# ارزشیابی در عصر هوش مصنوعی

**سید علی مسلمی[1]، رضا علم[2]**

۱. کارشناسی ارشد تکنولوژی آموزشی، دانشگاه علامه طباطبائی، تهران، ایران .(نویسنده مسئول).

۲. دانشجوی دکترای برنامه‌ریزی درسی، دانشگاه آزاد اسلامی، آزادشهر، گلستان، ایران.

فصلنامه ایده‌های نو در تعلیم و تربیت، دوره چهارم، شماره دوازدهم، پاییز ۱۴۰۴، صفحات ۱۸-۱

**چکیده**

در این مقاله، ما استدلال می‌کنیم که یک مجموعه خاص از مسائل شیوه‌های ارزیابی سنتی ممکن است برای مربیان طراحی و پیاده‌سازی آن‌ها دشوار باشد؛ تنها تصاویر لحظه‌ای گسسته از عملکرد را به‌جای دیدگاه‌های متفاوت یادگیری ارائه می‌دهد؛ بادانش، مهارت‌ها و زمینه‌های خاص شرکت‌کنندگان سازگاری نداشته باشید؛ برای آماده کردن دانش آموزان برای ورود به مدرسه، به‌جای فرهنگ مدرسه رفتن، فرهنگ مدرسه رفتن طراحی‌شده است؛ و مهارت‌هایی را که انسان‌ها به‌طورمعمول برای انجام آن‌ها از کامپیوتر استفاده می‌کنند، ارزیابی می‌کنند. ما روش‌های هوش مصنوعی موجود را بررسی می‌کنیم که حداقل تا حدی به این مسائل می‌پردازند و به‌طور انتقادی بحث می‌کنیم که آیا این روش‌ها چالش‌های بیشتری را برای روش ارزیابی ارائه می‌دهند.

**واژه‌های کلیدی:** هوش مصنوعی، ارزیابی، آموزش.

## مقدمه

ارزیابی‌های خوب طراحی‌شده برای تعیین این‌که آیا دانش‌آموزان آموخته‌اند یا خیر، ضروری هستند (آلموند، استاینبرگ و میسلاوی، ۲۰۰۲؛ میسلاوی، استاینبرگ و آلموند، ۲۰۰۳). شیوه‌های ارزیابی سنتی، مانند سؤالات چندگزینه‌ای، مقالات و سؤالات پاسخ کوتاه، به‌طور گسترده‌ای برای استنباط دانش و یادگیری دانش‌آموزان مورداستفاده قرارگرفته‌اند (کایپا، ۲۰۲۱). در این مقاله، ما استدلال می‌کنیم که این شیوه‌های سنتی مشکلات متعددی دارند. اول، طراحی و پیاده‌سازی آن‌ها می‌تواند برای مربیان طاقت‌فرسا باشد. دوم، آن‌ها ممکن است به‌جای دیدگاه‌های متفاوت یادگیری، تنها تصویری مجزا از عملکرد ارائه دهند. سوم، ممکن است یکنواخت باشند و درنتیجه با مهارت‌ها و زمینه‌های دانشِ خاصِ شرکت‌کنندگان سازگار نباشند. چهارم، ممکن است نامعتبر باشند، زیرا به‌جای آنکه بافرهنگ مدرسه هم‌سو باشند، برای آماده‌سازی دانش‌آموزان برای ورود به دانشگاه طراحی‌شده‌اند؛ و درنهایت، ممکن است قدیمی باشند و مهارت‌هایی را ارزیابی کنند که انسان‌ها امروزه معمولاً از ماشین‌ها برای انجام آن‌ها استفاده می‌کنند. پس از ترسیم این استدلال‌ها، ما چندین کاربرد هوش مصنوعی را توصیف می‌کنیم که دست‌کم تا حدی به این مسائل پرداخته‌اند. بااین‌حال، تأکید می‌کنیم که شیوه‌های ارزیابی سنتی به دلایل مشخصی توسعه‌یافته‌اند و تا حدودی برای درک و بهبود یادگیری دانش‌آموزان موفق و ارزشمند بوده‌اند. به‌این‌ترتیب، بحث را با بررسی چالش‌های منحصربه‌فردی که هوش مصنوعی ممکن است در ارزیابی ایجاد کند و نیز با اشاره به فرصت‌هایی برای پژوهش و توسعه‌ی مستمر، به پایان می‌رسانیم.

## الگوی ارزیابی استاندارد

میسلوی و همکاران (۲۰۱۲) استدلال می‌کنند که ارزیابی آموزشی اغلب در چارچوب الگوی ارزیابی استاندارد[1] قرار می‌گیرد. یک مجموعه از پیش تعریف‌شده از آیتم‌ها (به‌عنوان‌مثال، مشکلات یا سؤالات) برای استنباط ادعاها در مورد مهارت دانشجویان در یک یا چند ویژگی استفاده می‌شود. داده‌های مورداستفاده برای این استنتاج‌ها معمولاً پراکنده هستند و یادگیری دانشجو ممکن است تمرکز ارزیابی نباشد. موارد الگوی ارزیابی استاندارد شامل تکنیک‌های ارزیابی گسترده‌ای مانند سؤالات چندگزینه‌ای، مقالات و سؤالات پاسخ کوتاه است(کایپا، ۲۰۲۱). درحالی‌که روش‌هایی مانند این به‌طور گسترده مورداستفاده قرار می‌گیرند، مشکلات بالقوه متعددی دارند.

مشکل اول یک مشکل عملی است .ارزیابی‌ها در پارادایم استاندارد می‌تواند دشوار باشد .طراحی ارزیابی نیاز به موارد و تکنیک‌های با دقت ایجادشده برای ترجمه پاسخ‌های دانشجویی به ارزیابی عملکرد یا یادگیری چیزهایی مانند روبریک‌ها، کلیدهای پاسخ و به‌طور فزاینده مدل‌های آماری پیچیده دارد (میسلوی و همکاران، ۲۰۱۲). ارزیابی تنها بخشی از عملکرد یک معلم در زمینه کلاس درس است .آن‌ها همچنین فعالیت‌های یادگیری را برنامه‌ریزی و رهبری می‌کنند، بازخورد فراهم می‌کنند و به‌طور کلی‌تر، فرهنگ کلاس درس را مدیریت می‌کنند. بسته به تعداد دانشجویان، مسئولیت‌های دیگر مربی و میزان کمک آن‌ها، طراحی ارزیابی‌های دستی و استنتاج از آن‌ها می‌تواند دشوار و مستعد خطا باشد (سوتو، ناداش و بل، ۲۰۱۱).

دوم اینکه، این ارزیابی‌ها ممکن است گسسته باشند و تنها تصویری از آنچه دانشجویان می‌توانند در یک نقطه زمانی انجام دهند را ارائه دهند. اگرچه این عکس‌ها ممکن است در مورد کاری که دانش آموزان در یک‌زمان مشخص انجام می‌دهند و نمی‌دانند به ما بگویند، اما ممکن است در مورد یادگیری چیزی به ما نگویند. همان‌طور که دیگران استدلال کرده‌اند، یکی از اهداف شیوه‌های ارزیابی، پرورش یادگیری است. همان‌گونه که در علوم یادگیری درک شده است، یادگیری را می‌توان با تغییر تعریف کرد؛ برای مثال، تغییری در بازنمایی‌های ذهنی، تغییری از آنچه فرد می‌تواند با کمک دیگران انجام دهد به آنچه به‌تنهایی قادر به انجام آن است، یا فرآیند آشنا شدن با یک فرهنگ جدید (پرِه–کلرمون، ۱۹۸۰؛ ویگوتسکی و کول، ۱۹۷۸؛ لاو و وِنگر، ۱۹۹۱). بدون مقایسه‌ی «عکس‌های فوری» از عملکرد در طول زمان، هیچ درکی از تغییر و درنتیجه هیچ درکی از یادگیری حاصل نمی‌شود. این منطق، بنیان بسیاری از تحلیل‌های اساسی یادگیری است که دانش پیشین را کنترل می‌کنند. همان‌طور که ما در مورد نتایجی که تنها پس از آزمون گزارش می‌شوند و ادعا می‌کنند که یادگیری مشاهده‌شده است تردید داریم، ما باید نسبت به ارزیابی‌های مشابه محتاط باشیم.

به‌طور نسبی، تغییری در ادبیات مربوط به یادگیری، به‌ویژه در علوم یادگیری و یادگیری مشارکتی با پشتیبانی رایانه، رخ‌داده است؛ این تغییر استدلال می‌کند که فرآیندهای یادگیری، علاوه بر نتایج یادگیری، اهداف ارزشمندی برای مطالعه هستند (پونتامبکار و همکاران، ۲۰۱۱). به‌طور فزاینده‌ای، آشکار می‌شود که درک فرآیندهای یادگیری در طول زمان، هم برای پیشرفت دانش‌آموز و هم برای سؤالات اساسی در مورد

[1] standard assessment paradigm

چگونگی وقوع یادگیری، اهمیت دارد (لودج، ۲۰۱۸). ظرفیت دانش‌آموزان برای مشارکت در خودتنظیمیِ مؤثر یادگیری‌شان (پانادرو، ۲۰۱۷)، برای قضاوت دقیق در مورد پیشرفت آن‌ها بود، (اجاوی، داوسون و تای، ۲۰۱۸) و برای تغییر راهبردها در زمان موردنیاز (به‌عنوان‌مثال، آلتر، اوپنهایمر، ای، لیتون ایر، ۲۰۰۷)، حیاتی است؛ نه تنها برای کار در دسترس، بلکه برای یادگیری و توسعه‌ی بلندمدت. علاوه بر این، درک فرایندهایی که نشانگر یا پیش‌بینی کننده یادگیری هستند، می‌تواند به آگاه‌سازیِ بازخورد، مداخلات و دیگر اقدامات آموزشی که ممکن است تأثیر مثبتی بر یادگیری بگذارند، کمک کند (پونتامبکار و همکاران، ۲۰۱۱).

سوم اینکه ارزیابی‌ها در الگوی ارزیابی استاندارد ممکن است به این معنا یکنواخت باشد که وظایف یا آیتم‌های مشابه بدون توجه به دانش قبلی، توانایی‌ها، تجارب و پیشینه فرهنگی به هر دانش‌آموز داده می‌شود. این موضوع مربوط به موضوع اول است. اگر روش ارزیابی برای وضعیت فعلی دانش آموزان کالیبره نشود، در این صورت تنها با عملکرد در حال حاضر صحبت می‌کند و نه با یادگیری همان‌طور که برای تعریف آن آمده‌ایم. علاوه بر این، ارزیابی‌های مشاهده به‌عنوان یک اندازه – متناسب – همه ممکن است باعث ایجاد تعصب نسبت به ارزیابی شود به این معنا که همه دانش آموزان ممکن است فرصت‌های برابر برای نشان دادن یادگیری خود نداشته باشند (گیپس و استوبرت، ۲۰۰۹).

چهارم اینکه ارزیابی‌ها در الگوی ارزیابی استاندارد اغلب غیرمعتبر هستند. ارزیابی‌های مبتنی بر تحقیق را به‌عنوان‌مثال در نظر بگیرید. افرادی که نوشتن برای آن‌ها بخشی از حرفه آن‌ها است با کمک آن‌ها می‌نویسند آن‌ها تحقیق می‌کنند و از ایده‌های دیگران استفاده می‌کنند، پیش‌نویس‌ها را به اشتراک می‌گذارند، بازخورد می‌گیرند و تجدیدنظر می‌کنند؛ آن‌ها از ابزارهایی مانند واژه‌پردازها استفاده می‌کنند که املای، دستور زبان و کاربرد خود را تصحیح می‌کنند و گاهی متن را پیشنهاد می‌کنند. در مقابل، نوشتن برای ارزیابی‌ها ممکن است کاملاً متفاوت به نظر برسد. آزمون‌های پذیرش تحصیلات تکمیلی مانند آزمون مدارک تحصیلات تکمیلی از مردم می‌خواهند که در انزوا و بدون دسترسی به ابزارهایی بنویسند که در حال حاضر بخش استانداردی از تمرین نوشتن هستند (اتس، ۲۰۲۲). این عدم هماهنگی بین عملکرد معتبر و فرهنگ کلاس درس به‌طور گسترده‌تر بر ارزیابی تأثیر می‌گذارد .همان‌طور که براون، کالینز و دوگوید می‌گویند: زمانی که فعالیت‌های معتبر به کلاس درس منتقل می‌شوند، محیط آن‌ها به ناچار تغییر می‌کند: آن‌ها به وظایف کلاس درس و بخشی از فرهنگ مدرسه تبدیل می‌شوند .درنتیجه روش به آنچه به وظایف کلاس درس تبدیل‌شده است، اعمال می‌شود. سیستم یادگیری و استفاده (و البته آزمایش) پس‌ازآن به‌صورت ارثی در فرهنگ خود – تائید مدرسه می‌شود. درنتیجه، برخلاف هدف تحصیل، موفقیت در این فرهنگ اغلب تأثیر کمی بر عملکرد در جاه‌ای دیگر دارد (براون، کالینز و دوگوید، ۱۹۸۹).

درنهایت، ارزیابی‌ها در الگوی ارزیابی استاندارد اغلب منسوخ شده‌اند زیرا آن‌ها مهارت‌هایی را ارزیابی می‌کنند که به‌طور فزاینده‌ای منسوخ شده‌اند. همان‌طور که شافر و کاپوت (۱۹۹۸) استدلال می‌کنند، رسانه‌های محاسباتی مانند کامپیوترها، پردازش اطلاعات را بسیار شبیه به گزارش‌های کتبی، ممکن می‌سازند تا ذخیره اطلاعات را خارجی سازند. این تغییر برخی از وظایف شناختی را بر روی رسانه محاسباتی توزیع می‌کند، برای مثال، محاسبات در مورد انجام ریاضیات با ماشین‌حساب و ویرایش در مورد نوشتن با پردازشگر کلمه و آزاد کردن انسان‌ها برای انجام کاره‌ای دیگر. این وظایف دیگر ممکن است شامل درک مشکل، نشان دادن مشکل در انواع سیستم‌های پردازش خارجی و استفاده از نتایج این سیستم‌ها به روش‌های معنی‌دار به‌جای انجام خود فرایندهای واقعی باشد. درنتیجه، آن‌ها استدلال می‌کنند که در بسیاری از موارد، تعلیم و تربیت و ما ادعا می‌کنیم که ارزیابی باید بر انواع جدید وظایف و مهارت‌های ارائه‌شده توسط سیستم‌های پردازش خارجی تمرکز کند.

با وجود این‌که الگوهای ارزیابی استاندارد اغلب گسسته، یکنواخت، منزوی و قدیمی هستند، اما در فرهنگ آموزش و پرورش پایدار می‌مانند. بااین‌حال، پیشرفت‌های جدید در فن‌آوری و هوش مصنوعی به بسیاری از جنبه‌های زندگی انسان از نحوه کار ما، به محصولاتی که می‌خریم، به چگونگی صرف وقت آزاد ما، نفوذ کرده‌است. برخی کلاس‌ها نیز برای استفاده از هوش مصنوعی به‌عنوان بخشی از روال روزمره خود آمده‌اند (هوانگ، شیه، وا و گاسویک، ۲۰۲۰). این شامل فن‌آوری‌های نسبتا ثابت مانند نرم‌افزار درجه‌بندی مقاله خودکار و تست تطبیقی (ون در لیندن و گلاس، ۲۰۱۰)، در کنار توسعه اخیر ارزیابی پیوسته مبتنی بر داده دانشجویان در تعامل آنلاین با مواد یادگیری است (کی و نگ، ۲۰۱۹؛ شوت و رحیمی، ۲۰۲۱). همچنین علاقه فزاینده‌ای به چگونگی نظارت و دستکاری هوش مصنوعی دانشجویان درگیر با محیط‌های یادگیری آنلاین مانند بازی‌ها و شبیه‌سازی‌ها وجود دارد که می‌تواند ارزیابی معتبر مهارت‌ها و رفتارهای به نمایش گذاشته شده در محل را پشتیبانی کند. به‌طور خلاصه، همان‌طور که کوپ و همکاران استدلال می‌کنند: ارزیابی شاید مهم‌ترین حوزه فرصت ارائه‌شده توسط هوش مصنوعی برای تغییر تحول در آموزش و پرورش باشد. بااین‌حال، این ارزیابی در اشکال مرسوم درک شده آن نیست. ارزیابی مبتنی بر هوش مصنوعی از مصنوعات و

فرآیندهای بسیار متفاوت از ارزیابی‌های سنتی استفاده می‌کند. در واقع، هوش مصنوعی می‌تواند رهاشدگی و جایگزینی ارزیابی‌های سنتی را هجی کند و با این یک تحول در فرایندهای آموزش و پرورش (کوپ، کالانتزیس و سیرسمیت، ۲۰۲۱).

در بخش‌های بعدی، برخی از رویکردهای موجود هوش مصنوعی را مرور می‌کنیم که ممکن است به پرداختن به مسائل مربوط به ارزیابی در الگوی ارزیابی استاندارد کمک کنند.

## هوش مصنوعی برای ارزیابی: از طاقت‌فرسا تا امکان پذیر

تکنیک‌های مبتنی بر هوش مصنوعی برای خودکار کردن کامل یا جزئی بخش‌های شیوه ارزیابی سنتی توسعه داده شده‌اند. هوش مصنوعی می‌تواند وظایف ارزیابی را ایجاد کند، همتایان مناسب را برای نمره دهی به کار پیدا کند و به‌طور خودکار کار دانش‌آموز را نمره دهی کند. این تکنیک‌ها وظایف را از انسان به هوش مصنوعی منتقل می‌کنند و به ایجاد شیوه‌های ارزیابی امکان پذیر تر برای حفظ کمک می‌کنند.

### ساخت ارزیابی خودکار

یکی از مولفه‌های مهم طراحی ارزیابی، وظیفه مورداستفاده برای استخراج شواهد برای پشتیبانی از ادعاهای مربوط به یادگیری است. در سال‌های اخیر، تعداد کمی از مطالعات برای استفاده از تکنیک‌های هوش مصنوعی برای خودکار کردن تولید چنین وظایف ارزیابی مانند سؤالات چندگزینه‌ای و سؤالات پاسخ باز پیشنهاد شده‌اند. به‌طورمعمول، این مطالعات براساس تکنیک‌های هوش مصنوعی که توسط شبکه‌های عصبی عمیق هدایت می‌شوند، ساخته شده‌اند. به‌عنوان‌مثال، وو، سان، زوو و جیا (۲۰۲۰) برای بهبود کیفیت پرسش‌های تولید شده به روش دو مرحله‌ای پیشنهاد شده‌اند: نمایش متن ورودی با اعمال یک طرح برچسب گذاری پاسخ سریع و پاسخ کلید به دست می‌آید و سپس نمایش ورودی بیشتر توسط یک شبکه کانولوشن گراف پاسخ – محور برای گرفتن روابط بین جملات و درون جمله برای تولید سوال مورداستفاده قرار می‌گیرد.

موفقیت چنین رویکردهایی اغلب به در دسترس بودن مجموعه داده‌های بزرگ مقیاس و مرتبط مورداستفاده برای آموزش این مدل‌های شبکه عصبی عمیق بستگی دارد. هنگام استفاده از این مجموعه داده‌ها برای آموزش یک مولد سوال، سند منبع مربوط به هر سوال (به‌عنوان‌مثال، رونوشت یک ویدیو سخنرانی یا یک قطعه از مطالب خواندن) اغلب شامل چندین جمله است و هر جمله ارزش سوال ندارد. این نشان می‌دهد که جملات قابل سوال در یک مقاله باید ابتدا قبل از استفاده از آن‌ها به‌عنوان ورودی به مولد سوال شناسایی شوند. با توجه به این یافته‌ها، چن، یانگ و گسوییک (۲۰۱۹) به بررسی اثربخشی کل نه استراتژی انتخاب جمله در تولید سوال پرداختند و دریافتند که روش گراف محور تصادفی، لکسرنک، بیش‌ترین کارایی را در میان مجموعه داده‌های چندگانه دارد.

درحالی‌که تولید سؤالات خودکار می‌تواند یک ابزار قدرتمند برای ایجاد طراحی ارزیابی برای معلمان باشد، بدون محدودیت نیست. مجموعه داده‌های بزرگ مقیاس برای آموزش مدل‌هایی که سؤالات را تولید می‌کنند موردنیاز است. بااین‌حال تا جایی که ما می‌دانیم، بیشتر مجموعه داده‌های موجود ارتباط مستقیمی با آموزش و یادگیری ندارند، به جز RACE (لای، زی، لیو، یانگ، هووی، ۲۰۱۷) وLearningQ (چن، یانگ، هاف، هوبن، ۲۰۱۸). درحالی‌که معیارهایی برای ارزیابی کیفیت وظایف از نظر همپوشانی بین سؤالات تولید شده و سؤالات ساخته دست بشر وجود دارد،( برای مثال) بلیو N - (پاپین، روکوس، وارد، زوهو، ۲۰۰۲) و میتور را ببینید(این معیارها ارزش آموزشی و مناسب بودن سؤالات تولید شده را تضمین نمی‌کنند (هورباچ، الدابه، بسنت، دی لسل، مارتا الار، ۲۰۲۰).

### ارزیابی همتا به کمک هوش مصنوعی

نقش بازخورد با کیفیت بالا در نتایج یادگیرنده به خوبی در تحقیقات آموزشی تائید شده است. بااین‌حال، با افزایش اندازه کلاس‌ها، ارائه بازخورد غنی و به موقع برای مربیان چالش برانگیزتر می‌شود. ارزیابی نظیر به‌عنوان یک روش ارزیابی پایدار و توسعه‌ای شناخته شده است که می‌تواند به این چالش بپردازد. این روش نه تنها مقیاس خوبی برای اندازه‌های بزرگ کلاس، مانند آن‌هایی که در کلاس‌های آنلاین گسترده (موک ها ) قرار دارند (اشنایدر و پارکس، ۲۰۱۶) دارد، بلکه نشان داده‌است که سطح بالاتری از یادگیری را در مقایسه با ارزیابی یک طرفه مربی ارتقا می‌دهد (ار، دیمیتریادیس و گاسویک، ۲۰۲۰).

اگر چه برخی کارهای قبلی در مورد توانایی یادگیرندگان برای ارزیابی موثر منابع گزارش شده است (عبدی، خسروی، صدیق و دمارتینی، ۲۰۲۱؛ وایتهیل، آگربر و هیلاک، ۲۰۱۹). قضاوت دانشجویان به‌عنوان متخصص در آموزش نمی‌تواند کاملاً مورد اعتماد باشد که قابلیت اطمینان ارزیابی هم‌کار را به‌عنوان یک ابزار ارزیابی به خطر می‌اندازد .بااین‌حال، گام‌هایی را می توان برای افزایش قابلیت اطمینان برداشت .یک استراتژی مشترک که در بیشتر پلتفرم های ذکر شده در بالا استفاده می‌شود، تکیه‌بر خرد یک جمعیت به‌جای یک فرد با استفاده از یک استراتژی مبتنی

بر افزونگی و تخصیص همان وظیفه به چندین کاربر است. این مساله مشکل جدیدی را مطرح می‌کند که معمولاً به آن مساله اجماع گفته می‌شود: در غیاب حقیقت زمینی، چگونه می‌توانیم به‌طور بهینه تصمیمات اتخاذ شده توسط افراد متعدد را به سمت یک تصمیم نهایی دقیق ادغام کنیم (چنگ و همکاران، ۲۰۱۷).

یک روش ساده استفاده از آمار خلاصه مانند میانگین یا میانه است .بااین‌حال، آمار خلاصه از این فرض رنج می‌برد که همه دانشجویان دارای توانایی قضاوت مشابه هستند که نادرست بودن آن‌ها به اثبات رسیده‌است (عبدی و همکاران، ۲۰۲۱). یک روش جایگزین، استفاده از روش‌های توافقی پیشرفته است که از مدل‌های هوش مصنوعی برای استنباط قابلیت اطمینان هر ارزیاب استفاده می‌کنند (درویشی، خسروی و صدیق، ۲۰۲۰). استفاده از چنین مدل‌هایی به سیستم اجازه می‌دهد تا از یک تجمع وزنی استفاده کند که بر علائم ارائه‌شده توسط دانش آموزان قابل‌اطمینان‌تر تاکید دارد. یک خط تحقیقاتی مرتبط بر توسعه روش‌های بررسی نقطه تمرکز کرده‌است (وانگ و همکاران، ۲۰۱۸)؛ که به‌طور بهینه از حداقل در دسترس بودن مربیان برای بررسی بحث‌برانگیزترین موارد استفاده می‌کند (به‌عنوان‌مثال، مواردی که اعتماد الگوریتمی کم یا توافق بین ارزیابی کنندگان کم دارند) و توضیحاتی در مورد نتیجه برای یادگیرندگان فراهم می‌کند به‌طوری که آن‌ها می‌توانند بازخورد فردی ارزشمند دریافت کنند.

## نوشتن تجزیه و تحلیل

ارزیابی خودکار نوشته‌های دانشجویی از زمان حداقل ۱۹۶۶ یک حوزه تحقیقاتی غنی بوده‌است (پیج، ۱۹۶۶). درحالی‌که هم پاسخ‌های بلند مدت و هم پاسخ‌های کوتاه مورد بررسی قرار گرفته‌اند، موفق‌ترین رویکردها بر امتیاز دهی به کاره‌ای بلند مدت دانشجویی متمرکز شده‌اند. به‌عنوان‌مثال، چندین سیستم در عمل برای امتیاز دهی خودکار مقاله توسعه داده‌شده و مورداستفاده قرار گرفته‌اند که در میان آن‌ها MI Write یک نماینده است(گراهام، هربرت و هریس، ۲۰۱۵).

MI Write یک سیستم تعاملی مبتنی بر وب را برای دانشجویان جهت تمرین و بهبود مهارت‌های نوشتن آن‌ها ارائه می‌دهد. MI Write برای هر مقاله، یک دانش‌آموز را با یک امتیاز کلی برای مقاله و شش امتیاز ویژگی (به‌عنوان‌مثال، توسعه ایده‌ها، سازمان دهی، سبک، انتخاب کلمه، روانی جمله و کنوانسیون‌ها) برای دانش‌آموز فراهم می‌کند تا بر جنبه‌های خاص مقاله تمرکز کند .مطالعات متعددی نشان داده‌اند که ابزارهای نمره دهی خودکار مقاله مانند MI Write می‌توانند به دانش آموزان کمک کنند تا انگیزه نوشتن، خودکارآمدی نوشتن و مهارت‌های نوشتن خود را بهبود بخشند و به معلمان کمک کنند تا تمرینات خود را تسهیل کنند و به‌طور موثر بر انگیزه نوشتن و استقلال دانش آموزان تأثیر بگذارند (ویلسون و کزیک، ۲۰۱۶؛ ویلسون و روسکو، ۲۰۲۰؛ پالرمو و تامسون، ۲۰۱۸).

یک بررسی مفید از امتیازدهی مقاله خودکار توسط نگ و کی (۲۰۱۹) ارائه شد که انواع مختلف تکنیک‌های هوش مصنوعی توسعه‌یافته و به کار رفته در مساله را توصیف می‌کنند. به‌طورمعمول، این تکنیک‌های هوش مصنوعی، کار امتیاز دهی را به‌عنوان (الف) یک کار رگرسیون که هدف آن پیش‌بینی مستقیم یک نمره از یک مقاله و اغلب تکنیک‌های به کار گرفته‌شده مانند رگرسیون خطی (کراسلی، آلن، اسنو، مک نامارا، ۲۰۱۵) و رگرسیون بردار پشتیبان (کلبانوف، مدنی، بورستین، ۲۰۱۳)؛ (ب) یک وظیفه طبقه‌بندی که هدف آن طبقه‌بندی یک مقاله به یکی از مقوله‌های عددی (رادنر و لیانگ، ۲۰۰۲) و (ج) یک کار رتبه‌بندی با هدف مقایسه مقالات با توجه به کیفیت آن‌ها و اغلب تکنیک‌های به کار گرفته‌شده مانند ماشین‌های بردار پشتیبان و لامبدا ماری(یاناکوداکیس و بریسکاو، ۲۰۱۲؛ چن و هه، ۲۰۱۳). ابزارهای دیگر بیشتر بر ارائه بازخورد به دانشجویان تمرکز دارند تا یک ارزیابی کلی. برای مثال، آکادمی نویسندگان، پردازش زبان طبیعی و تطبیق الگو را برای شناسایی حضور و عدم حضور حرکات بلاغی خاص ترکیب می‌کند و بازخورد مربوطه را فراهم می‌کند (نایت و همکاران، ۲۰۲۰).

یک خط تحقیقاتی دیگر که ارتباط نزدیکی با امتیاز دهی خودکار مقاله دارد نرم‌افزار کشف سرقت ادبی است، به‌عنوان‌مثال، تورنیتین (هکلر، رایس و هابسن برایان، ۲۰۱۳). هدف از تورنیتین متفاوت از سیستم‌های مورداستفاده برای امتیازدهی خودکار مقاله، مقایسه ارائه از یک دانشجو در برابر مجموعه بزرگی از اسناد مرتبط است که ممکن است شامل مطالب ارسالی از دیگر دانشجویان، مقالات آنلاین و انتشارات دانشگاهی باشد. در مقایسه، ترنیتین گزارشی تهیه می‌کند تا نشان دهد که آیا بخش قابل‌توجهی از متن ارسالی با منبع دیگری مطابقت دارد یا خیر که مربیان می‌توانند از آن برای تعیین اینکه آیا یک مورد سرقت ادبی است یا خیر استفاده کنند. مرور ادبیات سیستماتیک اخیر(فولتی‌نک، موشکه و گیپ، ۲۰۱۹) پیشرفت قابل‌توجهی در کشف سرقت ادبی با افزایش استفاده از تکنیک‌های هوش مصنوعی به‌طور خاص، روش‌های تحلیل متن معنایی (به‌عنوان‌مثال، تجزیه و تحلیل معنایی نهفته و تعبیه کلمه)و الگوریتم های یادگیری ماشین.

## از گسسته تا پیوسته

درحالی‌که شیوه‌های ارزیابی سنتی ممکن است تصاویر مجددی از عملکرد بگیرند، چندین تکنیک هوش مصنوعی توسعه داده شده‌اند که نمای مستمرتری از عملکرد و درنتیجه بینش نسبت به یادگیری را فراهم می‌کنند برخی از این رویکردها از شیوه‌های ارزیابی سنتی مانند آزمون‌ها استفاده می‌کنند و آن‌ها را به محیط‌های دیجیتال منتقل می‌کنند، درحالی‌که برخی دیگر برای وظایف و شواهد ارزیابی کاملاً متفاوت به کار می‌روند.

## پلتفرم‌های ارزیابی الکترونیکی

در سال‌های اخیر، پلتفرم‌های ارزیابی الکترونیکی[1] که توانایی برگزاری امتحانات در داخل یا خارج از خط را فراهم می‌کنند، به‌طور فزاینده‌ای محبوب شده‌اند (لاماس–نیستال، فرناندز–ایگلسیاس، گونزالز–تاتو و میکیچ–فونته، ۲۰۱۳). مزایای کلیدی پلتفرم‌های ارزیابی الکترونیکی شامل ارائه توانایی برای ارائه سوالاتی است که ارسال آن‌ها بر روی کاغذ دشوار یا غیر ممکن خواهد بود، مانند سوالاتی که حاوی سؤالات چند رسانه‌ای با نظم از پیش تعیین‌شده یا تصادفی هستند و همچنین توانایی ارائه بازخورد سریع و شخصی به فراگیران است (دنیک، ویلکینسون و پورسل، ۲۰۰۹). همان‌طور که پلتفرم‌های ارزیابی الکترونیکی تکامل‌یافته اند، داده‌های استخراج‌شده از هر قسمت آزمون پیچیده‌تر شده‌اند و امکان بررسی فراتر از تکنیک‌های سنتی مانند تجزیه و تحلیل آیتم را فراهم می‌کنند. این داده‌ها ممکن است شامل مهره‌ای زمانی برای هر اقدام و پاسخ ایجادشده توسط یک امتحان شونده در سراسر آزمون آن‌ها باشد. این تصاویر نه تنها می‌توانند برای کشف اشکالات نرم‌افزاری و بررسی سو رفتار دانشگاهی مورداستفاده قرار گیرند، بلکه به‌طور فزاینده‌ای برای درک بهتر رفتار یادگیرندگان مورداستفاده قرار می‌گیرند. به‌طور خاص، تحقیقات قبلی بررسی کرده‌اند: اندازه‌گیری و طبقه‌بندی تلاش برای گرفتن آزمون(وایز و گاو، ۲۰۱۷)؛ پاسخ دادن و اصلاح رفتار در طول امتحانات (پاگنی و همکاران، ۲۰۱۷)؛ تنظیم فراشناختی استراتژی و پردازش شناختی (گلد همِر و همکاران، ۲۰۱۴)؛ اعتبار سنجی تفسیر نمره آزمون (انگل‌هارت و گلد همِر، ۲۰۱۹)؛ کشف رفتارهای سریع حدس زدن و پیش آگاهی (توتون و ماینز، ۲۰۱۹)؛ مدل‌سازی دقت، سرعت و بازدیدهایی را مورد بررسی قرار می‌دهد (بیرهان، فون داویر، یوجین گرابوسکی، ۲۰۲۱)؛ مدل‌سازی دانشجویان در زمان واقعی به هنگام انجام یک خودارزیابی (پاپامیتسیو و ایکونومیدس، ۲۰۱۷)؛ و درک عملکرد دانش آموزان در زمینه‌های مختلف مانند حل مساله پیچیده (گریف، اشتادلر، سون‌لایتنر، ولف و مارتین، ۲۰۱۵).

## ارزیابی داشته ها

تکنیک‌های ارزیابی پنهانی به‌طور نسبی داده‌هایی را جمع‌آوری می‌کنند که فراتر از این می‌روند که آیا دانش آموزان به سادگی سؤالات را به درستی پاسخ داده‌اند یا خیر .عبارت ارزیابی پنهانی توسط شوت و ونتورا (۲۰۱۳) برای روشی ابداع شد که در آن آن‌ها از داده‌های جمع‌آوری‌شده به‌طور خودکار از زبان آموزان در حین بازی دیجیتال استفاده می‌کردند. آن‌ها با جمع‌آوری داده‌های تولید شده در یک بازی فیزیک دیجیتال که معمولاً در مدارس استفاده می‌شود، معیارهای وظیفه‌شناسی، خلاقیت و توانایی فیزیک را توسعه دادند. آن‌ها مدل‌هایی از مسیر مورد انتظار رفتار در بازی را با افزایش توانایی دانش آموزان که یک نقشه ساختاری نامیده می‌شود، ساختند (ویلسون، ۲۰۰۵). سپس داده‌ها برای قرار دادن هر یادگیرنده بر روی این نقشه مورداستفاده قرار گرفت و یک ارزیابی پویا از افزایش قابلیت یادگیرنده در حین بازی ایجاد شد.

همان طور که در ابتدا تصور می‌شد، ارزیابی پنهانی چهار مولفه حیاتی دارد: (الف) طراحی ارزیابی مدرک محور، (ب) ارزیابی سازنده و بازخورد برای حمایت از یادگیری، (ج) حمایت از تصمیمات آموزشی و (د) استفاده از مدل‌های یادگیرنده که ممکن است شامل اطلاعات شناختی یا غیر شناختی باشد (میسلوی و همکاران، ۲۰۰۳؛ شوت، ۲۰۱۱). به‌طورمعمول، ارزیابی پنهانی به دنبال نمونه شووت شامل ثبت بدون مانع آثار رفتار یادگیرنده در محیط‌های بازی دیجیتال و مدل‌سازی فراگیران از طریق روش‌هایی مانند شبکه‌های بیزی است (پرل، ۱۹۸۸).

درحالی‌که ارزیابی پنهانی به یک رویکرد طراحی ارزیابی خاص اشاره دارد، عناصر آن به‌طور گسترده در استفاده از محیط‌های یادگیری دیجیتال به‌طور کلی مورداستفاده قرار گرفته‌اند .با استفاده از تکنیک‌های مشابه، گریفین و کر (۲۰۱۵) از داده جریان لگاریتمی تولید شده از بازی‌های دیجیتال دو بازیکن برای ارزیابی عملکرد دانش‌آموز در حل مساله مشارکتی استفاده کردند.

ویلسون و اسکالیس (۲۰۱۲)از رویکردی مشابه با داده‌های جریان لگاریتمی تولید شده از وظایف آنلاین انجام‌شده توسط دانشجویان برای تولید معیارهای توانایی دانش‌آموز برای یادگیری در محیط‌های دیجیتال شبکه‌ای استفاده کردند. هر یک از این مطالعات از وظایف دیجیتالی سفارشی

[1] electronic assessment platforms

برای تولید داده‌ها استفاده کردند. میلیگان و گریفین ( ۲۰۱۶ ) این روش را برای استفاده از داده فرآیند به‌دست‌آمده در پلتفرم های باز، با استفاده از داده جریان لگاریتمی موک ها برای تولید ارزیابی‌های نمایندگی یادگیرنده گسترش دادند. روش‌های پنهانی در حال حاضر به‌طور مکرر در بازی‌ها و پلتفرم‌های تجاری برای یادگیری مورداستفاده قرار می‌گیرند (شوت و همکاران، ۲۰۲۱).

## تخمین دامنه دانش

یک جز کلیدی هم پلتفرم‌های ارزیابی الکترونیکی و هم ارزیابی پنهانی، توانایی ردیابی مداوم اقدامات دانشجویی و ترکیب این اقدامات در مدل‌های عملکرد و یادگیری است. یک تکنیک هوش مصنوعی که به‌طور گسترده برای تولید این نوع مدل‌ها مورداستفاده قرار می‌گیرد، تخمین دانش پنهان است (کوربت و اندرسون، ۱۹۹۴). دلیل این امر به‌عنوان نهفته در این حقیقت نهفته‌است که دانش را نمی توان به‌طور مستقیم مشاهده کرد. چیزی که می توان مشاهده کرد این است که آیا یک یادگیرنده می‌تواند یک جز دانش را در برخی زمینه‌ها به کار گیرد یا خیر. این امر در سیستم‌های آموزشی هوشمند برای جمع‌آوری داده‌ها در مورد اقدامات یادگیرندگان برای فرصت‌های یادگیری خاص و این‌که آیا آن‌ها می‌توانند به درستی مولفه‌های دانش متمایز را به کار ببرند یا خیر، استفاده می‌شود (دسمارایس و بیکر، ۲۰۱۲). این نشان می‌دهد که یادگیرندگان می‌توانند یک نقطه داده باینری را برای هر فرصت یادگیری که آن‌ها در استفاده از مولفه‌های دانش موفق یا ناموفق بودند، تولید کنند.

ردیابی دانش بیزی[1] بهترین روش برای تخمین دانش نهفته است (کوربت و اندرسون، ۱۹۹۴). این تکنیک از چهار پارامتر برای تخمین اینکه آیا یک یادگیرنده می‌تواند یک مولفه دانش را به کار ببرد استفاده می‌کند، از جمله (الف) احتمال این‌که یادگیرنده در حال حاضر یک مولفه دانش را مدیریت می‌کند، (ب) احتمال یادگیری یک مولفه دانش پس از یک فرصت یادگیری، (ج) احتمال استفاده صحیح از یک مولفه دانش حتی زمانی که یادگیرنده بر آن تسلط نیافته است (حدس زدن) و (د) احتمال استفاده نادرست از یک مولفه دانش اگرچه آن‌ها آن را (لغزش) می‌دانند. درحالی‌که ردیابی دانش بیزی به‌طور گسترده محبوب شده است، تکنیک‌های دانش جدید اخیرا براساس پیشرفت‌هایی در یادگیری عمیق (گروت، کودینگر، اشنایدر و میچل، ۲۰۲۰)؛ از جمله استفاده از شبکه‌های عصبی بازگشتی و ترانسفورماتورها پیشنهاد شده است (پیچ و همکاران، ۲۰۱۵؛ شین و همکاران، ۲۰۲۱). ردیابی دانش همچنین به‌عنوان پایه‌ای برای توسعه یک لحظه تکنیکی با یادگیری دانش استفاده شده است (بیکر، گلدشتاین و هفرنان، ۲۰۱۱؛ ۲۰۱۳) که می‌تواند لحظه دقیق تسلط یک یادگیرنده بر یک مهارت خاص را استنتاج کند. این تکنیک نه تنها برای یادگیری در مورد موضوعات خاص به کار گرفته شده است، بلکه همچنین برای تخمین چگونگی خود تنظیمی یادگیرندگان (مولنار، هوروزر، کورنت بیکر، ۲۰۲۱)و ارائه تجسم شخصی نیز به کار رفته‌است (مولنار، ۲۰۲۲).

## فرآیندهای یادگیری

شیوه ارزیابی سنتی بر قضاوت در مورد یک محصول تولید شده توسط یادگیرنده، مانند یک مقاله، یک گزارش آزمایشگاهی یا یک برگه بررسی کامل تمرکز کرده‌است. دلیل اصلی دشوار بودن، اگر غیر ممکن نباشد، پی‌گیری فرآیندهای یادگیری این است که زمان و منابع زیادی وجود دارد. نظارت مداوم بر پیشرفت و جمع‌آوری مداوم شاخص‌هایی که اجازه استنباط فرآیندهای شناختی و فراشناختی را می‌دهند، موردنیاز است. این داده‌ها می‌توانند شامل خود گزارش دهی، رفتاری، روان – فیزیولوژی و دیگر داده‌ها باشند. جمع‌آوری و تجزیه و تحلیل این داده‌ها تا به امروز دشوار بوده‌است و نیازمند تجهیزات تخصصی، آزمایشگاه‌ها و تجزیه و تحلیل است. براساس رویکردهایی مانند ارزیابی پنهانی که قبلا بحث شد، هوش مصنوعی می‌تواند برای درک بهتر روندها در فرآیندهای یادگیری مورداستفاده قرار گیرد.

پیشرفت‌های اخیر در جمع‌آوری داده‌های چند وجهی، تجزیه و تحلیل یادگیری و هوش مصنوعی فرصت‌هایی را برای بهبود ارزیابی فرآیندها فراهم می‌کند. به‌عنوان‌مثال، استفاده از داده‌های چندکاناله مانند کلیجریان، حرکات ماوس و ردیابی چشم (آزودو و گاشوویچ، ۲۰۱۹)؛ (یارولاً، مالمبرگ، هاتاجا، سوبوسینسکی و کرشنر، ۲۰۲۰) همراه با ابزارهای پیشرفته محیط‌های یادگیری مانند استفاده از هایلایت ها (ون در گراف و همکاران، ۲۰۲۱)؛ (یوانوویچ، گاشوویچ، پاردو، داوسون و وایتلاک–وینرایت، ۲۰۱۹)؛ (ژو و وینی، ۲۰۱۲) می‌توانند گزارش‌های تجربی در مورد فرایندهای مرتبط با انگیزش، عاطفه، شناخت و فراشناخت ارائه دهند. مسیرهای مناسب برای ارزیابی فرآیندهای یادگیری با تجزیه و تحلیل داده‌های چند کاناله با هوش مصنوعی و تکنیک‌های یادگیری ماشینی مختلف مانند یادگیری عمیق، کاوش فرآیند و تجزیه و تحلیل شبکه توسعه داده می‌شوند (احمد اوزیر، گاشوویچ، ماچا، یوانوویچ و پاردو، ۲۰۲۰)؛ (فان، سنت، سینگ، یوانوویچ و گاشوویچ، ۲۰۲۱)؛ (سنت، گاشوویچ، ماچا، اوزیر و پاردو، ۲۰۲۰).

[1] Bayesian knowledge tracing

## از یکنواخت تا سازگار

به‌جای دادن یک کار ارزیابی یک‌سان برای همه دانش آموزان، تکنیک‌های هوش مصنوعی توسعه داده شده‌اند که کار را با توانایی‌های دانش آموزان تنظیم می‌کنند و به آن‌ها تجربیات ارزیابی متناسب می‌دهند.

سیستم‌های تست تطبیقی کامپیوتری[1] آزمایشی را با استفاده از مجموعه‌ای از سؤالات به‌طور متوالی اجرا می‌کنند تا دقت تخمین فعلی سیستم از توانایی دانش‌آموز را به حداکثر برسانند .پنج مولفه فنی متصل به هم برای ساخت یک تست تطبیقی کامپیوتری وجود دارد: (۱)مجموعه‌ای از آیتم‌های کالیبره شده با داده‌های قبل از آزمایش؛ (تامپسون، ۲۰۰۷)، (۲)یک نقطه شروع خاص برای هر امتحان شونده؛ (۳)یک الگوریتم انتخاب آیتم برای انتخاب آیتم بعدی؛ (۴)یک الگوریتم امتیاز دهی برای تخمین توانایی امتحان کنندگان و (۵)معیار پایانی برای امتحان.

نظریه سوال – پاسخ[2]، یک تکنیک روان‌سنجی رایج است که در بسیاری از تست های تطبیقی کامپیوتری ها برای درجه‌بندی آیتم‌ها استفاده می‌شود (امبرتسون و ریس، ۲۰۱۳). یکی از ویژگی‌های کلیدی نظریه سوال پاسخ که آن را برای تست تطبیقی کامپیوتری مناسب می‌سازد این است که توانایی امتحان شونده و سطح دشواری آیتم‌ها را در همان معیار قرار می‌دهد که به الگوریتم انتخاب آیتم کمک می‌کند تصمیم بگیرد که کدام آیتم باید در مرحله بعد اجرا شود .به‌طور نظری، یک امتحان شونده زمانی که آیتم‌های تست نه خیلی سخت و نه خیلی آسان هستند، به‌طور بسیار موثری اندازه‌گیری می‌شود .با توجه به اینکه نظریه سوال پاسخ شرکت‌کنندگان امتحان و آیتم‌هایی را بر روی یک معیار قرار می‌دهد، می‌تواند آیتمی را شناسایی کند که با توانایی فعلی کاربر مطابقت داشته باشد. درنتیجه، اگر امتحان شونده به یک آیتم به درستی پاسخ دهد، مورد بعدی انتخاب‌شده باید سخت‌تر باشد؛ اگر پاسخ اشتباه باشد، سوال بعدی باید ساده‌تر باشد.

برای عملیاتی کردن تست انطباقی، اندازه مخزن آیتم باید به اندازه کافی بزرگ باشد تا الگوریتم انتخاب بتواند یک آیتم مناسب را براساس توانایی فعلی امتحان شونده مدیریت کند. یک عامل مهم در تست تطبیقی کامپیوتری نقطه شروع است. اگر سیستم اطلاعاتی در مورد امتحان شونده داشته باشد، می‌تواند نقطه شروع به توانایی آن‌ها را بهینه کند؛ در غیر این صورت، می توان فرض کرد که آزمونگر از توانایی متوسط برخوردار است. هنگامی که یک آیتم دریافت می‌شود، تست تطبیقی کامپیوتری تخمین خود از سطح توانایی امتحان شونده را به روز می‌کند. این کار معمولاً با به روز رسانی تابع پاسخ آیتم با استفاده از برآورد حداکثر احتمال و تخمین بیزی (سورل، بارادا، د لا توره و آباد، ۲۰۲۰) یا سیستم‌های رتبه‌بندی مانند رتبه‌بندی الو (عبدی، خسروی، صدیق و گاشوویچ، ۲۰۱۹؛ فرشوور، برگر، موزر و کلاینتیس، ۲۰۱۹) انجام می‌شود. درنهایت، آزمون معمولاً زمانی پایان می‌یابد که سیستم توانایی دانش‌آموز را با سطح اعتمادی که از آستانه تعیین‌شده توسط کاربر فراتر می‌رود، تخمین بزند. نشان داده شده است که تست های تطبیقی کامپیوتری توانایی کوتاه کردن آزمون تا ۵۰ ٪ را دارد درحالی‌که قابلیت اطمینان بالاتر را در مقایسه با آزمون‌های معمولی حفظ می‌کند (کلارز و سسیلیو–فرناندس، ۲۰۱۹).

## از غیر اصیل به غیر اصیل

ارزیابی‌های معتبر، یادگیری را با استفاده از وظایفی که اعضای واقعی برخی از انجمن‌های عمل را شبیه‌سازی می‌کنند، اندازه‌گیری می‌کنند (ریوز و اوکی، ۱۹۹۶). تکنیک‌های هوش مصنوعی در حال حاضر برای تقویت کاره‌ای شبیه‌سازی شده و تجزیه و تحلیل شواهد مربوط به آن‌ها مورداستفاده قرار می‌گیرند.

در هر دو محیط یادگیری مجازی و فیزیکی، هوش مصنوعی نقش مهمی ایفا می‌کند. برای مثال، در شبیه‌سازی‌های مجازی که کارآموزی مجازی نامیده می‌شود، یادگیرندگان در یک شرکت داستانی که در آن در تیم‌هایی برای طراحی یک محصول کار می‌کنند، شرکت می‌کنند (شفر، ۲۰۰۶ الف، ۲۰۰۶ ب). هدف از دوره‌های کارآموزی مجازی این است که به فراگیران تجربه عملی کارهایی را بدهد که متخصصان واقعی انجام می‌دهند، مانند هدایت تحقیقات زمینه‌ای، برگزاری جلسات طراحی، گزارش به ناظران و توسعه و آزمایش نمونه‌های اولیه. در شبیه‌سازی‌های آفلاین مانند شبیه‌سازی‌های مورداستفاده در بهداشت و درمان، دانشجویان و متخصصان، دانش بالینی حیاتی را در موقعیت‌های نزدیک به زندگی واقعی به کار می‌برند (به‌عنوان‌مثال، پرداختن به واکنش آنتی‌بیوتیک، شبیه‌سازی جراحی) (سالیوان و همکاران، ۲۰۱۸؛ اچوریا، مارتینز–مالدونادو و باکینگهام شوم، ۲۰۱۹). فضاهای یادگیری فیزیکی از نزدیک آن فضاهایی که دانش آموزان در آینده تجربه خواهند کرد را تقلید می‌کنند.

[1] Computerised adaptive testing systems
[2] Item-response theory

شبیه‌سازی برای یادگیری برای کمک به یادگیرندگان برای انجام کارهایی که متخصصان انجام می‌دهند، طراحی‌شده است؛ اما در دنیای واقعی ممکن است بسیار دشوار، گران یا خطرناک باشد که اجازه این کار را به آن‌ها بدهیم. مهم‌تر اینکه، آن‌ها لزوما فاقد تخصص برای انجام این کار هستند. به هر حال این تخصص چیزی است که آن‌ها سعی می‌کنند یاد بگیرند. برای پرداختن به این موضوع، دوره‌های کارآموزی مجازی از هوش مصنوعی برای ایجاد محیطی استفاده می‌کنند که در آن برای دانشجویان امکان پذیر، ایمن و موثر است تا مانند حرفه‌ای‌ها عمل کنند. این کار از طریق ابزارهای حرفه‌ای شبیه‌سازی شده، پیام‌های خودکار از همکاران و ناظران و بازخورد خودکار در مورد محصولات کاری انجام می‌شود. به‌طور مشابه، پرستاران و پزشکان آینده‌نگر با بیماران واقعی در شبیه‌سازی‌های مراقبت‌های بهداشتی فیزیکی کار نمی‌کنند. در برخی موارد، آن‌ها با بیماران شبیه‌سازی شده‌ای کار می‌کنند که از هوش مصنوعی برای رفتار مانند بیماران واقعی استفاده می‌کنند، برای مثال، آن‌ها علائم خاصی را در زمان‌های خاص نشان می‌دهند (اچوریا و همکاران، ۲۰۱۹).

علاوه بر افزایش وظایف ارزیابی و محیط، هوش مصنوعی ممکن است داده‌ها را از ارزیابی‌های معتبر جمع‌آوری، ارائه و ارزیابی کند. با توجه به اینکه ارزیابی‌های معتبر ممکن است شامل افراد یا گروه‌های متعددی باشد که وظایف پیچیده و بد تعریف را انجام می‌دهند، می‌تواند برای مربیان چالش برانگیز باشد که از تمام اتفاقاتی که در طول شبیه‌سازی در حال رخ دادن است آگاه باشند و بازخورد دقیقی، به خصوص برای گروه‌های بزرگ فراهم کنند (مورفی، فاکس، فریمن و هیوز، ۲۰۱۷). همانند ارزیابی پنهانی و ارزیابی‌های مبتنی بر هوش مصنوعی از فرآیندهای یادگیری، هوش مصنوعی راهی برای پرداختن به پیچیدگی این شرایط ارزیابی از طریق جمع‌آوری و مدل‌سازی داده یکپارچه است.

به‌عنوان‌مثال، در دوره‌های کارآموزی مجازی، پلت فرم آنلاین به‌طور خودکار پیام‌های چت دانشجویی را گزارش می‌کند. برای مرتبط کردن این شواهد با ادعاهای مربوط به یادگیری، یک الگوریتم یادگیری ماشین نظارت شده برای طبقه‌بندی خودکار چت ها به‌عنوان شواهدی از عناصر یک چارچوب معرفتی و تحلیل شبکه شناختی استفاده می‌شود (شفر، کولیر و رویس، ۲۰۱۶) که برای شناسایی روابط بین این عناصر استفاده می‌شود. داشبورد این تکنیک‌ها را در نمایش‌های زنده شبکه‌های معرفتی ادغام می‌کند که مربیان می‌توانند از آن‌ها برای نظارت بر تعامل گروه و مداخلات برنامه‌ریزی در زمان واقعی استفاده کنند (هردر و همکاران، ۲۰۱۸).

در شبیه‌سازی‌های آفلاین، تجزیه و تحلیل یادگیری چند وجهی در حال توسعه است تا میلیون ها نقطه داده از جمله نمودارهای سیستم، مختصات موقعیت، گفتار و مسیرهای فیزیولوژیکی در فضاهای فیزیکی و در زمان نسبتا کوتاهی را به دست آورد. ممکن است هوش مصنوعی برای عملکرد این سنسورها همانند ابزارهای رونویسی خودکار ضروری باشد. برای در دسترس قرار دادن این داده‌ها برای آموزگاران، یک رویکرد که اتخاذ شده است استفاده از اصول داستان‌سرایی داده برای ایجاد واسطهایی است که در آن داستان‌ها از داده‌های چند وجهی پیچیده برای تمرکز بر روی یک هدف یادگیری یا انعکاس در یک‌زمان استخراج می‌شوند. به‌عنوان‌مثال، اچورریا و همکاران (۲۰۲۰) بر ایجاد داستان‌های داده مربوط به خطاهای رایج انجام‌شده توسط دانشجویان پرستاری براساس ارزیابی‌های خودکار توالی و به موقع بودن اقدامات ثبت‌شده آن‌ها تمرکز کردند.

## از باستانی تا مدرن

رسانه‌های محاسباتی مانند کامپیوتر، ماشین‌حساب و نرم‌افزار امکان پردازش اطلاعات را به روش‌های جدید و قدرتمند میسر می‌سازند. درحالی‌که رسانه‌های محاسباتی در حوزه‌های مختلف وجود دارند، در اینجا به‌طور خلاصه بر روی برخی از آن‌هایی که برای نوشتن وظایف به‌عنوان‌مثال توسعه داده شده‌اند، تمرکز می‌کنیم.

پردازشگرهای دیجیتال کلمه حداقل از دهه ۱۹۷۰ مورداستفاده قرار گرفته‌اند (برگین، ۲۰۰۶). علاوه بر ضبط و ذخیره‌سازی ساده متن، وظیفه اصلی آن‌ها تخلیه وظایف نوشتاری معمول، مانند ویرایش، از انسان به کامپیوتر بوده‌است. پردازشگرهای دیجیتال کلمه معمولاً شامل تکنیک‌های خودکار برای بررسی هجی کردن، دستور زبان و استفاده هستند. همان‌طور که این ابزارها توسعه می‌یابند، به‌طور فزاینده‌ای برای تکمیل وظایف پیچیده‌تر به هوش مصنوعی تکیه می‌کنند.

امروزه پردازشگرهای کلمه دیجیتال مانند مایکروسافت ورد و گوگل شامل تکنیک‌های هوش مصنوعی هستند که تکمیل کلمه و جمله را نشان می‌دهند (مایکروسافت، ۲۰۲۲). ابزارهای تجاری دیگر، مانند گراممارلی، شامل هوش مصنوعی هستند که لحن و سبک را القا می‌کند (گرامرلی، ۲۰۲۲). در حال حاضر ابزارهای مبتنی بر هوش مصنوعی مانند سودووریت وجود دارند که بخش‌های کاملاً جدیدی از متن را براساس چند خط نمونه ایجاد می‌کنند (مارشه، ۲۰۲۱). از آنجا که این ابزارها ممکن است توسط یادگیرندگان و متخصصان در فعالیت‌های روزمره شان

مورداستفاده قرار گیرند، طرح‌های ارزیابی ممکن است آن‌ها را با هم ترکیب کنند. استفاده از ابزار برای انجام وظایف پیچیده و انسانی دارای مفاهیم مهمی برای ارزیابی است که برخی از آن‌ها را در زیر مورد بحث قرار می‌دهیم.

## چالش‌های هوش مصنوعی و ارزیابی

تا کنون، ما مجموعه‌ای از مسائل با الگوی ارزیابی استاندارد را برجسته کرده‌ایم و برخی از رویکردهای مبتنی بر هوش مصنوعی را که در این مسائل وجود دارند مرور کرده‌ایم. درحالی‌که بخش‌های بالا نشان می‌دهند که هوش مصنوعی می‌تواند الگوی ارزیابی استاندارد را بهبود بخشد ما تصدیق می‌کنیم که این پارادایم دارای یک تاریخ طولانی و مسلما موفق است. پس ارزش آن را دارد که در مورد چیزی که ممکن است از دست بدهیم یا مشکلات دیگری که ممکن است با معرفی هوش مصنوعی به این پارادایم معرفی کنیم، تامل کنیم.

## کنار گذاشتن تخصص حرفه‌ای

بسیاری از محققان به دنبال توسعه فن‌آوری‌های هوش مصنوعی هستند که تصمیم‌گیری معلمان را حمایت و هدایت می‌کند، معلمان را از وظایف و تصمیمات معمولی و بدون نزاع رها می‌کند درحالی‌که به تعویق انداختن قضاوت و نظارت نهایی معلمان ادامه می‌دهد (هردِر و همکاران، ۲۰۱۸). در این مفهوم، این اطمینان بخش است که تصور کنیم ارزیابی مبتنی بر هوش مصنوعی، انسان در حلقه را حفظ خواهد کرد، با معلمانی که قادر به نظارت و برتری بر هر گونه تصمیم‌گیری خودکار در زمان مشاهده تناسب هستند.

بااین‌حال، یک خطر بالقوه تصمیم‌گیری خودکار، کنار گذاشتن تخصص حرفه‌ای است، یعنی، محاسبات و خروجی‌های ماشین به تعویق می‌افتد یا به‌طور خودکار به‌عنوان صحیح در نظر گرفته می‌شود .یک مثال فرضی از این موضوع را می توان با نرم‌افزار سرقت ادبی در موسسات آموزشی مشاهده کرد. در گذشته، معلمان در مورد اینکه آیا مطالب ارسالی دانش آموزان بسیار شبیه به یکدیگر هستند یا منابع موجود تصمیم‌گیری می‌کردند. بااین‌حال، با توجه به حجم منابع ممکن و پیشرفت‌ها در پردازش زبان طبیعی، هوش مصنوعی اکنون می‌تواند این کار را در بسیاری از زمینه‌ها انجام دهد .با توجه به دشواری این کار و کارایی الگوریتم های موجود، ممکن است و شاید برای مربیان آسان باشد که خروجی خود را به‌عنوان یک تصمیم صحیح به‌جای یک پیشنهاد آزمایشی در نظر بگیرند. برای به چالش کشیدن خروجی‌های این سیستم‌ها، یک معلم بااعتماد به نفس و غنی از زمان نیاز است. به‌این‌ترتیب، نگرانی‌های قابل درکی وجود دارد که ما با چشم‌انداز ظرفیت تصمیم‌گیری معلمان مواجه هستیم که "به‌عنوان سیستم‌های ارزیابی خودکار، فاصله بین تصمیمات آن‌ها و فرایندهای جمع‌آوری شواهد که آن تصمیمات باید بر آن تکیه کنند را برجسته می‌کند".

برای جلوگیری از این تخلیه، محققان شروع به طراحی سیستم‌هایی کرده‌اند که در آن فرآیندهای تصمیم‌گیری قابل توضیح به معلم هستند (روزه، مک‌لاکلین، لیو و کوئدینگر، ۲۰۱۹؛ خسروی و همکاران، ۲۰۲۲). درحالی‌که این یک مسیر امیدوار کننده است، برای درک بهتر تعادل بین هوش مصنوعی و تصمیم‌گیری معلم که برای تدریس، یادگیری و ارزیابی بهتر است، کار بیشتری موردنیاز است.

## جعبه سیاه پاسخگویی

درحالی‌که بسیاری از محققان ممکن است استدلال کنند که قصد آن‌ها برای انجام این کار نیست (بیکر، ۲۰۱۶)، بیرون کشیدن معلمان انسان از حلقه ارزیابی به احتمال زیاد یک چشم‌انداز جذاب برای بسیاری از ذینفعان کلیدی درگیر در مدرسه و آموزش دانشگاه است. موسسات آموزشی ممکن است از ظرفیت تولید قابل‌اطمینان و به موقع داده‌های ارزیابی در مقیاس اجتناب کنند و از تناقض‌های ناشی از نشانه‌گذاری یا تاخیر ناشی از عدم انجام به موقع آن اجتناب کنند. به‌طور مشابه، بسیاری از معلمان ممکن است خوشحال شوند که مسئولیت را به تعویق انداخته و از انجام کار ناشیانه نمره دهی شخصی به دانش آموزان که آن‌ها به‌طور خاص با توجه به تمایلات فعلی دانش آموزان برای درخواست تجدیدنظر و اعتراض به نمرات و حتی شروع اقدامات قانونی در مورد سو نمره دادن به آن‌ها آشنا شده‌اند، طفره بروند.

دانش آموزان نیز ممکن است از این گزینه استقبال کنند که مجبور نباشند خود را در معرض آسیب‌پذیری قضاوت معلمان، مدارس یا دیگر موسسات اجتماعی نزدیک به خانه قرار دهند، به عبارت دیگر، اختلافات ارزیابی شدن توسط افرادی که واقعا آن‌ها را می‌شناسند.

بااین‌حال، ارزیابی مبتنی بر هوش مصنوعی یک مورد ساده از تعویق قضاوت‌های آموزشی به نگاه بی‌طرف، عینی و قابل‌اعتماد ماشین نیست. چیزی به‌عنوان ارزیابی بی‌طرف و غیر انسانی وجود ندارد (مایفیلد و همکاران، ۲۰۱۹؛ شونمان، ۱۹۷۹). در عوض، ارزیابی مبتنی بر هوش مصنوعی را می‌توان با دقت بیشتری به‌عنوان انتقال آن تصمیمات به برنامه‌نویسان، مهندسان یادگیری، طراحان آموزشی، فروشندگان نرم‌افزار و دیگر انسان‌ها توصیف کرد که هیچ دانش مستقیمی از ارزیابی دانشجویان، زمینه‌های محلی آن‌ها و یا حتی لزوماً سیستم‌های آموزشی که آن‌ها در آن مطالعه می‌کنند ندارند؛ بنابراین، همانند هر شکلی از ارزیابی، ارزیابی فعال شده با هوش مصنوعی یک فرآیند عینی – جزیی است.

همان‌طور که هنسورث و همکارانش گفته‌اند: مهم نیست که ساختارها و فرآیندها در جای خود قرار گیرند، ارزیابی‌ها توسط انسان‌ها طراحی و ارزیابی می‌شوند، با تمام پیش‌زمینه‌های اجتماعی – فرهنگی پیچیده، تجربیات آموزشی و ارزش‌های فکری و شخصی شان.
در مورد ارزیابی مبتنی بر هوش مصنوعی، مسئولیت مدل‌سازی و اجرای ارزیابی آموزشی به دیگران دور از دسترس (برنامه نویسان، مهندسان یادگیری)به تعویق می‌افتد. از یک طرف، می توان از این امر به‌عنوان فاصله گرفتن تصمیمات ارزیابی از تعصبات و فرضیات معلمان کلاس درس استقبال کرد .بااین‌حال، از سوی دیگر، این امر نگرانی‌هایی را مطرح می‌کند که باید به‌طور جدی‌تر در مورد این‌که چگونه ارزیابی هوش مصنوعی شکل‌گرفته فراگیر را در معرض تعصبات، ارزش‌ها، فرضیات افراد دیگری قرار می‌دهد که در غیر این صورت هیچ دانش یا سرمایه‌گذاری شخصی در کسانی که ارزیابی می‌شوند ندارند.
حداقل، در شرایط عملی، این نگرانی‌ها نیاز مبرم به نظارت دقیق بر هر ارزیابی با هوش مصنوعی و ایجاد خطوط روشن پاسخگویی برای تصمیماتی که این سیستم‌ها و نرم‌افزار تولید می‌کنند و همچنین خطوط روشن پاسخگویی برای اینکه چگونه خروجی‌های نرم‌افزار سپس توسط موسسات آموزشی به نمرات نهایی ترجمه می‌شوند را افزایش می‌دهد.

## محدود کردن نقش آموزشی ارزیابی

در میان شور و شوق فعلی برای ارزیابی فعال شده با هوش مصنوعی، شناخت کمی از نقش آموزشی ارزیابی وجود دارد. این مربوط به این ایده است که ارزیابی آموزشی تنها موضوع سنجش آنچه دانش‌آموز آموخته‌است نیست (ویلیام، ۲۰۱۱). در عوض، هنگام در نظر گرفتن پیامدهای افزایش استفاده از ارزیابی‌های مبتنی بر هوش مصنوعی، مهم است که در نظر بگیریم چگونه این امر ممکن است بر توانایی معلمان برای تعامل با ارزیابی به‌عنوان یک عمل آموزشی تأثیر بگذارد.
به‌عنوان‌مثال، در سطح شخصی، معلمان اغلب از اشکال سنتی ارزیابی نمره معلم برای انگیزه دادن، حمایت و تشویق دانش آموزان استفاده می‌کنند (کالی و مک‌میلان، ۲۰۱۰؛ هارلن، ۲۰۱۲). زمانی که معلم احساس می‌کند یک دانش‌آموز از تشویق شدن و مشاهده موفقیت بهره‌مند خواهد شد، این امر ممکن است شامل ملایمت و ملایمت باشد. از طرف دیگر، این ممکن است شامل تنبیه بیشتر باشد که در آن یک معلم احساس می‌کند که یک دانش‌آموز ممکن است از یک مداخله بهره‌مند شود. در هر دو مورد، عمل ارزیابی ریشه در روابط شخصی و دانشی دارد که یک معلم بادانش‌آموز خود ایجاد کرده‌است.
بسیاری از معلمان نیز توجه زیادی به آنچه از هر اقدام ارزیابی یاد گرفته می‌شود، دارند. این امر در برخی از مربیان به‌طور ضمنی از اشکال جایگزین ارزیابی استفاده می‌کنند. به‌عنوان‌مثال، افزایش محبوبیت ارزیابی همتا در درجه اول به‌عنوان ابزاری برای تشویق به خوداندیشی در میان دانشجویان در کار خود ریشه دارد (چو و چو، ۲۰۱۱؛ تاپینگ، ۲۰۱۸). روند اجازه دادن به خود ارزیابی دانش‌آموز محور به‌طور مشابه براساس نیت توسعه مشورت دانش‌آموز در شیوه‌های یادگیری خود فرد است. به‌طور مشابه، علاقه رو به رشد به استفاده از "ارزیابی عدالت اجتماعی به دنبال حمایت از تعامل دانشجویان با دیدگاه‌های متعدد و مورد اعتراض و تعامل با تغییرات ناشی از تفاوت‌های متنی، جنبه‌های تاریخی و هنجاری شخصی است.
به‌عنوان‌مثال، این امر ممکن است مستلزم این باشد که به دانش آموزان اجازه دهیم نقش رهبری را در تصمیم‌گیری جمعی در مورد ماهیت و شکل ارزیابی آن‌ها داشته باشند. در تمام موارد، هدف حمایت از دانشجویان برای انعکاس فرآیندها و شیوه‌های آموزشی به‌جای تولید یک معیار عینی یادگیری است.
نگرانی‌ها را می‌توان مطرح کرد که برخی از ارزیابی‌های هوش مصنوعی مانع از استفاده معلمان از ارزیابی در این روش‌های جایگزین می‌شوند. بااین‌حال، چنین مثال‌هایی ماهیت ارزش محور چگونگی انجام ارزیابی آموزشی یک جنبه را برجسته می‌کنند که در بسیاری از بحث‌های ارزیابی با هوش مصنوعی برجسته نشده است. ایده ارزیابی عدالت اجتماعی قطعا مجموعه‌ای متمایز از ارزش‌ها را در مورد اینکه آموزش چیست و آموزش برای چیست، منتقل می‌کند. این امر به نوبه خود پرسش‌هایی را در مورد ارزش‌های ضمنی و اصول ایدئولوژیک ارزیابی فعال شده توسط هوش مصنوعی مطرح می‌کند. محققان همچنین شروع به پرداختن به این موضوع، حداقل به‌طور ضمنی کرده‌اند. به‌عنوان‌مثال، چندین محقق خواستار نقش برجسته تری برای نظریه آموزشی و یادگیری در توسعه روش‌های هوش مصنوعی شده‌اند (راجرز، گاسویچ و داوسون، ۲۰۱۶) این نظریه‌ها موضعی در مورد آنچه با توجه به یادگیری ارزشمند است اتخاذ می‌کنند و ممکن است به‌طور قابل‌توجهی متفاوت از روش‌های انتقال محور باشند، در عوض، برای مثال، تمرکز بر ارتقا ایده‌آل‌های جوامع خاص عمل (شفر، ۲۰۰۶ الف، ۲۰۰۶ ب) یا توانایی تنظیم یادگیری یک فرد (آزودو و گاسویچ، ۲۰۱۹؛ مولنار، ۲۰۲۲).

## ارزیابی اشکال محدود یادگیری

گسترش ایده ارزیابی هوش مصنوعی به‌عنوان محدود کردن اشکال مختلف تدریس، نگرانی‌هایی در مورد اشکال محدود یادگیری ضمنی در استفاده از ارزیابی هوش مصنوعی است. البته، یکی از وعده‌های اصلی ارزیابی فعال شده با هوش مصنوعی، ظرفیت تشخیص و پاسخ به تمام اشکال یادگیری رایج در عصر دیجیتال برای دانستن چیزهایی در مورد آنچه در غیر این صورت ناشناخته باقی می‌ماند، می‌باشد. بااین‌حال، این وعده ارزیابی جامع یادگیری در تمام اشکال آن، هر شکلی از ارزیابی را مبهم می‌سازد که آنچه را که با یادگیری در هر سیستم آموزشی درک می‌شود، مشخص و مشخص می‌کند (مِسیک، ۱۹۹۴). در این مفهوم، نگرانی‌ها را می‌توان مطرح کرد که بسیاری از اشکال ارزیابی مبتنی بر هوش مصنوعی، جهت گیری شیوه های ارزیابی سنتی فعلی را به سمت مهارت‌های تاکید، تفکر منطقی و رفتارها، در کنار ارزش‌های غالبا سفید، مرد، طبقه متوسط، ارزش‌های غربی عینیت و فردگرایی تداوم می‌بخشند (هینسوُرث و همکاران، ۲۰۱۹). در موارد دیگر، اهمیت تکنولوژی‌هایی مانند ردیابی چشم، خطرات عمل ارزیابی با قابلیت هوش مصنوعی برای تقویت و مزیت گرایی و به ویژه مدل‌های عصبی معمول یادگیری و این‌که چه معنایی برای نشان دادن رفتارهای مرتبط با یادگیری دارد را برجسته می‌کند (سواگر، ۲۰۲۰). به‌طور کلی، استدلال‌های قوی را می توان ایجاد کرد که ارزیابی فعال شده با هوش مصنوعی ممکن است به خوبی تغییر کند اما لزوما فرم‌های یادگیری که ارزیابی می‌شوند را توسعه ندهد؛ بنابراین، مکالمات در جامعه تحقیقاتی نیاز به کشف این بحث دارند که ارزیابی فعال شده با هوش مصنوعی یک سایت خنثی نیست که در آن هر شکلی از یادگیری شناسایی و ارزیابی شود. به‌عنوان‌مثال، همانند هر شکلی از ارزیابی، می توان استدلال کرد که هر نمونه‌ای از ارزیابی با هوش مصنوعی به ناچار هنجارهای خاص فرهنگی، رشته‌ای و فردی، سیستم‌های ارزشی و سلسله‌مراتب دانش را تدوین خواهد کرد. علاوه بر این، ممکن است این هنجارها، ارزش‌ها و سلسله‌مراتب دانش را در دانشجویان جا به جا کند. دانش آموزان یاد خواهند گرفت که به روشی که به‌صورت الگوریتمی ارزیابی پذیر و به‌صورت الگوریتمی پاداش داده می‌شود، عمل کنند؛ به عبارت دیگر، آموزش تست لزوما با استفاده از ارزیابی فعال شده با هوش مصنوعی اجتناب نمی‌شود (پاپهام، ۲۰۰۱).

در عین حال، نیاز به بحث در مورد ارزیابی فعال شده با هوش مصنوعی برای تصدیق بهتر اشکال مختلف یادگیری وجود دارد که هنوز قابل کشف، اندازه‌گیری و مدل‌سازی توسط غیر انسان‌ها نیستند. نرم‌افزار هوش مصنوعی در کشف معنا در زبان یا تصاویر بسیار محدود است، زیرا توسعه ساده یک استدلال منطقی برای ظرافت و برجستگی مانند کنایه و کنایه است. برای مثال، تکنولوژی پردازش زبان طبیعی ممکن است یک ظرفیت تقریبا نامحدود برای تشخیص واژگان داشته باشد اما برای ظرافت دوبله شدن زبان، تلمیحات، بومی محلی، تن و زیرمتن، کری آهنگ باقی می‌ماند.

به‌طور مشابه، ارزیابی فعال شده با هوش مصنوعی ممکن است به‌طور قابل‌درک در ظرفیت خود برای تشخیص (چه برسد به ارزیابی)نمونه‌های بداهه‌سازی، خلاقیت، شعر، اخلاق، یا اخلاق محدود باقی بماند. ممکن است فضای کمی برای تشخیص (و پاداش دهی)به وضوح متفاوت، غیر منتظره و شاید منحصر به فرد در مورد یک کار یادگیری وجود داشته باشد که در آن دانشجویان درگیر اصالت واقعی می‌شوند و " خارج از چارچوب فکر می‌کنند که یک ارزیاب انسانی خوب قادر به درک آن خواهد بود (حتی اگر آن‌ها هرگز خودشان به آن فکر نکرده باشند). به‌طور خلاصه، جنبه‌هایی از یادگیری وجود دارند که برای انسان‌ها قابل‌درک هستند اما نه ماشین‌ها. به‌این‌ترتیب، بحث در مورد ارزیابی فعال شده با هوش مصنوعی باید بیشتر در شناخت آنچه فن‌آوری نمی‌تواند (و ممکن است هرگز) قادر به ارزیابی باشد، مطرح شود.

## آموزش نظارت

در یک معنا، ارزیابی فعال شده با هوش مصنوعی براساس برخی منطق‌های متمایز از "تبادل اطلاعات در آموزش و پرورش "، مانند ایده تولید داده‌های پیوسته و جامع مربوط به تعامل مداوم فرد با یک محیط یادگیری آنلاین ایجاد می‌شود. این امر نوید ارزیابی مستمر را می‌دهد که لزوما توسط دانشجویان به‌عنوان ارزیابی به رسمیت شناخته نمی‌شود درنتیجه بر مشکلات "اضطراب امتحان " غلبه می‌کند و اجازه می‌دهد تمام جنبه‌های یادگیری یک فرد قابل‌مشاهده باشد (کولول، ۲۰۱۳). بااین‌حال، این وعده‌ها در زمینه نظارت مستمر بر داده‌ها را می توان به‌عنوان شرایط نظارت در نظر گرفت. به‌این‌ترتیب، وعده محیط‌های آموزشی داده‌محور برای قابل‌مشاهده ساختن آنچه در غیر این صورت ممکن است پنهان یا از دست برود (بین و همکاران، ۲۰۲۰).

به‌عنوان‌مثال، باید در بحث‌های ارزیابی مبتنی بر هوش مصنوعی در مورد این‌که چگونه این وضعیت نظارت مستمر خود را به فرایندهای کنترل و انطباق برای مثال، نظارت برای نشانه‌های سو استفاده و دیگر اشکال تقلب نیز قرض می‌دهد، تصدیق بیشتری وجود داشته باشد. البته، بسیاری از جنبه‌های موسسات آموزش رسمی مانند مدارس و دانشگاه‌ها به‌طور سنتی براساس "آموزش نظارت " نه حداقل مجموعه سنتی کلاس یا

سالن امتحان با صندلی‌های مرتب شده در ردیف، رو به جلوی کلاس و نظارت معلم بر بدنه دانش آموزان است (لوک، ۲۰۰۳؛ مک‌لارن و لئوناردو، ۱۹۹۸). بااین‌حال، آموزش آنلاین (و به‌طور ضمنی، ارزیابی با قابلیت هوش مصنوعی) دامنه این نظارت را تا تمام زمان‌ها و تمام فضاهای روز مدرسه یا تجربه دانشگاه گسترش و تقویت می‌کند.

در این مفهوم، می توان استدلال کرد که ارزیابی فعال شده با هوش مصنوعی به‌جای نگاه آموزشی، یک نگاه اجرایی را تشکیل می‌دهد که شرایط شکننده اعتماد را که بیشتر مربیان به‌عنوان زیربنای رابطه معلم / دانش‌آموز می‌بینند، تحت‌تاثیر قرار می‌دهد:

در محیط‌های آموزش عالی، فرهنگ نظارت، تسهیل و تشدید شده توسط تکنولوژی، خطرات ایجاد شرایطی را به وجود می‌آورد که بسیار ریسک‌گریز هستند و موجب تخریب اعتماد می‌شوند. ساختارهای فن‌آوری که برای ایجاد اعتماد از طریق نگاشت عملکرد معرفی شده‌اند، ممکن است به‌طور مستقیم این اهداف را تضعیف کنند (بین و همکاران، ۲۰۲۰).

همچنین مهم است که پیامدهای پیوسته نظارت بر دانشجویان از نظر خطوط آموزشی را بهتر در نظر بگیریم. به‌طور خاص، الگوی ارزیابی استاندارد همچنین انتقال می‌دهد که یادگیری می‌تواند به بهترین نحو در جایی که هیچ ارزیابی وجود ندارد رخ دهد. این در تضاد با مزایای نظارت و ارزیابی مستمر و جامع تعامل تحصیلی دانش آموزان است. به‌این‌ترتیب، درحالی‌که به هیچ وجه کامل نیست، اشکال فعلی آموزش به روش‌هایی ایجاد می‌شوند که از یادگیری و پیشرفت در طول اپیزودها که هیچ ارزیابی وجود ندارد، حمایت می‌کنند. تدریس خوب طراحی‌شده لحظات زیادی از تمرین را برای تشخیص آسیب‌پذیری یادگیری ارائه می‌دهد و به دانش آموزان فرصت کافی برای یادگیری خصوصی، شرکت در کارهای مقدماتی، آزمایش، اشتباه کردن و شکست را می‌دهد؛ به عبارت دیگر، عدم ارزیابی به‌عنوان بهترین شرایط برای یادگیری و پیشرفت دیده می‌شود.

بااین‌حال، اهمیت عدم وجود ارزیابی نیز متناقض به نظر می‌رسد. چگونه ما می‌دانیم که آیا برای یادگیری خوب است اگر ما نتوانیم بگوییم که آیا یادگیری رخ می‌دهد؟ شاید یک راه در مورد این مساله، ادامه تغییر تمرکز ارزیابی از ارزیابی یا قضاوت به توسعه باشد. در این دیدگاه، نظارت مستمر بر فرآیندهای دانشجویی وسیله‌ای برای تعیین این نیست که آیا کسی کار درست یا غلط انجام می‌دهد، بلکه در عوض، بر فرصت‌هایی برای ارائه بازخورد و بهبود یادگیری نظارت می‌کند (ویلیام، ۲۰۱۱)؛ بنابراین، آنچه موردنیاز است، لزوما یک تغییر در چگونگی انجام ارزیابی‌ها نیست، بلکه یک تغییر مفهومی در معنای آن‌ها و آنچه برای آن هستند، می‌باشد.

## مدل‌های ارزیابی توزیع‌شده

نگرانی نهایی مربوط به تغییراتی است که ابزارهای محاسباتی برای ارزیابی به کار می‌برند. یکی از راه‌های توصیف ارزیابی استدلال از شواهد است (مِسیک، ۱۹۹۴). برای مثال، در طراحی ارزیابی مبتنی بر شواهد، این بحث شامل یک مدل دانشجویی است که ویژگی‌ها، مهارت‌ها، یا توانایی‌های مورد ارزیابی را توصیف می‌کند؛ یک مدل کاری که فعالیت‌های دانش آموزان را توصیف می‌کند، برای نشان دادن این‌که آن‌ها این ویژگی‌ها را دارند، انجام خواهد داد؛ و یک مدل شاهد که به توصیف متغیرها و تکنیک‌هایی می‌پردازد که برای ارتباط شواهد با صفات مورداستفاده قرار می‌گیرند. یکی از نتایج ابزارهای محاسباتی مبتنی بر هوش مصنوعی این است که آن‌ها هر یک از این مدل‌ها را پیچیده می‌کنند.

از نظر مدل دانشجویی، حضور هوش مصنوعی نشان می‌دهد که ما باید ویژگی‌ها، مهارت‌ها و توانایی‌های ارزیابی‌شده را طوری تنظیم کنیم که به‌جای آن که هوش مصنوعی بتواند به تنهایی انجام دهد، به نفوذ انسان نیاز داشته باشد. از نظر مدل کار، هوش مصنوعی پیشنهاد می‌کند که ما باید به دانش آموزان اجازه دهیم تا از ابزارهای محاسباتی مبتنی بر هوش مصنوعی در طول ارزیابی استفاده کنند و از نظر مدل شواهد، حضور هوش مصنوعی نشان می‌دهد که ما باید این حقیقت را در نظر بگیریم که یک تیم هوش مصنوعی انسانی می‌تواند شواهد ارزیابی را تولید کند.

بسته به پیچیدگی هوش مصنوعی، این می‌تواند به معنی تلاش برای جدا کردن سهم انسان و هوش مصنوعی، محاسبه رابطه بین این سهم، یا رفتار با آن‌ها به گونه‌ای که گویی از یک منبع هستند، باشد .درحالی‌که تلاش‌هایی برای ادغام نظریه طراحی ارزیابی با هوش مصنوعی صورت‌گرفته است (میسلوی و همکاران، ۲۰۱۲) تا به امروز، آن‌ها عمدتا بر کاربردهای هوش مصنوعی برای مدل شواهد و کم‌تر بر مدل‌های وظیفه و دانش‌آموز تمرکز کرده‌اند.

## نتیجه‌گیری

ما استدلال کرده‌ایم که چندین مساله به الگوی ارزیابی استاندارد لطمه می‌زنند. اول، ارزیابی‌ها در این پارادایم می‌تواند برای مربیان برای طراحی و پیاده‌سازی دشوار باشد. دوم، آن‌ها ممکن است به‌جای دیدگاه‌های متفاوت یادگیری، تنها تصویری مجزا از عملکرد ارائه دهند. سوم، آن‌ها ممکن است یکنواخت باشند و درنتیجه با مهارت‌ها و زمینه‌های دانش خاص شرکت‌کنندگان سازگار نباشند. چهارم، آن‌ها ممکن است نا معتبر باشند، به‌جای اینکه به مدرسه رفتن در فرهنگ‌ها پایبند باشند، به فرهنگ مدرسه وفادار بمانند تا اینکه دانش آموزان را برای عضویت در مدرسه آماده کنند و درنهایت، آن‌ها ممکن است قدیمی باشند و مهارت‌هایی را ارزیابی کنند که انسان‌ها به‌طورمعمول از ماشین‌ها برای انجام آن‌ها استفاده می‌کنند. درحالی‌که رویکردهای هوش مصنوعی موجود تا حدی به مسائل بالا می‌پردازند، آن‌ها یک درمان کامل نیستند. همان‌طور که بحث ما برجسته می‌کند، این رویکردها مجموعه جدیدی از چالش‌ها را باخود به همراه می‌آورند که در طراحی و اجرای ارزیابی‌ها باید مد نظر قرار داد. امیدواریم که این مقاله هم مسائل مربوط به الگوی ارزیابی استاندارد و هم چالش‌های مربوط به هوش مصنوعی و ارزیابی را وارد بحث عمیق تری کند که درنهایت شیوه‌های ارزیابی را به‌طور عمومی‌تر بهبود خواهد بخشید.

## منابع


- Abdi, S., Khosravi, H., Sadiq, S., & Demartini, G. (2021). Evaluating the quality of learning resources: A learner sourcing approach. IEEE Transactions on Learning Technologies, 14(1), 81–92.
- Abdi, S., Khosravi, H., Sadiq, S., & Gasevic, D. (2019). A multivariate ELO-based learner model for adaptive educational systems. In Proceedings of the 12th international conference on educational data mining (pp. 462–467).
- Ahmad Uzir, N., Gaˇsevi´c, D., Matcha, W., Jovanovi´c, J., & Pardo, A. (2020). Analytics of time management strategies in a flipped classroom. Journal of Computer Assisted Learning, 36(1), 70–88.
- Almond, R. G., Steinber, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. The Journal of Technology, Learning, and Assessment, 5.
- Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. Journal of Experimental Psychology: General, 136, 569–576. https://doi.org/10.1037/0096-3445.136.4.569
- Azevedo, R., & Gaˇsevi´C, D. (2019). Analyzing multimodal multichannel data about selfregulated learning with advanced learning technologies: Issues and challenges. Computers in Human Behavior, 96, 207–210.
- Baker, R. S. (2016). Stupid tutoring systems, intelligent humans. International Journal of Artificial Intelligence in Education, 26(2), 600–614.
- Baker, R. S., Goldstein, A. B., & Heffernan, N. T. (2011). Detecting learning moment-bymoment. International Journal of Artificial Intelligence in Education, 21(1–2), 5–25.
- Baker, R. S., Hershkovitz, A., Rossi, L. M., Goldstein, A. B., & Gowda, S. M. (2013). Predicting robust learning with the visual form of the moment-by-moment learning curve. The Journal of the Learning Sciences, 22(4), 639–666.
- Bayne, S., Evans, P., Ewins, R., Knox, J., Lamb, J., Macleod, H., et al. (2020). The manifesto for teaching online. MIT Press.
- Bergin, T. J. (2006). The origins of word processing software for personal computers: 1976-1985. IEEE Annals of the History of Computing, 28(4), 32–47.
- Bezirhan, U., von Davier, M., & Grabovsky, I. (2021). Modeling item revisit behavior: The hierarchical speed–accuracy–revisits model. Educational and Psychological Measurement, 81(2), 363–387.
- Boud, D., Ajjawi, R., Dawson, P., & Tai, J. (2018). Developing evaluative judgement in higher education: Assessment for knowing and producing quality work. Abingdon, UK: Routledge.
- Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. Educational Researcher, 18(1), 32–42.
- Carless, D. (2022). From teacher transmission of information to student feedback literacy: Activating the learner role in feedback processes. Active Learning in Higher Education. https://doi.org/10.1177/1469787420945845 (in press).
- Cauley, K. M., & McMillan, J. H. (2010). Formative assessment techniques to support student motivation and achievement. The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 83(1), 1–6.
- Chen, G., Yang, J., & Gasevic, D. (2019). A comparative study on question-worthy sentence selection strategies for educational question generation. In Proceedings of the 20th international conference on artificial intelligence in education (pp. 59–70). Cham: Springer.
- Chen, G., Yang, J., Hauff, C., & Houben, G. J. (2018). LearningQ: A large-scale dataset for educational question generation. In Proceedings of the 12th international AAAI conference on web and social media (pp. 481–490). AAAI.
- Chen, H., & He, B. (2013). Automated essay scoring by maximizing human-machine agreement. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1741–1752).
- Cho, Y. H., & Cho, K. (2011). Peer reviewers learn from giving comments. Instructional Science, 39(5), 629–643.
- Collares, C. F., Cecilio-Fernandes, D. (2019). When I say computerised adaptive testing. Medical Education, 53(2), 115–116.
- Colwell, N. M. (2013). Test anxiety, computer-adaptive testing and the common core. Journal of Education and Training Studies, 1(2), 50–60.
- Cope, B., Kalantzis, M., & Searsmith, D. (2021). Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies. Educational Philosophy and Theory. https://doi.org/10.1080/00131857.2020.1728732

- Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 4(4), 253–278.
- Couldry, N. (2020). Recovering critique in an age of datafication. New Media & Society, 22(7), 1135–1151.
- Crossley, S., Allen, L. K., Snow, E. L., & McNamara, D. S. (2015). Pssst... textual features... there is more to automatic essay scoring than just you. In Proceedings of the fifth international conference on learning analytics and knowledge (pp. 203–207).
- Darvishi, A., Khosravi, H., & Sadiq, S. (2020). Utilising learner sourcing to inform design loop adaptivity. In Proceedings of the 14th European conference on technology-enhanced learning (pp. 332–346). Springer.
- Darvishi, A., Khosravi, H., & Sadiq, S. (2021). Employing peer review to evaluate the quality of student generated content at scale: A trust propagation approach. In Proceedings of the eighth ACM conference on learning@ scale (pp. 139–150).
- De Alfaro, L., & Shavlovsky, M. (2014). Crowdgrader: A tool for crowdsourcing the evaluation of homework assignments. In Proceedings of the 45th ACM technical symposium on computer science education (pp. 415–420).
- Denkowski, M., & Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the ninth workshop on statistical machine translation (pp. 376–380).
- Dennick, R., Wilkinson, S., & Purcell, N. (2009). Online eAssessment: AMEE guide no. 39. Medical Teacher, 31(3), 192–206.
- Desmarais, M. C., & Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. User Modeling and User-Adapted Interaction, 22(1), 9–38.
- Echeverria, V., Martinez-Maldonado, R., & Buckingham Shum, S. (2019). Towards collaboration translucence: Giving meaning to multimodal group data. In Proceedings of the 2019 CHI conference on human factors in computing systems (pp. 1–16).
- Educational Testing Service. (2022, March 1). What to expect during the GRE general test". https://www.ets.org/gre/revised_general/test_day/expect.
- Embretson, S. E., & Reise, S. P. (2013). Item response theory. Psychology Press.
- Engelhardt, L., & Goldhammer, F. (2019). Validating test score interpretations using time information. Frontiers in Psychology, 10, 1131.
- Er, E., Dimitriadis, Y., & Ga˘sevi´c, D. (2020). A collaborative learning approach to dialogic peer feedback: A theoretical framework. Assessment & Evaluation in Higher Education, 46(4), 586–600.
- Fan, Y., Saint, J., Singh, S., Jovanovic, J., & Ga˘sevi´c, D. (2021). April). A learning analytic approach to unveiling self-regulatory processes in learning tactics. In LAK21: 11th international learning analytics and knowledge conference (pp. 184–195).
- Foltýnek, T., Meuschke, N., & Gipp, B. (2019). Academic plagiarism detection: A systematic literature review. ACM Computing Surveys, 52(6), 1–42.
- Gervet, T., Koedinger, K., Schneider, J., & Mitchell, T. (2020). When is deep learning the best approach to knowledge tracing? Journal of Educational Data Mining, 12(3), 31–54.
- Gipps, C., & Stobart, G. (2009). Fairness in assessment. In Educational assessment in the 21st century (pp. 105–118). Dordrecht: Springer.
- Glassman, E. L., Lin, A., Cai, C. J., & Miller, R. C. (2016). Learner sourcing personalized hints. In Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing (pp. 1626–1636).
- Goldhammer, F., Naumann, J., Stelter, A., T˙oth, K., R¨olke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. Journal of Educational Psychology, 106(3), 608–626. https://doi.org/10.1037/a0034716
- Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. The Elementary School Journal, 115(4), 523–547.
- Greiff, S., Stadler, M., Sonnleitner, P., Wolff, C., & Martin, R. (2015). Sometimes less is more: Comparing the validity of complex problem solving measures. Intelligence, 50,100–113.
- Griffin, P., & Care, E. (2015). Assessment and teaching of 21st century skills: Methods and approaches (Vol. 2). Dordrecht: Springer.
- Griffiths, S. (2021). Families to sue over 'wrong' marks given by teachers. The Times. Retrieved from https://www.thetimes.co.uk/article/families-to-sue-over-wrongmarks-given-by-teachers-g2qjjc8x7.
- Hanesworth, P., Bracken, S., & Elkington, S. (2019). A typology for a social justice approach to assessment. Teaching in Higher Education, 24(1), 98–114.
- Harlen, W. (2012). The role of assessment in developing motivation for learning. InJ. Gardner (Ed.), Assessment and learning (2nd ed., pp. 61–80). SAGE Publications.
- Heckler, N. C., Rice, M., & Hobson Bryan, C. (2013). Turnitin systems: A deterrent to plagiarism in college classrooms. Journal of Research on Technology in Education, 45(3), 229–248.
- Herder, T., Swiecki, Z., Fougt, S. S., Tamborg, A. L., Allsopp, B. B., Shaffer, D. W., et al.(2018). Supporting teachers' intervention in students' virtual collaboration using a network based model. In Proceedings of the 8th international conference on learning analytics and knowledge (pp. 21–25).
- Horbach, A., Aldabe, I., Bexte, M., de Lacalle, O. L., & Maritxalar, M. (2020). Linguistic appropriateness and pedagogic usefulness of reading comprehension questions. In Proceedings of the 12th language resources and evaluation conference (pp. 1753–1762).
- Hwang, G. J., Xie, H., Wah, B. W., & Ga˘sevi´c, D. (2020). Vision, challenges, roles and research issues of Artificial Intelligence in Education. Computers & Education: Artificial Intelligence, 1, Article 100001.
- J¨arvel¨a, S., Malmberg, J., Haataja, E., Sobocinski, M., & Kirschner, P. A. (2020). What multimodal data can tell us about the students' regulation of their learning process. Learning and Instruction, 45, Article 100727.
- Jia, X., Zhou, W., Sun, X., & Wu, Y. (2020). EQG-RACE: Examination-Type question generation. arXiv preprint arXiv:2012.06106.

- Jovanovi´c, J., Ga˘sevi´c, D., Pardo, A., Dawson, S., & Whitelock-Wainwright, A. (2019). Introducing meaning to clicks: Towards traced-measures of self-efficacy and cognitive load. In Proceedings of the 9th international conference on learning analytics & knowledge (pp. 511–520). New York: ACM.
- Kaipa, R. M. (2021). Multiple choice questions and essay questions in curriculum. Journal of Applied Research in Higher Education, 13(1), 16–32. https://doi.org/10.1108/JARHE-01-2020-0011
- Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In Proceedings of the 28th international joint conference on artificial intelligence (pp.6300–6308).
- Khosravi, H., Conati, C., Martinez-Maldonado, R., Knight, S., Kay, J., Chen, G., et al. (2022). Explainable AI in education. Computers & Education: Artificial Intelligence. In this issue.
- Khosravi, H., Kitto, K., & Williams, J. J. (2019). Ripple: A crowdsourced adaptive platform for recommendation of learning activities. arXiv preprint arXiv:1910.05522.
- Klebanov, B. B., Madnani, N., & Burstein, J. (2013). Using pivot-based paraphrasing and sentiment profiles to improve a subjectivity lexicon for essay data. Transactions of the Association for Computational Linguistics, 1, 99–110.
- Knight, S., Shibani, A., Abel, S., Gibson, A., Ryan, P., Sutton, N., et al. (2020). Acawriter: A learning analytics tool for formative feedback on academic writing. Journal of Writing Research, 12(1), 141–186.
- Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). Race: Large-scale reading comprehension dataset from examinations. arXiv preprint arXiv:1704.04683.
- Lave, J., & Wenger, E. (1991). Situated learning: Legitimate peripheral participation. Cambridge University Press. van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). Elements of adaptive testing. New York, NY: Springer.
- Llamas-Nistal, M., Fern´andez-Iglesias, M. J., Gonz´alez-Tato, J., & Mikic-Fonte, F. A. (2013). Blended e-assessment: Migrating classical exams to the digital world. Computers & Education, 62, 72–87.
- Lodge, J. M. (2018). A futures perspective on information technology and assessment. In J. Voogt, G. Knezek, R. Christensen, & K.-W. Lai (Eds.), International handbook of information technology in primary and secondary education (2nd ed., pp. 1–13). Berlin: Springer.
- Luke, C. (2003). Pedagogy, connectivity, multimodality, and interdisciplinarity. Reading Research Quarterly, 38(3), 397–403.
- Marche, S. (2021). The computers are getting better at writing. https://www.newyorker.com/culture/cultural-comment/the-computers-are-getting-better-at-writing.
- Mayfield, E., Madaio, M., Prabhumoye, S., Gerritsen, D., McLaughlin, B., Dixon-Rom´an, E., et al. (2019, August). Equity beyond bias in language technologies for education. In Proceedings of the 14th workshop on innovative use of NLP for building educational applications (pp. 444–460).
- McArthur, J. (2016). Assessment for social justice. Assessment & Evaluation in Higher Education, 41(7), 967–981.
- McLaren, P., & Leonardo, Z. (1998). Deconstructing surveillance pedagogy. Studies in the Literary Imagination, 31(1), 127–147.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23(2), 13–23.
- Microsoft. (2022, March 1). Microsoft Editor checks grammar and more in documents, mail, and the web. https://support.microsoft.com/en-us/office/microsoft-editor-checks-grammar-and-more-in-documents-mail-and-the-web-91ecbe1b-d021-4e9e-a82e-abc4cd7163d7.
- Milligan, S. K., & Griffin, P. (2016). Understanding learning and learning design in MOOCs: A measurement-based interpretation. Journal of Learning Analytics, 3(2),88–115.
- Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. Journal of educational data mining, 4(1), 11–48.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. Measurement: Interdisciplinary Research and Perspectives, 1, 3–62.
- Molenaar, I. (2022). The concept of hybrid human-AI regulation: Exemplifying how to support young learners' self-regulated learning. Computers & Education: Artificial Intelligence. In this issue.
- Molenaar, I., Horvers, A., & Baker, R. S. (2021). What can moment-by-moment learning curves tell about students' self-regulated learning? Learning and Instruction, 72, Article 101206.
- Murphy, V., Fox, J., Freeman, S., & Hughes, N. (2017). Keeping it real": A review of the benefits, challenges and steps towards implementing authentic assessment. All Ireland Journal of Higher Education, 9(3), 1–13.
- Page, E. B. (1966). The imminence of... grading essays by computer. Phi Delta Kappan, 47(5), 238–243.
- Pagni, S. E., Bak, A. G., Eisen, S. E., Murphy, J. L., Finkelman, M. D., & Kugel, G. (2017). The benefit of a switch: Answer-changing on multiple-choice exams by first-year dental students. Journal of Dental Education, 81(1), 110–115.
- Palermo, C., & Thomson, M. M. (2018). Teacher implementation of self-regulated strategy development with an automated writing evaluation system: Effects on the argumentative writing performance of middle school students. Contemporary Educational Psychology, 54, 255–270.
- Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. Frontiers in Psychology, 8, 883–928. https://doi.org/10.3389/fpsyg.2017.00422
- Papamitsiou, Z., & Economides, A. A. (2017). Student modeling in real-time during selfassessment using stream mining techniques. In Proceedings of the 17th IEEE international conferenceon advanced learning technologies (pp. 286–290). IEEE.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the association for computational linguistics (pp. 311–318).

- Pearl, J. (1988). Probabilistic reasoning in intelligent systems: Networks of plausible inference. San Mateo, CA: Kaufmann.
- Perret-Clermont, A. N. (1980). Social interaction and cognitive development in children. Academic Press. Piech, C., Spencer, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., et al. (2015). Deep knowledge tracing. arXiv preprint arXiv:1506.05908.
- Popham, W. J. (2001). Teaching to the test? Educational Leadership, 58(6), 16–21.
- Puntambekar, S., Erkens, G., & Hmelo-Silver, C. (Eds.). (2011). Analyzing interactions in CSCL. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4419-7710-6.
- Purchase, H., & Hamer, J. (2018). Peer-review in practice: Eight years of Arop¨a. Assessment & Evaluation in Higher Education, 43(7), 1146–1165.
- Reeves, T. C., & Okey, J. R. (1996). Alternative assessment for constructivist learning environments. In B. G. Wilson (Ed.), Constructivist learning environments: Case studies in instructional design (pp. 191–202). Englewood Cliffs, NJ: Educational Technology Publications.
- Rogers, T., Ga˘sevi´c, D., & Dawson, S. (2016). Learning analytics and the imperative for theory driven research. The SAGE Handbook of E-Learning Research.
- Ros´e, C. P., McLaughlin, E. A., Liu, R., & Koedinger, K. R. (2019). Explanatory learner models: Why machine learning (alone) is not the answer. British Journal of Educational Technology, 50(6), 2943–2958.
- Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. The Journal of Technology, Learning, and Assessment, 1(2).
- Saint, J., Ga˘sevi´c, D., Matcha, W., Uzir, N. A. A., & Pardo, A. (2020). Combining analytic methods to unlock sequential and temporal patterns of self-regulated learning. In Proceedings of the tenth international conference on learning analytics & knowledge (pp.402–411). New York: ACM.
- Saltman, K. (2020). Artificial intelligence and the technological turn of public education privatization. London Review of Education, 18(2), 196–208.
- Scheuneman, J. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 143–152.
- Shaffer, D. W. (2006a). Epistemic frames for epistemic games. Computers in Education, 46 (3), 223–234.
- Shaffer, D. W. (2006b). How computer games help children learn. New York, NY: Palgrave. Shaffer, D. W., Collier, W., & Ruis, A. R. (2016). A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. Journal of Learning Analytics, 3(3), 9–45.
- Shaffer, D. W., & Kaput, J. J. (1998). Mathematics and virtual culture: An evolutionary perspective on technology and mathematics education. Educational Studies in Mathematics, 37, 97–119.
- Shin, D., Shim, Y., Yu, H., Lee, S., Kim, B., & Choi, Y. (2021). Saint+: Integrating temporal features for ednet correctness prediction. In Proceedings of the 11th international learning analytics and knowledge conference (pp. 490–496).
- Shnayder, V., & Parkes, D. C. (2016). Practical peer prediction for peer assessment. In Proceedings of the fourth AAAI conference on human computation and crowdsourcing (pp. 199–208).
- Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. Computer games and instruction, 55(2), 503–524.
- Shute, V. J., & Rahimi, S. (2021). Stealth assessment of creativity in a physics video game. Computers in Human Behavior, 116, Article 106647.
- Shute, V., & Ventura, M. (2013). Stealth assessment: Measuring and supporting learning in video games. Cambridge, MA: MIT Press.
- Shute, V., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C. P., … Sun, C. (2021). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity and learning supports in educational games. Journal of Computer Assisted Learning, 37(1),127–141.
- Sorrel, M. A., Barrada, J. R., de la Torre, J., & Abad, F. J. (2020). Adapting cognitive diagnosis computerized adaptive testing item selection rules to traditional item response theory. PLoS One, 15(1), Article e0227196.
- Sullivan, S. A., Warner-Hillard, C., Eagan, B., Thompson, R., Ruis, A. R., Haines, K., et al. (2018). Using epistemic network analysis to identify targets for educational interventions in trauma team communication. Surgery, 163(4), 938–943.
- Suto, I., N´adas, R., & Bell, J. (2011). Who should mark what? A study of factors affecting marking accuracy in a biology examination. Research Papers in Education, 26(1),21–51.
- Swauger, S. (2020). Our bodies encoded: Algorithmic test proctoring in higher education. In J. Stommel, C. Friend, & S. Morris (Eds.), Critical digital pedagogy. Press Books.
- Taras, M. (2008). Assessment for learning. Journal of Further and Higher Education, 32(4), 389–397.
- Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. Practical Assessment, Research and Evaluation, 12(1), 1.
- Topping, K. J. (2018). Using peer assessment to inspire reflection and learning. Routledge. Toton, S. L., & Maynes, D. D. (2019). Detecting examinees with pre-knowledge inExperimental data using conditional scaling of response times. Frontiers in Education, 4, 49.
- Van Der Graaf, J., Lim, L., Fan, Y., Kilgour, J., Moore, J., Bannert, M., … Molenaar, I. (2021). April). Do instrumentation tools capture self-regulated learning?. In LAK21: 11th international learning analytics and knowledge conference (pp. 438–448).
- Verschoor, A., Berger, S., Moser, U., & Kleintjes, F. (2019). On-the-Fly calibration in computerized adaptive testing. In Theoretical and practical advances in computer-based educational measurement (pp. 307–323). Cham: Springer.
- Vygotsky, L. S., & Cole, M. (1978). Mind in society: Development of higher psychological processes. Harvard University Press.
- Wang, W., An, B., & Jiang, Y. (2018). Optimal spot-checking for improving evaluation accuracy of peer grading systems. In Proceedings of the 32nd AAAI conference on artificial intelligence (pp. 833–840).

- Whitehill, J., Aguerrebere, C., & Hylak, B. (2019). Do learners know what's good for them? Crowdsourcing subjective ratings of oers to predict learning gains. In Proceedings of the 12th international conference on educational data mining (pp.462–467). IEDMS.
- Wiliam, D. (2011). What is assessment for learning? Studies In Educational Evaluation, 37 (1), 3–14.
- Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. Computers & Education, 100, 94–109.
- Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. Journal of Educational Computing Research, 58(1), 87–125.
- Wilson, J., Ahrendt, C., Fudge, E. A., Raiche, A., Beard, G., & MacArthur, C. (2021). Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. Computers & Education, 168, Article 104208.
- Wilson, M. (2005). Constructing measures: An item response modeling approach. New York: Taylor & Francis Group.
- Wilson, M., & Scalise, K. (2012). Assessment of learning in digital networks. In P. Griffin, & E. Care (Eds.), Assessment and teaching of 21st century skills: Methods and approach (pp. 37–56). Dordrecht: Springer.
- Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. Applied Measurement in Education, 30(4), 343–354.
- Wright, J. R., Thornton, C., & Leyton-Brown, K. (2015). Mechanical TA: Partially automated high-stakes peer grading. In Proceedings of the 46thACM technical symposium on computer science education (pp. 96–101).
- Yannakoudakis, H., & Briscoe, T. (2012). Modeling coherence in ESOL learner texts. In Proceedings of the seventh workshop on building educational applications using NLP (pp.33–43).
- Zheng, Y., Li, G., Li, Y., Shan, C., & Cheng, R. (2017). Truth inference in crowdsourcing: Is the problem solved? Proceedings of the VLDB Endowment, 10(5), 541–552.
- Zhou, M., & Winne, P. H. (2012). Modeling academic achievement by self-reported versus traced goal orientation. Learning and Instruction, 22(6), 413–419.