



AMERICAN
PSYCHOLOGICAL
ASSOCIATION

ESSENTIAL SCIENCE CONVERSATION:

AI IN PSYCHOLOGICAL ASSESSMENT: OPPORTUNITIES, RISKS, AND RESPONSIBLE PRACTICE (SEPTEMBER 24, 2026)

SESSION DESCRIPTION

Artificial Intelligence (AI) is rapidly transforming psychological and educational assessment, supporting tasks such as scoring, summarizing, reporting, and pattern recognition. When applied thoughtfully, AI has the potential to enhance efficiency, improve consistency, and expand access to high-quality assessment services. At the same time, its use raises important concerns related to bias, accuracy, and transparency. The APA Committee on Psychological Tests and Assessment (CPTA) has recently released new recommendations outlining domain-specific considerations and practical resources for AI use in assessment. Join the authors and leading experts for an engaging discussion on how to navigate opportunities and risks, make informed decisions about AI adoption, and uphold best professional practice.

Any opinions, findings, and conclusions or recommendations expressed by the presenters in this webinar do not necessarily reflect the views, policies or positions of the American Psychological Association.

UNOFFICIAL TRANSCRIPT

Shandol Hoover: Hello, everyone, and welcome. Thank you so much for joining today's program, *AI and Psychological Assessment: Opportunities, Risks, and Responsible Practice*. This webinar is part of APA's *Essential Science Conversations* series, which creates space for open dialogue between experts and the broader community on emerging topics in psychological science. I'm Shandol Hoover, Senior Director of APA Science Special Projects and Implementation.

Before we get started, I'd like to share a few quick announcements. Today's session is being recorded, and within two to three weeks, everyone who registered will receive a link to the recording, along with a transcript and additional resources. Please note that any opinions, findings, and conclusions or recommendations expressed by the presenters for today's webinar do not necessarily reflect the views, policies, or positions of the American Psychological Association.

We encourage you to visit apa.org/science to learn more about how to engage with APA Science. You can subscribe to our free *Science Spotlight* newsletter to stay current on news and opportunities for psychological scientists and to our *Editor's Choice* newsletter for curated articles from across APA's journals. We also invite you to explore the Response Center for updates on policy developments and access to helpful resources and support.

We have an exciting lineup of webinars coming this fall, and you can explore more than 70 on-demand sessions available for free and on our website.

Thank you to everyone who submitted questions during today's registration. We'll answer as many as time allows. You can also submit questions throughout today's session using the Q&A feature on your Zoom toolbar. Please note that chat will be reserved for panelists to share information and resources. If you experience any technical issues, let us know via Q&A as well. With that, I'm

pleased to turn things over to Dr. Ludmila Nunes, APA Senior Director for Science, Knowledge, Expertise, and Innovation, who will introduce and moderate today's conversation.

Ludmila Nunes, PhD: Thank you, Shandol. Welcome, everyone, to today's APA Essential Science Conversation, *AI in Psychological Assessment: Opportunities, Risks, and Responsible Practice*. Artificial intelligence, either we like it or not, is transforming psychological assessment. It's expanding efficiency and access while raising critical questions about bias, accuracy, and transparency. In this APA Essential Science Conversation, we are excited to talk with leading experts about how to leverage AI's potential while upholding best professional practice.

We have with us Danielle McNamara, who is the Executive Director of the Learning Engineering Institute and a Regents Professor in the Department of Psychology at Arizona State University, where her work advances scientific knowledge about learning through interdisciplinary research. An internationally recognized scholar, she specializes in cognitive science, AI in education, and educational technology with a strong focus on improving learning outcomes at scale.

We also have with us Liberty Munson, Director of Psychometrics at Microsoft, where she leads efforts to design and validate high-quality assessments that measure technical skills, with expertise in assessment science and data-driven measurement, she has played a key role in advancing innovative approaches to credentialing and evaluation. Liberty serves on the APA Committee on Psychological Tests and Assessment and co-authored APA's report on the *Responsible Use of AI in Assessment*.

She did this with our other guest, Adam Natoli, an Assistant Professor of Quantitative and Clinical Psychology at Montclair State University. His research investigates all aspects of measurement and personality functioning, striving to develop, evaluate, and apply advanced measurement and statistical methodologies to the empirical study of psychometrics, culture's role in personality disorders, multi-method assessment, and applications of artificial intelligence in these domains. Adam, like I said, serves on the APA Committee on Psychological Tests and Assessment and co-authored APA's report on the *Responsible Use of AI in Assessment* with Liberty.

Finally, we have with us Lauren White, who is a Senior Research Manager in Assessment Research and Innovation at WestEd, where she leads projects focused on advancing the design, validity, and use of educational assessments. Her work centers on developing innovative evidence-based approaches to measuring, learning, and supporting equitable outcomes across diverse educational settings. With expertise in assessment design, research methodology, and applied measurement, she contributes to initiatives that bridge research and practice in education systems.

Thank you so much, everyone, for being here. Let's get started with a question that I think all of us want to know the answer. I'm going to maybe ask this to you, Danielle. How is AI fundamentally changing psychological assessment and what shifts should we be paying the most attention to?

Danielle McNamara, PhD: Thank you. Thanks to everyone for being here. I want to particularly do a shout-out to Adam and Liberty for writing the report that they wrote. It's really a signal of how we need to be very careful in the age of AI. I really appreciated their report. Before answering this question, I thought it would be a good idea to ground how we're defining both AI and assessment. I'll just give my definitions or how I think of it, but the panelists may have other definitions and, of course, other people too.

I think it's important to distinguish, in terms of AI, the difference between-- well, let me just say not all AI is created equal. AI is being used as a bucket term to refer to many different things.

We've been using AI in assessment for many years in terms of using machine learning, so that's the ability to take information and create predictive algorithms using nonlinear and linear quantitative methods. Then, also, the ability to extract semantic meaning from discourse, so semantic models. We've been working with those.

Those are more probabilistic, deterministic, stable. They're not going to change over time. That is quite different from what generative AI is. Generative AI is the ability to not only extract meaning from discourse and even visual information now but then generate new information or generate information and responses. Both have changed the picture of assessment across the years.

We've been working on how to use predictive algorithms and how to use the ability to extract meaning from text for quite some time, across my career, at least. Then talking about what do we mean by assessment. We can mean assessment for mental health. We can be referring to assessment for education. We can talk about evaluations that are more summative, formative, or predictive. It's important to think about those distinctions as we think about how will AI fundamentally change psychological assessment.

For me, the door that AI opens up is the personalization and how assessment can be more adaptive and dynamic and responsive to different kinds of modes of assessment. For example, we've used AI for the past 20 years for things like assessing the quality of essays. You extract the language from the essay. You extract the semantic meaning from the essay, and then you can understand various things about the writer and also give indicators to improve writing. The same thing with, for example, conversations with people for various types of reasons. Taking those conversations and extracting meaning, sorting them, clustering them, et cetera.

One of the things I want to say is that AI changes what you can do at scale. We can, one-on-one, personalize assessment. We do that all the time. As an educator, as a therapist or clinical psychologist, you're listening to the person. You are trained to take that information. What AI does is opens up the door to be able to do that at scale because you can create algorithms that allow to extract information.

What AI does is it essentially expands what we can count as assessment evidence. It can move beyond isolated responses and include things like writing, explanations, dialogue, problem-solving processes. Also important is to think about the difference between a episodic assessment. We're quite used to an assessment that is done at a specific time, an event, a survey, an exam.

However, what I see as particularly important is the ability to connect the dots. Connecting these assessments across time to develop a fuller picture of the development of the individual or groups of individuals, change over time, and think and be able to understand patterns better.

Another thing is that AI changes the ability to adapt. We are better able to-- essentially, it is the essential ingredient for personalization or adaptability at scale. In terms of education, if we fundamentally believe that people come in with different things and that they need different things, then a standardized response only gets us part of the way. AI opens up the ability to dynamically adapt to the individual. Personalization is really what I'm talking about.

The types of evaluations, types of assessments, as well as what we do with the assessments. Then you also asked, "What should we be paying the most attention to?" That is going to be a lot of the conversation today about validity and reliability and the false coherence of the potential output, where it may look like it is valid and accurate, but we really have to pay a huge amount of attention

to developing the ability and protocols and pipelines to continuously assess the validity and reliability of these assessments.

I truly think it should be an entire training within psychology and data science, that we really need to prepare students who can specialize in this field and serve our community in that way. I expect that the other panelists can talk a lot more about those topics.

Ludmila: Yes. Thank you. I'm actually going to ask that question to Lauren. Could you expand on how AI is challenging these traditional standards of validity and reliability?

Lauren White, PhD: Sure. I'll do it in the context that Danielle presented, which is the capability that AI opens up to be able to do personalized assessment at scale. When we think about that, and we think about how that relates to validity and reliability, we have to take a step back and think about the assessment context. In what context are you assessing individual or a group of individuals, and what does that mean for your validity and reliability evidence?

For example, if we're talking about measurement in general, a lot of the general thought leadership in the measurement space is having comparability. If we have the capability of AI to personalize across students or across individuals, what does that mean about what comparability means now? I think there's a lot of conversation in the thought leadership space about what is comparability now. Are we changing how we think about what comparability is? Are we moving away from the need for comparability, and what does that mean for the field of measurement at large?

Another thing to consider is what we now define in this space as what is construct-relevant versus construct-irrelevant variance. If we're personalizing assessment experiences at the individual level, how do we define what is relevant to the construct that we're attending to measure, and how do we define what is not? That is an open conversation that I know is happening in a lot of spaces, specifically among the measurement community.

I'm very interested to know what the joint standards for educational and psychological assessment, the new standards that are due to come out hopefully soon, are going to give perhaps guidance in that space as it relates to AI because it'll be really interesting to develop these pieces of guidance and best practice over time.

Also, consequential validity. As we know, AI can sometimes introduce or have pervasive effects on bias that already exist within AI system training sets. What does that mean for consequential validity across groups of individuals of differing lived experiences? Also, what new pieces of validity evidence will we have to have as a part of being in the new space of AI and assessment? What does that mean for our traditional displays of validity evidence that we include in our technical reporting? What other pieces of evidence might we have to display as well?

One of the things that is being talked about in a lot of the spaces where I talk with colleagues is making sure that you're being as transparent in a layman's perspective to the individuals whose data will be used in AI spaces, making that a new piece of validity evidence that we have not necessarily explored before. Then what does that look like?

Those are just a few things that we have to really think about as a community. Then, of course, that validity and reliability are related. If we're having challenges to all of these aspects of validity, are we reliably measuring the things that we are intending to measure? I think that's still an open question and a question that's still being researched.

Ludmila: Yes. These questions are very important for people who are studying these tools and people who are developing them. I have a question for Adam about practitioners. Given all of these possible issues with reliability and validity, how can practitioners evaluate whether an AI they're trying to use for assessment is valid, reliable, and ultimately appropriate?

Adam Natoli, PhD: Yes. I like this question because we're facing a lot of rethinking about what we mean by validity in the era of AI. What do we mean by reliability? The cool thing about this question is that I don't think we really have to rethink anything. I think the answer is actually deceptively simple. I think we need to evaluate AI tools the same way you would evaluate any measurement tool. The same way you would evaluate any measurement instrument you're considering for a client or a study. We address this a bit in our responsible use of AI in assessment document and offer some kind of standardized resources to work through those questions so I won't get into the weeds.

I want to focus on the core question. That core question is whether there is adequate support for using the AI tool given the context for which it will be used. For example, just like you'd ask whether groups your client belongs to were appropriately represented in a measures normative sample before you use those norms. Those are things that are already going through our head automatically. You'd want to know whether those populations were represented in the data used to train the AI tool.

We don't really have to think differently, so to speak. We just have to generalize it to AI. Similarly, it's also important to weigh the risks and benefits of using the tool. Some of these, I think, are newer. What risks are possible, and what gains can be guaranteed for the client or person being assessed? Then considering those in context.

For instance, say you have an AI tool that will summarize your client's health records. You scan in their records or you download the electronic health record and it summarizes. It could do this at the bare minimum of acceptable. This would make your job easier and give you more time to see clients face-to-face, which I think are two valuable goals for everyone. However, you will have less familiarity with your client's health history, possibly interfering with a deeper understanding of your client's experiences and risking a less personalized case conceptualization.

Well, using the AI tool is arguably acceptable if the assessment's objective has little to do with that health history. Perhaps you've been asked to assess acute violence risk. You don't need a deep understanding of their health history to do that well. It's probably inappropriate if health history-- or not probably. It is inappropriate if health history is meaningfully relevant, such as a diagnostic clarification meant to inform treatment. That context also matter. The objective of the assessment also matters for that appropriate use piece.

My recommendation really is just rely on your expertise in psychometrics and instrument selection, the things that you're already doing well, to guide your evaluation of AI tools. Then supplement that with just good available resources to have a little bit of structure in some of the questions that you're asking and have some of the language to understand how different AI tools might be described. Just think the way you've always been thinking about assessment instruments for AI.

Ludmila: Yes, thank you. We have been talking about the statement on AI in assessment. The statement organizes responsible AI use in psychological assessment around eight principles; transparency and accountability, validity and reliability, privacy and confidentiality, informed consent, bias and fairness, competence and training, impact on applied and clinical work, and continuous improvement.

Liberty and Adam, if you want to chime in as well, as part of the APA Committee on Psychological Test and Assessment, you helped develop this statement. Do you think that these eight principles extend to ongoing AI case management, or do they apply only to formal testing like scoring, summarizing, and reporting? I know, Adam, you partially responded to this, [chuckles] but I think this is an important question for practitioners.

Liberty Munson, PhD: Okay. I'll start, and then if Adam wants to chime in, he can. I think the short answer is I think this is more than just around assessment because key question is not where in the workflow the AI is being used. The key question is really about is AI influencing the assessment information, its interpretation, the consequential decisions? If it is, then these principles are important regardless of what you're doing in the assessment world.

That being said, though, these principles are unlikely to be a complete framework, even when we think about it within assessment. I mean, Adam and I did spend a lot of time thinking through what the key principles should be, and I think we have maybe 90% of the space covered with those. It's entirely possible there's more to think about, especially if you look outside the assessment lens, because these were specifically written from an assessment perspective.

There might be other ethical, clinical, legal, organizational considerations that may come into play as AI becomes more deeply embedded in ongoing care and case management, as the example you suggested here. My view is that the principles are not limited to formal testing activities, but they should be seen as a foundation, not the entire governance framework, because it's entirely possible there's other important principles to take into consideration, given the context of the work that you're doing in the assessment space. They're a really good start.

Adam: I completely agree. I think that some of the specifics within the document might not generalize to different practices within psychologists' scope of practice, but I do believe that the core sentiments of these eight principles can be extended. Like Liberty said, I don't think we cover everything in all different applications of AI in psychology, but I think we're covering a good deal of what is probably very important foundational ideas that, again, the core sentiments can translate to a lot of places.

Ludmila: Yes, and related to these, how-- and I'm going to ask you this, Liberty. How do you think psychologists should be thinking about validation, for example, when these AI systems continue to learn and evolve over time? They're always changing and they're becoming different.

Liberty: Danielle actually hinted at this a little bit because we have traditionally operated on this release and maintain model, this one and done, and then just maintain it as it makes sense. You validate a tool, you document its properties, and you review it periodically. The AI systems really challenge that model because the technology data and how we're using stuff is evolving super rapidly. The way we were thinking about AI even two weeks ago has changed pretty dramatically. Being at Microsoft, maybe I'm closer to the [chuckles] fire when it comes to this kind of stuff and how fast these things are moving.

I think as you think about the assessment tools, the future really requires us thinking about a continuous evaluation. Instead of asking only whether a system was valid at launch, we need mechanisms to monitor performance, fairness, drift, unintended consequences over time. Really, if you think about the future of responsible AI, it depends not only on better algorithms, and they will continue to improve over time, but we need better governance, and it needs to be ongoing.

I'm going to call it oversight, but there has to be this continuous check-in to make sure you're still validating what you think you're validating with your assessment tool and that there isn't any drift, that bias hasn't been introduced in unexpected ways. I love the idea of thinking through the education process and how we reimagine what the process is to validate tools when we think about the education of PhD and doctoral students and things like that because it needs to change. We can't just do this, release and maintain a model that we've always fallen for. It needs to be more continuous. Figuring out what that looks like, I think, right now is a question that we still need to figure out what that means.

Ludmila: Danielle, we as psychologists, we can probably help shape this algorithm by which AI provides information regarding assessment. How can we do that?

Danielle: Well, thank you for that. In many ways, it's not any different than what it was in the sense of in psychology, we're quite sensitive to many things such as defining the constructs clearly and linking those to the behavior and defining what should be measured, what counts as measurement, thinking about uncertainty and variance, and then also considering what decisions the system is allowed to support as well, as Adam was talking about.

In many ways, it doesn't change. Developing transparent rubrics, decision rules, testing performance, making sure that we are testing the outcome continuously, just as Liberty just described so well. We really do need to think about how we change our paradigms. If you think about an assessment, you would work for years to develop the assessment. It would be you would validate it, you would publish it, and you've got an assessment, and then people can use this assessment to make decisions.

Now you might be sent or find online an assessment that's automated and gives you all this magic power, and it might actually do that, but the question, if you have a more traditional assessment, there's very little under the hood. It has face validity. You can see what it's doing. You can read what the theoretical motivation was or its framework.

With AI, it's under the hood if it's really an automated system. There needs to be-- and I don't want to put the onus on the person using it. I'm going to put the onus on the people who are developing and marketing these kinds of assessment. There needs to be clear transparency on, for example, in AI systems, it may be a single agent, there may be multiple agents. How does it work? What is the knowledge base? What is actually informing and guiding this AI? Which LLM is it, and does that LLM model change over time?

All of these questions are the kinds of things that need to be asked by the user, but they need to be transparent by the person who is developing and providing these assessments, thinking about not only the one-and-done reliability and validity, but continually providing information and giving really transparent information about what is under the hood.

Ludmila: Does anyone would like to add anything?

Adam: I would just add that I completely agree, especially with the transparency piece. I think transparency has to be a little bit more. It also has to be accessible. A lot of these models, transparently, we could see the billions of features, but that's not-- understanding how it works is not accessible. I think there has to be a good balance between accessibility and understanding how it's working and allowing these things to work in their very, very complex ways. I would just tack that on to everything you said about transparency, which I completely agree with.

Ludmila: Would we be talking also about a little bit of literacy on the side of psychologists using these tools, and how can we achieve that?

Danielle: Yes, I talk about that a lot. I was just thinking about it earlier that we hadn't really talked about the need to understand AI in the education world. We talk about that a lot in terms of students using AI, faculty using AI. It's so important to understand how AI works and also to use it yourself. Understanding how it works and what it does, what it does well, what it does very poorly, what it will get right, what it will get wrong, and how you can use it to your own advantage. There's all sorts of ways to increase your knowledge of AI, how it works, and then also just use it. Just get a feel for what could be under the hood and what, like I said, what it can do and what it can't do. It's very, very important.

Liberty: I just want to tack on to that a little bit. I've been-- Okay. Luxury of being at Microsoft. I get to play with a lot of these different models. There is something to be said. Some of these models are much better than other models to solve different types of problems. I think there's a lot to be said to just experiment with your AI companion of choice and figure out where it works well and where it doesn't and experiment with other models to the extent that you can because you will find that there are definite differences in what models work well to solve different types of problems.

Actually, as you get familiar with that, that might become a question that you consider as you think about the different solutions from an assessment perspective is, what is the model that's powering the whole thing, and was it the right model to solve that problem?

In the last couple of weeks, I've been really playing around with different models, and I'm seeing noticeable differences. Since we're talking about this, I just want to mention that experimentation is really important to understanding, and you will find a difference in the models. I think that just provides a better framework for thinking about some of these questions from an assessment perspective.

Ludmila: Yes. Thank you for that. We've been talking about validity, transparency, important things, but also a very important thing or principle when we're evaluating these tools to be used in assessment has to do with bias and fairness. Lauren, I wanted to ask you, what are the biggest risks of bias in AI-driven assessments for historically underrepresented or marginalized groups, and what can we do to mitigate those risks?

Lauren: All right. I'm going to frame this question more so in the realm of considerations. Of course, we've talked a lot about making sure your training sets are representative, your normative sets are representative. I would also go a step further to say you probably should oversample certain groups that tend to have smaller sample sizes in the grand population, just to make sure that you're not in danger of treating populations as monoliths. The more you can oversample some of your smaller communities, the more you're gathering more variants within their individual communities, which I think is really important when we're talking about bias in these AI models.

I also want to take a moment to bring to the thought space that it's very important that when we're talking about transparency, it's very important that we have open and honest communication, specifically with communities that have been either marginalized historically, currently marginalized, or underrepresented in the past. Many of these communities have really complicated histories with institutions and systems commodifying their data, stealing their data, making money off [chuckles] their data, or even using that data against those communities to further perpetuate harm.

It's really important that we come as practitioners into the space with a spirit of having a co-design approach with these communities so that they feel like if they are going to engage with AI, it's being done in partnership with them and not being done to them. A lot of times that's going to involve developing trust because there is a lot of mistrust in a lot of these communities around technology, specifically AI, and some of the cautions that are very widely published about AI's capabilities and some of the risks and harms.

I think it's really important for practitioners not only be literate enough to understand what these AI tools are doing, but to be able to explain it to varying communities that they're hoping to work alongside in the AI and assessment space. I think that's the best way to understand, number one, what the needs of the community are. This might be maybe more specific, maybe not necessarily, but maybe more specific to assessment and education space specifically because there is so much federal, state-level, local-level accountability involved in assessment.

Making sure that the communities really feel like they're a part of the conversation when we're talking about introducing AI tools, both in assessment and at the classroom level, specifically when children are going to be engaging with these agents, and ensuring that the AI tools, for a lot of communities, human interaction is paramount to their education, training, and well-being. I think that's true for all humans, but specifically for some communities, it's even more paramount.

We need to make sure that we're not using these AI tools to supplant human interaction, which could present harmful or negative impacts to students who are trying to learn or individuals who are trying to progress in a certain space when we're talking about AI and assessment.

These are all things that, in essence, we have to just really be culturally literate of the impacts that AI can present in differing lived experiences and communities and be sure that we're being responsible with how we're engaging with those communities as we bring them on board or explore the possibility of bringing them on board in that space. I think I answered the question, but if not, just feel free to ask me more. [chuckles]

Ludmila: No, completely. Thank you. Thank you. You also made me think of how can we actually design these AI assessments and implement them, not just to prevent these issues, but to actively reduce existing inequities across different populations? I know Adam has some work on these as well and opinions about these, so if both of you could talk a little bit about this?

Adam: Yes. I think at the very start, we obviously should be following best practices for designing and implementing any tool or technique in our field. That doesn't change here. I do think there are a few other considerations that warrant extra attention when it comes to AI and assessment. One that I tend to think about a lot is equitable AI systems and models are great, but they'll only serve those they're designed to help if training and the use of them is also implemented equitably.

Setting serving the most vulnerable populations are often under-resourced, and they might not have funding to support trainings for their clinicians. We can design and develop AI tools intended to reduce existing inequities, but we wouldn't have clinicians trained to use them effectively in those situations, and that's an issue. Doing more things like this webinar to offer these trainings more equitably is going to be important.

Then AI systems can suffer the same shortcomings as our measurement instruments. Certain populations might not be adequately represented in the data used in the development. Other factors can bias these models outputs in ways that reinforce existing inequities or introduce new issues, just to name a few things. It will really be important for us to use everything we know about the

ways assessment and psychology more broadly can reduce rather than reinforce inequity and apply that knowledge for the use of AI and assessment.

I also think that research has a burden here. I think this is a crucial area for research that we need to aim at understanding the limitations of these systems and their use because those findings are going to really help us target factors and practices underlying inequities. For example, testing whether group-specific or universal models are better performing, for whom and when.

Then the next step to that is doing on-the-ground research to see how do we best balance that with feasibility. You could arguably say that a group-specific model is always going to do better because it's specific to that group. It was trained specifically for that group. How feasible is it to do that for certain populations that maybe are not presenting in certain settings very frequently? At what point can we justify using a universal model? At what point do we need to use group-specific models? That's just one arm of this research that I think is really foundational to equitable applications of AI.

Then echoing what Lauren discussed just a moment ago, we also need to do better at listening to our clients' concerns and the experiences of different populations to get a better idea of where we're falling short and how things can be improved. We're up here in our ivory towers. We're in our echo chambers and everything, but we really have to go to the community to understand what are your experiences that we may not be seeing and how can we do better? Ultimately, that's who we're serving. Asking, I think, is a great start.

Ludmila: Lauren, do you--? Yes.

Lauren: Oh, sorry.

Ludmila: I was going to ask you to-- Yes, I know. I was just going to ask you to chime in on this question.

Lauren: Yes. I was going to expand on something that Adam said. I thought it was very, very important that there is a lot of burden on research and researchers in this specific intersection of multiple industries. One of the things that in conversations that I have with other practitioners or people who are ed policy or in the classroom space is they don't have a lot of accessible information, tools, et cetera, to understand how to evaluate the instruments that are being presented to them.

If you work in these spaces, they're being inundated with, "Try this tool," or "This tool can do all the things that you ever wanted it to do and more." They don't really-- and a lot of times, have the expertise or access to the experts to be able to determine whether or not these tools are valid for use as they're being marketed, whether these tools can introduce harm in the spaces where they're trying to be applied, whether these tools are actually going to do the things that they are purporting that they're going to do.

A lot of my research recently has been being able to do research to develop publicly-accessible tools for individuals to just be able to evaluate AI that they're interacting with on a daily basis. It is a lot of work and it definitely-- research goes slower than [chuckles] AI advances. I think we're definitely behind the curve, but I think it's really important for not only for us to research some of the more philosophical and theoretical thought spaces but also to produce readily accessible tools that laypeople can use to be able to evaluate AI in their own individual spaces.

Ludmila: I really appreciate you and Adam mentioning research because that is one question that I had for the whole group, which is what research is needed to design better and more human-centered AI assessment tools? I know you already provided some clues about this, but I would invite everyone to answer.

Adam: I'll just expand on some of the things I was saying because I think about this question a lot. First, I think it's important, and we've said this before, that AI systems and models relevant to assessment are trained on human data. They can only be as good as the data used in their training. Actually, I think a crucial focus of research needs to be on the tools and methods we use to obtain those training data.

Just because AI is new and exciting, we absolutely cannot reduce our scientific efforts in creating psychometrically-sound measurement instruments. In fact, I think we need to do even more in this area because everything we've been talking about today and the vast majority of biases in these instruments or in these tools comes from and assumes that these systems and models are being trained on good data, and we know that's not always true.

I think related to that is that important target to understand the shortcomings and biases of the AI that we're using. It's great to know what AI can do, but I think it's more important to clearly understand its limitations. Basically, all the questions we commonly answer about assessment measures through psychometric research are more or less the same questions I think we should be asking about AI.

How is this tool collecting information, and what steps does it take to derive an output that we're interpreting? What errors and biases are present at what frequencies and under what conditions? Essentially, is the way it works aligned with the way we think it works? I know, Danielle, you were talking a bit about that and how that alignment is so important, and the transparency is so important.

Developments and breakthroughs are going to come. I don't think we need to advocate for research on the development front. Instead, I think we need to be very attentive to research that facilitates the robustness of the data behind the tools, that whole garbage-in, garbage-out issue, but also strives to understand the limitations of these systems because these systems are going to be biased. They are going to have these limitations.

The only way we can mitigate the harm from those biases is to understand them fully. That's where if I were to shine a light on the research that needs to be done, it would be any research that helps us have better data and any research that helps clarify where there's bias in these tools.

Lauren: I'd like to extend on what Adam said because I think that's really important. The training data behind the systems is key, especially when marketing to varying populations. If your training data isn't appropriate for that population, it probably shouldn't be marketed to that population, or if you are marketing it to the population, this is where I wanted to extend on Adam's point. It should be very plainly stated the populations for which this instrument is most appropriate given the training data.

If there are instances where maybe there are limitations in the training data, I feel like the documentation and the marketing for that instrument should plainly state what those limitations are. I think that's part of what Adam was saying is that being very transparent about what the limitations of the system are instead of overpromising what the system can actually do.

I think a lot of times, if you dig hard enough, you can find it, but a lot of times, people consuming these tools don't know where to dig or even if they did know where to dig, they don't know what they're finding. I think making it very plain about those limitations is really important.

Liberty: I just want to shine a lens or a light. I can't remember the word Adam used on one other place that I think we need research when it comes to this, and that's the interaction between people and AI. As the more we think about AI becoming more personalized, adaptive, conversational, we need, first of all, evidence that these different paths still lead to "comparable." I'm air quoting that in fair conclusions because comparability, standardization, all of that changes in the face of AI.

We also need better research around human oversight, what it means. It's been traditionally called human in the loop. Now I'm calling it SME in the loop, Subject Matter Expert in the loop, because I think it's expert in the loop that's more important. We need research about how people are potentially misusing AI.

In my space, because I'm doing certification exams, we have all sorts of interesting experiences around people using AI to cheat on our certification exams. There's got to be some corollary, you know what the word I'm saying, in your space, too. If people are using AI to take advantage of something about your assessment process so they can show themselves in a particular way, we need to understand how people are maybe misusing AI and things like that.

I think there's a lot of research around where people and AI come together and what that means for how we interpret the results we get, how they're using it, and all that other stuff. This is where I spend a lot of my time focused on in the work that I do because I do want to create more personalized adaptive assessments and experiences because I think there's a lot to be said for that power that AI brings to the assessment conversation, but there's a lot of unknown, especially from a measurement perspective. [silence]

Ludmila: Danielle, do you want to add something about research as well?

Danielle: Oh.

Ludmila: Sorry. I saw that you were unmuted.

Danielle: The only thing I would add, those are extremely thorough and coherent ideas and responses. The only thing I might add is from a learning engineering perspective is really a call to the developers of the assessments in terms of human-centered design. In many spaces, what we have is the development of solutions without problems and assessments without need.

Thinking about in a learning engineering perspective, we really center on the co-design and involving the stakeholders from the very beginning and not develop-- We often have these great ideas. "Oh, we should have this thing," or "We should have that thing," and we think that we're right, and then we develop this thing.

Learning engineering perspective, we really focus on solving problems, thinking about the context, involving, for example, the educators, learners, families, communities, psychologists in the process. It's not easy, but as we are developing these assessments, it's really important to consider the humans and involve them in that whole process and thinking about different contexts and different needs and developing what people actually need and not what you think they need.

Ludmila: Thank you. We are really approaching the end of this session. I would like to look at the live questions we've been getting. There are a lot of very good questions. There is something we haven't talked much about. Someone asks, "What changes need to be made in graduate programs' training to help develop necessary skills to use AI assessment tools?" Anyone who would like to take it?

Adam: Yes, I'll say just something quick because I go back and forth arguing with myself about when AI should be brought in. I do believe that AI is a foundational competency at this point for us, but there's also this work on de-learning or unlearning, where I think for us to use AI well, we still have to be good psychologists.

If we immediately start using AI in a lot of these areas, it's possible that our students may not develop the critical thinking, the theory, the-- It's the same argument of why your professors might make you hand score a WAIS first before you start using the automated scoring so you understand how and why things are working the way they are.

I argue with my-- I don't have an answer, I guess, is what I'm saying. I argue with myself of we have to prepare them to use these tools, but maybe that preparation looks like not using those tools for a while. I just don't know when is the right time to introduce them. That's my confusion that I'm sharing with all of you.

Danielle: I think about this a lot in terms of two things. One is how do we need to adapt to the availability of AI and rethinking how we are approaching our pedagogical goals and thinking more deeply about assessments that we use. For example, writing an essay is probably not going to be effective anymore if it ever was. However, breaking that down and really deeply thinking about what your pedagogical goals are and considering involving AI in that process. For example, if you want the students to see different models of responses or one really fantastic way of using AI is as a conversational agent.

Practicing, we have a number of systems where students are practicing with clients in a therapeutic scenario. We have conversational agents in the medical realm diagnosing patients. There are all sorts of ways that we can use them for practice and also just thinking about critical thinking. Using the AI in this process to really get much deeper into what our goals are pedagogically.

The other angle that I think a lot about is how we need to train our quantitative psychologists and data scientists to be able to support our assessment of the reliability and validity and use of these AI systems.

Ludmila: Thank you. Thank you so much. We are actually one minute over time. I want to thank everyone for this conversation, and I hope we continue having this discussion. I also want to thank everyone who was here today, and I hope you enjoyed this conversation.

You should receive a one-minute survey to complete after today's broadcast. We value your feedback. We also want to hear from you on topics you'd like us to cover. Please email us at science@apa.org with recommendations. We'll send a link to the recording and the transcript in a couple of weeks. Thank you so much. We hope to see you at future events.