building on Golden’s point, D’Mello pointed out that “our context is a bit different as we are helping students’ collaborative learning in classrooms through inquiry-based approaches where kids are working with physics sensors and measuring various things. In this context LLMs are essentially useless as you have to develop the AI [models] from scratch. The LLMs, in such contexts, are useful in a sense to just get representations of contextual semantic words, but then you have to ground language in action and gesture [conducted by the learners].” D’Mello continued by bringing up a point about assessments validity and indicated that “one thing we always think about when evaluating the validity of our assessments is the purpose of the assessment. In our context, the purpose of some of the assessments we use is to promote learning. There are multiple dimensions to students’ learning. For example, collaboration, in our context, is a [learning] outcome itself. Sometimes success in collaboration is not about accomplishing the task, but it is about how you, [as the learner], felt about the group dynamics and how you felt about yourself. We are trying to grapple with these issues and address the question of how do you measure collaboration that is authentic with the intent of the collaborators (i.e., what they value and honoring their perception of what is a successful collaboration).”

The discussion followed by a question about issues of fairness and equity. Specifically, as we are moving to integrate LLMs and AI in general into educational technologies, what are the challenges regarding fairness, bias, and equity? Baker addressed this question by saying that “... Open AI has put a lot of effort into Chat GPT to make it bland, pleasant, and unbiased, but that does not mean that those issues don’t creep in in various places. I think the research on how biased these technologies are has not happened yet. We still don’t entirely know how these tools work both in English and other languages. We have some real questions around fairness and equity that I have not seen people to look at.” Shute added that “going back to the stealth assessment work that we did, we have looked at differences between genders and ethnicities and we found that while various groups started in different positions (regarding learning outcomes) they ended up not different in terms of learning and enjoyment. This positive result, we think, has to do with careful design decisions we made every step of the way preventing bias and providing an equitable learning experience for all students.”

von Davier started by saying that “for us, the biggest challenges are related to technology. We want our test takers to be able to reach the high quality, rich items from anywhere they are. For example, we created an item which included an interactive video and our pilot test showed that less than one percent of the test takers had issues with downloading the video and completing the task. We had to think of creative solutions for this problem to resolve the issue and make the item accessible to everyone. This is an example of the type of effort we are putting into creating our tests to make them fair and accessible.”

D’Mello discussed the fairness and equity issues from the perspective of their project in iSAT. He pointed out that “we have been working on a framework to test ML bias in the context of psychological assessment. In the context of psychological ML models we need to check how the models work for different groups of people. One of the challenges we have is non-normative talk in classrooms. Youth of color have to do a lot of code switching when they come into the education system (classrooms) and this might introduce a lot of implicit bias that we don’t pick up on. If we use a model which is trained on colonial ideals about what good collaboration and conversations look like we will inherently miss out on the conversations that students’ of color bring to class without code switching. So, the question is not about if the algorithms are biased; rather it is about the institutions. Therefore, we started to ask ourselves how do we do assessment and learning outside of the institutions and how do we space the institutional gravity to design equitable learning environments. Bias goes beyond the machines; it also has to do with broader societal constructs of what education looks like.”

Next, the panels discussed how AI can be leveraged to assess and possibly improve hard-to-measure constructs (e.g., creativity, collaboration, critical thinking). D’Mello addressed this question by saying that “.. we have done some work in assessing collaboration remotely over Zoom and using speech in open-ended communications. When we took this to classrooms, things suddenly got very difficult as dealing with speech recognition is a huge problem with error rates as much as 80%. Then, integrating multimodal data is also very challenging because a lot of collaboration is occurring through gesture and pointing and we noticed this even more in middle

schoolers. We have been embracing transparent AI. We have been showing to kids this is what AI is saying; do you agree or disagree? What do you think? Let's have a conversation. This acceptance of reality while creating tools to embed other [multimodal data] into LLMs is what we have been doing for collaboration assessment."

Shute addressed the question regarding assessment of hard-to-measure constructs by providing an example about assessing persistence and said that "...when we have multiple indicators of a construct such as persistence we can amalgamate those to have a more accurate estimate of one's persistence. We usually think of persistence in a linear way (i.e., more is better). I have heard that [it is good to be persistent and tenacious] growing up. But in fact what we have been seeing is a quadratic relationship showing more persistence to be good and too much persistence (i.e., wheelspinning) to be not good." Baker added that "persistence is a great example. We often look at these behaviors in isolation, but they are related. One of my students had a paper where she showed that persistence could have positive or negative results also depending on what you did during the persistence. For example, if you persist really hard and use good strategies or even take breaks switching what you have been working on and coming back then you may end up learning more. But if you have been wheelspinning you may not get there. So, it is not just practicing persistence; it is deliberate practice of persistence [that can be helpful]. Related to affective states, we can see the same pattern. Boredom and off task behavior are related; boredom as gaming the system are related, frustration and gaming the system are connected. So, these things are complex and we have to understand them in their contexts or we end up with some over simplified models that say persist good."

The panelists, next, discussed how AI-powered models and technologies could be implemented in both summative and formative ways to improve learning. Almond pointed out that "I hope we get rid of many of the summative assessments we do and replace them with more summary of the many formative assessments we gather over the course of a period. I call it a ubiquitous assessment rather than a stealth assessment. That is, assessing the students all the time and if you are capturing a lot of information about the students you can picture a detailed picture about students' strengths and weaknesses. That type of information about the students then can be used to address their needs in a personalized manner. Not all students learn at the same rate."

D'Mello talked about feedback by saying that "there is a lot of work on feedback. For example, Val [Shute] has done a lot of work on formative feedback many years ago, yet in the current educational system that does seem to be forgotten. It is now good to rethink formative feedback with these new AI systems because some of these formative feedback can be automated using these systems which was hard to do previously. I think one mechanism regarding use of AI in supporting learning is through helping with feedback and reflection in learning. There is a concern, though, in what AI can measure which is in small levels (e.g., dialogue moves and discourse level). For example, we ask authentic, open-ended questions from students to elicit information about their learning. For instance, if I ask Alina what do you think about that book, how did it make you feel? If I don't have a preconceived notion about what she is saying, I will not know what she is saying. However, measuring things at low levels can lead to emulations at

low level (e.g., learners can mirror the open-ended questions and AI may not catch that). What is really needed is at the higher levels (i.e., what is my high-level philosophy of pedagogy?) which I think is an interesting area to pursue as we develop these AI technologies for teaching and learning.”

The panels related to AI in assessment and improving learning ended on an important note made by D’Mello. He asserted that “I would like to look at what AI should or could do and what humans could or should do. This is related to a human-centered view where the human is in the center and AI is in the periphery as opposed to AI-centered view where AI is in the center and humans are dancing around it. If you take a human-centered view, you should think about what it is that human beings are unique at doing. Nurturing and caring for each other is uniquely human. You can have a robot emulate that, but I’d delegate that to humans and not AI. What AI can do is to help teachers to be able to do the nurturing, caring, motivating the students. Teachers usually spend a lot of time doing things that can be automated using AI (e.g., classroom management, grading). AI can free up time and make room for humans to do what they can do best and not try to emulate that in machines.”

AI Policy Challenges and Scale

The last panel included Chris Dede from Harvard University, Danielle McNamara from Arizona State University, and Rene Kizilcec from Cornell University. The panel started by investigating practical ways to ensure students’ learning without violating their privacy. The panelists shared their perspectives while drawing a parallel between education and other areas (e.g., health and market). To address this issue, Dede highlighted the distinction between education and health. He pointed out that “it’s interesting the contrast between education and health because any health information is far more sensitive and dangerous than any educational information sixth graders do on algebra.” He addressed that in spite of the importance of privacy as a fundamental human right, in education, the data sharing is not sensitive and said “I think that scale is a way of solving some of the privacy problems because when we can deliver something at scale it becomes very difficult to narrow down a single individuals data in an enormous amount of say quick stream data that’s coming through so some of it I think is making the case and some of it is more powerful ways of anonymization of data.”

McNamara addressed the complex interplay of privacy in educational data, stating, “There’s a disconnect between how people consider their privacy in the context of something like Google and education.” She highlighted the challenge of navigating legal constraints, such as FERPA, and shared insights into the “Learning at Scale” initiative at ASU, emphasizing the need to consider student privacy with measures such as training and MOUs. McNamara expressed her aspiration, stating, “We dream about a means of putting the data behind a certain wall where you can do analysis and extract aggregate data but not actually see the private data.” Expressing ongoing frustration, McNamara pointed out the disproportionate focus on safeguarding educational data compared to personal information in other contexts. She underscored the necessity of using private data to improve education, stating, “At some point,

we're going to realize that we have to use private data to improve education.” Similarly, Rene stated that “In medicine, the electronic health record (EHR) was revolutionary, enabling large-scale analysis that changed the way we treat diseases and fueled the recent wave of AI applications. Distinguishing research-focused use cases is crucial, similar to education.” He highlighted the diverse stakeholders in education, noting, “While researchers use data for the science of learning, others, as Chris mentioned, may use it for marketing services or employers mining it for hiring decisions.” Rene stressed, “We need to separate those two use cases clearly and make a case, as Chris mentioned and Danielle too, for why it is important for researchers to have access to this data and the benefits that can provide to students.” Looking ahead, Rene emphasized the potential risks, stating, “If this kind of data leaks and becomes publicly available, employers may unfairly target individuals, leading to unintended consequences for students applying for employment.” Drawing parallels with medicine, he urged the need to separate use cases, acknowledging the importance of providing researchers access to the data for its potential benefits to students.

The panel addressed attendees’ questions, beginning with a question on universities’ efforts in generative AI and ethical applications. McNamara emphasized ASU’s commitment to creatively integrating AI into the learning process through innovative approaches. Rene highlighted that some universities employ task forces to handle ethical concerns in AI use for assessments and admissions, emphasizing fairness and equity. The second question delved into the concerns and aspirations surrounding AI in education. Dede emphasized the dual nature of AI as a potential double-edged sword, expressing apprehension about its current application. He underscored the risk of using AI merely as a faster means to achieve the wrong goals in education. Dede’s hope lies in reshaping educational outcomes by recognizing the uniqueness of human contributions over AI. McNamara expressed concern about the overwhelming volume of information and highlighted her hope for finding effective ways to manage it. Additionally, Rene emphasized the importance of guiding young individuals to contribute value in an era dominated by AI. He stressed the need for appropriate collaboration with AI and expressed hope that AI could alleviate humans from repetitive tasks, ultimately benefiting mental health.

The panel explored core competencies for students and instructors in the age of AI, posing the question of what skills are essential. McNamara shared her vision, stating, “I have a dream of developing a mindset or an entire discipline of learning engineers in K-12.” She acknowledged the challenge of expecting teachers to become technologists, emphasizing, “It’s not fair to ask teachers to be everything; they entered teaching to interact personally with students.” McNamara highlighted the need for a shift in mindset, drawing parallels to psychology, stating, “We need to think about education as a science, just as psychology transitioned from an intuitive to a scientific approach.” She proposed a model where teachers and learning engineers collaborate, recognizing the distinct specialties required for effective education in the age of AI.

To address this question, Dede expressed concern about the misuse of AI, stating, “There’s a lot of misuse of AI going on that doesn’t involve a lack of understanding of the

technical aspects.” He illustrated this with an example, likening the situation to requesting a recommendation letter through an intermediary. Dede emphasized the need for thoughtful consideration, stating, “This kind of differentiation between poor and good use of AI isn’t a technical issue; it’s about fundamental thought about the nature of the task.” He criticized proposals simplifying work in ways that are inhumane, urging a shift in mindset to ensure the ethical and humane application of AI in education and beyond.

Rene emphasized, “It’s less of a technical problem; it’s a question of how we use it and teach the next generation the right ways to use it.” Reflecting on the internet as an apparent equalizer, he highlighted disparities in using resources due to varying opportunities and contexts. Kizilcec questioned assumptions about building skills, citing the example of AI correcting spelling mistakes to level the playing field. He considered, “What are the places we want to teach students mastery skills, and where should we embed AI to level the playing field from the start?” Acknowledging the potential impact on social demographics, Kizilcec challenged the balance between teaching specific skills and integrating AI to bridge existing gaps. He identified an opportunity for universal access through technology, stating, “Anybody can use this technology to get help, leveling the playing field.” Kizilcec underscored the importance of embedding AI to advance opportunities for those historically underserved, ensuring it becomes a universal tool rather than contributing to disparities based on means.

Panel continued to answer the question about the ban on the fear of AI chat apps on some college campuses. Should we embrace them fully or set guidelines for adoptions? McNamara addressed concerns about the use of AI chat apps on college campuses, emphasizing the need for guidelines rather than an outright ban. She advocated for providing instructors with instructions on distinguishing between good and bad use, asserting that understanding how to use such technology prevents being its victim. Drawing an analogy, she likened the situation to not banning cars due to their potential danger but instead promoting responsible usage through guidelines. Hence, the approach should be to embrace AI chat apps while establishing clear guidelines for their adoption.

In their final thoughts, panelists shared their perspectives. Kizilcec expressed hope despite challenges, emphasizing that those at the forefront are carefully considering implications for students, communities, privacy, equity, and the future of work and learning. McNamara highlighted transformative works facilitated by institutions founded by NSF, stressing the importance of recognizing and appreciating these projects. Dede talked about the dilemma of educational innovations, questioning whether “they make things better or do better things.” He urged a shift in the use of AI from merely improving processes to creating better outcomes, particularly addressing the achievement gap by focusing on enrichment rather than remediation. Overall, the panelists conveyed optimism, thoughtful consideration of implications, and the need for transformative approaches in the realm of AI in education.

Panel Discussions' Key Takeaways

Exploring the diverse landscape of AI in education, panel discussions covered crucial themes ranging from human-centered AI, use of AI in assessment, AI education and AI literacy, and policy and challenges to privacy concerns and the transformative potential of AI. This workshop's esteemed panelists offered insightful perspectives on shaping the future of education, navigating legal constraints, and envisioning innovative approaches to learning. In this part (Table 1), we present key takeaways from the panel discussions, providing a comprehensive overview of the discussions and their implications for the field of education.

Table 1. *Key Takeaways from the Panel Discussions.*

Panel	Key Takeaways
Partnership of AI and Humans: Augmented Intelligence	• The definition of AI evolves with its use. • AI is a tool that augments human capabilities rather than an independent, conscious entity. • Despite the promise of AI, there are challenges to consider, such as the need for human oversight and the potential risks of relying solely on AI in self-directed learning environments. • Learner agency should be maintained and cultural sensitivity should be considered when implementing AI in education. • AI tools should be evaluated to ensure they meet the diverse needs of learners, including those with special needs.
AI Education	• AI literacy is essential for demystifying misconceptions and preventing the formation of new ones. • Educators need to broaden their understanding beyond popular AI technologies and address misinformation and disinformation amplified by AI. • Trust in AI depends on transparency and alignment with users' goals. Transparency in AI usage varies based on context and user needs. While complete transparency may not be necessary for end users, it is essential for technical populations. • Educating both developers and users about the capabilities and limitations of AI systems is crucial for building trust. • Educators should focus on short-term and long-term literacies related to AI. Short-term literacy includes understanding AI capabilities, limitations, and data literacy. Long-term literacy involves conceptualizing new roles for humans in an AI-driven world.

	AI technologies may exacerbate the digital divide due to economic barriers, lack of support for proper usage, and biases in AI models. Solutions include providing access to AI tools, supporting local productions of AI, and addressing biases in AI models.
Fundamental Research, Development, and Theory Building	- A national infrastructure for AI in education could facilitate the adoption of new technologies in school districts. - While LLMs offer significant opportunities, researchers are dependent on large tech companies for infrastructure and computing power. - Not all AI applications require such extensive resources, and there is a need for skilled individuals to leverage AI for educational purposes. - AI technologies have the potential to revolutionize educational research by enabling qualitative analyses, multimodal learning analytics, and faster assessment creation. - Researchers can now ask new questions and test hypotheses that were previously challenging or impossible to explore. - While AI models may not always align with traditional learning theories, they offer opportunities to predict learning outcomes and explore new constructs. - Hybrid approaches combining top-down and bottom-up methodologies can enhance AI models, while explainable AI aims to make AI systems more transparent and interpretable. - Despite ongoing efforts in XAI, progress has been slow, and there is a need for advancements in understanding the relationship between features in data and model predictions. - Industry advancements in AI capabilities are expected to drive improvements in model explainability, benefiting educational research and practice.
AI in Assessment and Improving Learning	- Validity and reliability should be carefully considered in AI-driven assessment models. - Methods such as stealth assessment and evidence-centered design could be used to ensure the validity and reliability of assessments. - Researchers should address bias and equity issues in AI-driven assessments using techniques such as expert review, fact-checking, and diverse representation in data to mitigate bias and ensure fairness in assessments. - Assessing hard-to-measure constructs such as creativity, collaboration, and persistence is challenging—AI can enhance our capabilities in this area.

	● It is important to consider multiple indicators and contextual factors in assessing constructs accurately. ● We need to have a human-centered approach to AI in education rather than an AI-centered approach. ● We should leverage AI to support teachers in tasks such as classroom management and grading, allowing them to focus on the unique aspects of human interaction and care in education.
AI Policy Challenges and Scale	● It is important to ensure students' privacy while leveraging data for educational purposes. ● AI in Education can learn from AI in Health (e.g., scale and anonymization to address privacy issues effectively). ● The panelists emphasized the need for measures such as training and Memorandums of Understanding (MOUs) to ensure compliance with privacy regulations. ● While recognizing AI's potential to enhance learning outcomes, AI can be misused, therefore ethical and responsible use AI should be promoted. ● There are essential skills and competencies for students and educators in the age of AI that we need to identify and establish. ● New roles such as "learning engineers" are needed to complement teachers' roles in personalizing learning experiences. ● When it comes to use of AI chat apps on college campuses, we should establish clear guidelines rather than outright bans. Guidelines that promote responsible usage and understanding how to leverage such technology effectively and ethically. ● AI technologies can be transformative in education, however, thoughtful consideration of its implications are necessary by all stakeholders. ● A shift in focus from merely improving processes to creating better educational outcomes is needed, particularly addressing equity and enrichment.

Conclusion

The AI in Education Workshop, held online in April 2023 at the University of Florida, gathered scholars to explore AI-driven education advancements and challenges. The workshop facilitated discussions on AI applications in education, from research to practical implementations and policy. Key accomplishments included identifying themes such as human-AI partnerships, AI literacy, infrastructure needs, and ethical implications. Additionally, discussions emphasized human oversight, cultural sensitivity, and addressing the digital divide.

Successful human-AI collaboration models, such as those demonstrated by ASSISTments and Khan Academy.

Since April 2023, researchers have developed guidelines for generative AI technologies. Now that these guidelines are established, it is time to refine them and implement them effectively. However, AI in Education scholars must remain vigilant to rapid advancements, particularly in industry. Looking ahead, the evolution of AI technologies, including advancements in quantum computing and computational power, along with the integration of vast datasets, promises to further revolutionize the educational landscape. These technological leaps could lead to more sophisticated AI models capable of providing deeper insights into learning processes and student engagement, thereby refining educational practices and outcomes at all levels. With quantum computing-powered AI on the horizon, educators face several open questions: “What educational problems can be addressed by such powerful AI technologies?” “What challenges may arise that we need to be prepared for?” Several next steps emerged from the workshop:

- Development of AI Literacy and Trust: Create AI literacy programs for students and educators, focusing on practical skills, ethical considerations (e.g., Akgun et al., 2022) , and understanding AI’s limitations and biases.
- National Infrastructure for AI in Education: Establish a robust national infrastructure, including hardware, software, and training programs, to support AI integration in school districts.
- Addressing the Digital Divide: Develop business models and smaller, accessible AI models to lower economic barriers and ensure equitable access to AI technologies.
- Integration of Learning Theories in AI Models: Focus research on incorporating learning theories into AI models through hybrid approaches, enhancing their educational value and explainability.
- Ethical and Responsible AI Use: Develop guidelines to promote ethical AI use in education, ensuring privacy, addressing bias, and considering societal implications.
- Ongoing Evaluation and Adaptation: Continuously evaluate and adapt AI tools to evolving educational needs, incorporating new findings and technologies.

Acknowledgments

We would like to thank the Institute for Advanced Learning Technologies (IALT), and the E-Learning, Technology and Communications teams at the University of Florida who helped in various stages of conducting this workshop. We also would like to thank everyone who helped in planning, conducting, and moderating the panels in this workshop: Tom Dana, Pasha Antonenko, Jason Dean Arnold, Albert Ritzhoup, Corrine Huggins-Manley, Walter Leite, Anthony Botelho, Jinnie Shin, Wanli Xing, Robe Moore, Kevin Coulson, Alexis Moore, Moxy Moczygemba, Hailey Kuang, Yukyeong Song, Andrea Ramirez-Salgado, Brian Abramowitz, Hyunju Oh, Hanxiang Du, Maedeh Agharazi, Woorin Hwang, Ray Opoku, Lissette Tolentino, Ziyang Li, Amber Hatch, and Priya Ganapathy.

References

- Akgun, S., & Greenhow, C. (2022). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI and Ethics*, 2(3), 431-440.
- Arieli-Attali, M., Ward, S., Thomas, J., Deonovic, B., & Von Davier, A. A. (2019). The expanded evidence-centered design (e-ECD) for learning and assessment systems: A framework for incorporating learning goals and processes within assessment design. *Frontiers in psychology*, 10, 441982.
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58, 82-115.
- Baral, S., Botelho, A. F., Erickson, J. A., Benachamardi, P., & Heffernan, N. T. (2021). Improving Automated Scoring of Student Open Responses in Mathematics. *International Educational Data Mining Society*.
- Berry, M. W., Mohamed, A., & Yap, B. W. (Eds.). (2019). *Supervised and unsupervised learning for data science*. Springer Nature.
- Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, 13(6), 4-16.
- Boyens, J., Smith, A., Bartol, N., Winkler, K., Holbrook, A., & Fallon, M. (2021). *Cyber Supply Chain Risk Management Practices for Systems and Organizations* (No. NIST Special Publication (SP) 800-161 Rev. 1 (Withdrawn)). National Institute of Standards and Technology.
- Brożek, B., Furman, M., Jakubiec, M., & Kucharzyk, B. (2023). The black box problem revisited. Real and imaginary challenges for automated legal decision making. *Artificial Intelligence and Law*, 1-14.
- Buchanan, B. G. (2005). A (very) brief history of artificial intelligence. *Ai Magazine*, 26(4), 53-53.
- Burstein, J., LaFlair, G. T., Kunnan, A. J., & von Davier, A. A. (2021). A theoretical assessment ecosystem for a digital-first assessment: The Duolingo English Test. *DRR-21-04*.
- Cao, J., Ganesh, A., Cai, J., Southwell, R., Perkoff, E. M., Regan, M., ... & D'Mello, S. (2023, June). A comparative analysis of automatic speech recognition errors in small group classroom discourse. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization* (pp. 250-262).
- Castelli, M., Manzoni, L., Mariot, L., Nobile, M. S., & Tangherloni, A. (2022). Salp swarm optimization: a critical review. *Expert Systems with Applications*, 189, 116029.

- Chassignol, M., Khoroshavin, A., Klimova, A., & Bilyatdinova, A. (2018). Artificial Intelligence trends in education: a narrative overview. *Procedia Computer Science*, 136, 16-24.
- Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *Ieee Access*, 8, 75264-75278.
- Chiu, T. K., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 4, 100118.
- Crossley, S., & Holmes, L. (2023). Assessing receptive vocabulary using state-of-the-art natural language processing techniques. *Journal of Second Language Studies*, 6(1), 1-28.
- Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higher Education*, 20(1), 22.
- D'Mello, S., Kappas, A., & Gratch, J. (2018). The affective computing approach to affect measurement. *Emotion Review*, 10(2), 174-183.
- Demszky, D., Bush, J. B., D'Mello, S. K., Jacobs, J., Hau, I., Hill, H., Liu, J., Loeb, S., Maples, B., Peppler, K., Pokorny, R., Rascoff, M., Robinson, J., Yeager, D., & Wentworth, L. (2023). *Empowering educators via language technology*. Stanford University.
- Dede, C., & Lidwell, W. (2023). Developing a next-generation model for massive digital learning. *Education Sciences*, 13(8), 845.
- DiCerbo, K. (2021, July). Why Not Go All-In with Artificial Intelligence?. In *International Conference on Human-Computer Interaction* (pp. 361-369). Cham: Springer International Publishing.
- Doherty, E., Spencer, C. A., Eloy, L., Kumar, N., Dickler, R., & Hirshfield, L. (2023, October). Using Speech Patterns to Model the Dimensions of Teamness in Human-Agent Teams. In *Proceedings of the 25th International Conference on Multimodal Interaction* (pp. 640-648).
- Erickson, J. A., Botelho, A. F., McAteer, S., Varatharaj, A., & Heffernan, N. T. (2020, March). The automated grading of student open responses in mathematics. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 615-624).
- Frank, M. R. (2019). The Evolution of AI Research and the Study of Its Social Implications.
- Garrison, D. R., Anderson, T., & Archer, W. (2010). The first decade of the community of inquiry framework: A retrospective. *The internet and higher education*, 13(1-2), 5-9.

- Gustafson, S. M., & Gustafson, D. A. (2016). *Artificial intelligence*. AccessScience. <https://doi.org/10.1036/1097-8542.243600>
- Hassan, Q. F., & Madani, S. A. (Eds.). (2017). Internet of things: Challenges, advances, and applications.
- Hmelo-Silver, C. E., Puntambekar, S., Glazewski, K. D., Lawrence, L., Rummel, N., Aleven, V., ... & Worsley, M. (2022). Artificial Intelligence and Ambitious Learning Practices. In *Proceedings of the 15th International Conference on Computer-Supported Collaborative Learning-CSCL 2022*, pp. 525-532. International Society of the Learning Sciences.
- Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education*, 57(4), 542-570.
- Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., ... & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4).
- Kahn, K., & Winters, N. (2021). Constructionism and AI: A history and possible futures. *British Journal of Educational Technology*, 52(3), 1130-1142.
- Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business horizons*, 62(1), 15-25.
- Kaplan, J. (2016). *Artificial intelligence: What everyone needs to know*. Oxford University Press.
- Kennedy, D. (2016). Cultural intelligence (CQ) as a mirror and path in pre-service teacher candidate preparation. *Teaching and Learning*, 29(2), 51-89.
- Khosravi, H., Denny, P., Moore, S., & Stamper, J. (2023). Learnersourcing in the age of AI: Student, educator and machine partnerships for content creation. *Computers and Education: Artificial Intelligence*, 5, 100151.
- Kim, M. K., Kim, N. J., Haddadian, G., & Heidari, A. (2023). A Test of Learning Progress Models Using an AI-Enabled Knowledge Representation System. In *Proceedings of the 17th International Conference of the Learning Sciences-ICLS 2023*, pp. 986-989. International Society of the Learning Sciences.
- Krstić, L., Aleksić, V., & Krstić, M. (2022). Artificial intelligence in education: A review.
- Staff, T. (2018). 10 Roles For Artificial Intelligence In Education. *Website* [Abruf [24.01. 2019] unter: <https://www.teachthought.com/the-future-of-learning/10-roles-for-artificialintelligence-in-education>].

- Langenfeld, T., Burstein, J., & von Davier, A. A. (2022, May). Digital-First learning and assessment systems for the 21st century. In *Frontiers in Education* (Vol. 7, p. 857604). Frontiers.
- Lester, J. C., Mott, B. W., Robison, J. L., Rowe, J. P., & Shores, L. R. (2013). Supporting self-regulated science learning in narrative-centered learning environments. *International handbook of metacognition and learning technologies*, 471-483.
- Lin, X., Rivenson, Y., Yardimci, N. T., Veli, M., Luo, Y., Jarrahi, M., & Ozcan, A. (2018). All-optical machine learning using diffractive deep neural networks. *Science*, 361(6406), 1004-1008.
- Luckin, R., & Holmes, W. (2016). Intelligence unleashed: An argument for AI in education.
- Luckin, R., & Cukurova, M. (2019). Designing educational technologies in the age of AI: A learning sciences-driven approach. *British Journal of Educational Technology*, 50(6), 2824-2838.
- McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, 27(4), 12-12.
- Memarian, B., & Doleck, T. (2023). A review of assessment for learning with artificial intelligence. *Computers in Human Behavior: Artificial Humans*, 100040.
- Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (2013). *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.
- Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, 1-66.
- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, 2003 (1), i-29.
- Niemi, H., Pea, R. D., & Lu, Y. (2023). *AI in learning: designing the future* (p. 344). Springer Nature.
- Prihar, E., Lee, M., Hopman, M., Kalai, A. T., Vempala, S., Wang, A., ... & Heffernan, N. (2023, June). Comparing different approaches to generating mathematics explanations using large language models. In the *International Conference on Artificial Intelligence in Education* (pp. 290-295). Cham: Springer Nature Switzerland.
- Reddy, Y. C. A. P., Viswanath, P., & Reddy, B. E. (2018). Semi-supervised learning: A brief review. *Int. J. Eng. Technol*, 7(1.8), 81.

- Roschelle, J., Feng, M., Murphy, R. F., & Mason, C. A. (2016). Online mathematics homework increases student achievement. *AERA open*, 2(4), 2332858416673968.
- Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson.
- Saravanan, K., Sreedevi, E., & Subhamathi, V. (2017). A review of Artificial Intelligence systems. *International Journal of Advanced Research in Computer Science*, 8(9).
- Trognon, A., Cherifi, Y. I., Habibi, I., Demange, L., & Prudent, C. (2022). Using machine-learning strategies to solve psychometric problems. *Scientific Reports*, 12(1), 18922.