AI in Assessment: Ethics, Innovation and Research

On-screen: [This recording has been modified to remove technical issues that occurred during the live broadcast.

Adrea Anderson, MD, FAAFP. Associate Chief, Div. of Family Medicine, GWU.]

Andrea: Hello, everyone. Welcome to the AI in Assessment Presentation. I'm Dr. Andrea Anderson, and last week I had the pleasure to sit down with Victoria Yaneva and Kimberly Swygert from the NBME to discuss the impact of using AI in assessment practices with medical education. Please enjoy this recorded conversation and we'll come back in about 40 minutes for our live Q&A part of the program where our presenters will address your questions.

Welcome to the AI in Assessment Presentation. I'm Dr. Andrea Anderson, an associate professor and the Associate Chief of the Division of Family Medicine at the George Washington University School of Medicine and Health Sciences in Washington DC. Here are a few of the things we will discuss during today's presentation. How AI fits into medical education and assessment. Some of NBME's innovative research into AI. Existing technologies to measure complex interactions in AI. And ethical considerations of AI's use in medical education. I'm excited to be here today with two noted experts in this field. One is Dr. Victoria Yaneva, NBME's manager of NLP Research, and Dr. Kimberly Swygert, NBME's Director of Test Development Innovations, to discuss some of the innovations, research, and ethical considerations of using AI in medical assessment and medical education.

Victoria, let's start with you. Why is NBME exploring the use of AI in assessment?

Victoria: Thank you, Dr. Anderson. At the moment, advances in NLP and machine learning, or as we more colloquially call them AI are shaping the landscape of medical education by enabling us to have innovative avenues to evaluate essential skills.

So one example that I can give is if we want to assess complex constructs such as clinical reasoning or communication skills, we may need to have item types beyond the multiple choice question. And so whatever the shape of or form of these innovative item types, it's highly likely that they would require open-ended responses, whether it is written text or a verbal response to an item. And this is where NLP becomes very helpful because it enables these new assessments to scale.

So let's maybe work with a couple of examples here. If we even have only 20 items on a test that require open-ended responses and we administer that test to let's say a thousand examinees, we very quickly find ourselves with a sample of 20,000 free text responses that we need to score. And we know from experience that the phrasing of these responses can vary greatly.

So what we can use NLP for is, to learn from how human experts and SMEs score these responses and have an AI algorithm do that scoring. And through doing this, we reduce the manual workload a lot while reserving those essentially human skills for evaluating those responses that are shown to be particularly challenging for an AI algorithm. Another example that I can give is if we want to support learner growth through provision of feedback. So if that feedback is not personalized and timely and accurate, learners may not benefit from it. So to summarize, at NBME, we use with the machine learning research and AI research to provide tailored applications to the needs of medical education assessment. And that is with the view of enabling the measurement of complex constructs. So I gave examples with clinical reasoning and communication skills. Kimberly, would you like to add anything to that?

Kimberly: Sure, so Victoria has already nicely introduced us to the concept of essential skills and the use of AI technology. And both of these fit very nicely into NBME's mission to produce state-of-the-art assessments. And within the field of assessment, there is an emerging body of literature about the technology Victoria just described, documenting how elements of AI can be used to increase the efficiency of scoring without sacrificing accuracy and perhaps support new item types as well. Some prominent assessment journals such as the "Journal of Educational Measurement" have devoted entire issues on the promising nature of AI applications to support more valid design of assessments, enhanced scoring and score reporting. So the projects you're going to hear about today are evidence that NBME is staying current with the field of psychometrics as well as the field of AI.

Andrea: Thank you. Those examples were really helpful. I enjoyed hearing how AI is helpful in things like feedback or assessing complex constructs like clinical reasoning or communication skills. Let's get a little more granular. Tell us about some of the recent developments in AI, NLP and machine learning that are directly applicable and have an impact on medical education, for example.

Victoria: Here would be a good place to introduce a few terms and the first term would be transformer models. And the reason I bring up this term is because a lot of the

advances that we see right with large language models, with generative AI, they're thanks to this one specific type of model called the transformer model. And if we bring that specifically to assessment in medical education, what these transformer models help us a lot with is applications such as automated scoring providing personalized feedback or content generation.

Andrea: Wow. So transformers are more than just an action figure. Tell us more about automated scoring.

Victoria: Sure, so because these transformer models they are pre-trained on a lot of data that exists in the world. This is digital data that is not necessarily part of our own data sets that we work with. So what we can do in this case is use a procedure called fine-tuning that helps us adapt these models already trained on that vast amount of data, specifically to our tasks. And we have a lot of empirical evidence that this leads to improved precision when we do automated scoring. So if I was to compare this with just a few years ago, if I was to train an automated scoring algorithm, I would just use the data we have available here at NBME. And so everything that the model learns is just confined to this relatively small sample of data. But because the model has now been pre-trained on other data, we can retain the knowledge of that larger data set, but still tune it to perform particularly well on the task that we care about in this case, automated scoring.

Andrea: Well, it seems like there's been a lot of advancement. Can you tell us about advancements in test development? What about creating item, new items, creating new items? Can AI be helpful there?

Victoria: So that's a great question. As you know, testing organizations always need new items. And when we saw the recent advances in generative AI over the past couple of years there was a lot of hope that this technology can be used to generate items in assessment. And we see, for example, from the field of language testing that this is not just you know, a mirage of the future. We see that, for example, in the Duolingo English test, they use GPT-3 to generate reading passages and then, you know, human experts evaluate these passages. And that's an excellent use of this technology to generate a coherent reading passage. But if we are to apply this to the field of medicine we see that we actually need you know, more than this. We need not just a coherent passage, but we need that, you know, the relationships between entities within medicine to be truthful, to be factual. And that's currently a big challenge. So many of the members of our audience would be familiar with the concept of hallucinations. So these models, you

know generate language that sounds coherent but is not factually correct. And that of course would be a problem if we apply this in the context of creating items in medical exams. Here at NBME, we have started an exploration of the use of generative AI for exam development and content development, but it is in its very early stages. So we don't yet have a sufficient body of empirical evidence that we can that we can provide on this topic.

Andrea: Thank you, I know as a former volunteer item writer that sounds really great in terms of the advancements that AI is offering in this area. So thank you for explaining that for our audience. You have outlined a lot of the uses of AI. With use comes the potential for misuse. I'm interested to know what is the broader AI community doing in terms of research or consideration of the potential for misuse of AI?

Victoria: I'm very glad you brought up this question because there's certainly a lot of work, happening within the AI community exactly to prevent, you know, potential misuse or even if there is no misuse, just behavior from these models that is unexpected and potentially harmful. And I think that the field has come really far in a very short period of time. You may remember in 2016 when Microsoft released the Twitter bot tie. I don't know how many of our audience members remember that, but that Twitter bot had to be taken down in less than 24 hours because internet users learned how to tamper with it and it learned from different interactions to essentially produce a lot of antisocial racist tweets. Fast forward just six years to 2022 when ChatGPT was first released in the market. And while imperfect, we certainly did not see this type of dangerous behavior with this AI agent. And the reason why we don't see this is because of you know, advances in machine learning and a specific type of research called alignment research that helps, you know, these models behave in a way that is aligned with human expectations and without understanding of appropriate use. So a huge amount of research is currently being published in the field of AI. We have, you know, hundreds, if not thousands of papers every year on topics such as factuality evaluation. That is doing research on how to use technology to detect, for example, hallucinations or, you know, content that is not factually correct. We have a lot of research happening in human-centered AI, which acknowledges that these tools, they don't exist in a vacuum. They are only as useful as you know, we as humans use them or we are all part of like a sociotechnical system. So how humans interact with these tools is of paramount importance. We have a field of NLP called Trustworthy NLP, that, for example, checks to what extent we can create automated benchmarks to evaluate the trustworthiness of

these systems. And the list keeps going. What I want to highlight here for our audience to take away from this segment is that the AI community is incredibly vibrant when it comes to that type of research.

Andrea: Kimberly, in your opinion, with this rapid pace of AI advancement, can we expect some of the newer AI models to include guardrails to improve protection against misuse in the future?

Kimberly: Yes, in fact, I think not only we can expect it, we should insist on it. So Victoria's mentioned AI is a booming area of research and development. And these advances in technology have been accompanied by urgent calls within the scientific community to balance AI's disruptive potential with guidelines for responsible use. In addition, as Victoria mentioned, ethics within AI is now a highly visible topic among the general public. And there's now competing chatbots with ChatGPT. There's Bard, there's Gemini, there's Grok, there's other engines that have been rolled out that the public are able to see and evaluate. And also one thing to keep in mind is that within the field of psychometrics, there's already a focus on responsible development and use of our existing assessments. We already have assessment guidelines and publications such as the Standards for Educational and Psychological Testing. And the more recent guidelines for technology-based assessment that contain specific language. For example, not only our responsibility for ethical guidelines with standard assessment, but also our responsibility to address the questions of say bias within AI-based scoring systems. So I think because of all this, while we acknowledge that AI is transformative, it's also in some ways just another tool that's available for psychometricians and test developers. And just as we have a responsibility to anticipate and address unintended consequences with a standard assessment, we also have the responsibility to consider the unintended consequences from the use of AI with an assessment and address these proactively by installing the guardrails that would help us mitigate them.

Andrea: Thank you, so what I hear you saying and what I think would be important for our audience to know is that although we refer to AI as artificial intelligence, it's really just augmented intelligence, right? It's helping us do the things that we were already doing as medical educators or as researchers, psychometricians, but it's helping us to do it in a more efficient manner, faster manner, and a better manner, would you say?

Kimberly: Yes, that'd be a good summary.

Andrea: Wow, you both are so passionate about your work. What's the most exciting part for you?

Victoria: If I have to talk about most exciting part, I would say that so much wonderful work is being done to develop new assessment types to assess new constructs such as we mentioned clinical reasoning, or communication skills. And I feel like what AI can enable us to do is to allow these assessments to scale to thousands of examinees and we want to be able to score them, you know, reliably and fairly. So probably if I had to speak from a personal point of view, that's the on part that I'm most excited about. And we are of course all excited about generative AI. You know, we are investigating how it can help us create new questions, but these are unchartered territories and we need to do a lot more research and collect a lot of more robust evidence before we know how it can impact educational assessment in medicine. And my guess is that it will certainly impact it. I just don't know if it would be in the ways we currently imagine.

Andrea: What you've been talking about is so interesting. And previously we talked about how AI and NLP can help us with complex constructs like clinical reasoning. I know as a medical educator, clinical reasoning is so important when we work with learners. We want them to be able to think, to get to the correct answer and not just arrive at the correct answer by chance, for example. Can you tell us more about how AI can help with those type of complex topics?

Victoria: Certainly, so you know that there has been a lot of research in item types such as short answer questions or the so called sharp items that, you know, NBME has been investigating which have multiple components. I'm not going to list them all here, but one of them is a free text response. We also know that, you know, medical students are frequently examined on writing clinical patient notes, like was the case with the former Step 2 Clinical Skills exam. Right now NBME has the OSCE Creative Community whose goal is to reimagine OSCE assessment and also have students provide some sort of free text you know free text notes. And so what we do is we apply the transformer models that I mentioned earlier to score these. And what this process actually looks like is we first have subject matter experts, physicians, who go through a large sample of these responses and in the case of short answer responses, they might mark them as either correct or incorrect.

On-screen: [Screenshot of Scoring Example. NBME NLP page with scoring results.]

In the case of, let's say, clinical patient notes, they would highlight the specific text within a patient note that corresponds to a key concept from the rubric.

On-screen: [Feedback Provision Example. NBME Concept Identification, Patient notes highlighting responses that match key essentials.]

And so we collect all of this information from our subject matter experts and then we use that with the transformer models to fine tune the models and make them very acutely aware of the specific things that they need to look for in any given task. And one of the metrics that we use you know, to evaluate is called F score. I'm not going to go in a lot of detail here, but an F score of zero means that there is no agreement whatsoever between the predictions of the model and the subject matter experts. While a score of one means that they are perfectly in agreement.

On-screen: [Results explanation. Evaluation metric is F1 score. 0 = no overlap with SME judgement. 1 = Perfect agreement between ACTA and SMEs.

- Scoring short answers = .98.
- Scoring patient notes = .96.

Scoring communication vignettes: in preparation.]

And we see a lot of that, for example, we have for short answer responses, an F score of 0.98 in many cases. Also the way we envision the use of these systems is that borderline cases would always be reviewed you know, by human experts. And here is the place to mention that we have a whole evaluation roadmap for such scenarios, so we don't just measure how accurate these models are. We measure many other things. One example I'm going to give is what happens if an examinee doesn't know the correct response and they try to enter several responses, for example, hoping that one of them is correct. Or what happens if they enter words from the item stem, or just, you know, a generic summary of the item. So we know that such responses are not going to trick a human rater. So they shouldn't trick an automated scoring system. And so when we performed this test, we actually saw that, you know, these automated scoring systems are vulnerable to such gaming attempts. And this is not new, this is consistent with the literature on this topic. So by performing you know, this type of evaluation as one example, we were able to go back and make the models more robust to such attempts by using a technique called adversarial training. And so we see that these new adversarially trained models, I know that this sounds very hostile, but it's a term. They are a lot more robust to such attempts. So this is one way in which we consider this kind

of real world application of these models. So we are not restricted just to measuring accuracy, but we are measuring many different aspects of these models that are important for fair and reliable assessment.

Andrea: That was a wonderful explanation. And what I took from that is not only is the NBME looking at the range of what could be a correct answer, you also have to look at the range of what could be an incorrect answer and make sure that the automation and the machine learning, et cetera, is able to discriminate between the two.

Victoria: Absolutely. Yes.

Andrea: You mentioned earlier about feedback, which I know is very important for our audience to hear about. Is there a role for these technologies to assist with giving feedback in medical education and assessment?

Victoria: This is an area that we're actively researching at the moment, because when automated scoring approaches are able to identify specific phrases within a text written by examinee or a transcription of a text that an examinee spoke, we are able to match these phrases to key concepts. And then once we have a phrase matched to a concept, we can do additional work to present that as feedback that helps learn their growth. So one area that we are currently doing research on is feedback specifically on communication skills. And it is a challenging area because you know, you as a physician know how many different ways we can use to phrase things like validating the patient's feelings for one example. So detecting phrases that are so long and also varied in the way they are phrased is a challenge, but we are seeing some optimistic early results in this area.

Andrea: So what about trust? This is something that the medical community as a whole is grappling with and I would imagine the assessment community is no different. How do we know that these systems are trustworthy? How do we ensure transparency? How can the medical community know that these systems are rigorously vetted and that we can put our trust in them?

Victoria: Right, well, the best way to build trust is that we collect rigorous evidence and we present it to the public and we do that through peer reviewed research. So one important thing to know is that the teams who develop these you know, capabilities, they're multidisciplinary teams. So we have AI scientists, we have measurement scientists, we have experts in test development. And likewise, when we communicate our findings to the public we communicate them to different research communities.

Because you know, if we publish something in a medical education, they knew you know, the vast majority of questions would be about the implications to medical education not to the, for example, rigorousness of the AI approach, which is why we make sure that we also have these you know experiments and findings published within the AI community where a lot of the questions are around, you know, the technical aspects and likewise with publishing them, measurement and assessment communities. So that is one way to communicate with different stakeholders and to raise awareness and have our research vetted by making sure that the right questions are asked by the right expert.

Andrea: Well I'd like to pivot here and let's talk a little bit more about the ethical considerations when using AI. Kimberly, are there any guidelines towards the ethical use of AI that you can describe for us?

Kimberly: Yes. Thank you so much. So there are indeed plenty of ethical guidelines to which to refer. The field of ethical AI or responsible AI in general has grown rapidly because as you've heard from Victoria, the technological advances are moving rapidly as well. But at the center of all that we're talking about machines that extract data that create models and they make sense of data on a scale that can be challenging for humans to comprehend. But the responsibility for the ethical use of these applications comes entirely from the humans that are using them.

So there's a substantial set of science and industry led ethical guidelines that are intended for AI developers. And there are also some geopolitical and governmental publications such as the European Commission's Ethics Guidelines for Trustworthy AI. and the Biden Whitehouse's Executive Order on the Safe, Secure and Trustworthy Development and Use of Artificial Intelligence. Victoria has already raised some questions of AI alignment with the needs and goals of organizations, including her own, the explainability of the AI, the trustworthiness of the AI. And in doing so, she's already begun introducing you to some of the most common topics that you'll see within ethical AI. There can be variation when you look at sets of guidelines that are out there across location, across companies and across different use cases.

But when you see these published ethical guidelines, they almost always contain topics that fit into one of six buckets. And those are privacy, accountability, fairness, transparency, human control and oversight and promotion of human values. And these are distinctive, they're not necessarily mutually exclusive. So you will also see published guidelines that will be some combination of these.

On-screen: [Existing ethical guidelines are plentiful, though there is no one set of agreed-upon ethical standards across all contexts. There are, however, similarities in the high-level topics that emerge: privacy, fairness, accountability, human oversight, transparency, promotion of human values.]

So let's begin with privacy, which is usually interpreted in this context as data privacy. So that includes things such as data exploitation, persistence and repurposing of data. So for AI models to be feasible, the systems need a large amount of data. And most of us probably use applications in our daily lives that collect data, but it can be difficult to understand what kinds of data and how much data is being collected, how that's being processed, how that's being shared via devices and networks and platforms that we use. So because of this, a responsible deployment of AI will include assurances of clear rules of how personal data will be collected, how long it will be kept, and whether it's acceptable to use data for other purposes. So for the next topic there is fairness. One path to supporting fairness is to begin with the assumption that the use of AI is never neutral or impartial truly. And acknowledge that the use of AI, especially for things like automated decision making, could lead to discriminatory outcomes. So people could be misclassified, misjudged or otherwise negatively impact by AI systems and any errors or biases on the part of the system could disproportionately affect certain demographics or continue to perpetuate historical inequities. So because of this AI fairness efforts focus on mitigating such biases to ensure that AI decision making is not discriminatory towards certain groups. And fairness guidelines require quite a bit of detail to cover the potential uses that can introduce bias and how they can be avoided. So some of the most common threats to fairness could be bias in the training data or the algorithmic bias, especially if the goal is to model is to use past outcomes to predict future classifications. And really the question is not whether the training data set contains bias, but what bias going to be there with the types of data that an organization might be using and the models that they're using and what the best way would be to avoid it. Next we have transparency, which in the context of AI is about how much it's possible to understand about a system's inner workings in theory and how feasible it is to provide explanations of algorithmic models and decisions that are comprehensible to the user. So transparency involves the public perception and understanding of how AI works. And it can be notoriously difficult to translate algorithmically derived concepts into human understandable concepts. And this is also dependent on the recipient's degree of technical knowledge. So we might first need to clarify explainable to whom. So this is where we might see a trade-off in the AI models that we use as simpler models are

more explainable, but might sacrifice some accuracy in the process. And in scenarios where we do want to use highly complex algorithms that may not be as transparent, the concept of trustworthiness that Victoria has already introduced might be a more useful organizing principle. So if I can switch for a minute To a non-assessment example, we've all heard of self-driving automobiles. And for those automobiles drivers are not necessarily going to understand the inner workings of those cars when they drive them. But through a combination of training, the assurances offered by safety regulations, a manufacturer's reputation for safety and tacit knowledge acquired from using their vehicles, the user develops a working sense of the conditions under which they can trust those vehicles to get from point A to point B. Human oversight involves the or refers to the involvement of humans in the development, the deployment, and the monitoring of AI systems. And human oversight, it's crucial for many reasons that map onto everything that we have been talking about today. So humans are able to provide a level of judgment and discretion that the AI systems currently lack. The human oversight is important for ensuring that the AI systems are transparent and explainable or trustworthy. And then human oversight is important for ensuring that AI systems are used in a way that benefits society. So the trap we want to evolve falling into is any kind of say automation bias on our part where we assume that the AI models are accurate, that data's representative, we've done everything so well, and then we assume that just because the AI can act so efficiently and autonomously that it should be making certain decisions for us. That would be a mistaken assumption. Humans are always going to retain control of some of the decisions. And finally on the topic of promotion of human values, one of the most common approaches that should be very familiar to this audience is non-maleficence, which is first, do no harm. This concept comes from the ethical principles that we should evaluate the moral quality of an action based on its consequences. So an ethical standard of non-maleficence with the use of AI states that we should avoid causing both foreseeable and unintentional harm. And unfortunately, we've all seen examples of AI deployment that do not abide by these principles. So we've seen bad actors, we've seen examples online of deepfakes or other malicious use of AI. And obviously a great majority of the organizations who would like to use or deploy AI have no intention of deliberately causing harm. In that case, avoiding unintended consequences is paramount concern. In addition, I should point out that upholding these ethical principles also requires that we understand the scientific limitations of the AI technology and any potential resulting risks to NBME or to stakeholders. And this can include the risk of using these technologies, the risk of not

using the technology, and potentially risk to our mission from the existence of external AI capabilities that people may use. So because of this, one of the first things that NBME developed alongside the AI research agenda was an organization-wide AI risk scorecard that's regularly updated and evaluated to ensure that we are appropriately identifying and then finding ways to mitigate risks that could be truly harmful to any of our stakeholders.

Andrea: Wow, that was a really comprehensive overview of the six ethical principles that apply to AI. So just to be sure that I heard you correctly, and also for the benefit of our audience, you mentioned privacy and accountability, fairness, transparency, human oversight and human values, including do no harm, for example. Could you speak more to how these apply to test development?

Kimberly Yes, I will actually choose one scenario here we've already talked about and are all excited about, which is using AI to generate exam content. So for a little bit of background, the defensibility of the exams that we give rest on the defensibility of our content, which is why NBME devotes a substantial amount of resources on recruiting and training board certified physician item writers, developing and refining these MCQs and applying metadata to items and media to ensure that we're building high quality test forms. So the research Victoria has mentioned, is part of evaluating whether AI could be used to make gains in efficiency and scale in these areas. And there are clear ethical topics we would want to address as well. Although just to make sure at this point I would say we've identified some of the most crucial questions as opposed to actually having the answers. So I'll provide two ways in which we could think about this ethical framework with respect to the question of AI generated content. So as part of our current item writing and test construction process, we routinely review our items on an ongoing basis to ensure that the language is updated and that any outdated or stereotypical language is removed because this could contribute to the presence of bias in an assessment. So if we were to use generative AI to develop test items, that is a new technology, but the items would still need to meet the same stringent quality criteria as our current items. So we might need to develop new processes to detect and mitigate bias in that scenario, but the need to do so and the standards by which we judge the result would remain consistent. And then second, our current item development process relies on a great deal of human effort. So the physician item writers and medical editors invest a substantial amount of time and effort to create and refine our test items. Generative AI might provide a more efficient solution for some

tasks within item development, but the ethical guidelines would require that we evaluate each step in this process to determine which tasks should be reserved for item writers and our medical editors. So in this scenario, we need to identify what the AI does not do as well as humans.

Andrea: That really helped me to understand the process of how to apply those ethical principles. And I can say as a former item writer, there is a good amount of work that goes into it and really the work of the editors, really the work of the whole NBME team to test development. And it's really exciting to hear how AI can be applied to this. Thank you, Victoria and Kimberly for this innovative presentation about AI. We've gone over so many topics. We've defined AI, talked about new development, new research, and concluded with ethical considerations. This has been an amazing discussion. You've shared so much information and it's been a pleasure talking with you.

Kimberly: Absolutely.

Victoria: Thank you.

On-screen: [NBME. Copyright 2024 NBME. All Rights Reserved.]

Andrea: Hi all. Welcome back to the Q&A portion. That was a great presentation from our experts whom I'm joined here with. Kimberly Swygert and Victoria Yaneva. Thank you so much for your presentation and we're looking forward to you answering the questions from our audience. I want to remind everyone, please use the Q&A function on Zoom as the chat has been disabled. And please remember, we will provide you with an executive summary of all of the key points of this presentation next week. Let's dive right into the questions. And I believe, Kimberly, the first one does go to you. So how should we balance the benefits of advanced AI technologies with some of the very real concerns about privacy, data security, and you know, the potential for misuse of personal information?

Kimberly: Thank you, Dr. Anderson. This is a very good question and it's one that I think everyone in the world is thinking about at this point if they are thinking about the potential for using AI technology in their work, in their private life, and how it's being used with society. And I think that this audience actually is going to have some insight into this question, even if they don't realize it, because this question is framed in a risk benefit analysis framework. So we are considering that we have benefits of advanced AI technology that can be used to support medical education, to support assessment, to support perhaps use in medical technologies. But we know about concerns with privacy

and data security and potential for misuse. And with any AI effort it should start with some explicit recognition that there is potential harm to groups in terms of these particular issues and that responsible AI use would start by making these concerns explicit. And then to consider additional aspects such as privacy by design where you are building in the privacy to start and also with the idea that any AI function might need to have additional auditing and accountability if it's used. And a lot of this actually is already there if you think about perhaps medical research. There has always been the questions of patient privacy, of requiring patient consent for data to be used, the potential for misuse of that personal information. So I think it's important to remember that that's always been there and that's always been something to think about. And that responsible AI use would involve identifying additional risks upfront that would be specifically due to AI, such as whether or not there's patient data that would normally be shared with the research study that perhaps is being requested in AI use, or is there perhaps a need for new modified patient consent agreement. Will there be a need for additional auditing of the data and things like that.

Andrea: Well, thank you for that answer. Here's the next question. How can educators use AI to improve student learning, for example, with formative assessments? Victoria, I think this one goes to you.

Victoria: Thank you, Dr. Anderson. This is an excellent question. AI can help with many aspects of formative assessments starting from you know assistance with some of the questions, the development of the items, some of the scoring of the responses, but one specific aspect that I want to focus on because it's still in its early stages of exploration of providing feedback to learners. We know that for this feedback to be to be useful, you know, to among other different criteria that it needs to be personalized and it needs to be provided in a timely manner. And I think this is a very interesting application for AI because I see one way in which, one of several different ways in which it can be done, you know, very reliably and some ways in which it could be done in a very harmful way. So I want to maybe take a second to distinguish between these two ways. So I see a lot of comments around, for example, using ChatGPT or another large language model to give you feedback on the way you have communicated with a patient, for example. And so many of us are impressed with how fluent it is and how helpful it could be. However, this holds a trap there that we actually at the moment, the way we use ChatGPT for such purposes, for example, we have no way of verifying whether this feedback is in any way accurate and whether it might mislead a learner through this potential

inaccuracy. However, if we you know, use different approaches that can rely on first creating expert annotations from past experiences where we know what the scoring rubric is and we know how examinees or users have responded to some of these questions, we would first know how experts would provide this feedback. And then we can use AI to learn from that expertise, and we can also evaluate how aligned it is with the feedback that the experts would provide. And then we can provide that in a personalized manner, but by quantifying its accuracy by understanding where it does well, where it does not do well, and so on and so forth. So I like this question because it allows for a juxtaposition of using AI in one way that is risky and using it in a different way that can actually benefit assessment and it's a lot more safe and evidence-based.

Andrea: Great answer. Thank you for answering that. So the next question is, how can AI be used for blueprinting or curriculum mapping? Kimberly, I think this one goes to you.

Kimberly: Yes. Thank you so much. And thankfully, Victoria has already nicely set the stage for us because to some extent I can say ditto to what she just said about using AI for improving student learning in terms of understanding the capabilities of generative AI, understanding the risks, and understanding that there's a lot of human oversight still required to ensure that what you generate with it is accurate. However, perhaps for blueprint and curriculum mapping, that might be kind of a less risky, less sensitive area. So perhaps safer to play around when at this point, if you have access to ChatGPT or if you have access to Bing Chat or something like that, I would urge you to try it out and see what you get with the understanding, as Victoria just stated, that it's going to be up to you to verify the accuracy of what's there. You know, this technology certainly can assert misinformation as fact and you would need to verify, for example, if it comes up with citations that you would want to include with blueprinting or curriculum. It's going to be up to you to provide any additional context or any transparency into what it's doing because it's not necessarily going to tell you where its sources are from. However, I would say this is an area in which you're probably going to see a lot of research very soon. The K-12 EdTech space is moving rapidly on this type of thing. I'm not aware of any EdTech products. There's some that are advertised as generative AI for curriculum mapping, but they don't really seem to be anything other than a nicely tailored usable interface that's doing a backend call to ChatGPT or something like that. Absolutely use it as a creative tool for now to start. If you have public documents that you're comfortable sharing with the software such as, you know, public high level blueprints or

existing curriculum maps that you have, you could use that to refine the output. But I think you're going to see a lot that start to emerge in the literature, both for K-12, but I'm also seeing quite a few things in journals like Academic Medicine that are related to this as well of late. So rapidly growing area, test it out yourself and see what you think about it.

Andrea: Excellent advice. And moving on with that, what skills should we teach in the world of AI? Or are there new assessments that we should adopt? Kimberly, maybe this one could be for you.

Kimberly: Sure, that's a super interesting area because along with everything that we're seeing in terms of the capabilities that people can use, such as everyone having access to ChatGPT and things like that, we're starting to see articles, again, referring back to Academic Medicine, I feel like I've seen one recently in there, that are saying medical students and faculty should have coursework, should have expectations, should have curriculum related to safe and ethical and accurate use of AI. I think that could include a better understanding of what some of these more popular models are actually doing. How to safely use AI, how to think about ethical principles with use of AI, how to think about data privacy and some of the misuse that we've already referred to here. And I don't know that there are any assessments that assess your skills on how well you do that at this point, which I think what the question is about. I think that's also something that you may see start happening with some you know, various educational programs, including medical education programs where they are going to say our students are going to be using AI either as part of their education or possibly as part of their practice in the future. How can we get them best prepared for that?

Andrea: Thanks for that. And I really know that our audience is really benefiting from your advice to them. We're going to move to our last question. What is your guidance, and this can be for both of you, what is your guidance on the use of generative AI for creating assessments?

Victoria: Thank you, I'm aware of time so I'm going to give a very brief response here and because it can be a very long response to a big question like this, but I would say briefly that because we put the human in the center we need to make sure that we look at the whole process and that we are not thinking that just by automating certain parts and having it being AI assisted, that we are helping with the overall process. So I'm just going to give a couple of examples just because we can suggest distractors to item writers that are relevant to the question. It doesn't mean that these distractors were not

immediately obvious to them when they were writing these questions, so we are not actually really adding value even though, you know, the distractor is relevant. Or if we are generating samples of questions that we want to have human reviewed, you know, we should ask the question of, is it beneficial to go through 100 questions that are not so great to find five that are great as opposed to writing five questions from scratch? Or if we are creating variations of questions, are they too similar and is it easier to create, you know, these five non-similar questions, let's say, from scratch? So I'm going to leave it here and give a chance to Kimberly to respond as well.

Kimberly: Yeah, thank you. I know we're very close to time, so I do just want to say that the potential is great, but nothing we're seeing right now suggests that this replaces the human effort, especially when you're talking about the contribution that experts make to the content validity evidence for any kind of assessment. So there's no evidence that we suddenly can create great items that don't require humans to review them and edit them and check them and approve them. That part of the process is still there. So it's going to be a very interesting future as we use the new technology to see what we can automate while ensuring that we still have human expertise in the loop where it needs to be.

Andrea: Thank you. Well, the future is bright for AI and we really have enjoyed your presentation and your expertise. I want to thank the audience for joining us this afternoon for this AI in assessment presentation. NBME will be sharing a link of this recording along with the Q&A session and an executive summary of the key points discussed today for your review. Thank you so much.

On-screen: [NMBE. Thank you for joining us!]