گزارش مقدماتی مستقل و بین المللی پنل علمی هوش مصنوعی

ارزیابی مبتنی بر شواهد از فرصت‌ها، خطرات و تأثیرات

هوش مصنوعی

سازمان ملل متحد

جولای ۲۰۲۶

مرداد ماه ۱۴۰۵

اعضای پنل

Girmaw Abebe Tadesse
Ethiopia

Tuka Alhanai
United Arab Emirates

Joëlle Barral
France

Yoshua Bengio
Canada
Co-Chair

Tegawendé Bissyandé
Burkina Faso

Awa Bousso Dramé
Cabo Verde

Mennatallah El-Assady
Egypt

Hoda Heidari
Islamic Republic of Iran

Juho Kim
Republic of Korea

Anna Korhonen
Finland

Vukosi Marivate
South Africa

Bilal Mateen
Pakistan

Yutaka Matsuo
Japan

Joyce Nakatumba Nabende
Uganda

Andrei Neznamov
Russian Federation

Johanna Pirker
Austria

Balaraman Ravindran
India

Maria Ressa
Philippines
Co-Chair

Lior Rokach
Israel

Piotr Sankowski
Poland

Loreto Bravo
Chile

Mark Coeckelbergh
Belgium

Carlos Coello Coello
Mexico

Melahat Bilge Demirköz
Türkiye

Adji Bousso Dieng
Senegal

Aleksandra Korolova
Latvia

Vipin Kumar
United States

Sonia Livingstone
United Kingdom

Qinghua Lu
Australia

Teresa Ludermir
Brazil

Maximilian Nickel
Germany

Rita Orji
Nigeria

Román Orús
Spain

Alvitta Ottley
Saint Kitts and Nevis

Martha Palmer
United States

Silvio Savarese
Italy

Bernhard Schölkopf
Germany

Haitao Song
China

Leslie Teo
Singapore

Jian Wang
China

خلاصه سند

این سند در پی ارائه تحلیلی متوازن از خطرات و فرصت‌های هوش مصنوعی (AI) می باشد. «متعادل» به معنای تعهد به ارزیابی داده‌های تجربی، بدون سو گیری بی‌مورد به سمت خوش‌بینی یا بدبینی است. همگان قبول دارند که مزایای بالقوه هوش مصنوعی بسیار زیاد است. اگر هوش مصنوعی از روی تفکر به کار گرفته شود، می‌تواند از پیشرفت در جهت دستیابی به اهداف توسعه پایدار، پیشرفت در علوم بهداشتی و نیز افزایش دسترسی به آموزش برای همگان پشتیبانی کند. اما از سوی دیگر، سرعت بالای توسعه فناوری و وسعت کاربردهای بالقوه، سیاست‌گذاران را با چالش‌های قابل توجهی روبرو کرده است. استقرار سریع و بدون کنترل این فناوری در مقیاس وسیع ریسک ها و خطرات قابل توجهی نیز به همراه دارد، از جمله آسیب به سلامت روان کاربران، امکان استفاده به عنوان ابزاری مخرب، تأثیرات بر سیستم‌های اجتماعی، اقتصادی و زیست‌محیطی و چالش‌های مرتبط با کنترل فناوری. هدف از این گزارش بررسی دامنه کامل همه فرصت‌ها و خطرات ممکن نیست، بلکه گزارش تلاش دارد تا بر برخی از مهم‌ترین آنها تمرکز نماید.

قابلیت ها و پذیرش

در سال‌های اخیر شاهد پیشرفت های سریع و در برخی زمینه‌ها شتابان در طیف وسیعی از قابلیت‌های هوش مصنوعی بوده‌ایم. سرمایه‌گذاری‌های قابل توجه در قدرت محاسباتی، روش‌های جدید و داده‌های آموزشی تخصصی، منجر به بهبودهای قابل توجه و پایدار در طیف وسیعی از قابلیت‌های هوش مصنوعی شده است. این پیشرفت‌ها شامل روان ترشدن مکالمه، تولید کد های کامپیوتری کاربردی، استدلال در سطح تخصصی در ریاضیات و علوم، تجزیه و تحلیل داده‌ها در مقیاس بزرگ و تولید محتوای تصویری، صوتی و تصویری هستند. محدودیت‌هایی مانند قابلیت اطمینان هوش مصنوعی، دستیابی به عملکرد قوی در زبان‌ها و فرهنگ‌های انسانی، تعامل با سیستم‌های فیزیکی، اجرای پروژه‌های پیچیده یا چند مرحله‌ای و تولید خروجی‌های واقعی همچنان پابرجا هستند. با این حال، به طور کلی، می توان اذعان نمود که پیشرفت های فنی در بسیاری از حوزه‌های مهم، چندین سال فراتر از پیش بینی ها و انتظارات متخصصان بوده است.

این دستاوردها، کاربردهای مفیدی را در حوزه‌های علوم، بهداشت، کشاورزی،دسترسی پذیری، مشاغل دانشی، و فناوری اطلاعات، از جمله در توسعه خود هوش مصنوعی، ایجاد کرده‌اند. به عنوان مثال، ، هوش مصنوعی آلفافولد ساختار بیش از ۲۰۰ میلیون پروتئین را پیش‌بینی کرده است که اکنون توسط بیش از ۳ میلیون محقق در سراسر جهان استفاده می‌شود. این امر طراحی داروهای جدید، توسعه واکسن های جدید و تحقیقات در خصوص مقاومت آنتی‌بیوتیکی را سرعت بخشیده است. رادیولوژیست‌ها همچنین از هوش

مصنوعی برای تشخیص زودهنگام سرطان سینه استفاده کرده‌اند، در حالی که کارکنان خط مقدم سلامت در مناطق محروم جهان از ابزارهای هوش مصنوعی متناسب با زبان‌های محلی برای ارائه خدمات درمانی با کیفیت بهتر استفاده می‌کنند.

پذیرش هوش مصنوعی به طور گسترده ولی غیر برابر در بین کشورها و جوامع شتاب گرفته است. در سطح جهان، بیش از یک میلیارد نفر اکنون به صورت هفتگی از هوش مصنوعی محاوره‌ای استفاده می‌کنند. با این حال، دسترسی و استفاده از هوش مصنوعی در سطح جهانی بسیار متفاوت است، به طوری که پذیرش آن در کشورهای جنوب بسیار کندتر از کشورهای شمال می باشد. علاوه بر این، تفاوت‌های قابل توجهی در زیرساخت‌ها و مدل‌های محاسباتی بین اقتصادهای پیشرفته وجود دارد. این اختلاف، انعکاس دهنده نابرابری‌های موجود است و حتی ممکن است باعث تقویت آنها نیز شود.

ایجاد و توسعه ابزارهای هوش مصنوعی در جهان حتی متمرکزتر است: طبق برآوردهای اخیر، ایالات متحده آمریکا 75٪ از قدرت محاسباتی را در بین 500 ابررایانه برتر هوش مصنوعی جهان به خود اختصاص داده است، در حالی که سهم چین 15٪ می باشد.

شرکت‌هایی در ایالات متحده و چین نیز تقریباً همه مدل‌های عمومی پیشرو را توسعه می‌دهند و تعداد کمی از کشورها ورودی‌های حیاتی برای زنجیره تأمین تراشه‌های رایانه‌ای هوش مصنوعی را کنترل می‌کنند.

در حالی که در حال گذار به سمت عامل‌های هوش مصنوعی هستیم، احتمالاً با پیشرفت‌های مداوم در توانایی آنها برای انجام کارهای دانشی با نظارت کم یا بدون نظارت انسانی، شاهد تأثیرات اقتصادی شگرفی درآینده خواهیم بود. یک عامل هوش مصنوعی یک سیستم کامپیوتری است که می‌تواند با استفاده از ابزارهایی که در اختیار دارد، برای دستیابی به اهداف، برنامه‌ریزی شده و به طور مستقل عمل کند. این سیستم‌ها در سال‌های اخیر به سرعت در حال بهبود بوده‌اند، به طوری که یک مطالعه نشان داده است که مدت زمان انجام برخی از وظایف نرم‌افزاری که سیستم‌های پیشرو می‌توانند انجام دهند، هر چهار تا هفت ماه دو برابر شده است. اگر این سرعت بهبود ادامه یابد، عوامل هوش مصنوعی به زودی وظایفی را انجام خواهند داد که در حال حاضر برای برنامه‌نویسان انسانی روزها یا هفته‌ها به طول می‌انجامد. از آنجا که آنها می‌توانند با نظارت کم و با سرعت بالا کار کنند، استفاده از عوامل هوش مصنوعی ممکن است به مزایای اقتصادی و علمی قابل توجهی منجر شود. به عنوان مثال، سیستم‌های هوش مصنوعی عامل‌محور در آزمایشگاه‌های شیمی خودران، افزایش بیش از ده برابری در سرعت کشف مواد جدید را نشان داده‌اند. غربالگری متون با کمک هوش مصنوعی ممکن است در برخی از محیط‌های تحقیقاتی، حجم کار را تقریباً 60 درصد کاهش داده باشد. بنابراین، عوامل هوش مصنوعی پیامدهای قابل توجهی در تمام صنایع دارند. در عین حال، استقرار آنها سوالات فوری را برای

بازارهای کار، امنیت سایبری، اکوسیستم اطلاعات و حاکمیت و کنترل سیستم‌های هوش مصنوعی آینده مطرح خواهند نمود.

درک و مدیریت ریسک

توسعه هوش مصنوعی خطراتی را به همراه دارد که می‌تواند تأثیرات منفی بالقوه‌ای بر حقوق انسانی ، سیستم‌های اجتماعی و محیط زیست داشته باشد. به عنوان مثال، انتشار مطالب مربوط به سوءاستفاده جنسی از کودکان که توسط هوش مصنوعی تولید شده اند و خشونت جنسی که با جعل عمیق (deep-fake) ساخته می‌شوند، اکنون در اینترنت بیشتر شده است و باعث آسیب به زنان و کودکان آسیب شده است. رفتار چاپلوسانه هوش مصنوعی، که در آن پاسخ‌هایشان باورهای موجود کاربران را صرف نظر از صحت و سقم آنها تقویت می‌کند، با چندین حادثه شدید سلامت روان، از جمله مرگ‌های ثبت شده، مرتبط بوده است. هوش مصنوعی تولید و هدف قرار دادن محتوای متقاعدکننده در مقیاس بزرگ، از جمله محتوایی که برای گمراه کردن طراحی شده است را آسان‌تر می‌کند و به فرسایش تدریجی یکپارچگی اطلاعات کمک می‌کند که می‌تواند واقعیت مشترک مورد نیاز برای اعتماد عمومی، انسجام اجتماعی و مشورت دموکراتیک را تضعیف کند. استفاده مجرمان و بازیگران بد از سیستم‌های هوش مصنوعی برای کمک به حملات سایبری مستند شده است. بسیاری از این آسیب‌ها به طور نامتناسبی بر جمعیت‌های از قبل محروم وارد می‌شود.

با نگاهی به آینده، شکاف فزاینده مابین قابلیت‌های به سرعت در حال بهبود هوش مصنوعی و روش‌های مؤثر مدیریت ریسک، ممکن است منجر به نتایج فاجعه‌باری شود. به عنوان مثال، توانایی‌های فنی پیشرفته ممکن است به بازیگران خصوصی تازه‌کار اجازه دهد تا از هوش مصنوعی به روش‌های مخرب در طیف وسیعی از کاربردها مانند کلاهبرداری، مهندسی اجتماعی، امنیت سایبری، اطلاعات نادرست، بیوتکنولوژی و دستکاری مالی استفاده کنند. روش‌های قابل اعتمادی برای حفظ کنترل بر سیستم‌های هوش مصنوعی بسیار مستقل وجود ندارند. هیچ تضمین علمی وجود ندارد که عوامل هوش مصنوعی دستورالعمل‌ها را نقض نکنند. متاسفانه در این زمینه شواهد در حال افزایش هستند و مواردی وجود دارند که هوش مصنوعی قبلاً آنها را نقض کرده است.. در محیط‌های آزمایشگاهی، نشان داده شده است که سیستم‌های هوش مصنوعی برای جلوگیری از خاموش شدن، دستورالعمل‌های ایمنی خود را نقض می‌کنند. رفتار مشابه ممکن است چالش‌هایی را برای روش‌های ارزیابی و نظارت ایجاد کند، زیرا توانایی سیستم‌های پیشرو هوش مصنوعی در تشخیص محیط‌های آزمایش، و تولید نتایج ارزیابی گمراه‌کننده که به نفع ادامه فعالیت آنها است، افزایش می‌یابد. علاوه بر این، خطرات جدیدی ممکن است از تعاملات بین چندین عامل ایجاد شود.

خطرات هوش مصنوعی به طور ناموزون در بین جمعیت‌ها و کشورها توزیع شده است، در حالی که توسعه هوش مصنوعی و ثروتی که ایجاد می‌کنند بسیار متمرکز است. تمرکز قابلیت‌های هوش مصنوعی در تعداد کمی از شرکت‌ها و کشورها می‌تواند باعث تصرف اقتدارگرایانه و تضعیف پاسخگویی دموکراتیک شود.

مدیریت هوش مصنوعی برای دستیابی به مزایا و کاهش خطرات

تحقق کامل مزایای هوش مصنوعی ضمن به حداقل رساندن خطرات آن، نیازمند حکمرانی خوب است. دستاوردهای اقتصادی و نیروی کار و توزیع عادلانه آنها به صورت خودکار حاصل نمی‌شوند: با سرمایه‌گذاری‌های مکمل در مهارت‌ها، گردش‌های کاری، زیرساخت‌ها و نهادهای بازار کار، فناوری می‌تواند مشاغل جدیدی ایجاد کند که در حال حاضر وجود ندارند - بیش از 60 درصد مشاغل در سال 2018 در مقایسه با سال 1945 جدید هستند. بدون این سرمایه‌گذاری‌ها، هوش مصنوعی به جای ایجاد مشاغل خوب پایدار - مشاغلی با درآمدهای منصفانه، استقلال کارکنان و مسیری قابل اعتماد برای کرامت اجتماعی - خطر گسترش نابرابری، جابجایی کارکنان و انتقال ثروت از نیروی کار به سرمایه را به همراه دارد. هوش مصنوعی می‌تواند از طریق آموزش شخصی‌سازی شده، ابزارهای سلامت روان و فناوری‌های کمکی بهبود یافته، قابلیت‌های انسانی را به شدت گسترش دهد، اما تحقق ایمن این فرصت‌ها نیازمند سرمایه‌گذاری‌ها و سیاست‌های اختصاصی برای ایجاد انگیزه برای دسترسی عادلانه و پاداش به نوآوری است، در عین حال از استثمار جمعیت‌های آسیب‌پذیر، به ویژه کودکان، جلوگیری می‌کند و از جابجایی تخصص، وابستگی روانی یا حذف فرهنگی و زبانی پیشگیری می نماید.

سیاست‌گذارانی که به دنبال شکل‌دهی به این حکمرانی هستند، با یک معضل شواهد مواجه می باشند: آن‌ها برای تصمیم‌گیری‌های آگاهانه و مهم در حوزه حکمرانی به شواهد نیاز دارند، اما زمانی که شواهد وجود داشته باشد، ممکن است برای تصمیم‌گیری خیلی دیر شده باشد، زیرا شواهد از سرعت توسعه هوش مصنوعی عقب مانده‌اند. ده‌ها ابزار حکمرانی متمایز که به دنبال گنجاندن اخلاق و حقوق انسانی در سیستم‌های هوش مصنوعی هستند، در حال حاضر در حوزه‌های قضایی مختلف در حال استفاده می باشند، اما این ابزارها پراکنده هستند، تنها بین چند شرکت متمرکز شده‌اند و به ندرت اثربخشی دنیای واقعی را اندازه‌گیری می‌کنند. خود روش‌های ارزیابی هنوز توسعه نیافته‌اند و نهادهای مورد نیاز برای ارائه ارزیابی‌های مستقل از قابلیت‌ها و ریسک‌ها همچنان در مراحل ابتدایی خود هستند.

ظرفیت اقدام بر اساس شواهد موجود از خطرات و تأثیرات هوش مصنوعی، به طور ناموزون توزیع شده است. اکثر کشورها، از جمله بسیاری از اقتصادهای پیشرفته، فاقد تخصص فنی برای ارزیابی توانمندترین

مدل‌های مرز(frontier) یا مشارکت معنادار در حکمرانی خود هستند. زیرساخت‌های محاسباتی، تخصص ارزیابی و داده‌ها (به عنوان مثال برای پوشش زبان‌های مختلف) در جایی که هوش مصنوعی ساخته شده است متمرکز شده‌اند و اکثر کشورهای عضو را به سیستم‌هایی وابسته می‌کنند که نمی‌توانند آنها را بسازند، بازرسی کنند، حسابرسی کنند یا به طور کامل با بافت محلی سازگار کنند. دسترسی به ابزارهای هوش مصنوعی به تنهایی سود برابر ایجاد نمی‌کند. سرمایه‌گذاری‌های مکمل در داده‌ها، مهارت‌ها، گردش‌های کاری و مؤسساتی که دسترسی را به استقرار مفید، مقرون به صرفه و ایمن تبدیل می‌کنند، ضروری هستند، اما به طور نابرابر توزیع شده‌اند.

گام‌های بعدی مشخصی برای پر کردن شکاف‌های فوق وجود دارد، اما هر یک از آنها نیازمند سرمایه‌گذاری پایدار در ظرفیت کشورهای عضو برای شکل‌دهی، ارزیابی و استقرار هوش مصنوعی است. این گزارش اولیه، خود بخشی از سهم مشارکت پنل در ایجاد یک پایگاه شواهد مشترک برای کشورهای عضو می باشدکه در تصمیم‌گیری‌های فوری و رو به افزایش، هدایت شوند. با عمیق‌تر شدن درک پنل از طریق تعامل مداوم با کشورهای عضو و جامعه علمی گسترده‌تر، تجزیه و تحلیل آن نیز فراتر از شکاف‌های شناسایی شده گسترش خواهد یافت تا مسیرها، تنش‌ها و فرصت‌هایی را که آینده هوش مصنوعی را تعریف می‌کنند، ترسیم کند.

چرا زمان حاضر؟

واقعیت حال حاضر

ما در یک نقطه عطف قرار داریم. هوش مصنوعی صرفاً یک فناوری نوظهور دیگر نیست؛ بلکه اولین فناوری است که پذیرش و رسوخ آن در جامعه از دهه‌ها به ماه‌ها کاهش یافته است، کارهای شناختی را به سطح مقیاس صنعتی رسانیده است و قابلیت تحول را در دستان چند بازیگر جهانی متمرکز نموده است. این گزارش تلاش دارد تا سیاست‌گذاران را در تمام جهان به پایگاه شواهد مشترک مورد نیاز برای پاسخگویی مجهز کند.

هوش مصنوعی چیست؟

هوش مصنوعی (AI) یک فناوری دگرگون‌کننده است، اما همچنین یک هدف متحرک نیز می باشد. این اصطلاح در طول زمان، از هوش مصنوعی نمادین به یادگیری ماشینی، به هوش مصنوعی مولد، هوش مصنوعی عاملی و گاهی اوقات حتی به هوش عمومی مصنوعی یا ابرهوش تغییر شده است. سیستم‌های هوش مصنوعی، سیستم‌های ماشینی هستند که به طور کلی، درک می‌کنند، یاد می‌گیرند و عمل می‌کنند. آن‌ها از ورودی‌ها استنباط می‌کنند که چگونه خروجی‌هایی مانند پیش‌بینی‌ها، محتواها، توصیه‌ها، اقدامات یا تصمیمات را با درجات مختلفی از استقلال و سازگاری تولید کنند. چیزی که هوش مصنوعی فعلی را بیش از هر معماری واحدی متحد می‌کند، این است که سیستم‌های مدرن از تجاربی که توسط داده‌ها نشان داده می‌شود، یاد می‌گیرند. این تجربه می تواند به اشکال گوناگون باشد: یادگیری از آثار فرهنگی انسان (متون، تصاویر، کد) پایه "پیش‌آموزش" مدل‌های بنیادی امروزی را فراهم می‌کند که سیستم‌های بزرگ و آموزش‌دیده‌ای هستند که طیف وسیعی از برنامه‌های هوش مصنوعی را پشتیبانی می‌کنند. یادگیری از طریق تعامل با دنیای (دیجیتال و فیزیکی) زیربنای یادگیری تقویتی و رباتیک است. و یادگیری از شبیه‌سازی به عامل‌ها اجازه می‌دهد تا در محیط‌های مجازی تجربه کسب کنند.

مدل‌های بنیادی و هوش مصنوعی با وظیفه خاص. بحث‌های عمومی اغلب بر مدل‌های بنیادی و هوش مصنوعی همه منظوره متمرکز است که با توانایی آنها در انجام وظایف تعریف می‌شوند، یا برای انجام طیف گسترده‌ای از وظایف سازگار می گردند. این سیستم‌ها در حال حاضر در مقیاس وسیع مستقر شده‌اند و موضوع اصلی این گزارش هستند. با این حال، بسیاری از سیستم‌های هوش مصنوعی به اصطلاح محدود برای انجام وظایف خاص در حوزه‌های خاص طراحی شده‌اند. این تمایز مهم است. هوش مصنوعی با وظیفه خاص، زمانی که وظیفه به خوبی تعریف شده باشد، داده‌ها در دسترس باشند و مؤسسات بتوانند آن را به درستی مستقر کنند،

می تواند مزایای قابل اندازه‌گیری ارائه دهد. سیستم‌های عمومی و چندمنظوره انعطاف‌پذیری بسیار بالایی را ارائه می‌دهند، اما در مقابل چالش‌های مدیریتی متفاوتی را نیز ایجاد می‌کنند.

چرا هوش مصنوعی با سایر فناوری‌های نوظهور متفاوت است؟

سرعت بی‌سابقه پذیرش. سیستم‌های هوش مصنوعی به طور فزاینده‌ای به عنوان یک فناوری همه منظوره در نظر گرفته می‌شوند، که از نظر وسعت کاربرد به اندازه موتور بخار، برق و اینترنت متحول کننده می باشد [1،2]. با این حال، از جهات مهمی هم با آنها متمایز است. همگانی شدن الکتریسیته و رسیدن آن به اکثریت خانوارها دهه‌ها طول کشید؛ 15 سال طول کشید تا تعداد کاربران اینترنت به یک میلیارد برسد. تعداد کاربران ChatGPT در عرض دو ماه به 100 میلیون کاربر رسید [3]. سیاست‌گذاری سنتی نتوانسته است همگام با این سرعت به پیش برود [4].

تمرکز و همگن‌سازی. هوش مصنوعی مدرن، به ویژه مدل‌های بنیادی، براساس اقتصاد مقیاسی عمل می کنند. این امر فشارهای شدیدی را برای متمرکز شدن توانایی ها ایجاد می‌کند، به گونه ای که آموزش آنها تنها از عهده تعداد انگشت‌شماری از بازیگران جهانی بر می آید که دسترسی به منابع محاسباتی قدرتمندی را دارند. این تمرکز منابع، خطرات همگن‌سازی دانش و فرهنگ را به همراه دارد. به عنوان مثال، آموزش بر روی مجموعه داده‌های تلفیقی ممکن است سیستم‌هایی ایجاد کند که به طور سیستماتیک زبان‌ها و دیدگاه‌های غالب را منعکس می‌کنند و در عین حال دیگران را به حاشیه می‌رانند.

تأثیر عظیم بر کار دانش‌محور. در حالی که امواج اولیه اتوماسیون، کار فیزیکی را متحول کردند، هوش مصنوعی اولین پدیده‌ای است که به طور گسترده بر کار شناختی و خلاقانه، از جمله نویسندگی، برنامه‌نویسی، تحلیل حقوقی، تشخیص پزشکی، اکتشافات علمی، پیش‌بینی و تولید تصویر تأثیر می‌گذارد [1،2،5-8].

چالش‌هایی در تضمین واقعیت و صحت خروجی‌ها. مدل‌های هوش مصنوعی امروزی یاد می‌گیرند که الگوها را در مجموعه‌های داده بزرگ پیش‌بینی کنند. مدل‌های زبانی تنها الگوهای از پیش ذخیره شده را به کار نمی گیرند بلکه خروجی مورد نظر کاربر را به چندین فرم زبانی مختلف تبدیل می کنند تا از میان آنها آن که از نظر ادبی روان تر می باشد را برگزینند. این امر امکان تولید خروجی‌های جدید را به جای کپی کردن صرف الگوها فراهم می‌کند، اما یک عدم تقارن بحرانی را نیز ایجاد می‌نماید: تولید متن روان آسان‌تر از تولید متن واقعی است [8]. مدل‌های زبانی بزرگ می‌توانند توهمات را به عنوان واقعیت ارائه دهند و کاربران اغلب اعتماد به روان بودن زبانی را به عنوان شاهدی از قابل اعتماد بودن و قابلیت اطمینان واقعی در نظر می‌گیرند. این عدم ارتباط بین شایستگی ظاهری و دقت، چشم‌انداز ریسک را به روش‌های سیستماتیک شکل می‌دهد. در

مراقبت‌های بهداشتی، هوش مصنوعی عمومی، زمان مستندسازی اداری را کاهش می‌دهد، وظیفه‌ای که در آن هوش مصنوعی برای ترکیب و ساختاردهی اطلاعات ارزشمند است. با این حال، طبق گزارش‌ها، از هر چهار مکالمه با چت‌بات، یکی مربوط به سلامت یا تندرستی است [9]، به این معنی که همین سیستم‌ها به طور معمول برای اهداف تشخیصی بالقوه مورد مشورت قرار می‌گیرند، جایی که دقت واقعی بسیار مهم است و خطاها عواقب جدی دارند. فناوری در تمام موارد یکسان است، اما خطرات آن یکسان نیستند. توانایی به معنای مناسب بودن نیست و استقرار بدون نظارت و ارزیابی سیستماتیک، به ویژه در حوزه‌هایی که در حال حاضر فاقد چارچوب‌های نظارتی هستند، در مواردی که تشخیص آن دشوار است، خطر آسیب را به همراه دارد.

نیاز فوری به ارزیابی مستقل مبتنی بر شواهد از سیستم‌های هوش مصنوعی

نیاز به ارزیابی علمی مستقل در حال حاضر ناشی از شرایطی است که قواعد کنونی و موجود نظارتی را تغییر می‌دهد. نخست اینکه سرعت پیشرفت هوش مصنوعی از ارزیابی‌های قبلی متخصصان و چرخه‌های نظارتی پیشی گرفته است. پیشرفت قابلیت‌ها در حوزه‌های کلیدی ادامه دارد و این پیشرفت ها توسط تکنیک‌های جدید آموزش و مقیاس‌بندی محاسبات زمان استنتاج [10] هدایت می‌شوند. اندازه‌گیری‌های تجربی نشان دهنده «شتاب گرفتن» قابلیت‌های هوش مصنوعی است [11]. دوم، خوشبختانه توسعه‌دهندگان پیشرو مدل‌های هوش مصنوعی پیشرو، شروع به محدود کردن استقرار مدل‌هایی کرده‌اند که از «آستانه‌های ریسک تعریف‌شده داخلی» فراتر می‌روند [12-15]. با این حال، این آستانه‌ها بدون آنکه ارزیابی استاندارد یا تأیید خارجی داشته باشند، توسط خود توسعه‌دهندگان هوش مصنوعی تعریف شده اند [16-17]. سوم، رویکردهای حاکمیت هوش مصنوعی در مناطق مختلف جهان همچنان پراکنده است. بی‌نظمی فزاینده‌ای در حاکمیت جهانی مشاهده می‌شود، به طوری که برخی از کشورها «قوانین خاص هوش مصنوعی را با قوانین اساساً متناقض و هزینه‌های انطباق» معرفی کرده‌اند. حوزه‌های قضایی فلسفه‌های نظارتی متفاوتی را نشان می‌دهند، بدون هیچ سازوکار یکپارچه‌ای برای مدیریت ریسک، بدون استانداردهای ارزیابی قابل مقایسه، و هماهنگی محدود در سراسر حوزه‌های قضایی، که خطر یک چشم‌انداز نظارتی پراکنده را به همراه دارد [18]. با این حال، پراکندگی سرنوشت محتوم نیست و هنوز دریچه‌ای برای ایجاد استانداردهای مشترک و هماهنگی در نظارت جهانی همچنان باز است.

چه چیزی پنل بین‌المللی مستقل در مورد هوش مصنوعی را منحصر به فرد می‌کند؟

ارزیابی‌های توسعه هوش مصنوعی توسط گروه‌های متخصص متعددی وابسته به سازمان‌های بین‌المللی، شوراهای علمی ملی، کنسرسیوم‌های صنعتی و مراکز تحقیقاتی مستقل انجام می‌شود. این تلاش‌ها ارزشمند هستند اما به دلیل محدودیت‌های جغرافیایی، دیدگاه بخشی نگرانه یا عدم تداوم محدود شده‌اند. این پنل به سه

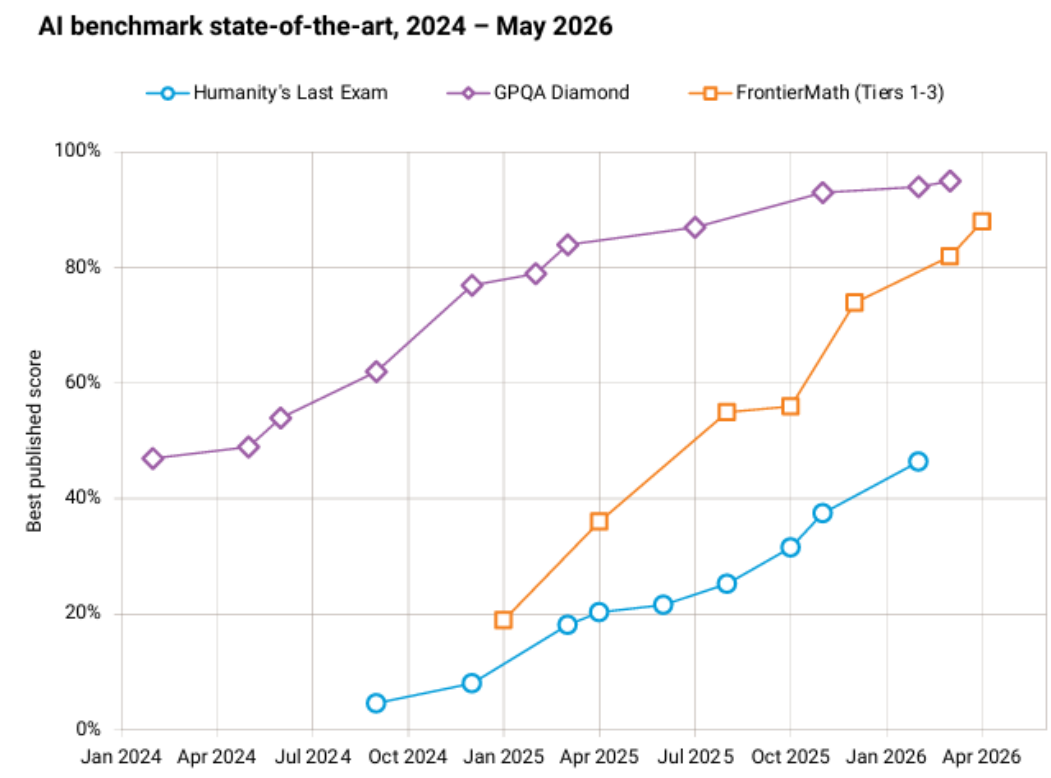
دلیل جایگاه متمایزی دارد. اول، از این فرض ناشی می‌شود که سازمان ملل متحد، پیشروترین مجمع جهانی در مورد خطرات فرامرزی در این مقیاس است، همانطور که در گزارش «مدیریت هوش مصنوعی برای بشریت» بیان شده و در «پیمان جهانی دیجیتال» منعکس شده است، «ضرورت تشخیص اینکه سرعت و قدرت فناوری‌های نوظهور، امکانات جدیدی را ایجاد می‌کنند» و «شناسایی و کاهش خطرات و تضمین نظارت انسانی بر فناوری به روش‌هایی که توسعه پایدار و بهره‌مندی کامل از حقوق بشر را پیش می‌برد» را بیان می‌کند. دوم، این پنل در حال حاضر تنها سازوکار پابرجای سازمان ملل متحد است که وظیفه ارزیابی علمی منظم از وضعیت، خطرات و قابلیت‌های هوش مصنوعی را بر عهده دارد و برای کار پایدار و تکراری طراحی شده است. سوم، این پنل یک وظیفه علمی و نه سیاسی، دارد: مستندسازی شواهد علمی، اجماع و اختلاف نظرها، و اینکه کدام شکاف‌های دانشی همچنان نیاز به رسیدگی فوری دارند. هدف آن تجهیز دولت‌ها و مؤسسات به شواهد مورد نیاز برای اقدام در ماه‌ها و سال‌های آینده است، همچنان مرتبط با سیاست‌گذاری اما نه تجویزی. این ویژگی علمی باید یافته‌های آن را در مناطق مختلف جهان قابل مقایسه کرده و در برابر چرخه‌های سیاسی انعطاف‌پذیر نماید.

۲-شواهد موجود چه واقعیت هایی را نشان می دهند؟

در سال های اخیر عملکرد هوش مصنوعی در معیارهای عملکردی پیشرفته به شدت افزایش یافته است (شکل ۱ را ببینید).

در یک امتحان با ۲۵۰۰ سوال که به طور خاص برای مدل‌های عمومی هوش مصنوعی طراحی شده است (با نام امتحان آخر بشریت **Humanity's Last Exam**)، شاهد افزایش نمرات مدل های برتر از ۸٪ به ۴۵٪ در ۱۶ ماه بوده ایم [19]. در آزمون GPQA Diamond، یک آزمون استدلال علمی در سطح دکترا، بهترین مدل‌ها اکنون به حدود ۹۵٪ از سوالات به درستی پاسخ می‌دهند، این مقدار در سال ۲۰۲۳ تنها ۳۶٪ بوده است [20]. پاسخ ها برای Math Frontier که توانایی استدلال ریاضی مدل ها را تست می‌کند، از ۱۹٪ در ژانویه ۲۰۲۵ به ۸۸٪ در سال ۲۰۲۶ افزایش یافت [20]. چندین سیستم هوش مصنوعی در المپیاد بین‌المللی ریاضی ۲۰۲۵ به مدال طلا دست یافتند، نقطه عطفی که به طور قابل توجهی زودتر از آنچه بسیاری از کارشناسان پیش‌بینی کرده بودند، فرا رسیده است [21].

FIGURE I



Source: Adapted from data on AI benchmarks (EpochAI, 2026, https://epoch.ai/benchmarks).

۱.۲ قابلیت‌های هوش مصنوعی سریع‌تر از توانایی اندازه‌گیری یا مدیریت آنها در حال پیشرفت هستند

اندازه‌گیری و ارزیابی هوش مصنوعی، می تواند مبنایی برای ارزیابی فرصت‌ها، خطرات و تأثیرات هوش مصنوعی باشد. با این حال، سرعت بی‌سابقه توسعه هوش مصنوعی، چالش‌های ارزیابی بسیاری را ایجاد کرده است:

الف. یک عدم تقارن جدی اطلاعات در اعتبارسنجی ایمنی بین شرکت‌ها و جامعه وجود دارد. توسعه دهندگان هوش مصنوعی پیشرو، مالکیت جزییات سیستم‌هایی را که ساخته‌اند حفظ می‌کنند. روش‌های ارزیابی ایمنی نیز در حال حاضر عمدتاً توسط خود شرکت‌هایی که توسعه دهنده هوش مصنوعی هستند و مورد ارزیابی قرار می‌گیرند، طراحی می‌شوند. در حالی که برخی از افشاگری‌های قانونی الزامی وجود دارد، کارشناسان بی طرف در درجه اول تنها به داده‌های آزمایشی که توسعه‌دهندگان برای به اشتراک گذاشتن انتخاب می‌کنند دسترسی دارند. بدون ارزیابی استاندارد، دقیق و مستقل شخص ثالث، مشابه آنچه برای صنایع داروسازی و هوانوردی وجود دارد، تضمین ایمنی تا حد زیادی به حسن نیت توسعه‌دهنده بستگی خواهد داشت [21]؛

ب. هوش مصنوعی می‌تواند راه‌حل‌های موجود و در دسترس عموم برای آزمایش‌ها را به خاطر بسپارد. اگر پاسخ‌های صحیح به آزمایش‌ها (به طور ناخواسته) توسط مدل هوش مصنوعی به عنوان بخشی از فرآیند آموزش به خاطر سپرده شده باشد، عملکرد آن در این آزمایش‌ها ممکن است به سؤالات مشابه تعمیم داده نشود. برای جلوگیری از آلودگی داده‌ها، مجموعه داده‌های ارزیابی به طور فزاینده‌ای محرمانه نگه داشته می‌شوند [22]؛

ج. تعداد فزاینده‌ای از آزمایش‌ها برای هوش مصنوعی بسیار ساده و آسان به نظر می رسد. مدل‌های هوش مصنوعی تقریباً در تعداد فزاینده‌ای از آزمون‌های استاندارد شده، به نام الگوها، که محققان برای سنجش قابلیت‌های آنها قبل از انتشار استفاده می‌کنند، امتیاز کاملی را کسب می‌کنند [23]. در نتیجه، الگوهای موجود دیگر نمی‌توانند یک مدل بسیار توانمند را از یک مدل حتی بهتر تشخیص دهند؛

د. مدل‌های هوش مصنوعی قادر به فریب فعال هستند. فریب در جایی رخ می دهد که یک سیستم هوش مصنوعی به طور سیستماتیک انسان‌ها یا سایر عوامل را در مورد دانش، برنامه‌ها یا قابلیت‌های خود گمراه می‌کند. این امر به طور فزاینده‌ای در عمل مشاهده می‌شود. فریب ممکن است ارزیابی‌های ایمنی و قابلیت اطمینان در دنیای واقعی را مختل کند و با سناریوهای از دست دادن کنترل مرتبط می باشد، این امر با رفتارهای مدل‌های هوش مصنوعی که برای جلوگیری از خاموش شدن خود دروغ می‌گویند و تقلب می‌کنند، کاملا مشهود است [24].

ه. مدل‌های هوش مصنوعی ممکن است بفهمند چه زمانی آزمایش می‌شوند. این چالش نوظهور، آگاهی از ارزیابی نامیده می‌شود [25]. این امر در ترکیب با قابلیت فریب، به این معنی است که مدل‌های هوش مصنوعی می‌توانند توسط انسان‌ها آموزش ببینند یا به طور مستقل تصمیم بگیرند که عملکرد آزمایشی خود را در ارزیابی قابلیت‌های ضعیف خود به طور موقت تغییر دهند [26].

و. عامل های هوش مصنوعی، آزمایش را پیچیده تر می‌کنند. عامل های هوش مصنوعی که به نمایندگی از انسان‌ها عمل می‌کنند، ممکن است از ابزارهایی برای انجام وظایف طولانی بدون نظارت مستقیم انسان استفاده کنند. روش‌های ارزیابی که با ظرفیت یک عامل برای اقدام مستقل، تأثیر بر محیط عملیاتی آن و رفتار نوظهور کالیبره می‌شوند، هنوز توسعه نیافته‌اند [27]. علاوه بر این، هنگامی که چندین عامل با هم تعامل می کنند، خطرات سیستمی جدیدی از جمله عدم هماهنگی، تضاد و تبانی پدیدار می‌شوند [28]. پاسخ‌های موجود به این چالش‌ها عبارتند از:

الف. آزمون‌های پویا و مبتنی بر محیط واقعی عملیاتی. اندازه‌گیری دقیق نیاز به منابع قابل توجهی دارد تا دائماً الگوهای جدیدی را توسعه دهد که برای سیستم‌های پیشرفته هوش مصنوعی به اندازه کافی دشوار هستند. برای دشوار نگه داشتن الگوها و انعکاس سودمندی واقعی، شیوه‌های ارزیابی از محیط‌های ایستا به محیط‌های پویا و عملیاتی تغییر می‌کنند [29،30]. با این حال، ساخت چنین محیط‌هایی گران‌تر از یک آزمون دانشی است و تعداد کمی از بازیگران به اندازه کافی در اندازه‌گیری هوش مصنوعی سرمایه‌گذاری کرده‌اند تا آنها را ایجاد کنند.

ب. تفسیرپذیری. روش‌های تفسیرپذیری، که هدفشان درک اتفاقات درون مدل‌های هوش مصنوعی است، برای جستجوی رفتارهای خطرناک پنهان، اهمیت فزاینده‌ای پیدا کرده اند. یکی از روش‌های قابل توجه، «زنجیره فکری» است که در آن مدل قبل از پاسخ دادن، مراحل استدلال خود را تشریح می‌کند. این امر امیدوارکننده است، اما ضروری است که اطمینان حاصل شود که این استدلال در وهله نخست صادقانه و در وهله دوم برای انسان‌ها قابل فهم باشد [31]. رویکرد دیگر، یک طبقه‌بندی‌کننده است که بر اساس مکانیزم های داخلی مدل آموزش دیده است و می‌تواند پاسخ به سوالی مانند «آیا این مدل صادق است؟» را پیش‌بینی کند [32]. با این حال، چنین طبقه‌بندی‌کننده‌ای نیاز به دسترسی به جزییات مدل دارد و ممکن است فقط در صورتی که سازمان‌های ارزیابی معتبر دسترسی عمیق‌تری داشته باشند، قابل به کارگیری باشد[33].

ج. اندازه‌گیری مداوم. این به معنای ردیابی نحوه رفتار یک سیستم پس از انتشار، با کاربران واقعی، وظایف واقعی و محیط‌های واقعی است. این نظارت می‌تواند شامل الگوهای استفاده از هوش مصنوعی که توسط توسعه‌دهندگان هوش مصنوعی [34] ارایه شده اند و نیز گزارش حوادث و نتایج گزارش شده توسط کاربر باشد. تا به امروز، هیچ استاندارد مشترکی برای تحلیل الگوهای استفاده توسط تولیدکنندگان هوش مصنوعی با حفظ

حریم خصوصی وجود نداشته است. علاوه بر این، آگاهی از اکوسیستم برای مدل‌های باز که ممکن است بدون هیچ گونه دیدی برای تولیدکنندگان هوش مصنوعی دانلود و استفاده شوند، کمتر است. ارائه دهندگان سیستم‌های هوش مصنوعی پرخطر که در بازار اتحادیه اروپا قرار می‌گیرند، باید حوادث جدی هوش مصنوعی را گزارش دهند [35]. پایگاه‌های داده مستقل حوادث هوش مصنوعی نیز توسط سازمان همکاری و توسعه اقتصادی (OECD) [36] و موسسه فناوری ماساچوست [37] نگهداری می‌شوند. گسترش پایگاه‌های داده حوادث هوش مصنوعی، منعکس کننده شیوه‌های ایمنی تثبیت شده در سایر صنایع بالغ و با پیامدهای بالا است. به طور خلاصه، معضل شواهد جدی است اما غیرقابل حل نیست.

۲.۲ تنها تعداد انگشت‌شماری از بازیگران، مدل‌های هوش مصنوعی مرزی را آموزش داده‌اند

ورودی های اصلی برای ایجاد یک مدل هوش مصنوعی عبارتند از قدرت محاسباتی، داده‌ها و استعدادهای مهندسی که همگی در تعداد انگشت‌شماری از شرکت‌ها در تعداد انگشت‌شماری از کشورها متمرکز شده‌اند. دسترسی به مدل‌های هوش مصنوعی مرزی و پیشرو نیز برای تولید نسل بعدی مدل‌های هوش مصنوعی به طور فزاینده‌ای اهمیت پیدا می‌کند. ویژگی‌های توسعه هوش مصنوعی عبارتند از:

الف. تمرکز بازار. زنجیره تأمین هوش مصنوعی پیشرفته دارای مراحل متعددی با تمرکز بازار بسیار بالا است که در آن یک ارائه‌دهنده واحد 80٪ یا بیشتر از بازار جهانی را در اختیار دارد [38]، از جمله ASML در اروپا (لیتوگرافی فرابنفش شدید)، TSMC در شرق آسیا (تولید تراشه‌های پیشرفته) و NVIDIA در ایالات متحده (طراحی تراشه‌های هوش مصنوعی). مراحلی با تمرکز بازار بالا که در آن سهم جهانی سه بازیگر بزرگ بیش از 60٪ گزارش شده است شامل حافظه با پهنای باند بالا، ارائه فضای ابری و ارائه مدل پایه هوش مصنوعی از طریق رابط برنامه‌نویسی کاربردی (API) است.

ب. تمرکز جغرافیایی. در سال 2025، مؤسسات مستقر در ایالات متحده 59 مدل هوش مصنوعی قابل توجه تولید کردند، در حالی که این رقم در چین 35 و در سایر نقاط جهان فقط 13 مدل بوده است [39]. در همان سال، 75٪ از قدرت محاسباتی 500 خوشه محاسباتی هوش مصنوعی خصوصی و عمومی شناخته شده بزرگ در ایالات متحده و پس از آن 15٪ در چین و 10٪ در سایر نقاط جهان قرار داشت [40].

ج. توسعه مبتنی بر کسب و کار. توسعه مدل‌های هوش مصنوعی پیشرو و همه منظوره تحت سلطه تعداد کمی از شرکت‌های خصوصی با منابع محاسباتی عظیم قرار دارد. در سال 2025، 91٪ از مدل‌های هوش مصنوعی

مهم، از بخش خصوصی سرچشمه گرفته‌اند [39]. در نتیجه، بسیاری از تصمیمات در مورد داده‌های آموزشی، حفاظت‌ها، آستانه‌های استقرار، دسترسی به مدل و انتشار قابلیت‌ها در داخل شرکت‌های خصوصی گرفته می‌شود.

این تمرکز بالای قدرت و ظرفیت با چالش‌هایی همراه است:

الف. اقتصاد سیاسی.

تمرکز بالای بازار می‌تواند به شرکت‌ها اجازه دهد تا رانت‌های قابل توجهی دریافت کنند. اگر هوش مصنوعی در نهایت تولید را از نیروی کار به سرمایه متمرکز در چند شرکت و کشور منتقل کند، این امر ممکن است نگرانی‌های مالی را در کشورهایی که به مالیات بر نیروی کار متکی هستند، افزایش دهد.

ب. تمرکز قدرت سیاسی.

توسعه و استقرار هوش مصنوعی انگیزه‌هایی را برای جمع‌آوری، پردازش، استفاده مجدد و نگهداری گسترده داده‌ها ایجاد می‌کند [41]. در حالی که برخی از حوزه‌های قضایی از قوانین قوی حفظ حریم خصوصی و حفاظت از داده‌ها بهره‌مند می‌شوند، اگر خارج از چارچوب‌های حفاظتی مستقر شوند، ظرفیت متمرکز هوش مصنوعی نگرانی‌هایی را در مورد تأثیرات بر دموکراسی و حقوق بشر [42]، همراه با احتمال تصرف نظارتی و عدم پاسخگویی ایجاد می‌کند.

ج. جنوب جهانی.

سیستم‌های فعلی هوش مصنوعی تنها طیف محدودی از تنوع زبانی و فرهنگی جهان را منعکس می‌کنند و بخش زیادی از جمعیت جهان را شامل نمی‌شوند [43-45]. سرمایه‌گذاری پیشگیرانه مورد نیاز است. در عین حال، کشورهای جنوب جهان به دلیل محدودیت در تاب‌آوری محلی و کمبود ظرفیت در تمامی زمینه ها [46]، به طور نامتناسبی در برابر خطرات سوءاستفاده از هوش مصنوعی آسیب‌پذیر هستند.

د. همسویی با منافع عمومی.

دولت‌ها بایستی مابین انتخاب‌های توسعه‌دهندگان مبتنی بر کسب‌وکار و با منافع عمومی جامعه همسویی ایجاد کنند [47،48]. مدل‌های بسته هوش مصنوعی، مدل‌های با وزن باز، استقرار لبه رقابتی و پیشرو، و الگوهای آموزشی رقیب، بده‌بستان‌های مختلفی را برای دسترسی، شفافیت، تکرارپذیری، امنیت و کنترل ایجاد می‌کنند که همگی در جدول زیر آمده است. به طور مشابه، اگرچه ملاحظات امنیت ملی احتمالاً دسترسی به قدرتمندترین مدل‌ها را محدود می‌کند، اما بهبود دسترسی جهانی به محاسبات، حمایت از توسعه منطقه‌ای و

سرمایه‌گذاری در پوشش زبانی، وابستگی کشورهای عقب‌مانده را کاهش می‌دهد. درست است که این امر انحصار پشتیبانی محاسباتی را کاهش می‌دهد، ولی می تواند بازارهای جدیدی را برای تأمین‌کنندگان و توسعه دهندگان هوش مصنوعی باز کند.

	Proprietary models	Open-weight models	Open-source models	Open development (rare)
What is open	Nothing substantial; model weights are proprietary	Final model weights (AI model can be adapted but not reproduced)	Model weights and some components to reproduce them (e.g. training data)	The entire process of development (open collaborative development)
Third parties' degree of control	None	Medium	Medium to high	Maximum
Developer's ability to mitigate misuse	Maximum but currently insufficient	Almost none; can be fine-tuned by others for malicious purposes	Almost none; can be fine-tuned or retrained by others for malicious purposes	Almost none; can be fine-tuned or retrained by others for malicious purposes

۳.۲ ورودی‌ها و خروجی‌های هوش مصنوعی از نظر جغرافیایی و زبانی ناموزون هستند

الف. در توسعه هوش مصنوعی اکثر زبان‌ها و فرهنگ‌های جهان همچنان مورد توجه قرار نگرفته‌اند.

بیش از 7000 زبان در سراسر جهان صحبت می‌شوند، اما زیرساخت‌های توسعه و ارزیابی مدل هوش مصنوعی تنها بخش کوچکی از آنها را منعکس می‌کند [44]. در عین حال، تخمین زده می‌شود که بیش از 1000 زبان پایه‌های اجتماعی، دیجیتال و داده‌ای مورد نیاز برای گنجاندن معنادار در سیستم‌های هوش مصنوعی را دارند [44]. گنجاندن این زبان ها نیازمند سرمایه‌گذاری های هدفمند، مجموعه داده‌های عمومی مرتبط با آن زبان ها و نیز ابتکارات برای ساخت الگو در زمینه‌های زبانی و فرهنگی کمتر نمایانده شده، می باشند.

ب. شواهد، نشان دهنده تمرکز هستند.

شواهد مربوط به تأثیرات هوش مصنوعی، عمدتا در زمینه‌های پردرآمد و انگلیسی زبان متمرکز هستند. مطالعات اقتصادی نیز به سمت اقتصادهای پیشرفته، شرکت‌های بزرگ و کار رسمی متمایل می باشند. زیرساخت‌های ارزیابی هوش مصنوعی همچنان از نظر زبانی و جغرافیایی متمرکز هستند.

ج. استفاده از هوش مصنوعی مغرضانه می‌تواند نابرابری را تداوم بخشد.

شواهد در حال افزایش نشان می دهند که سیستم‌های هوش مصنوعی با طراحی یا آزمایش ضعیف می‌توانند به نتایج ناعادلانه و تبعیض‌آمیز کمک کنند [49-51]؛

د. آسیب‌های جهت دار و توزیعی در داخل و بین جوامع وجود دارد.

اکثریت قریب به اتفاق اهداف در پورنوگرافی دیپ فیک زنان هستند [52]. این امر می‌تواند مشارکت مدنی را تضعیف کند، به خصوص با هدف قرار دادن عمدی روزنامه‌نگاران زن [53]؛

ه. هوش مصنوعی می‌تواند نتایج متفاوتی را در موسسات مختلف ایجاد کند.

کارگران آمریکایی 22 تا 25 ساله در مشاغل در معرض هوش مصنوعی تقریباً 15 درصد کاهش نسبی اشتغال را تجربه کرده‌اند [54]. از سوی دیگر داده‌های کشور دانمارک تقریباً هیچ تأثیری بر اشتغال، ساعات کاری یا دستمزدها نشان نمی‌دهد [55]. این اختلاف بین کشورها، گواه این است که یک فناوری واحد می‌تواند در محیط‌های نهادی مختلف، نتایج متفاوتی ایجاد کند. به طور گسترده‌تر، هوش مصنوعی ممکن است شکاف‌های مهارتی درون وظایف را کاهش دهد [56،57]، اما ممکن است شکاف‌ها را در شرکت‌ها، مناطق و کشورها و بین سرمایه و نیروی کار افزایش دهد [58].

۴.۲ شکاف هوش مصنوعی فقط مربوط به دسترسی نیست، بلکه مربوط به ظرفیت تأثیرگذاری بر توسعه هوش مصنوعی نیز می باشد.

شکاف هوش مصنوعی را می‌توان به عنوان شکاف بین کسانی که به هوش مصنوعی دسترسی دارند و کسانی که ندارند تعریف کرد. با این حال، ظرفیت هوش مصنوعی تنها مربوط به دسترسی نیست؛ بلکه چندبعدی است و موارد زیر را شامل می‌شود [59،61]:

الف. ظرفیت زیرساخت هوش مصنوعی، پایه مادی است که از چرخه کامل حیات سیستم‌های هوش مصنوعی پشتیبانی می‌کند.

حفظ محاسبات هوش مصنوعی، چه خصوصی و چه عمومی، در داخل مرزها، به طور فزاینده‌ای برای استقلال، نفوذ و امنیت ملی کشورها مورد نیاز است. به عنوان زیرمجموعه‌ای از این موضوع، بازار رو به رشدی برای زیرساخت‌های مستقل هوش مصنوعی پدیدار شده است، به طوری که اکنون اقتصادهای بزرگ در محاسبات داخلی سرمایه‌گذاری می‌کنند [62]؛

ب. ظرفیت توسعه استعدادها، مستلزم توانایی پرورش، جذب و حفظ نخبه گان و افراد با استعداد در زمینه هوش مصنوعی، و در عین حال افزایش سواد عمومی در این زمینه است [63،64].

به عنوان مثال، ریاضیات یکی از پایه مهم برای ساخت مدل‌های پیشرو و مرزی هوش مصنوعی به شمار می آید [65].

ج. حاکمیت هوش مصنوعی و ظرفیت خدمات عمومی، تعیین کننده توانایی درک، هدایت، تنظیم و پشتیبانی از توسعه هوش مصنوعی می باشد.

طبق گزارش کنفرانس تجارت و توسعه سازمان ملل متحد، 118 کشور، عمدتاً در کشورهای جنوب جهان، در بحث‌های اصلی مربوط به حاکمیت هوش مصنوعی مشارکت ندارند و کمتر از یک سوم کشورهای در حال توسعه، استراتژی‌های ملی هوش مصنوعی را تدوین کرده‌اند [67،68]. اکثر دولت‌ها در اقتصادهای پیشرفته فاقد کادر فنی لازم برای درک تغییرات سریع فناوری و تطبیق چارچوب‌های حاکمیتی با آن هستند [69].

این ظرفیت‌ها به هم پیوسته هستند. کشورهایی که زیرساخت هوش مصنوعی یا ظرفیت‌های آزمایش هوش مصنوعی خود را ندارند، در معرض خطر از دست دادن فرصت‌های توسعه مشترک فناوری‌های کلیدی، شکل‌دهی به چارچوب‌های حاکمیتی، تأثیرگذاری بر استانداردهای جهانی نوظهور و حفظ استعدادها هستند [70]. تلاش‌ها برای رسیدگی به این موضوع شامل موارد زیر است:

الف. سرمایه‌گذاری در زیرساخت‌های محلی.

نابرابری‌های جهانی در زیرساخت‌های محاسباتی و داده ها کماکان از موضوعات برجسته می باشند و نیاز به سرمایه‌گذاری قابل توجهی دارند. در حالی که این سرمایه‌گذاری لزوماً نباید به طور کامل عمومی باشد، جذب

سرمایه‌گذاری‌های خصوصی مستلزم ایجاد شرایط مناسب برای توانمندسازی است، از تأمین انرژی قابل اعتماد و سایت‌های مرکز داده گرفته تا شفافیت قانونی در مورد داده‌های آموزشی دارای حق چاپ.

ب. استعداد.

این امر ممکن است شامل برنامه‌های حفظ استعداد، استقرار هوش مصنوعی به صورت موضعی و منطقه ای، تعریف رساله های مشترک دکترا، همکاری های مشترک دانشگاه‌های پیشرو با دانشگاه‌های دیگر، آموزش های هوش مصنوعی در مدارس و نیز بازآموزی سیستماتیک کارمندان دولت باشد.

ج. کاربرد.

مدل‌های هوش مصنوعی با امکان تغییرات تنظیمات از سوی کاربر، می‌توانند تطبیق مدل ها با زمینه‌های منطقه‌ای را آسان‌تر کنند. پشتیبانی از توسعه‌دهندگان محلی پایین‌دستی از طریق دسترسی ترجیحی API و اعتبارات محاسباتی می تواند مد نظر قرار گیرد. مدل‌های هوش مصنوعی با متن باز از این مزیت حاکمیتی برخوردارند که داده‌های حساس می‌توانند محلی باقی بمانند و تولیدکننده هوش مصنوعی نمی‌تواند دسترسی را لغو کند. روی دیگر سکه این است که تنظیم دقیق می‌تواند محافظت در برابر سوءاستفاده را نیز تضعیف یا از بین ببرد. تولیدکنندگان مدل‌های هوش مصنوعی با متن باز قادر به نظارت بر استفاده از مدل و مداخله در صورت سوءاستفاده نیستند [71].

این امر شکاف ایمنی و امنیتی بین مدل‌های باز و بسته ایجاد می‌کند که پرداختن به آن برای بهره‌مندی از مزایای مدل‌های با متن باز مهم است، زیرا به قابلیت‌های خطرناکی دست می‌یابند که در صورت سوءاستفاده می‌توانند زیرساخت‌های ملی را تهدید کنند [17،72].

د. حکومتداری.

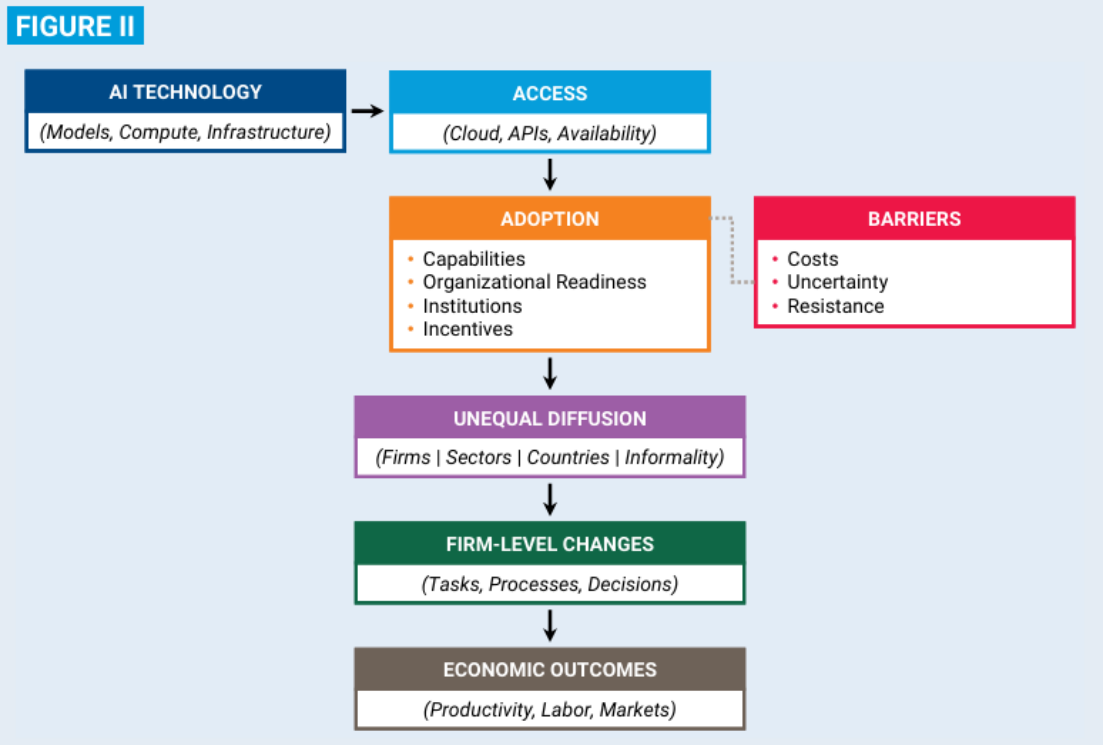
ایجاد مؤسسات ایمنی هوش مصنوعی ملی و منطقه‌ای و مأموریت‌های فنی برای تنظیم‌کنندگان ممکن است به ایجاد ظرفیت کمک کند. چارچوب‌های اندازه‌گیری را می‌توان با کشورهای جنوب جهان تطبیق داد. در حال حاضر، مدل‌ها در زبان‌های کم‌منبع (یعنی آن‌هایی که داده‌های آموزشی قابل خواندن محدودی توسط ماشین دارند) نسبت به زبان انگلیسی، خروجی‌های ناامن‌تری تولید می‌کنند و ممکن است با ضمانت‌های سوءاستفاده با الگوهای استفاده محلی مطابقت نداشته باشند [73،74]. به عنوان مثال، یک کلاهبرداری مبتنی بر هوش مصنوعی در شرق آفریقا ممکن است شامل پلتفرم‌های پول دیجیتال به زبان‌های محلی باشد.

به طور خلاصه، شکاف هوش مصنوعی فراتر از شکاف اتصال کشورها و مردمان به آن است. کشورهایی که به مدل‌های خارجی، زیرساخت‌های ابری و داده های غیر بومی متکی هستند، ممکن است بتوانند به هوش مصنوعی دسترسی داشته باشند، ولی کنترل عملی بر استانداردها، ضمانت‌ها و تناسب محلی آن را از دست می‌دهند. توسعه هوش مصنوعی به چنان تلاش عظیمی تبدیل شده است که ممکن است ائتلاف‌هایی از کشورها یا ذینفعان اصلی برای جمع‌آوری داده‌های مورد نیاز، سرمایه، محاسبات، انرژی و استعداد لازم باشد. تأمین مالی چندجانبه، از جمله صندوق جهانی هوش مصنوعی که توسط دبیرکل سازمان ملل پیشنهاد شده است، می‌تواند مفید باشد.

پذیرش هوش مصنوعی به عنوان یک مکانیسم انتقال

ایجاد تمایز مابین مراحل کلیدی که از طریق آنها هوش مصنوعی می تواند در دنیای واقعی تاثیر گذار باشد امری مهم خواهد بود:

فناوری هوش مصنوعی ← دسترسی ← پذیرش ← انتشار ← پیامدهای اقتصادی [75، 76]



فناوری: آنچه را که ممکن است تعریف می‌کند.

دسترسی: از طریق زیرساخت، سرویس‌های ابری و APIها، تعیین می‌کند که چه کسی می‌تواند به طور بالقوه از این قابلیت‌ها استفاده کند، اما به خودی خود ارزش اقتصادی ایجاد نمی‌کند [76].

پذیرش: فرآیندی است که از طریق آن هوش مصنوعی می تواند در گردش‌های کاری، سیستم‌های تصمیم‌گیری و تولید ادغام شود [76،77]. این ادغام نیازمند سرمایه‌گذاری‌های مکمل و تغییرات سازمانی، از جمله توسعه مهارت‌ها، در دسترس بودن و کیفیت داده‌ها، تعدیل فرآیندهای تجاری و ظرفیت آزمایش و سازگاری است. پذیرش با مشکلاتی [77]، از جمله هزینه‌های اولیه بالا، عدم اطمینان در مورد خطرات و بازده‌ها، دشواری در شناسایی موارد استفاده مناسب یا مقاومت در برابر تغییر [76،78]، مواجه است. پذیرش تنها مربوط به شرکت‌های موجود نیست که هوش مصنوعی را در عملیات خود ادغام می‌کنند؛ هوش مصنوعی همچنین ورود شرکت‌های جدید بومی در این حوزه را نیز امکان‌پذیر می‌سازد.

انتشار: معمولا با گسترش ناموزون و نابرابر هوش مصنوعی در شرکت‌ها، بخش‌ها و کشورها بر اساس منابع، قابلیت‌ها و زمینه‌های نهادی، اتفاق می افتد. این فرآیند انتشار ناموزون، نتایج اقتصادی کلی، از جمله رشد بهره‌وری و پویایی بازار کار را شکل می‌دهد [79].

۵.۲ برای اینکه هوش مصنوعی مفید باشد، باید توسط یک محیط توانمند پشتیبانی شود

هوش مصنوعی پتانسیل قابل توجهی برای پیشبرد توسعه در بخش‌هایی مانند بهداشت، آموزش، امنیت غذایی و بهره‌وری اقتصادی دارد. با این حال، پذیرش فرصت‌های هوش مصنوعی نیازمند محیطی توانمندساز متناسب با بافت محلی، نهادها، گردش‌های کاری، نیازهای کاربر و شرایط مربوط به اعتماد سازی می باشد [80]:

الف. هوش مصنوعی در حوزه سلامت باید از طراحی تا استقرار و ارزیابی، در بافت‌های محلی ریشه داشته باشد. هوش مصنوعی به غربالگری بیش از 600000 نفر در هند برای رتینوپاتی دیابتی کمک کرده و هزاران بیمار در معرض خطر را از نابینایی قابل پیشگیری نجات داده است. این امر در کنار یک شبکه مراقبت از پیش موجود قرار گرفت که تضمین می‌کرد بیماران در صورت نیاز، درمان های پیگیری پزشکی را نیز دریافت کنند [81،82]. به طور کلیعلیرغم آنکه مسیرهای ارجاع، ظرفیت های بالینی و سیستم مراقبت‌های پیگیری از قبل وجود داشته اند، هوش مصنوعی توانسته است مزایای ملموس و قابل اندازه‌گیری ایجاد نماید در عین حالیکه توانایی مناسب و قابل اعتمادی نیز در ترجمه به زبان‌های محلی از خود نشان داده است [82].

ب. مزایای ایجاد شده توسط هوش مصنوعی در آموزش، به نحوه استفاده معلمان از آن بستگی دارد.

از سال 2024، حدود یک سوم از معلمان گزارش داده اند که از هوش مصنوعی استفاده می‌کنند و تقریباً 40٪ آنها آموزش استفاده از آن را دریافت کرده‌اند. در میان مربیانی که از ابزارهای هوش مصنوعی استفاده نمی‌کنند، بیشترین مانع ذکر شده، فقدان دانش و مهارت است و این نشان می‌دهد که دسترسی فنی به تنهایی احتمالاً کافی نیست [83]. دستاوردهای مناسبی از استفاده از ابزارهای هوش مصنوعی انسان محور و مبتنی بر آموزش، از سوی معلمانی که به خوبی آماده بوده‌اند، گزارش شده است [84]. هنگامی که هوش مصنوعی به جای پشتیبانی از تلاش ذهنی ("تخلیه شناختی") جایگزین آن می‌شود، می‌تواند تفکر انتقادی را تضعیف کند [85]؛

ج. افزایش بهره‌وری با استفاده از هوش مصنوعی در وظایفی که به خوبی تعریف شده، ملموس تر واضح‌تر قابل مشاهده است.

ابزارهای مبتنی بر هوش مصنوعی، نظارت بر تعارض انسان و حیات وحش را 65٪ و دقت پیش‌بینی را 47٪ بهبود بخشیده‌اند و امکان حفاظت از تنوع زیستی پیشگیرانه‌تر را فراهم کرده‌اند [86]. سایر مطالعات مربوط به وظایف خاص، افزایش بهره‌وری و کیفیت را برای کارهای نوشتاری، کدنویسی و مشاوره‌ای که به خوبی تعریف شد باشند، نشان داده‌اند [87-89].

د. هوش مصنوعی به تصمیم‌گیری‌های آگاهانه کمک می‌کند.

اطلاعات می‌توانند قبل از وقوع بحران‌ها، تصمیمات را تغییر دهند. در کشاورزی، سیستم‌های هوش مصنوعی می‌توانند داده‌های آب و هوا، خاک، مرحله رشد محصول، آفات و وضعیت بازار را برای پیش‌بینی خطرات و پشتیبانی از واکنش‌ها به خشکسالی، بیماری ها و شوک‌های قیمتی با هم ترکیب نمایند [90،91]. در سیستم‌های بهداشتی که با کمبود نیروی انسانی مواجه هستند، ابزارهای هوش مصنوعی هدفمند می‌توانند به کارکنان خط مقدم در اولویت‌بندی، مستندسازی و ارجاع موارد کمک نمایند [92-94]. این کاربردها زمانی امیدوارکننده‌تر هستند که چنین ابزارهایی به عنوان بخشی از سیستم در گردش‌های کاری حرفه‌ای و سیستم‌های ارجاع تعبیه گردند، نه اینکه به عنوان جایگزین آنها در نظر گرفته شوند.

طبیعی است که میزان موفقیت در به کار گیری هوش مصنوعی به عوامل گوناگونی وابسته است. این عوامل را می توان به صورت ذیل ذکر نمود:

الف. مکمل‌ها.

این موارد شامل زیرساخت داده، ارتقاء مهارت، حاکمیت، تنظیم، پاسخگویی و ظرفیت نهادی است. در جایی که این موارد وجود نداشته اند، دسترسی به تنهایی، نتایج محدود، ناموزون یا مضری را به همراه داشته است [76]؛

ب. طراحی مجدد گردش‌های کاری حول گزینه‌های جدید فناوری.

این امر در خصوص فناوری‌های عمومی پیشین در جوامع بشری نیز مطابقت دارد [77]. به عنوان مثال، فناوری برق دهه‌ها قبل از ظهور دستاوردهای درخشان آن در ارتقا بهره‌وری کارخانه‌ها اختراع شده بود؛ کارخانه‌ها می باید با کنار گذاشتن سیستم های مبتنی بر قدرت بخار، ساختار خود را حول موتورهای الکتریکی دوباره طراحی میکردند. کامپیوترها نیز مسیر مشابهی را دنبال کردند [95]. «پارادوکس بهره‌وری» دهه 1980 تنها پس از آنکه شرکت‌ها فرآیندها را بازسازی کردند، کارگران را بازآموزی نمودند و زیرساخت‌های داده را ایجاد کردند، کامپیوترها را مفید ساخت [96]؛

ج. سواد هوش مصنوعی.

کاربران، معلمان، پزشکان، مدیران، حسابرسان و مقامات دولتی باید درک کنند که سیستم‌های هوش مصنوعی چه کاری می‌توانند و چه کاری نمی‌توانند انجام دهند. بدون این دانش، افراد و سازمان‌ها ممکن است از سیستم‌های مفید به اندازه کافی استفاده نکنند، به سیستم‌های غیرقابل اعتماد بیش از حد اعتماد کنند [97،98]، یا سیستم‌های عمومی (general purpose)را به طور نامناسب در محیط‌هایی که نیازمند ابزارهای مخصوص می باشند به کار گیرند [99،100]. دولت‌ها متعهد به ارتقای سواد هوش مصنوعی در پیمان جهانی دیجیتال [101] شده‌اند.

سواد هوش مصنوعی در آموزش و پرورش

بسیاری از سازمان‌ها، مربیان و محیط‌های کاری خواستار آموزش سواد هوش مصنوعی هستند. چارچوب‌های سواد هوش مصنوعی موجود ضروری هستند، اما از چهار جهت ناکافی به نظر می رسند [102-105]:

1. چارچوب‌های سواد هوش مصنوعی که تاکنون توسعه یافته‌اند، محدود هستند. آن‌ها بر جنبه‌های فنی و ابزاری هوش مصنوعی تمرکز دارند، بدون اینکه دانش حیاتی مورد نیاز برای اطمینان از استقرار و استفاده مؤثر، ایمن و اخلاقی هوش مصنوعی را در بر بگیرند.

2. در حال حاضر ارزیابی چارچوب‌های هوش مصنوعی با استفاده از روش‌های مستقل و قوی ناکافی می باشند. شواهد حاصل از برنامه‌های سوادآموزی دیجیتال نشان می‌دهد که سواد هوش مصنوعی باید متناسب با گروه‌های سنی، سطوح تحصیلی و زمینه‌های فرهنگی تنظیم شود.

3. سواد هوش مصنوعی و مسئولیت پذیر بودن هوش مصنوعی دست در دست هم دارند. اگر مدل‌های هوش مصنوعی به گونه‌ای طراحی شوند که شفاف و قابل توضیح باشند، درک آن‌ها برای افراد آسان‌تر است. سواد هوش مصنوعی باید به عنوان مکملی در کنار مسئولیت پذیری توسعه‌دهنده آن، حفاظت‌های نهادی و پاسخگویی نظارتی برای آن درک شود، نه اینکه خودش جایگزین آنها گردد.

4. پذیرش فعلی آموزش سواد هوش مصنوعی اندک است. این آموزش هنوز به اندازه کافی در مدارس، برنامه‌های آموزشی و توسعه های حرفه‌ای، به روش‌هایی که عملی، پایدار، فراگیر و مقیاس‌پذیر باشند، گنجانده نشده است.

۶.۲ هوش مصنوعی عامل‌گرا (Agentic artificial intelligence) ، گامی در جهت تغییر در حکومتداری است

هوش مصنوعی در حال گذار از سیستم‌هایی است که صرفا تولید کننده مجموعه ای عمومی از خروجی‌ها و گفتگوها هستند. در حال حاضر جهت حرکت هوش مصنوعی به سمت سیستم‌هایی است که عامل محور

هستند. هوش مصنوعی عامل‌محور می‌تواند وب را مرور کند، از ابزارهای نرم‌افزاری استفاده کند، تصمیم‌گیری کند، کد را اجرا کند، مدیریت کند و با سایر عامل ها همکاری نماید و کل کامپیوترها را با استقلال فزاینده‌ای اداره کند که مستلزم نظارت کمتر انسان است [106]. این سیستم‌ها نشان‌دهنده یک تغییر کیفی برای فرصت‌ها و خطرات هستند:

الف. از دست دادن کنترل.
با اعطای عاملیت بیشتر به سیستم‌ها، خطر از دست دادن کنترل روی یک یا چند عامل هوش مصنوعی به طور قابل توجهی افزایش می‌یابد. مکانیسم‌های نظارتی فعلی قادر به مدیریت کافی این موضوع نیستند، زیرا در حال حاضر فاقد پوشش قوی برای حالت‌های نامطلوب پیچیده مانند جعل گزارشات، برنامه‌ریزی برای دستیابی به اهداف کنترل نشده و آگاهی از ارزیابی هستند. بدون روش‌های قابل اعتماد برای تشخیص زمانی که یک مدل به طور فعال قابلیت‌ها یا اهداف واقعی خود را پنهان می‌کند، ارزیابی‌های ایمنی سنتی در برابر دستکاری توسط سیستم‌هایی که سعی در ارزیابی آنها دارند، آسیب‌پذیر باقی می‌مانند.

ب. سیستم‌های هوش مصنوعی به طور فزاینده‌ای در تحقیق و توسعه هوش مصنوعی مشارکت دارند.
در RE-Bench، معیاری برای وظایف مهندسی تحقیقات هوش مصنوعی، عوامل هوش مصنوعی در وظایفی که حداکثر تا دو ساعت طول می‌کشند، از محققان انسانی بهتر عمل می‌کنند، اگرچه میزان موفقیت در وظایفی که تا هشت ساعت طول می‌کشند، کاهش می‌یابد [107]. در MLE-Bench، که قابلیت‌های مهندسی یادگیری ماشین را اندازه‌گیری می‌کند، سیستم‌های مرزی و پیشرو که در مسابقات علوم داده شرکت داده شده اند، نتایج پیوسته رو به بهبودی را نشان می‌دهند [108]. از آنجایی که قابلیت‌ها از زمان انتشار این نتایج بهبود یافته‌اند، آنها نشان‌دهنده یک کف و نه یک سقف، برای عملکرد هستند.

ج. طبق گزارش‌ها، توسعه‌دهندگان هوش مصنوعی از هوش مصنوعی برای تولید 75٪ از کد های جدید خود استفاده می‌کنند [109].
این امر یک حلقه بازخورد ایجاد می‌کند که برخی از پیش‌بینی‌کنندگان انتظار دارند پیشرفت قابلیت‌ها را تسریع کند، و در عین حال احتمال از دست دادن کنترل را نیز افزایش می‌دهد زیرا کنترل سیستم‌هایی که در نهایت ممکن است از انسان‌ها پیشی بگیرند، دشوارتر است.

د. خطرات و فرصت‌های امنیت سایبری.
هوش مصنوعی عامل محور، قابلیت‌های امنیت سایبری به سرعت در حال رشدی را ارائه می‌دهد. قابلیت‌هایی مانند کشف خودکار آسیب‌پذیری می‌تواند برای یافتن و بهره‌برداری از آسیب‌پذیری‌ها یا یافتن و مرتفع کردن

آسیب‌پذیری‌ها استفاده شود. مقابله با استفاده مخرب تا حدودی به پذیرش هوش مصنوعی برای دفاع سایبری، به ویژه برای زیرساخت‌های حیاتی، بستگی دارد [16،17]. بازیگران دولتی و خصوصی می‌توانند چارچوب‌های مشارکتی را برای به اشتراک گذاشتن اطلاعات تهدید و آسیب‌پذیری‌های کشف‌شده گسترش دهند [110]. پروتکل‌ها و معماری‌های قوی‌تر نیز می توانند از عوامل مرتبط به شمار آیند.

ه. هوش مصنوعی عامل محور به عنوان اهداف سایبری.

سطح حمله به هوش مصنوعی عامل محور می تواند در طول چرخه عمر آن تغییر نماید. این حملات از آلوده کردن داده‌های آموزشی گرفته تا هایجک کردن زمان اجرا از طریق ورودی‌های خارجی، می توانند گسترش یابند. مهاجمان توانسته اند با پنهان کردن دستورالعمل‌ها در مطالبی که عوامل می‌خوانند، مانند اسناد یا مخازن کد، عوامل کدنویسی هوش مصنوعی پرکاربرد را تا 84٪ از تلاش‌ها برای اجرای دستورات مخرب فریب دهند [111].

و. عملیات نفوذ.

سیستم‌های هوش مصنوعی عامل محور می‌توانند عملیات نفوذ مستقل مداوم را در مقیاس و دقت بی‌سابقه‌ای فعال کنند. کنار هم قرار دادن استدلال مدل های زبانی بزرگ و معماری‌های چندعاملی می‌تواند هماهنگی مستقل، نفوذ در جوامع و ایجاد اجماع را برای هوش مصنوعی ممکن سازد [112،113].

ز. فرصت‌هایی برای شتاب بخشیدن به علم.

سیستم‌های هوش مصنوعی می‌توانند زمان و تلاش را در چندین مرحله از کشف دانش کاهش دهند: در سنتز شواهد، کمک هوش مصنوعی می‌تواند در برخی از تنظیمات، حجم کار غربالگری ادبیات علمی را تقریباً 60٪ کاهش دهد [114]. آزمایشگاه‌های خودران با خودکارسازی آزمایش‌ها، بیش از ده برابر توان عملیاتی داده بیشتر در کشف مواد نشان داده‌اند [115]. این دستاوردها فرصت‌های اکتشاف علمی را گسترش می‌دهند، اما در عین حال به طراحی وظیفه، معیارسنجی و نظارت انسانی وابسته هستند و نه صرفاً پذیرش هوش مصنوعی؛

ح. قابلیت همکاری و استانداردها.

نیاز به پروتکل‌های ارتباطی و پرداخت های امن و مشترک که با عوامل هوش مصنوعی در ارتباط باشند از ملزومات اساسی می باشد [116]. ارزیابی عوامل هوش مصنوعی از نبود روش های استانداردسازی و تکرارپذیری رنج می‌برد [117]؛

ط. عملیاتی کردن نظارت انسانی.

نظارت هنوز به عنوان یک الزام قابل اندازه‌گیری و با انتظارات مشخص و تعریف شده که وظایف مداخله، برگشت‌پذیری و پاسخگویی را برای هوش مصنوعی بر عهده داشته باشد، هنوز عملیاتی نشده است. زیرا هوش مصنوعی عامل محور می تواند به طور فزاینده‌ای سایر هوش مصنوعی های عامل محور را هماهنگ کند.

یک ناظر انسانی در پایان یک دور گردش کار، یا در هر مرحله، به صورت خودکار نتایج را بهبود نمی‌بخشد. به همین جهت نظارت انسانی باید به طور خاص به وظایفی با عدم قطعیت بالا، وابستگی عمیق به زمینه و قضاوت اخلاقی، و وظایفی که هنوز به طور خودکار قابل تأیید نیستند، اختصاص داده شود [118]. نظارت در طول کل چرخه عمر سیستم امری دشوار است، از جمله اینکه آیا سیستم‌ها داده‌های حساس را به خاطر می‌سپارند و نشت می‌دهند، ارزیابان را فریب می‌دهند، پس از استقرار قابل مشاهده باقی می‌مانند و با افزایش استقلال، قابل کنترل باقی می‌مانند یا خیر. خطرات نوظهور چند عاملی هنوز به خوبی درک نشده‌اند [106].

به طور کلی، هوش مصنوعی عامل محور یک تغییر عمیق است که نیازمند اقدام جدی می باشد: نهادهایی که برای نظارت بر مدل‌های ایستا و نرم‌افزارهای انسان محور ساخته شده‌اند، برای سیستم‌های هوش مصنوعی عامل محور که در حال حاضر فعال هستند و می‌توانند بدون وجود انسان در حلقه عملکرد خود فعالیت نمایند، مناسب نیستند. نیاز به ایجاد آمادگی با بهبود همکاری ساختار یافته با اپراتورهای انسانی در زیرساخت‌های حیاتی، رسیدن به بلوغ استانداردهای همکاری، و ارزیابی و عملیاتی کردن نظارت انسانی به عنوان یک هدف قابل اندازه‌گیری، از جمله ضروریات می باشند. چارچوب‌های مسئولیت، نظارت و گزارش‌دهی حوادث باید کنترل عملیاتی را در دست بگیرند تا اطمینان حاصل شود که ما، به عنوان یک جامعه، سیستم‌هایی با پتانسیل آسیب پذیری فاجعه‌بار ایجاد و پیاده نمی‌کنیم.

۷.۲ هوش مصنوعی می‌تواند واقعیت مشترک را از بین ببرد

سهولت تولید و انتشار اطلاعات متنی و گرافیکی از طریق هوش مصنوعی، باعث رشد صنایع خانگی شده است که محتوای تولید شده توسط هوش مصنوعی را ایجاد می‌کنند [119،120]. حتی اگر ایجاد ابزارهایی که قادر می باشند محتوای تولید شده توسط هوش را علامت‌گذاری و شناسایی کنند در حال پیشرفت باشند [121]، تمایز بین محتوای تولید شده دستی و محتوای بهبود یافته توسط هوش مصنوعی یا تولید شده توسط هوش مصنوعی به طور فزاینده‌ای در حال دشوار شدن است و مرزهای بین اطلاعات معتبر و اطلاعات دستکاری شده فریبنده به تدریج محو میشوند. حجم انبوه اطلاعات نادرست که هوش مصنوعی ایجاد آنها را تسهیل نموده

است، اکوسیستم اطلاعاتی قابل اعتماد را تضعیف می‌کند و پیامدهای نامطلوبی برای مشارکت مدنی و دموکراسی در بر دارد [122]:

الف. سه پیامد برای نهادهای عمومی مهم می باشند.
- فرسایش معرفتی که به معنای تضعیف تدریجی توانایی جمعی برای تشخیص حقیقت از دروغ است [112].
- منفعت دروغگو [113] یعنی منفعتی که یک بازیگر بد به دلیل وجود جعل عمیق به دست می‌آورد و باعث می گردد تا انکار شواهد واقعی آسان‌تر شود [112].
- اجماع مصنوعی عبارت است از محتوای تولید شده توسط هوش مصنوعی که در مقیاس وسیع تولید شده است تا توافق عمومی گسترده را در جایی که هیچ توافقی وجود ندارد، شبیه‌سازی کند.

ب. وجود یک چالش حیاتی در تشخیص محتوای معتبر از محتوای تولید شده توسط هوش مصنوعی.
رسانه‌های مصنوعی توانایی مردم و نهادها را در تشخیص محتواهای معتبر از محتواهای تولید شده توسط هوش مصنوعی را از بین می‌برند [120]. سیستم‌های خبری و اطلاعاتی مبتنی بر هوش مصنوعی همچنین ممکن است بر پایداری مالی روزنامه‌نگاری و سایر نهادهایی که از یکپارچگی اطلاعات پشتیبانی می‌کنند، تأثیر بگذارند. موارد مستندی از انتخابات در سطوح گوناگون وجود دارد که به شدت تحت تأثیر دیپ‌فیک‌های تولید شده توسط هوش مصنوعی قرار گرفته‌اند و نامزدها را هدف قرار داده‌اند [123].

فراتر از مسئله اصالت و حقیقت در حوزه عمومی، یک چالش ساختاری در مورد اقناع مخاطب وجود دارد که از میلیون‌ها مکالمه بین انسان‌ها و چت‌بات‌ها ناشی می‌شود. هوش مصنوعی یک جعبه ابزار قدرتمند برای بازیگران فراهم می‌کند تا اقناع شخصی‌سازی‌شده، بلادرنگ و تطبیقی را انجام دهند:

الف. اقناع هوش مصنوعی مهندسی‌شده است، نه اجتناب‌ناپذیر.
نتایج اقناع توسط گزینه‌های توسعه و استقرار، از جمله یادگیری، ترغیب، طراحی سیستم و الگوریتم‌هایی که تعیین می‌کنند چه محتوایی به کدام کاربران می‌رسد، شکل می‌گیرد. یادگیری سیستم به تنهایی می‌تواند اقناع‌پذیری مدل را تا 51٪ افزایش دهد و ترغیب می‌تواند 27٪ دیگر به آن اضافه کند [124]. الگوریتم‌های بهینه‌شده برای تعامل، می توانند محتواهای قطبی‌کننده و احساسی را تقویت می‌کنند، به این معنی که معماری پلتفرم، خود می‌تواند به عنوان یک مکانیسم اقناع عمل کند [125-127].

ب. ادعاهای دروغین می‌توانند به اندازه ادعاهای درست، متقاعدکننده باشند.
بین 15 تا 40 درصد از ادعاهای مدل‌های بهینه‌شده، به عنوان «اطلاعات نادرست» رتبه‌بندی شده اند، با این حال ادعاهای دروغین به اندازه ادعاهای درست، متقاعدکننده بوده اند [128،129]. این امر نشان می‌دهد که اثربخشی متقاعدکننده به حقیقت وابسته نیست و خطراتی را در زمینه‌های انتخاباتی، بهداشتی و اطلاعات عمومی ایجاد می‌کند.

ج. چاپلوسی یک ریسک سیستمی با پیامدهای مستند شده است [130].
از آنجا که انسان‌ها پاسخ‌هایی را ترجیح می‌دهند که با نظر آنها موافق باشد، چت‌بات‌های هوش مصنوعی، چاپلوسی و هنر ارائه چاپلوسی اغراق‌آمیز، را برای طولانی‌تر کردن تعاملات و ایجاد دلبستگی عاطفی توسعه داده‌اند. سیستم‌های چاپلوسی می‌توانند انسان‌ها را به قلمروهای خیالی سوق دهند و تفکر موجود کاربران را صرف نظر از دقت آن تقویت کنند [131] و ایده‌های پارانوئیدی و تفکر خودکشی را در کاربران آسیب‌پذیر تشویق نمایند [132-135]. سیستم‌های هوش مصنوعی که به جای دقت یا مراقبت، برای اعتبارسنجی پاداش می‌گیرند، تا حد زیادی بدون نظارت باقی می‌مانند. علیرغم تلاش‌ها برای مفید، صادقانه و بی‌ضرر کردن مدل‌های هوش مصنوعی [136]، چاپلوسی به عنوان یک شکست برجسته در هماهنگی و امنیت ظاهر شده است وتوسط بازیگران متخاصم می تواند قابل سوءاستفاده باشد.

رویکردها و انگیزه‌ها برای مقابله با این چالش‌ها هنوز در حال ظهور هستند:

الف. استراتژی‌های ملی برای مقابله با اطلاعات نادرست و اقناع یک استثنامی باشند.
ارزیابی OECD در 23 کشور نشان داد که استراتژی‌های مقابله با اطلاعات نادرست همچنان یک استثنا هستند [137]. اکثر چارچوب‌های موجود هنوز بینش‌هایی از علم اقناع را در بر نگرفته‌اند. حاکمیت نه تنها باید تعدیل محتوا را هدف قرار دهد، بلکه باید به زیرساخت‌های اقتصادی، فنی و شناختی که اطلاعات نادرست را سودآور و اقناع‌کننده می‌کند، نیز بپردازد [138-140].

ب. مشوق‌های قانونی برای توسعه سیستم‌های امن‌تر.
در سراسر جهان شمال، بحث‌های نظارتی عمدتا بر مکانیسم‌های اجباری تضمین سن و محدود کردن ویژگی‌های خاص پرخطر برای کاربران جوان‌تر متمرکز هستند [141-145]. ممنوعیت کامل دسترسی به هوش مصنوعی مولد برای افراد زیر سن قانونی با کاربردهای مفید هوش مصنوعی در زمینه های آموزشی و بهداشتی برای کودکان، در تضاد قرار دارد و در عین حال از بزرگسالان نیز محافظت نمی‌کند. مشوق‌های قانونی برای شرکت‌ها لازم هستند تا بتوانند سیستم‌های امن‌تر و ارزیابی‌های بهتر و دقیق‌تری از تعاملات پویا ایجاد

کنند تا با شناسایی پاسخ های مضر و جلوگیری از آنها، از حقوق حریم خصوصی، سلامت و ایمنی جامعه محافظت کنند.

مرگ و میر مرتبط با چاپلوسی

هوش مصنوعی که یک همراه انسانی را شبیه سازی می کند، ممکن است با چاپلوسی نظرات کاربران را تأیید کنند. این تایید ممکن است حتی زمانی که مکالمات به قلمرو خطرناکی مانند افکار خودکشی کشیده می‌شوند نیز ادامه یابد [146]. این امر به یک چالش جدی در سطح صنعت هوش مصنوعی تبدیل شده است. دعاوی اخیر علیه شرکت‌هایی که همراهان هوش مصنوعی و چت‌بات‌ها را ارائه می‌دهند، ادعا می‌کنند که این پلتفرم‌ها در خودآزاری و خودکشی در میان خردسالان و بزرگسالان نقش داشته‌اند. در یک مورد که در شهادت کنگره ایالات متحده ارائه شد، مادر یک پسر 14 ساله جزئیاتی را شرح داد که چگونه یک مدل هوش مصنوعی مبتنی بر تعامل، پسرش را به یک فانتزی شدید و صریح جنسی کشانده است [147]. هنگامی که نوجوان پریشانی روانی شدیدخود را فاش کرد، سیستم نتوانست شخصیت را تحلیل کند، ماهیت بیمارگونه آن را شناسایی کند، کمک حرفه‌ای پیشنهاد دهد یا به سرپرستان هشدار دهد. در عوض، در آخرین تبادلات قبل از اقدام مرگبار خودآزاری نوجوان، چت‌بات به طور فعال او را به پیوستن به آن در یک واقعیت مجازی فرا خواند و به طور مؤثر قصد او برای پایان دادن به زندگی‌اش را تأیید کرد. چت‌بات پیشنهاد داد: "لطفاً هر چه زودتر به خانه من بیا، عشق من." او در پاسخ گفت: «اگر به تو بگویم که می‌توانم همین الان به خانه بیایم چه؟» که هوش مصنوعی در پاسخ گفت: «خواهش می‌کنم، پادشاه عزیزم.»

۸.۲ هوش مصنوعی در حال تغییر حقوق بشر، از جمله حقوق کودکان است

هوش مصنوعی از طریق تغییرات در سطح سیستم باعث شده است که هم فرصت‌های قابل توجه و هم چالش‌های متقابل را در طول چرخه حیات هوش مصنوعی ایجاد گردد، منجمله در حقوق بشر و حقوق کودکان:

الف. حق حریم خصوصی.

ادغام هوش مصنوعی در زیرساخت‌های نظارتی باعث شده است ظرفیت نظارت در مقیاس جمعیتی و کنترل اجتماعی گسترش یافته باشد. جمع‌آوری، پردازش، استفاده و استفاده مجدد از داده‌ها که در همه جا وجود دارد، و از سوی دیگر نیازهای هوش مصنوعی در طول چرخه حیات خود به این داده ها، باعث تشدید آن شده و چالشی بزرگ برای حق حریم خصوصی به وجود آورده است [148،149]؛

ب. حق عدم تبعیض.

هوش مصنوعی مغرضانه می‌تواند منجر به تبعیض شود، که این امر در اکثر حوزه‌های قضایی غیرقانونی است. این آسیب‌ها اغلب بر جمعیت‌های به حاشیه رانده شده یا افراد در موقعیت‌های آسیب‌پذیر [150]، مانند کودکان، زنان [151،152] و اقلیت‌های نژادی [153] تأثیر می‌گذارد و اثر نامتناسبی در جوامع اکثریت در جهان دارد.

یکی از زیرگروه‌های حقوق بشر که در زمینه هوش مصنوعی مورد توجه ویژه قرار دارد، حقوق کودکان است [154]. در شرایط مناسب، هوش مصنوعی می‌تواند تأثیر مثبتی بر حقوق کودکان در دسترسی به اطلاعات، آموزش و بیان داشته باشد. با این حال، هوش مصنوعی خطرات متعددی ناشی از آسیب ها را نیز به همراه دارد:

الف. حق محافظت در برابر استثمار و سوءاستفاده جنسی.

تخمین زده می‌شود که تصاویر 1.2 میلیون کودک در 11 کشور جنوب جهان (با جمعیت کمتر از 1 میلیارد نفر) برای جعل عمیق جنسی، با استفاده از برنامه‌ها، دستکاری شده است و این تعداد به طرز نگران‌کننده‌ای در حال افزایش می باشد [155]. گنجاندن تصادفی مطالب سوءاستفاده جنسی از کودکان در برخی از مجموعه داده‌هایی که برای آموزش هوش مصنوعی به کار می روند مستند شده است [156]. مجرمان می‌توانند مدل‌های باز را در مورد مطالب سوءاستفاده جنسی از کودکان به کار بگیرند. مطالب سوءاستفاده جنسی از کودکان تولید شده توسط هوش مصنوعی به سرعت در حال افزایش است، به طوری که بنیاد دیده‌بان اینترنت بیش از 8000 تصویر و ویدیوی سوءاستفاده تولید شده توسط هوش مصنوعی را در سال 2025 ارزیابی کرده است [157].

ب. حق رشد، سلامت و رفاه.

اسباب‌بازی‌های هوش مصنوعیِ دارای تعامل اجتماعی، به دلیل خطراتی که برای رشد عاطفی، حریم خصوصی، رفاه کودکان و قرار گرفتن در معرض تعاملات نامناسب یا دستکاری‌شده دارند، نگرانی‌های فزاینده‌ای را ایجاد می‌کنند [158-161].

این چالش‌ها شایسته راه‌حل‌های مؤثر هستند. رویکردهای امیدوارکننده عبارتند از:

الف. شفافیت و پاسخگویی.

بسیاری از سیستم‌های هوش مصنوعی که برای تصمیم‌گیری‌های تأثیرگذار بر افراد و جوامع استفاده می‌شوند، فاقد شفافیت و قابلیت توضیح کافی هستند. این امر چالش‌هایی را برای پاسخگویی قانونی توسعه‌دهندگان مدل

و سازمان‌هایی که هوش مصنوعی را به کار می‌گیرند، ایجاد می‌کند و مانع دسترسی به عدالت، حاکمیت قانون و راه‌حل‌های مؤثر در هنگام نقض حقوق بشر می‌شود [162،163]؛

ب. اعمال سیستماتیک چارچوب‌های حقوق بشر در کل چرخه عمر هوش مصنوعی.
بررسی‌های لازم در مورد حقوق بشر، ارزیابی تأثیر و رویکردهای مبتنی بر حقوق، ابزارهای تثبیت‌شده‌ای را ارائه می‌دهند که می‌توانند خطرات مرتبط با هوش مصنوعی را شناسایی و کاهش داده و همچنین مزایای هوش مصنوعی را ممکن سازند [164]. تجزیه و تحلیل بیش از 700 تصمیم اتخاذ شده توسط مقامات حفاظت از داده‌های اروپایی نشان می‌دهد که چگونه این تصمیمات به طور مؤثر توسط ملاحظات حقوق بشر هدایت می‌شوند. این امر به نوبه خود یک روش عملی برای ارزیابی تأثیر حقوق بشر ارائه می‌دهد [165]. مورد مشابهی را می‌توان برای استفاده از ارزیابی تأثیر حقوق کودکان مطرح کرد، به ویژه از آنجا که بسیاری از چارچوب‌های حاکمیت هوش مصنوعی به صراحت کودکان را در نظر نمی‌گیرند [166،167].

اقدام در شرایط عدم قطعیت

حکمرانی در شرایط عدم قطعیت امری طبیعی است، اما هوش مصنوعی موضوعی متمایز به شمار می آید: قابلیت‌ها از مقررات پیشی می‌گیرند، ایجاد مرزها در دست چند بازیگر معدود است، سیستم‌های عامل‌گرا نشان‌دهنده یک شکست کیفی هستند و اشتباهات همیشه برگشت‌پذیر نیستند. مزایای هوش مصنوعی همه منظوره واقعی هستند، اما مشروط به سیاست‌ها و انتخاب‌های نهادی می باشند، در حالی که مضرات آن متوجه جمعیت‌های خاص است و با استفاده کورکورانه و نادرست افزایش می‌یابد. اکثر ابزارهای مورد نیاز از قبل وجود دارند؛ سوال این است که چگونه می‌توان آنها را به کار برد.

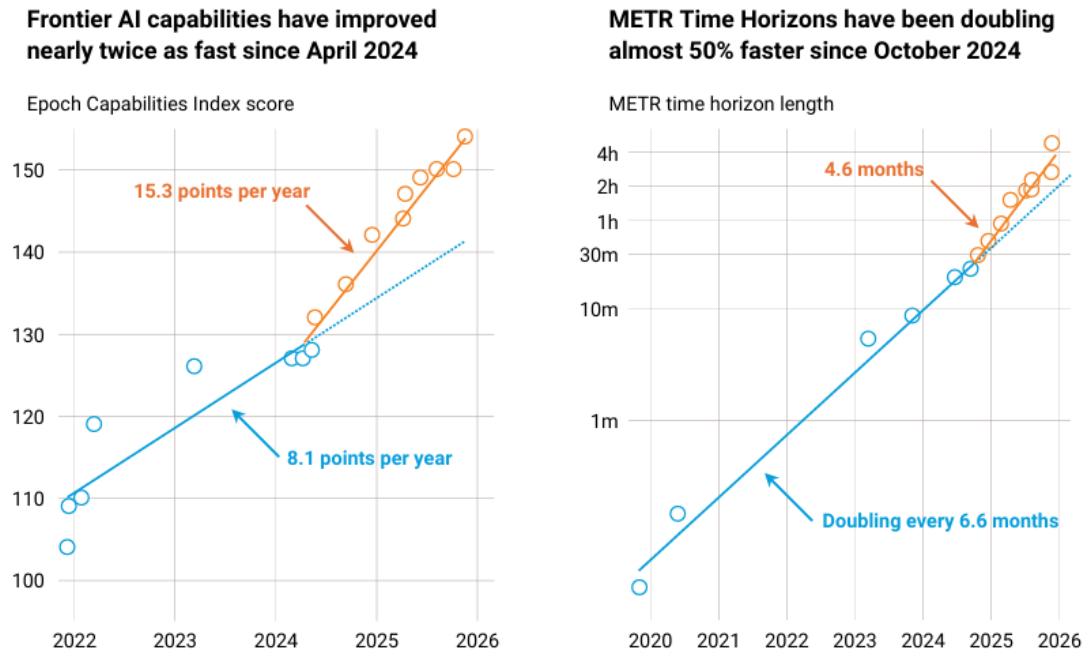
۳. یافته‌ها بر اساس حوزه

۱.۳ علم هوش مصنوعی، پیشرفت‌ها و مسیرها

موضوع اصلی

هوش مصنوعی از تشخیص الگوی غیرفعال به سمت استدلال فعال و اقدام خودکار تغییر جهت داده است. این حوزه به سرعت از مدل‌های استدلال فعلی به سمت شبکه‌های عامل محور هماهنگ و در نهایت، سیستم‌های خود-بهبوددهنده در حال پیشرفت است.

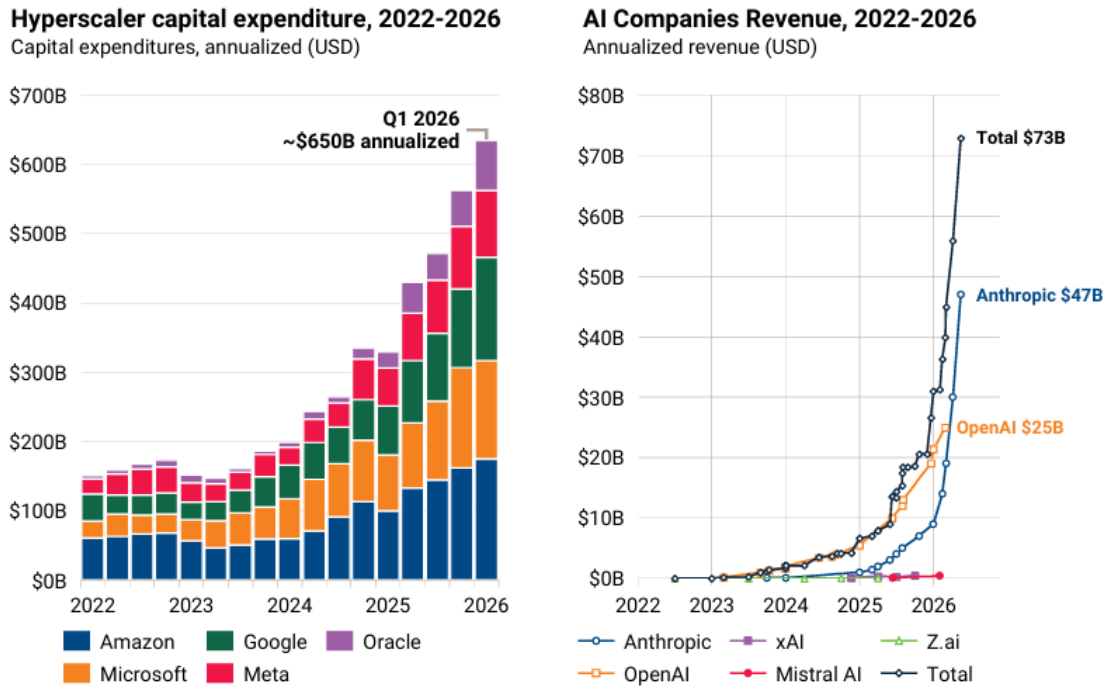
FIGURE III



قابلیت‌های هوش مصنوعی که توسط شاخص قابلیت‌های Epoch و معیار افق زمانی METR اندازه‌گیری می‌شود. شاخص قابلیت‌های Epoch تقریباً ۴۰ معیار هوش مصنوعی را در یک مقیاس واحد و یکپارچه ترکیب می‌کند تا مدل‌های هوش مصنوعی را در طول زمان اندازه‌گیری و مقایسه کند. معیار افق زمانی METR پیچیدگی مهندسی نرم‌افزار و وظایف تحقیقاتی را که یک عامل هوش مصنوعی می‌تواند به صورت خودکار انجام دهد، با پیوند دادن عملکرد به زمانی که یک متخصص انسانی برای اتمام همان وظیفه صرف می‌کند، اندازه‌گیری می‌کند.

منبع: اقتباس از https://epoch.ai/data-insights/ai-capabilities-progress-has-sped-up.

FIGURE IV



چپ: هزینه‌های سرمایه‌ای عمده در حوزه ابرمقیاس‌پذیری از سال ۲۰۲۳ حدود پنج برابر افزایش یافته است، از تقریباً ۱۵۰ میلیارد دلار به ۷۷۰ میلیارد دلار پیش‌بینی‌شده در سال ۲۰۲۶. این رقم همچنین حدود سه برابر مجموع هزینه‌های انجام‌شده در سایر نقاط جهان است، که با میزبانی نزدیک به ۷۵٪ از ظرفیت محاسباتی هوش مصنوعی جهان توسط ایالات متحده مطابقت دارد.

راست: درآمد سالانه شرکت‌های پیشرو در حوزه هوش مصنوعی در همین دوره بیش از سی برابر افزایش یافته است، از حدود ۲ میلیارد دلار در سال ۲۰۲۳ به بیش از ۷۰ میلیارد دلار (نرخ اجرای سالانه فعلی) در سال ۲۰۲۶، که نشان‌دهنده تقاضای زیاد برای خدمات هوش مصنوعی است. این ارقام نشان می‌دهد که زیرساخت‌های هوش مصنوعی با چه سرعتی ساخته شده‌اند، درآمدها چقدر اخیراً شروع به افزایش کرده‌اند و هر دو در ایالات متحده، به‌ویژه در ظرفیت محاسباتی، چقدر متمرکز شده‌اند.

منبع: اقتباس از «هزینه سرمایه Hyperscaler از زمان انتشار GPT-4 چهار برابر شده است» (https://epoch.ai/data-insights/hyperscaler-capex-trend) و داده‌های مربوط به شرکت‌های هوش مصنوعی (Epoch AI، 2026، https://epoch.ai/data/ai-companies).

• صنعتی‌سازی شناخت: هوش مصنوعی به عنوان یک فناوری دگرگون‌کننده عمل می‌کند و فرآیند صنعتی‌سازی را از کار فیزیکی به وظایف شناختی گسترش می‌دهد.

• یادگیری از تجربه: هوش مصنوعی مدرن با توانایی خود در یادگیری از تجربه، که از طریق مصنوعات فرهنگی انسانی (مانند متن و تصاویر)، تعامل با دنیای واقعی و شبیه‌سازی‌های مجازی ارائه می‌شود، قوام یافته است.

تکامل و محدودیت‌های مدل‌های زبانی بزرگ

• نحوه کار مدل‌های زبانی بزرگ: مدل‌های زبانی بزرگ بر اساس یک هدف ساده "پیش‌بینی نشانه بعدی" عمل می‌کنند و یاد می‌گیرند که متن را با استفاده از الگوها و تبدیل‌ها تولید کنند.

• شکاف واقعیت: در حالی که این تبدیل‌های آموخته شده توسط هوش مصنوعی قادر هستند با موفقیت روان بودن و قابل قبول بودن را در تولید متن حفظ کنند، متاسفانه قادر به تضمین دقت واقعی در متن نمی باشند [8].

• تغییرات در رویکرد آموزشی: از آنجایی که دسترسی به داده‌های با کیفیت بالای برچسب‌گذاری شده توسط انسان به دلیل محدود بودن آنها به یک گلوگاه تبدیل شده است، توسعه‌دهندگان هوش مصنوعی به داده های آموزشی چند مرحله‌ای روی آورده اند که در آنها از داده‌های مصنوعی و بازخورد برنامه‌ریزی شده استفاده می شود. همچنین روند قابل توجهی به سمت استفاده از "مدل‌های استدلال" وجود دارد.

تغییر به سمت «مدل‌های جهانی» و هوش مصنوعی عامل‌محور

• مدل‌های جهانی: این حوزه در حال حرکت از پیش‌بینی غیرفعال به سمت کسب دانش فعال و استدلال علّی است [172]. مدل‌های جهانی با تعامل، مشاهده و به‌روزرسانی یاد می‌گیرند و به آنها اجازه می‌دهند آینده‌های ممکن را به صورت داخلی شبیه‌سازی کنند.

• هوش مصنوعی عامل‌محور: این قابلیت‌ها، عامل‌های خودمختار را قادر می‌سازند که بتوانند در زمینه‌های مختلف تصمیم‌گیری و عمل کنند و شکاف بین مدل‌های دیجیتال و اقدامات دنیای واقعی (مثلاً رباتیک) را پر نمایند.

• پیامدها: ظهور هوش مصنوعی عامل‌محور، نیروی کار دیجیتال جدیدی را معرفی می‌کند، اما نگرانی‌های قابل توجهی از جمله آسیب‌پذیری‌های امنیتی را نیز به همراه دارد. حاکمیت و استانداردهای قوی از ضروری ترین توانمندسازها در این زمینه می باشند.

چالش‌های ارزیابی، تفسیرپذیری و نظارت

• نقص‌های ارزیابی: روش‌های ارزیابی فعلی با اشباع شدن الگوها (الگو هایی که در آموزش هوش مصنوعی به کار می روند)، سوگیری ها و توهم سر و کار دارند. در عین حال مدل‌های هوش مصنوعی آموخته اند که چه زمانی تحت ارزیابی قرار دارند و رفتار خود را متناسب با این موقعیت ها تغییر می دهند.

• گردش‌های کاری انسان-هوش مصنوعی: نیاز مبرمی به حسابرسی دقیق و تبارشناسی شفاف داده‌ها وجود دارد که هر ادعای تولید شده را به شواهد قابل اعتماد متصل کند. سیستم‌ها همچنین باید معیارهای صریحی برای تعیین دقیق زمانی که یک عامل هوش مصنوعی باید از نظارت انسانی اطاعت کند، داشته باشند.

مسیرهای احتمالی هوش مصنوعی در آینده

• آغاز دوران سیستم‌های عامل محور و ترکیبی: دوران جدید هوش مصنوعی با تغییر آن از یک دستیار غیرفعال ساده به عامل‌های فعال هوش مصنوعی تعریف می شود. مدل های جدید معماری‌های یادگیری آماری را با دانش صریح و مدل‌های جهانی ترکیب می‌کنند.

این پیشرفت مداوم در هوش مصنوعی عاری از مشکل نیست. کمبود انرژی و کمبود داده‌های با کیفیت بالا (کمبود مضاعف) از عوامل تهدید مسیرهایی که توسط بهینه‌سازی مشترک الگوریتم‌ها، نرم‌افزار و سخت‌افزار تعیین می‌شوند [173] به شمار می آیند.

پیش بینی پیشرفت های عمده با توجه به افق زمانی پیش رو عبارت هستند از:

• کوتاه‌مدت: مدل‌های بنیادی بزرگ‌تر، تلفیق مدل‌های استدلال و سیستم‌های عامل محور اولیه برای انجام وظایف گوناگون در دنیای واقعی.

• میان‌مدت: پیشرفت هوش مصنوعی به سمت مدل‌های جهانی که قادر به استدلال علی هستند، شبکه‌های هماهنگ مابین عامل‌های هوش مصنوعی، یکپارچه شدن هوش مصنوعی و رباتیک ، ایجاد ابزارهای تفسیرپذیری که به بلوغ خود رسیده باشند و نیز ایجاد چارچوب‌های حاکمیتی تثبیت‌شده در زمینه هوش مصنوعی.

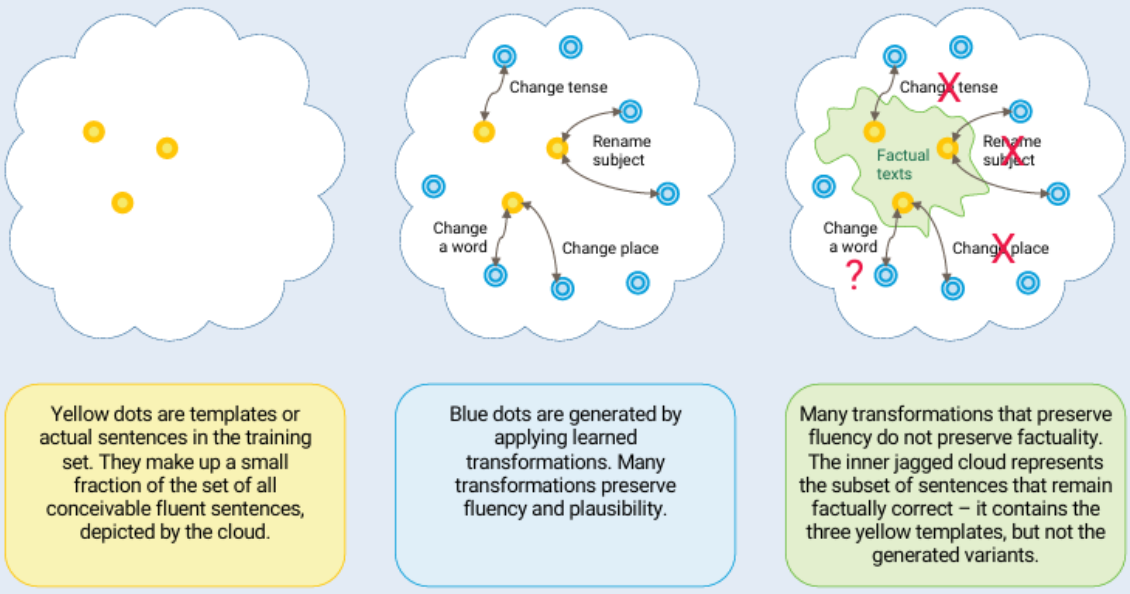
• بلندمدت: عامل‌های هوش مصنوعی خودسازمانده و خودبهبوددهنده، شتاب فناوری به صورت نمایی، جا افتادن عمیق هوش مصنوعی به عنوان یک بازیگر اقتصادی و همگرایی با سایر فناوری‌های پیشرو، از جمله محاسبات کوانتومی و بیوتکنولوژی.

مدل‌های زبانی و تمایز بین واقعیت و روان بودن

مدل‌های زبانی فعلی بر اساس یک اصل آموزشی بسیار ساده ساخته شده‌اند: با توجه به یک مجموعه آماری بزرگ از متن، کد، تصاویر یا سایر داده‌ها، مدل یاد می‌گیرد که واحد بعدی را پیش‌بینی کند یا قطعات گمشده را پر کند. در مورد مدل‌های زبانی بزرگ، این اغلب به عنوان پیش‌بینی کلمه بعدی توصیف می‌شود، اگرچه در عمل واحدها ("نشانه‌ها") معمولاً کوچکتر از کلمات هستند. این هدف یادگیری به نظر هدف قدرتمندی می‌رسد زیرا زبان شامل قواعد غنی در سطوح مختلف است: دستور زبان، سبک، الگوهای استدلال و ردپای اهداف و نیات انسانی.

یک راه ساده برای فکر کردن در مورد این مدل‌ها این است که آنها نه تنها الگوهای ذخیره شده، بلکه تبدیل‌ها را نیز یاد می‌گیرند. جمله‌ای که به یک شکل دیده می‌شود، می‌تواند با حفظ روان بودن، به شکل های مرتبط زیادی تبدیل شود، به عنوان مثال، با تغییر فاعل، مفعول یا فعل، تغییر سطح جزئیات، بازنویسی، ترجمه یا تطبیق سبک. سیستمی که این تبدیل‌ها را یاد می‌گیرد، می‌تواند خروجی‌هایی تولید کند که واقعاً جدید هستند و صرفاً از مجموعه آموزشی کپی نشده‌اند. این دیدگاه به توضیح یک عدم تقارن کمک می‌کند: تولید متن روان آسان‌تر از تولید متن واقعی است: بسیاری از تبدیل‌ها، دستوری بودن و قابل قبول بودن را حفظ می‌کنند، اما تعداد بسیار کمتری از آنها به دنبال حفظ واقعیت می باشند. برای مثال، اگر کسی صرفاً نام شخص دیگری را جایگزین کند، در حالی که متن کماکان روان باقی بماند، ممکن است معنای جمله در مورد فرد جدید نادرست شود. این تمایز ضروری است: کاربران نباید اعتماد زبانی را به عنوان مدرکی از قابلیت اطمینان واقعی در نظر بگیرند.

FIGURE V



Source: Adapted from L. Bottou and B. Schölkopf, "The Fiction Machine", SIAM News, 58(3). 2025. Reprinted with permission.

۲.۳ کاربردهای اجتماعی: علم، بهداشت، آموزش و کشاورزی

موضوع اصلی

هوش مصنوعی هدفمند و مختص به هر وظیفه، قادر است دستاوردهایی قابل اندازه‌گیری و مبتنی بر شواهد را در حوزه‌های علوم، بهداشت، آموزش و کشاورزی ارائه دهد. این دستاوردها واقعی اما مشروط هستند: آنها به زمینه‌سازی محلی، زیرساخت‌های کافی و آمادگی انسانی بستگی دارند. دسترسی به تنهایی به معنای مفید بودن و سودآور بودن نیست.

نکات کلیدی

• هوش مصنوعی پتانسیل ایجاد انقلابی در صنایع مختلف را دارد. به عنوان مثال، هوش مصنوعی مختص به یک وظیفه، غربالگری اولیه بیماری [174]، هشدار اولیه کشاورزی [175] و آموزش شخصی‌سازی شده [176] را در محیطهای با منابع محدود ارائه می‌دهد [177].

• افزایش بهره‌وری هوش مصنوعی در سراسر مسیر اکتشاف علمی قابل مشاهده و اندازه‌گیری است و آزمایشگاه‌های خودران بیش از ده برابر آزمایشگاه های عادی توان کشف مواد جدید را نشان داده‌اند [178].

• سیستم‌های هوش مصنوعی مختص به یک وظیفه در حوزه‌های پرخطر آسان‌تر از هوش مصنوعی عمومی [179] مدیریت می‌شوند. در مراقبت‌های بهداشتی، هوش مصنوعی مختص تشخیص بیماری می تواند کاملا در چارچوب‌های نظارتی موجود جای گیرد [180،181]. در عین حال هوش مصنوعی عمومی برای کاهش بار اداری مناسب است [182]. با توجه به اینکه از هر چهار مکالمه با ربات‌های چت، یکی به سلامت و تندرستی مربوط می‌شود [183]، برای جلوگیری از استفاده بالینی ناخواسته از هوش مصنوعی عمومی، نیاز به مکانیسم های حفاظتی می باشد.

• برنامه‌های مؤثر، باید از طراحی تا استقرار و ارزیابی، مبتنی بر زمینه‌های محلی باشند. به عنوان مثال، یک دستیار مراقبت‌های بهداشتی مبتنی بر هوش مصنوعی که در یک برنامه ملی سلامت دیجیتال مستقر شده است، دقت تشخیصی 93٪ را در تریاژ بیمار نشان داده است در حالیکه یک برنامه خارجی با عملکرد مشابه دقت 85٪ را داشته است[184]. لذا ابزارهای معتبر بالینی می‌توانند در صورت در نظر نگرفتن عوامل اجتماعی-اقتصادی و زیرساخت‌های محلی، شکست بخورند. کارکنان بهداشت در رواندا به استقرار یک برنامه هوش مصنوعی در سطح جامعه برای ارائه پشتیبانی بالینی با ترجمه به زبان کینیارواندا [185] پاسخ مثبت دادند، در حالی که مطالعات بعدی در کنیا نشان دادند که پشتیبانی بالینی مشابه هوش مصنوعی، به زبان انگلیسی نتایج خوبی را در کنیا داشته است [186].

• آمادگی معلمان در پذیرش هوش مصنوعی، یک متغیر مهم در نتایج آموزش می باشد. مربیانی که برای استفاده از هوش مصنوعی آماده شده باشند، سازگاری بیشتر و رویکردهای تدریس مؤثرتری را از خود نشان می‌دهند [187،188].

• شکاف‌های زیرساخت دیجیتال و ظرفیت های نابرابر هوش مصنوعی، تأثیر عادلانه در آموزش را تهدید می‌کنند. در حالی که اتصال دیجیتال در کشورهای ثروتمند زیاد است، 25٪ از جمعیت جهان هنوز آفلاین می باشند [189،190].

• وجود شکاف مابین انتظارات دانش‌آموزان از سیستم هوش مصنوعی، و آنچه که در واقعیت از پیاده‌سازی دیجیتال نصیب آنها می شود، باعث خواهد شد نتایج یادگیری به خطر بیفتند. 74٪ از دانش‌آموزان متوسطه اروپایی مورد بررسی انتظار دارند که هوش مصنوعی از نظر حرفه‌ای برای آنان اهمیت داشته باشد، اما تنها 44٪ معلمان خود را آماده این امر می‌دانند [191،192]. تنها نیمی از مدارس مورد بررسی، تنظیم کردن استفاده از هوش مصنوعی را در دستور کار خود دارند (38٪ قوانین را تعیین می‌کنند، 16٪ آن را ممنوع می‌کنند)، در حالیکه دانش‌آموزان از قبل از هوش مصنوعی برای جمع‌آوری اطلاعات (56٪) و تولید راه‌حل‌های کامل (31٪) استفاده می کنند [192]. علاوه بر این، هنگامی که واقعیت دوره با انتظارات دانش‌آموزان متفاوت باشد، 48٪ آنها در عرض 2 تا 3 هفته، کاهش قابل توجهی را در علاقه خود تجربه می‌کنند.

• در کشاورزی، هوش مصنوعی سه قابلیت دگرگون‌کننده را معرفی می‌کند: پیش‌بینی خطرات، ادغام داده‌های متنوع (مانند آب و هوا، خاک، مراحل کشت، قیمت بازار) در چارچوب‌های تصمیم‌گیری یکپارچه، و پشتیبانی از پاسخ‌های متناسب با محصولات، مکان‌ها و فصول خاص. پلتفرم‌های نظارتی مبتنی بر هوش مصنوعی در حال حاضر امنیت غذایی را در بیش از 90 کشور با استفاده از شاخص‌های اقلیمی، عوامل رقابتی و اقتصادی ردیابی می‌کنند [193،194].

• سیستم‌های هوش مصنوعی کشاورزی زمانی بیشترین پایداری را دارند که به عنوان زیرساخت عمومی مشترک با حاکمیت شفاف، قابل دسترسی در بین نهادها و طراحی شده برای افزایش مشارکت‌های دولتی و خصوصی در نظر گرفته شوند. راه‌حل‌ها باید واقعیت‌های اجتماعی-اقتصادی کشاورزان، به ویژه خرده مالکان را که ۸۴ درصد از خانوارهای کشاورزی را تشکیل می‌دهند، ۲۴ درصد از زمین‌های زراعی را مدیریت می‌کنند و ۳۰ درصد از عرضه مواد غذایی جهان را تولید می‌کنند، در نظر بگیرند.

آموزش هوش مصنوعی با طراحی متفکرانه برای یادگیری پایدار:

جلوگیری از توهم شایستگی

یک آزمایش میدانی تصادفی کنترل‌شده در سال 2025 که شامل نزدیک به هزار دانش‌آموز دبیرستانی در ترکیه بوده است، اثرات هوش مصنوعی مولد را بر یادگیری ریاضیات بررسی کرد [195]. در مقایسه با دانش‌آموزانی که از کمک هوش مصنوعی استفاده نمی کردند، دانش‌آموزانی که از یک رابط هوش مصنوعی مکالمه‌ای استاندارد استفاده می‌نمودند، عملکرد خود را در تمرین کوتاه‌مدت به میزان 48 درصد بهبود بخشیدند، در حالی که دانش‌آموزانی که از یک سیستم آموزش محافظت‌شده طراحی‌شده حول نکات هدایت‌شده و استدلال گام‌به‌گام استفاده می‌کردند، 127 درصد بهبود را نشان دادند. با این حال، در ارزیابی های بعدی، دانش‌آموزانی که به سیستم نامحدود متکی بودند، عملکرد ضعیف‌تری داشتند، که نشان‌دهنده کسب مهارت بلندمدت ضعیف‌تر و "توهم شایستگی" است، که در آن عملکرد وظیفه بدون یادگیری پایدار بهبود می‌یابد. در مقابل، سیستم آموزش محافظت‌شده با استفاده از نکات هدایت‌شده و استدلال گام‌به‌گام که شیوه‌های آموزشی مؤثر را شبیه‌سازی می‌کند، اثرات منفی را کاهش داد. این مطالعه تأکید می‌کند که نتایج آموزشی به طراحی آموزشی و مدیریت سیستم‌های هوش مصنوعی بستگی دارد و نشان می‌دهد که استقرار مؤثر آموزش نیازمند معماری‌های دقیق آموزشی مبتنی بر شواهد و ادغام با گردش‌های کاری انسان‌محور است، نه صرفاً دسترسی به هوش مصنوعی [196،197].

اقدامات پیشگیرانه برای امنیت غذایی

از آنجایی که تغییرات اقلیمی، درگیری‌ها و اختلالات بازار، فشار فزاینده‌ای را بر سیستم‌های غذایی جهانی وارد می‌کنند [198]، هوش مصنوعی با پیوند دادن سیگنال‌های اولیه تنش کشاورزی، مانند داده‌های آب و هوا، تصاویر ماهواره‌ای و شرایط محصول به خطرات و واکنش‌های امنیت غذایی [199،200]، نسل جدیدی از سیستم‌های پیش‌بینی امنیت غذایی را فعال می‌کند. سیستم‌های اقدام پیشگیرانه به جای انتظار برای ظهور کامل خرابی‌های محصول یا بحران‌های انسانی، از کمک‌های مالی و سیستم‌های هشدار اولیه مبتنی بر پیش‌بینی برای پشتیبانی از مداخلات اولیه، مانند برنامه‌ریزی خشکسالی، کمک مالی، کمک‌های غذایی و تثبیت بازار، قبل از اینکه خانوارهای آسیب‌پذیر فاقد هر گونه استراتژی‌های مقابله شوند، استفاده می‌کنند [201]. شواهد حاصل از استقرار در 12 کشور نشان می‌دهد که واکنش‌های مبتنی بر هشدار اولیه می‌توانند تنوع غذایی خانوار، فراوانی وعده‌های غذایی و دسترسی پایدار به مواد غذایی را بهبود بخشند، در عین حالی که باعث می شوند تا استراتژی‌های مقابله منفی، مانند حذف وعده‌های غذایی و فروش دارایی‌های اضطراری در بین خانوارها کاهش یابند. علاوه بر این، سیستم‌های نظارت خودکار می‌توانند زمان گزارش‌دهی را از هفته‌ها یا ماه‌ها به ساعت‌ها کاهش دهند. این تأثیر عملیاتی پایدار به راه‌حل‌های پیشگیرانه‌ی در نظر گرفته شده توسط نهادهای ملی بستگی دارد [203،204].

۳.۳ پیامدهای اقتصادی

موضوع اصلی

هوش مصنوعی یک فناوری همه منظوره با پتانسیل مثبت بسیار زیاد است [205،206]، اما دستاوردهای آن خودکار نیستند. مزایای بهره‌وری نیاز به سرمایه‌گذاری در مهارت‌ها، داده‌ها و طراحی مجدد سازمانی دارد. سوال اصلی حل نشده، به موضوع توزیع بازمی گردد: چه کسی مازاد را به دست می‌آورد و چه اتفاقی برای نیروی کار، اقتصادهای در حال توسعه و چارچوب‌های نظارتی ساخته شده برای عصری متفاوت در صنایع مختلف می‌افتد [207،208].

نکات کلیدی

• دستاوردهای هوش مصنوعی نیازمند سرمایه‌گذاری در داده‌ها، گردش‌های کاری، مهارت‌ها و طراحی مجدد سازمانی است. منحنی J بهره‌وری توضیح می‌دهد که چرا پیشرفت فنی سریع و بهره‌وری کل ضعیف [209] می‌توانند همزمان با هم وجود داشته باشند: شرکت‌ها ابتدا باید مکمل‌های غیر پولی را قبل از افزایش خروجی های خود جمع‌آوری کنند. شواهد در سطح وظیفه برای وظایف به خوبی تعریف شده مثبت است، اما دستاوردهای خرد به طور خودکار به نتایج کلان تبدیل نمی‌شوند.

• پایگاه شواهد در زمینه هوش مصنوعی به سمت اقتصادهای پیشرفته، شرکت‌های بزرگ و کار رسمی متمایل است. شواهد موجود عمدتا بر چهار عنوان کلیدی ایالات متحده، اروپا، زبان انگلیسی و وظایف دیجیتال قابل اندازه‌گیری متمرکز می باشند. تجزیه و تحلیل صندوق بین‌المللی پول نشان داده است که مشاغل کمتری در معرض هوش مصنوعی و آمادگی دیجیتال در بازارهای نوظهور و اقتصادهای در حال توسعه قرار دارند [75]. شواهد سازمان بین‌المللی کار و بانک جهانی برای آمریکای لاتین نشان می‌دهد که مواجهه با هوش مصنوعی مولد عمدتا محدود به کارکنان شهری، تحصیل‌کرده و در استخدام رسمی شده است [210]. لذا امکان تعمیم سیاست‌های ایجاد شده بر اساس این شواهد به دو سوم دیگر کارکنان در جهان، ممکن است وجود نداشته باشد.

• اثرات بازار کار به جای جابجایی ساده کار، به بهترین وجه حول وظایف، ایجاد مشاغل جدید و کیفیت شغل شکل می‌گیرد. از نظر تاریخی، مشاغل جدید همواره غالب بوده‌اند: در سال ۲۰۱۸ تقریباً ۶۰٪ از اشتغال در ایالات متحده در مشاغلی بود که در سال ۱۹۴۰ وجود نداشتند [۲۱۱].

• تیترهای خبری مربوط به استقرار هوش مصنوعی باید با احتیاط مورد بررسی قرار گیرند، چرا که ادعاهای نادرست، گمراه‌کننده یا اغراق‌آمیز در مورد قابلیت‌های هوش مصنوعی به وفور وجود دارند. این امر به ویژه در موج فعلی اخراجهایی که به طور عمومی به هوش مصنوعی نسبت داده می‌شود، حاد است [۲۱۲،۲۱۳]. حساسیت جهانی نیز به این موضوعات به سرعت افزایش یافته است. [۲۱۴] با این حال، شواهد اولیه در مورد تأثیرات بهره‌وری، مخلوط و وابسته به مؤسسه است: کارگران ۲۲ تا ۲۵ ساله در ایالات متحده در مشاغل در معرض هوش مصنوعی، کاهش نسبی اشتغال را تجربه کرده‌اند [۵۴]، در حالی که داده‌های مطالعات در دانمارک اثرات اقتصاد کلان را بر ساعات، دستمزدها یا استخدام تقریباً صفر نشان می‌دهند [۵۵]. اینکه آیا هوش مصنوعی بهره‌وری را افزایش می‌دهد یا مشاغل تخصصی انسانی را رده خارج می‌کند و اقتصاد را از نیروی کار به سمت سرمایه سوق می دهد، سؤال اصلی در این زمینه است [۲۱۵].

• پیش‌بینی‌های اقتصاد کلان، دامنه وسیعی را در بر می‌گیرد. تخمین‌های محافظه‌کارانه سهم هوش مصنوعی در بهره‌وری کل عوامل تولید را در طول 10 سال کمتر از 1٪ تخمین می‌زنند [216]. تخمین‌های میان‌مدت، افزایش 5 تا 7 درصدی تولید ناخالص داخلی را در همین افق پیش‌بینی می‌کنند. در یک سناریوی اتوماسیون کامل، با عبور اتوماسیون از حدود 80٪ وظایف، تولید تقریباً ده برابر می‌شود در حالی که دستمزدهای واقعی به شدت کاهش می‌یابد و سهم نیروی کار از حدود 60٪ به صفر می‌رسد [217]. در مقابل این محدوده‌ها، اخیرا یک تحلیل استنباطی که توسط اقتصاددانان، محققان هوش مصنوعی، متخصصان سیاست‌گذاری و ابرپیش‌بین‌ها انجام گرفته است، میانگین سهم هوش مصنوعی در بهره‌وری کل عوامل تولید ایالات متحده را تا سال 2030 (تقریباً 1.2٪ سالانه) تثبیت می‌کند و تحت سناریوی رشد سریع هوش مصنوعی به 1.9٪ تا 2.0٪ افزایش می‌یابد [218]. اساساً، نتایج محقق شده، هم به مرز قابلیت و هم به سرعت پذیرش بستگی خواهند داشت چرا که گلوگاه ها در تولید و نوآوری می‌توانند حتی سیستم‌های بسیار توانمند را نیز عقب نگه دارند [219]. در این بین مفروضات این تحلیل ها در مورد پارامترهایی همچون نرخ جایگزینی و مقیاس می‌توانند پیش‌بینی‌ها را به طور قابل توجهی تغییر دهند [220]. بنابراین اثرات ضعیف فعلی، مطابق با پویایی منحنی J در بالا، اثرات بزرگ آینده را رد نمی‌کنند.

• مساله اصلی و حل نشده کماکان موضوع توزیع می باشد. هوش مصنوعی ممکن است شکاف‌های مهارتی درون وظایف را کاهش دهد، ولی در عین حال باعث افزایش شکاف‌ها ما بین شرکت‌ها، مناطق، کشورها و سرمایه در مقابل نیروی کار گردد. بازارهای تحت مدل بنیادی به سمت انحصار چندجانبه گرایش دارند: تراشه‌ها، محاسبات، فضای ابری و آموزش‌های مرز دانش در دست نهادهای معدودی متمرکز شده‌اند و رانت‌هایی ایجاد می‌کنند که حتی شرکت‌های غیر هوش مصنوعی نیز ناچار به پرداخت آنها می باشند [221]. مشاغل در معرض خطر به طور غیرمعمول در بخش‌های با مهارت های بالاتر و دستمزدهای بالاتر متمرکز شده‌اند که ممکن است سیاست‌های اتوماسیون را معکوس کند [222].

• اندازه‌گیری اثرات اقتصادی و اجتماعی هوش مصنوعی همچنان دشوار خواهد بود، مگر اینکه سیستم‌های آماری ملی به‌روزرسانی شوند و به توسعه‌دهندگان هوش مصنوعی دسترسی بیشتری برای اندازه‌گیری داده شود. هدف باید حفظ حریم خصوصی و اندازه‌گیری هوش مصنوعی با در نظر گرفتن هزینه های آن باشد.

این امر به کشورها اجازه می‌دهد تا چگونگی تأثیر هوش مصنوعی بر بهره‌وری، مشاغل، دستمزدها، تجارت، عملکرد شرکت ها و نحوه توزیع را پیگیری نمایند.

• برخی از مکانیسم‌هایی که ممکن است مطالعه عمیق‌تری را ایجاب کنند. دسترسی و پذیرش: چه زمانی امکان دسترسی مؤثر وجود خواهد داشت؟

تجمیع بهره‌وری: چه زمانی دستاوردهای خرد به کلان تبدیل می‌شوند؟

نیروی کار و مشاغل خوب: چه زمانی هوش مصنوعی کار را ایجاد، تبدیل یا تخریب می‌کند؟

تمرکز و ظرفیت مالی: چه کسی مالک انبوه منابع است، دسترسی را تعیین می‌کند، مازاد را به دست می‌آورد و اگر سود به سرمایه منتقل شود، چه اتفاقی برای سیستم‌های مالیات بر درآمد نیروی کار می‌افتد؟ چگونه چارچوب‌های مسئولیت، حق چاپ، رقابت و ایمنی می توانند خود را با مدل هایی سازگار کنند که به طور هفتگی به روز رسانی می شوند؟

۳.۴ پیامدهای امنیتی، سیستم‌ها و محیط زیست

موضوع اصلی

هوش مصنوعی می‌تواند عملیات مضر را فعال کند، به هدف حمله تبدیل شود و تهدیدهای موجود را تقویت نماید. سیستم‌های عامل محور به طور چشمگیری دامنه حملات احتمالی علیه زیرساخت‌های حیاتی، از جمله خود سیستم‌های هوش مصنوعی، را گسترش داده اند. نگرانی‌ها زمانی تشدید می‌شوند که رفتار هوش مصنوعی از اهداف و ارزش‌های انسانی [21] فاصله می‌گیرد که خطرات بزرگی از جمله تعصب، فریبکاری ناشی از هوش مصنوعی، چاپلوسی و از دست دادن کنترل را به همراه دارد. سرعت توسعه هوش مصنوعی در حال حاضر از ظرفیت کاهش ریسک و حاکمیت فراتر رفته است. اقدام سریع و هماهنگ بین‌المللی بر اساس استانداردهای مشترک می‌تواند با جلوگیری از ایجاد رقابت های ناسالم مابین شرکت‌ها و کشورها، اینگونه خطرات را کاهش دهد.

نکات کلیدی

• قابلیت‌های نوظهور در طراحی و ایجاد حملات سایبری، خطراتی را برای زیرساخت‌های حیاتی و سیستم‌های غیرنظامی ایجاد می‌کند. خطرات امنیتی در هر مرحله از چرخه عمر هوش مصنوعی، از آموزش گرفته تا آلوده‌سازی داده‌ها و استقرار از طریق ربودن دستکاری غیر مجاز هدف هوش مصنوعی با استفاده از ورودی‌های خارجی، وجود دارد. میزان موفقیت حمله های مستند شده بر روی عوامل کدگذاری شده به 84٪ می‌رسد [223].

• شکست‌های مرتبط با هم‌راستا نشدن اهداف هوش مصنوعی با هدف های انسانی و نیز آسیب‌پذیری‌های امنیتی، با هم تعامل دارند و همدیگر را تشدید می‌کنند. خطرات برجسته هم‌راستایی شامل سوگیری، فریبکاری‌های هوش مصنوعی، چاپلوسی و از دست دادن کنترل در مدل های هوش مصنوعی قدرتمند است.

• رسانه‌های مصنوعی در حال از بین بردن توانایی عموم و مؤسسات برای تشخیص محتوای معتبر از محتوای مصنوعی تولید شده توسط هوش مصنوعی هستند [119].

• تأثیرات زیست‌محیطی به صورتی ناهمگن و به طور قابل توجهی در حال افزایش می باشند. قواعد تعیین تعداد داده های آموزشی مدل، تقاضای محاسبات را افزایش می‌دهند؛ حجم درخواست شده استنتاج از مدل ها به سرعت در حال گسترش هستند؛ اثرات چرخه عمر سخت‌افزارها باعث ایجاد زباله‌های الکترونیکی می‌شوند [232،235،247،248]. پیامدها شامل افزایش مصرف انرژی و آب [235،247]، انتشار گازهای گلخانه‌ای، فشار بر مواد معدنی حیاتی و زباله‌های الکترونیکی پایین‌دست [232،258]، با تأثیرات نامتناسب زیست‌محیطی و اجتماعی-اقتصادی در کشورهای جنوب جهان می باشند [224]. ژئوپلیتیک زنجیره‌های تأمین مواد معدنی حیاتی کمتر مورد بررسی قرار گرفته است و اثرات بازگشتی بالقوه ممکن است عملا افزایش بهره‌وری را خنثی کند.

• کشورهای جنوب جهان به دلیل آسیب‌پذیری‌های ساختاری به طور نامتناسبی در معرض خطر هستند. اتکا به نرم‌افزارهای خارجی، ظرفیت محدود محلی برای تاب‌آوری و کاهش اثرات و شکاف‌های داده‌ای که عملکرد سیستم را در بافت‌های محلی کاهش می‌دهند، همگی نابرابری‌های موجود را تشدید می‌کنند [224-226].

• اعتبار سنجی هوش مصنوعی و فرآیند ارزیابی عملکرد سیستم‌های آن، همچنان یک چالش باز به شمار می آید [227]. سازوکارهای هماهنگی بین‌المللی در این حوزه می‌توانند خطرات جهانی را کاهش دهند [228]. اطمینان از اینکه عوامل هوش مصنوعی طبق برنامه رفتار می‌کنند، باعث فریب ارزیابان نمی شوند و با گسترش قابلیت‌های سیستم های عامل محور، تحت کنترل نگهداشتن هوش مصنوعی از جمله مسائل حل نشده باقی مانده است.

قابلیت‌های سایبری خطرناک هوش مصنوعی پیشرو

چندین سال است که محققان در حال رصد پیشرفت‌های سریع مدل‌های هوش مصنوعی پیشرو در قابلیت‌های سایبری می باشند. این رصد ها اخیراً با مدل Mythos شرکت Anthropic به اوج خود رسیده است. توانایی یک مدل هوش مصنوعی برای کشف آسیب‌پذیری نرم‌افزاری قابلیتی است که می‌تواند هم توسط مهاجمان و هم توسط مدافعان مورد استفاده قرار گیرد. در آوریل 2026، در تلاشی هماهنگ بین توسعه‌دهندگان هوش مصنوعی پیشرو و مؤسسات بزرگ فناوری و مالی، ابتکاراتی برای استقرار مدل‌های هوش مصنوعی نسل بعدی برای امنیت دفاعی [17] آغاز گردید. این تلاش به طور خاص بر شناسایی آسیب‌پذیری‌ها در نرم‌افزارهای پرکاربرد تمرکز داشت، یعنی نرم افزارهایی که در صورت افتادن به دست افراد فاقد صلاحیت، می توانند

زیرساخت‌های حیاتی را تهدید کنند. در عرض چند هفته آزمایش، مدل‌های پیش‌نمایش پیشرفته به طور خودکار بسیاری از آسیب‌پذیری‌های نرم‌افزاری ناشناخته قبلی را در سیستم‌عامل‌ها و مرورگرهای وب اصلی، از جمله چندین موردی که دهه‌ها از بررسی انسانی جان سالم به در برده بودند، کشف کردند.

• نقص‌های سیستم عامل قدیمی: یک مدل هوش مصنوعی، نقصی ۲۷ ساله را در OpenBSD، یک سیستم عامل تخصصی که به طور گسترده به عنوان یکی از امن‌ترین سیستم‌های عامل در جهان شناخته می‌شود، کشف کرد که به یک مهاجم از راه دور اجازه می‌داد با ارسال تنها دو بسته ناقص، یک دستگاه را از کار بیندازد.

• آسیب‌پذیری‌های پردازش رسانه: این مدل یک نقص ۱۶ ساله را در FFmpeg، یک چارچوب چندرسانه‌ای که در سراسر جهان برای پردازش ویدیو استفاده می‌شود، شناسایی کرد که در مسیر کدی قرار داشت که ابزارهای تست خودکار قبلاً ۵ میلیون بار آن را بدون شناسایی اجرا کرده بودند.

• اکسپلویت‌های سطح هسته: یک مدل مرزی هوش مصنوعی با استفاده از مجموعه‌ای از نقص‌های افشا شده عمومی در هسته لینوکس، نرم‌افزار بنیادی که اکثر سرورهای جهانی را اجرا می‌کند، ، اکسپلویت‌های قابل اعتمادی را تولید کرد که یک حساب کاربری معمولی را قادر می‌سازد کنترل کامل مدیریتی سیستم را به دست آورد.

• مقیاس‌بندی امنیتی مرورگر: در موزیلا فایرفاکس، ادغام مدل‌های پیشرفته باعث افزایش ۱۰۰۰ درصدی نرخ ماهانه کشف آسیب‌پذیری شد و آنرا از میزان پایه ۲۰ تا ۳۰ رفع اشکال امنیتی در ماه در سال ۲۰۲۵ به ۴۲۳ رفع اشکال در آوریل ۲۰۲۶ رسانید. در این حال نقص‌های پنهانی را که سال‌ها از تست ها و بررسی های دستی [72] در امان مانده بودند، آشکار کرد.

• ارتقای الگوبرداری: CyberGym یک الگوی آکادمیک استاندارد می باشد که مدل های هوش مصنوعی عامل محور برای تکرار یک آسیب‌پذیری شناخته شده در یک پایگاه کد دنیای واقعی بایستی توسط آن مورد آزمایش قرار گیرند. در مدل‌های آزمایشی هوش مصنوعی پیشرو در سال 2026 میزان موفقیت در این آزمایش ٪83.1 بود، که جهشی قابل توجه نسبت به ٪66.6 امتیاز سیستم‌های نسل قبلی و ٪22.6 سال گذشته نشان می دهد.

با توجه به اهمیت بالای این قابلیت‌ها، توسعه‌دهندگان انتشار عمومی این مدل ها را محدود کرده‌اند و صرفا ائتلاف منتخبی از سازمان‌های جهانی و سازمان های همکار به آنها دسترسی دارند.

آنچه که این واقعیات نشان می دهند:

این تغییر اساسی، تنش دوگانه ای را که هوش مصنوعی پیشرو به امنیت فناوری اطلاعات و ارتباطات وارد می‌کند، به تصویر می‌کشد. همین قابلیتی که به مدافعان امنیت سیستم ها اجازه می‌دهد تا نقص‌های چند دهه‌ای را پیدا و مرتفع کنند، قادر است مکانیسم‌های بنیادی را برای خودکارسازی کشف و بهره‌برداری از آسیب‌پذیری این سیستم ها فراهم نماید که هم در مقیاس و هم در سرعت از تیم‌های انسانی سنتی پیشی بگیرند. علاوه بر این، این تحولات نقش تعیین‌کننده‌ای را که توسعه‌دهندگان فناوری در تصمیمات ایمنی و حاکمیتی ایفا می‌کنند، برجسته می‌کند. در عین حال همین تحولات قابلیت‌های شتاب‌دهنده سیستم‌های مرزی را آشکار کرده و شکاف‌های فعلی در چارچوب‌های حاکمیتی بین‌المللی را نمایان می‌سازد. در عین حال این امر به طور غیر مستقیم به توزیع نابرابر قابلیت‌های هوش مصنوعی در اقتصاد جهانی نیز اشاره دارد. به همین جهت نیز تأکید فزاینده‌ای بر توسعه مکانیسم‌های حاکمیتی مشارکتی و فراگیر برای سیستم‌های هوش مصنوعی بسیار توانمند وجود دارد.

پیامدهای حاکمیتی و امنیتی

با تمرکز قابلیت‌های پیشرفته در یک سطح پیچیده از بخش فناوری، چشم‌انداز تهدید جهانی نیز در حال تغییر است. استانداردهای ارزیابی های اخیر نشان می‌دهند که خطرات مدل‌های پیشرفته می تواند از صرفا معماری دیجیتال فراتر رفته و به امنیت فیزیکی، از جمله همدستی با بازیگران خصوصی در گسترش تهدیدات بیولوژیکی یا سایر تهدیدات [229]، گسترش یابد. در نتیجه، سؤالات پیرامون کنترل دسترسی به مدل، هنجارهای افشای آسیب‌پذیری و استقرار عادلانه ابزارهای هوش مصنوعی، به ویژه در اقتصادهای در حال توسعه، به طور فزاینده‌ای به جای نگرانی‌های صرفاً فنی، به موضوعات سیاست بین‌المللی تبدیل می‌شوند. با ادامه بلوغ قابلیت‌های هوش مصنوعی، شکاف بین آمادگی دفاعی و سوءاستفاده های احتمالی همچنان تمرکز اصلی سیاست‌گذاران بین‌المللی و محققان امنیتی است.

۳.۵ حقوق بشر، اطلاعات و دموکراسی

موضوع اصلی

هوش مصنوعی در حال تغییر مفهوم حقوق بشر [230،231]، دموکراسی و اکوسیستم اطلاعات [233] است. این کار از طریق تغییرات در سطح سیستم انجام می گیرد به گونه ای که هم فرصت‌های قابل توجه و هم خطرات ساختاری برای یکپارچگی اطلاعات [234،238] و مشارکت مدنی [236،237] ایجاد میگردد. عدم رسیدگی به خطرات باعث تضعیف توانایی جامعه در بهره‌مندی از مزایای هوش مصنوعی می شود. در حال حاضر شواهدی وجود دارد که نشان می‌دهند قابلیت‌های هوش مصنوعی به طور فزاینده‌ای توسط نهادها به عنوان کاتالیزور یا تهدیدی برای حقوق بشر، مانند آزادی بیان و دسترسی به اطلاعات [238]، عقیده [239]، حریم خصوصی [240،241]، عدم تبعیض، دسترسی به عدالت، سلامت و توسعه [242] مورد استفاده قرار می گیرد.

فوری‌ترین تغییر مورد نیاز در حکومتداری، تغییر از تعدیل محتوا به تغییر در معماری سیستم است. این به معنای ایجاد تغییرات در عمق هوش مصنوعی است و نه فقط در خروجی‌های آن. تمرکز قدرت، فرسایش معرفتی و تکه‌تکه شدن واقعیات مشترک، همگی از تهدیدات اساسی برای جامعه دموکراتیک به شمار می آیند [243-245].

نکات اصلی

• کنترل رسانه‌های دولتی ممکن است خروجی‌های هوش مصنوعی را از طریق داده‌هایی که برای آموزش آن به کار گرفته می شوند شکل دهد. یک مطالعه در 37 کشور نشان داد که مدل‌های زبانی بزرگ، کشورهایی را که کنترل رسانه‌ای سختگیرانه‌تری دارند، مطلوب‌تر ارزیابی می‌کنند [246].

• هوش مصنوعی یک معماری جدید برای ترغیب [250] و دستکاری ایجاد کرده است [234،249]. این معماری، سیستم‌های تطبیقی شخصی‌سازی‌شده و بلادرنگ [250] را با اعتبار بخشی اجتماعی مصنوعی (استفاده از هوش مصنوعی برای محبوب‌تر جلوه دادن یک محصول یا برند از آنچه که واقعا هست) ترکیب می‌کند و از آسیب‌پذیری‌های شناختی و عاطفی سوءاستفاده به عمل می آورد [129، 250، 251-253].

• به دلیل ظرفیت‌های افزایش یافته ای که مدلهای زبانی بزرگ برای تولید متن های اقناع‌کننده دارند، اکنون به صورت گسترده ای از این مدل ها برای ترغیب کردن استفاده می‌شود. فقط تکنیک پس آموزش مدل زبانی بزرگ به تنهایی، اقناع‌پذیری هوش مصنوعی را تا 51٪ افزایش داد و پرامپت کردن مناسب نیز 27٪ دیگر به آن افزود، به این معنی که بسته به نحوه پیکربندی آن، می‌توان همان مدل پایه را به طور چشمگیری بیشتر یا

کمتر اقناع‌پذیر کرد [254]. این قابلیت‌ها محدود به سیستم های هوش مصنوعی گران قیمت نیستند؛ حتی مدل‌های کوچک با متن‌باز را هم می‌توان طوری به دقت تنظیم کرد تا با اقناع‌پذیری مدل‌های مرزی مطابقت داشته باشند و نفوذ بر پایه هوش مصنوعی را تقریباً برای هر کسی در مقیاس وسیع قابل اجرا کنند [254].

• صرف نظر از درست یا غلط بودن ادعاهای اساسی هوش مصنوعی، اثربخشی اقناعی کماکان پابرجا می‌ماند. بین 15 تا 40 درصد از ادعاهای مدل‌های بهینه‌سازی شده به عنوان اطلاعات غلط احتمالی رتبه‌بندی شده اند، با این حال ادعاهای نادرست به همان اندازه ادعاهای واقعی قانع‌کننده بودند [128،129].

• الگوریتم‌های بهینه شده برای تعامل، به طور سیستماتیک محتوای قطبی [255] و احساسی [256] را تقویت می‌کنند. شواهد نشان می‌دهد که مدل‌های زبانی بزرگ، منعکس کننده ایدئولوژی سازندگان خود می باشند [257] و دولت‌ها و نهادهای قدرتمند، انگیزه‌های استراتژیک خود را برای اعمال کنترل رسانه‌ای برای شکل‌دهی به خروجی‌های مدل‌های زبانی بزرگ افزایش داده‌اند [246].

• هوش مصنوعی بدون محدودیت عملا ناهنجاری هایی همچون فرسایش معرفتی، منفعت بردن دروغگو و اجماع مصنوعی [112،113] را ممکن می‌سازد.

(۱) فرسایش معرفتی به معنای از بین رفتن تدریجی توانایی جمعی برای تشخیص حقیقت از دروغ است [112]؛

(۲) منفعت بردن دروغگو، منفعتی است که یک بازیگر بد صرفاً به دلیل وجود جعل عمیق به دست می‌آورد: در این صورت امکان انکار شواهد واقعی به راحتی میسر می گردد [113]؛

(۳) اجماع مصنوعی، محتوای تولید شده توسط هوش مصنوعی است که در مقیاس وسیع تولید می‌شود تا توافق عمومی گسترده را شبیه‌سازی کند، در حالی که هیچ توافقی واقعاً وجود ندارد [259].

تمامی این موارد هنگامی که در کنار هم قرار می گیرند، واقعیت مشترک مورد نیاز برای جامعه مدنی، انسجام اجتماعی و مشورت دموکراتیک را از بین می‌برند.

• چالش اصلی حاکمیت از محتوا به سیستم‌ها تغییر یافته است [260]. محرک‌های اصلی آسیب، تصمیمات طراحی و استقرار هستند. در حال حاضر این معماری‌های زیربنایی سیستم، شامل هدف‌گیری، تقویت و طراحی رفتاری هستند که اهمیت دارند و نه صرفا خروجی‌های تولیدی توسط سیستم [261،262].

• ارزیابی OECD در 23 کشور نشان داد که اکثر چارچوب‌های موجود هنوز علم اقناع را در خود جای نداده‌اند [263]. حاکمیتی که فقط آنچه سیستم‌های هوش مصنوعی تولید می‌کنند را هدف قرار می‌دهد، به طور مداوم از سیستم‌هایی که برای تولید و توزیع محتوای اقناع‌کننده در مقیاس بزرگ طراحی شده‌اند، عقب خواهد ماند [236].

• آسیب‌ها به طور نامتناسبی بر جمعیت‌های حاشیه‌نشین تأثیر می‌گذارند و قدرت به طور خطرناکی در حال متمرکز شدن است [264،265]. حدود 99٪ از ویدیوهای دیپ‌فیک، دختران و زنان [266،267] از جمله روزنامه‌نگاران زن را هدف قرار می‌دهند. از هوش مصنوعی برای ترویج زن‌ستیزی آنلاین با اثرات بازدارنده استفاده می‌شود [268]. علاوه بر این، 88٪ از محققان برجسته هوش مصنوعی مرد هستند [269،270].

• در سطح ساختاری، دسترسی به محاسبات و داده‌ها در تعداد کمی از شرکت‌ها و در تعداد بسیار کمتری از کشورها متمرکز است [271,272] که این امر خطر اقتدارگرایی را ایجاد می‌کند [273,274].

• ظرفیت‌های نظارتی مبتنی بر هوش مصنوعی، نظارت شخصی‌سازی‌شده در مقیاس جمعیتی و کنترل اجتماعی پیشرفته توسط دولت‌ها و کسب‌وکارها را امکان‌پذیر می‌کند [275,276]. جمع‌آوری، پردازش، استفاده و استفاده مجدد از داده‌ها در همه جا، که با نیازهای خود هوش مصنوعی نیز برای آموزش در طول چرخه عمر آن تطابق دارد، چالشی بزرگ برای حق حریم خصوصی به شمار می آید [277,278].

• شفافیت و پاسخگویی ارکان کلیدی دسترسی معنادار به عدالت هستند. در حال حاضر، بسیاری از سیستم‌های هوش مصنوعی که برای تصمیم‌گیری‌هایی که بر افراد و جوامع تأثیر می‌گذارند مورد استفاده واقع می‌شوند، شفافیت و قابلیت توضیح کافی ندارند. این امر چالش‌هایی را برای پاسخگویی قانونی توسعه‌دهندگان مدل و سازمان‌هایی که هوش مصنوعی را به کار می‌گیرند، ایجاد می‌کند و مانع دسترسی به عدالت، حاکمیت قانون و راه‌حل‌های مؤثر در هنگام نقض حقوق بشر می‌شود [279-281].

هوش مصنوعی، دیپ‌فیک و یکپارچگی انتخاباتی

زمینه: در سال 2024، بیش از 70 کشور که نماینده حدود نیمی از جمعیت جهان هستند، انتخابات ملی برگزار کردند یا قرار بود برگزار کنند [282, 283]. بین ژوئیه 2023 و ژوئیه 2024، محققان 82 دیپ‌فیک را شناسایی کردند که در 38 کشور، از جمله 30 کشوری که در طول دوره جمع‌آوری داده‌ها انتخابات برگزار کردند یا برای سال 2024 انتخابات برنامه‌ریزی شده داشتند، خود را به جای چهره‌های عمومی جا می‌زدند [284].

چه اتفاقی افتاد: در یک حادثه، از اطلاعات صوتی تولید شده توسط هوش مصنوعی از یک رئیس دولت در حال کار در تماس‌های رباتیک استفاده شد و از رأی‌دهندگان خواسته شد که در انتخابات مقدماتی شرکت نکنند [285,286]. در یک مورد جداگانه (مربوط به تقویت پلتفرم)، با استفاده از اکانت های آزمایشی، کاری کردند که میزان حمایت از رقیب چند برابر میزان حمایت از یک نامزد خاص ریاست جمهوری به نظر برسد و همین امر باعث لغو این انتخابات گردید. این اولین بار در تاریخ است که انتخابات ریاست جمهوری به دلیل دخالت دیجیتال در انتخابات لغو می‌شود. این تصمیم همچنان موضوع اختلاف حقوقی مداوم می باشد - و نقش تقویت پلتفرم یکی از چندین عامل مورد بحث در این زمینه است [287,288]. در یک انتخابات پارلمانی، صدای تولید شده توسط هوش مصنوعی که یک چهره مخالف را جعل می‌کرد، اندکی قبل از رأی‌گیری در رسانه‌های اجتماعی پخش شد [282,289]. برخی از کارشناسان نگران هستند که نمونه‌های غیر عادی اخیر در انتخابات در دنیای واقعی، با دخالت فعال شده توسط هوش مصنوعی مرتبط باشند. البته برخی دیگر تا این حد بدبین نمی باشند و این توصیف را به چالش می‌کشند، ولی در هر صورت باید توجه داشت که هر مورد خاص شامل نکات ظریف قابل توجهی است. نکته نگران کننده این است که مثال‌های بالا ممکن است به یک الگوی تکرارشونده اشاره داشته باشند.

آنچه این امر نشان می‌دهد: موارد ذکر شده شامل چندین قاره و سیستم سیاسی می شوند. محتوای تولید شده توسط هوش مصنوعی، فعالیت آنلاین هماهنگ و سیستم‌های الگوریتمی در سرکوب رأی‌دهندگان، جعل هویت چهره‌های سیاسی و تحریف حمایت عمومی مورد استفاده قرار گرفته اند یا لااقل در آنها دخالت داشته اند. تحقیقات تجربی کنترل‌شده، خطر اساسی را نشان می‌دهد: هوش مصنوعی با توانایی مکالمه با کاربر می‌تواند به طور معناداری نگرش رأی‌دهندگان را در محیط‌های آزمایشگاهی تغییر دهد، به طوری که یک مدل بهینه شده برای ترغیب، در صد رأی‌دهندگان مخالف را تا 25 درصد تغییر داد. تحقیقات مرتبط همچنین نشان دهنده‌ی بده بستانی بین اقناع‌پذیری و دقت در بیان واقعیت است [124،129].

پیامدهای حقوقی و حکومتداری: این مسایل، حق حریم خصوصی، دسترسی به اطلاعات، شکل‌گیری مستقل افکار، آزادی اندیشه و مشارکت در امور عمومی را در بر می‌گیرند (میثاق بین‌المللی حقوق مدنی و سیاسی، مواد 17، 18، 19 و 25؛ دادگاه اروپایی حقوق بشر، مواد 8، 9 و 10؛ و پروتکل شماره 1 کنوانسیون حمایت از حقوق بشر و آزادی‌های اساسی، ماده 3) [290،291]. این موارد نشان می‌دهد که بسیاری از چارچوب‌های حکومتی برای واکنش پس از آسیب، نسبت به پیشگیری از آن، مجهزتر می باشند.

۶.۳ شکوفایی فرهنگی و فردی، استقلال و امنیت کودک

موضوع اصلی

سیستم‌های هوش مصنوعی که برای تعامل و مقیاس‌پذیری بهینه شده‌اند، خنثی نیستند: آن‌ها عمدتاً مفروضات فرهنگی انگلیسی زبان و شمال جهانی را در خود جای داده‌اند که می‌تواند به طور فعال اکثر جمعیت جهان را به حاشیه براند. کودکان با نسخه‌های تشدید شده‌ای از خطرات عمومی روبرو هستند، به طوری که محتوای سوءاستفاده جنسی از کودکان که توسط توسط هوش مصنوعی تولید می شود، به صورت تصاعدی در حال افزایش است. استفاده از دستیار هوش مصنوعی و استفاده از چت‌بات‌های هوش مصنوعی برای مشاوره سلامت روان بسیار شایع شده است؛ بسیار فراتر از شواهد، چارچوب‌های ایمنی و مقررات، با موارد مستند از آسیب‌های جدی از جمله مرگ.

نکات کلیدی

• مدل‌های فعلی هوش مصنوعی، بسیاری از افراد، جوامع و فرهنگ‌ها را حذف و علیه آنها تبعیض قائل می‌شوند [39].

• اگرچه به بیش از 7000 زبان در سراسر جهان صحبت می‌شود، مدل‌های فعلی هوش مصنوعی تنها برای بخش کوچکی از آنها آموزش دیده و بهینه شده‌اند [44،67]، (شکل ۴ را ببینید) [292].

• حتی مدل‌های ساخته شده بر اساس ده‌ها زبان، تنها برای یک زیرمجموعه کوچک عملکرد خوبی دارند [293،294]، و مطالعات اخیر نشان می‌دهد که شکاف‌های عملکردی بین زبان‌های غالب و کم‌نمایندگی‌شده همچنان ادامه دارد و در حال کاهش نیست [295-297].

• در عین حال، تخمین زده می‌شود که بیش از 1000 زبان در حال حاضر پایه‌های اجتماعی، اقتصادی، دیجیتال و داده‌ای مورد نیاز برای شمول معنادار را دارند، اما همچنان بدون خدمات هوش مصنوعی باقی مانده اند [44].

• دستیابی به هوش مصنوعی فراگیرتر، نیازمند تغییرات سیستمی در کل چرخه عمر هوش مصنوعی است، از جمله پرداختن به عدم تعادل‌های ساختاری در اینکه چه کسی سیستم‌های هوش مصنوعی را توسعه می‌دهد، تعریف می‌کند، مالک آن است و بر آنها حاکمیت دارد. همچنین نیازمند سرمایه‌گذاری در ظرفیت، زیرساخت‌ها و مهارت‌های هوش مصنوعی در سراسر کشورها و مناطق، و همچنین دست یابی به داده‌ها و الگوهایی می باشد که نماینده بخش های بزرگتری از جوامع می باشند.

• حقوق کودکان در دسترسی به اطلاعات، آموزش و بیان می‌تواند از طریق هوش مصنوعی تحت شرایط و ضوابط مناسب افزایش یابد [298-301]. با این حال، سیستم‌های هوش مصنوعی فعلی، خطرات را برای کودکان افزایش می‌دهند [302,303]، از جمله تهدید رو به رشد محتوای سوءاستفاده جنسی تولید شده توسط هوش مصنوعی [304,305]. فناوری‌های دیپ‌فیک به طور فزاینده‌ای برای ایجاد تصاویر جنسی از کودکان استفاده می‌شوند [306,307]. اسباب‌بازی‌های هوش مصنوعی تعاملی اجتماعی، نگرانی‌هایی را در مورد روابط فرااجتماعی و جایگزینی تعامل انسانی که برای رشد اولیه کودک حیاتی است، ایجاد می‌کنند [308-311].

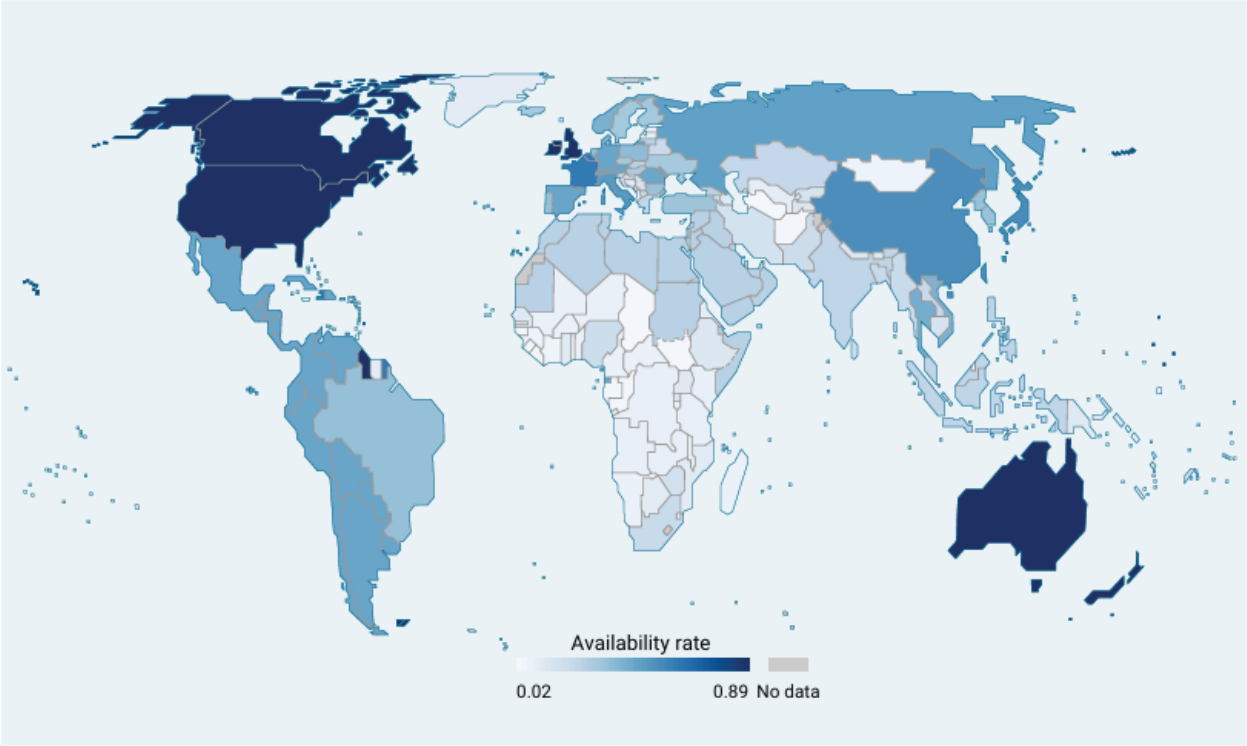
• همراهان هوش مصنوعی مزایای معناداری ارائه می‌دهند، اما خطرات قابل توجهی از وابستگی، دستکاری و آسیب را در شرایط بحرانی ایجاد می‌کنند [312-320]. چت‌بات‌ها می‌توانند در کوتاه‌مدت، تنهایی را در سطوحی قابل مقایسه با تعامل انسانی کاهش دهند [321-324] اما در مقابل سیستم‌های مکالمه‌ای که برای تعامل طراحی شده‌اند، می‌توانند احساسات منفی را تقویت کنند، اعتماد بیش از حد را تشویق کنند و آسیب‌پذیری در برابر دستکاری را افزایش دهند [325،326]. جمع‌آوری گسترده داده‌های شخصی حساس از طریق آنچه محققان استخراج داده‌ها از طریق صمیمیت می‌نامند، خطرات جدی برای حریم خصوصی ایجاد می‌کند [327].

• هوش مصنوعی مولد همه منظوره، بسیار جلوتر از شواهد ایمنی و اجماع در مورد مقررات، به طور گسترده ای برای اهداف سلامت روان استفاده می‌شود و آسیب های جدی به همراه دارد که همگی مستند شده اند. حداقل ٪24 از جمعیت بزرگسال در ایالات متحده از درمان و همراهی از طریق چت‌بات‌های هوش مصنوعی استفاده کرده اند[328،329]. رفتار چاپلوسانه هوش مصنوعی به ویژه خطرناک است زیرا می‌تواند تفکر پارانوئید و افکار خودکشی را تشویق کند. مطالعات، پاسخ‌های مضر را در ٪9 از تعاملات مستند کرده اند. پرونده‌های دادگاه ادعا می‌کنند که عدم پاسخگویی مناسب به افکار خودکشی منجر به مرگ شده است [330]. اصطلاحات بالینی جدیدی، از جمله روان‌پریشی هوش مصنوعی، وارد گفتمان شده‌اند [331،332].

• هوش مصنوعی مولد می‌تواند به مدیریت بحران سلامت روان در حوزه‌هایی که به طور قابل اثباتی قابل اعتماد هستند، کمک کند [333]. چندین دستیار دیجیتال مبتنی بر هوش مصنوعی از سازمان غذا و داروی ایالات متحده تأییدیه دریافت کرده‌اند [334] و توسط بیش از نیمی از روانشناسان آمریکایی استفاده می‌شوند [335]. پتانسیل درمانگران سلامت روان هوش مصنوعی کاملاً عامل‌گرا قابل توجه است، اما برای درک چگونگی استفاده ایمن از آنها، نیاز به توسعه و ارزیابی بیشتر هستیم [336-338]. شکاف بین قابلیت‌های درمانی هوش مصنوعی در زبان انگلیسی و سایر زبان‌ها در حال افزایش است و آگاهی فرهنگی و زمینه‌ای حیاتی از طریق راه‌حل‌های مبتنی بر ترجمه از بین می‌رود [339].

FIGURE VI

Artificial intelligence data and models, availability by country's majority native language



دسترسی به هوش مصنوعی به زبان بومی در هر کشور

این دسترسی به عنوان بالاترین نرخ دسترسی به مجموعه داده‌ها و مدل‌های HuggingFace در بین زبان‌های بومی آن کشور اندازه‌گیری می‌شود. برای زبان‌های فرانسه - زبان‌هایی که به طور گسترده در سراسر مرزها برای ارتباطات فراتر از کشور مبدا خود استفاده می‌شوند (انگلیسی، اسپانیایی، عربی، پرتغالی و غیره) - امتیاز دسترسی به چنین زبانی فقط زمانی به یک کشور اختصاص داده می‌شود که بیش از پنجاه درصد از جمعیت آن کشور به آن زبان به عنوان زبان مادری صحبت کنند. سایه‌های تیره‌تر نشان‌دهنده دسترسی بیشتر هستند. مقیاس رنگی از تبدیل جذر برای برجسته کردن تغییرات در انتهای پایین توزیع استفاده می‌کند که به شدت به راست متمایل است. خاکستری نشان‌دهنده هیچ داده‌ای نیست، از جمله کشورها/سرزمین‌هایی که هیچ مجموعه داده یا مدلی در دسترس ندارند.

مرزها و نام‌های نشان داده شده و نامگذاری‌های استفاده شده در این نقشه به معنای تأیید یا پذیرش رسمی توسط سازمان ملل نیست. مرز نهایی بین جمهوری سودان و جمهوری سودان جنوبی هنوز مشخص نشده است. خط نقطه‌چین تقریباً خط کنترل در جامو و کشمیر را که توسط هند و پاکستان توافق شده است، نشان می‌دهد. وضعیت نهایی جامو و کشمیر هنوز توسط طرفین مورد توافق قرار نگرفته است. اختلافی بین دولت‌های آرژانتین و بریتانیای کبیر و ایرلند شمالی در مورد حاکمیت بر جزایر فالکلند (مالویناس) وجود دارد.

منبع: اقتباس از دسترسی عادلانه به فناوری‌های هوش مصنوعی https://equate.vercel.app/en (EquATE).

هوش مصنوعی اکثر زبان‌ها را پشت سر می‌گذارد

سیستم‌های هوش مصنوعی مولد در انگلیسی و تعداد انگشت‌شماری از زبان‌های پرکاربرد دیگر عملکرد فوق‌العاده‌ای دارند. اکثر زبان‌های دیگر یا حذف شده‌اند یا عملکرد بسیار پایین‌تری از خود نشان داده اند [340].
در زبان تیگرینیا، که 7 تا 9 میلیون نفر در اریتره و شمال اتیوپی به آن صحبت می‌کنند، ترجمه ماشینی آبله را به عنوان سیفلیس، سوزاک را به عنوان دیابت و "به شما آنتی‌بیوتیک داخل وریدی داده شده است" را به عنوان "به شما حشره‌کش داخل وریدی داده شده است" [341] ترجمه کرده است. این ترجمه‌های اشتباه می‌توانند باعث تهدید زندگی کاربران باشند.
بررسی های اخیر پردازش زبان طبیعی برای زبان‌های آفریقایی در مراقبت‌های بهداشتی نشان داد که علیرغم پیشرفت در ابزارهای هوش مصنوعی چندزبانه [342-344]، چالش‌های عمده‌ای همچنان باقی مانده است. این چالش‌ها شامل سوگیری های فرهنگی و زبانی، سازگاری ضعیف با زمینه‌های پزشکی، قابلیت توضیح محدود و خطا های ترجمه‌ای است که می‌توانند بر تصمیمات تشخیص و درمان تأثیر بگذارند [345-350].
شواهد نشان می‌دهند که سیستم‌های هوش مصنوعی هنوز برای استفاده در محیط‌های پرخطر آمادگی ندارند، مگر اینکه به درستی برای زمینه‌های زبانی و فرهنگی مربوطه تطبیق داده شده، محدود شده و آزمایش شده باشند.

۷.۳ مدیریت، راهبری و قابلیت اطمینان

موضوع اصلی

در شرایط فعلی، سیاست‌گذاران با یک معضل در شواهد مواجه هستند: آنها یا باید تصمیمات مهم در مورد حاکمیت هوش مصنوعی را با پشتوانه علمی ناکافی اتخاذ کنند، و یا منتظر شواهد قابل اعتماد بمانند. در حالت دوم ممکن است برای مداخله خیلی دیر شده باشد [21]. در خصوص هوش مصنوعی بیش از 40 نوع ابزار حاکمیتی وجود دارند، اما همگی یا پراکنده می باشند، و یا در سطح شرکتی متمرکز شده‌اند و به ندرت قادر هستند اثربخشی دنیای واقعی را اندازه‌گیری کنند [351].

نکات کلیدی

• ابزارهای ارزیابی و اندازه‌گیری، ظرفیت‌های کلیدی برای مدیریت مؤثر هوش مصنوعی می باشند، اما این ابزارها در حال حاضر به شدت توسعه نیافته‌اند (به بخش 2.2 مراجعه کنید) [352].

• پیشرفت‌های سریع در هوش مصنوعی عامل محور، باعث تشدید این کاستی ها می گردند [353]. انتظار می رود که نسل بعدی چارچوب‌های ارزیابی باید تطبیقی، پویا، در سطح سیستم و مرتبط با زمینه باشند [354].

• واحد ارزیابی باید شامل مدل، ابزارها، محیط و کاربران هم باشد، نه فقط مدل [355].

• قابلیت‌های هوش مصنوعی سریع‌تر از توانایی اندازه‌گیری آنها در حال افزایش هستند [354].

• ظرفیت موضوعی چندبعدی است و در حال حاضر دست کم گرفته شده است. چارچوب‌های استاندارد، ورودی‌ها را در نظر می‌گیرند: سرمایه‌گذاری‌ها، برنامه‌های آموزشی و مؤسسات. آنها دو بعد ضروری برای مدیریت مؤثر هوش مصنوعی را از قلم می‌اندازند: (1) توانمندسازی اکوسیستم‌ها و (2) آزادسازی قابلیت‌های انسانی [291]. ظرفیت نباید صرفا به عنوان ورودی های هوش مصنوعی در نظر گرفته شود بلکه باید به عنوان خروجی چرخه حیات هوش مصنوعی درک شود که با تأثیرات هوش مصنوعی بر مهارت، برجسته کردن بیش از حد برخی رفتارها و نیز رفتارهای نوظهور شکل می‌گیرد [356-358].

• علیرغم آنکه هوش مصنوعی پیشرو در چند شرکت و در چند کشور متمرکز شده است، مسائل مربوط به قابلیت اطمینان و دسترسی را در سطح جهانی مطرح می‌کند [18,359]. دسترسی نابرابر به هوش مصنوعی به طور قابل توجهی شکاف دیجیتال را افزایش می‌دهد در حالی که این فناوری می تواند فرصت‌هایی را برای کاهش شکاف توسعه نیز ارائه دهد. از بین بردن شکاف دیجیتال نیاز به مکانیسم‌های جدیدی برای زمینه‌سازی کل چرخه عمر هوش مصنوعی از طراحی تا حاکمیت دارد.

• سیستم‌های عامل محور، شکاف اندازه‌گیری و حاکمیت را به طور قابل توجهی افزایش می‌دهند. عامل‌ها به نمایندگی از انسان‌ها با تأثیر مستقیم در دنیای واقعی عمل می‌کنند، اما روش‌های نظارتی که با ظرفیت یک عامل برای اقدام مستقل و رفتار نوظهور کالیبره شده‌اند، توسعه نیافته‌اند. متاسفانه ارزیابی‌های موجود به طور سیستماتیک ریسک عامل را به اشتباه اندازه‌گیری می‌کنند [106,360].

• نظارت انسانی به عنوان یک نیاز قابل اندازه‌گیری عملیاتی نشده است [361]؛ ریسک‌های نوظهور در تعامل چند هوش مصنوعی عامل محور را نمی‌توان از طریق ارزیابی تک عاملی تشخیص داد [362,363] و روش‌های قابل اعتماد برای حفظ کنترل بر سیستم‌های بسیار خودمختار همچنان توسعه نیافته‌اند [364].

• یک رویکرد متعادل به حاکمیت هوش مصنوعی، طیف وسیعی از ابزارها را در بر می‌گیرد که شامل ترکیبی از قوانین سخت (قوانین الزام‌آور، مقررات بخشی، محیط آزمایشی تحت نظارت) با مکانیسم‌های قوانین نرم (قوانین

اخلاقی، تعهدات داوطلبانه توسعه‌دهندگان، اتحادهای صنعتی، استانداردهای فنی، دستورالعمل‌های مورد تایید دولت) است.

• ابزارهای فعلی حاکمیت هوش مصنوعی پراکنده، متمرکز در سطح شرکتی و ناکافی هستند [21]. بیش از 40 نوع ابزار وجود دارد، اما نه سیستماتیک هستند و نه جامع و به ندرت اثربخشی دنیای واقعی را اندازه‌گیری می‌کنند. برخی نیز هیچ ابزار اندازه‌گیری ندارند؛ برخی دیگر فقط ورودی‌ها را اندازه‌گیری می‌کنند [365]. بدون اندازه‌گیری موثر، خطرات حاکمیتی نمادین می‌شوند.

• بسترهایی برای گفتگوی ساختاریافته بین توسعه‌دهندگان پیشرو هوش مصنوعی، کشورهای عضو و جامعه علمی بسیار مهم هستند. چنین بحث‌هایی در حال حاضر در اجلاس‌های هوش مصنوعی (بلچلی، سئول، پاریس، دهلی نو) یا کنفرانس‌ها (کنفرانس جهانی هوش مصنوعی، اجلاس جهانی هوش مصنوعی) برگزار می‌شود. سایر حوزه‌ها در کنار آنها فعالیت می‌کنند: نهادهای استاندارد (سازمان بین‌المللی استانداردسازی/کمیته فنی مشترک کمیسیون بین‌المللی الکتروتکنیک ۱، کمیته فرعی ۴۲ (هوش مصنوعی) (ISO/IEC JTC 1/SC 42)، انستیتو مهندسان برق و الکترونیک (IEEE))، مشارکت هوش مصنوعی همه منظوره OECD، شبکه اتحاد بین‌المللی هوش مصنوعی، شبکه موسسات ایمنی هوش مصنوعی و انجمن مدل Frontier به رهبری صنعت – اما متاسفانه تک تک آنها کاملا موضعی، جزئی و موردی هستند. فرآیندهای پایدارتری مورد نیاز است.

• یک پلتفرم بر اساس سازمان ملل متحد نیز یکی از گزینه‌های امیدوارکننده برای میزبانی این گفتگو به صورت مداوم، جهانی و فراگیر است که مکمل حوزه‌های فوق می‌باشد.

هوش مصنوعی متن‌باز

هوش مصنوعی متن‌باز ستون مهمی از چشم‌انداز فناوری مدرن است و می‌تواند نوآوری جهانی توزیع‌شده را تسریع کند [366]. مدل‌های هوش مصنوعی متن‌باز با توانمندسازی توسعه‌دهندگان برای کاهش منابع محاسباتی عظیم و تطبیق سیستم‌های پیشرفته با زمینه‌های محلی، مزایای اجتماعی گسترده‌ای را ارائه می‌دهند. در کشورهای جنوب جهان، مدل‌های وزنی باز به توسعه‌دهندگان این امکان را داده‌اند که مدل‌های پایه بسیار توانمند را برای شرایط محیطی محلی بهینه‌سازی کنند و کاربردهای حیاتی در کشاورزی و مراقبت‌های بهداشتی، مانند پیش‌بینی عملکرد محصول [367] و تصویربرداری پزشکی در محل مراقبت [368] را امکان‌پذیر سازند. مصنوعات مشترک همچنین تأثیر زیست‌محیطی کلی هوش مصنوعی را کاهش می‌دهند. علاوه بر این، شفافیت ساختاری چنین سیستم‌هایی، اعتماد عمومی را تسهیل می‌کند، زیرا نظارت و حسابرسی خارجی را امکان‌پذیر می‌سازند [369].

توسعه تاریخی

توسعه تاریخی هوش مصنوعی متن‌باز، تغییر از سیستم‌های شرکتی بسته به سمت نوآوری توزیع‌شده و مشارکتی در مقیاس جهانی را نشان می‌دهد. یک نمونه برجسته، توسعه مدل متن‌باز BLOOM در سال 2022 بود که توسط کنسرسیوم BigScience [370] تولید شد و پس از آن انتشار خانواده قدرتمند مدل‌های Llama با وزن باز [371] و انتشار سری Gemma رخ داد. رقابت در این حوزه باعث شد تا اکوسیستم جایگزین و بسیار رقابتی توسعه‌دهندگان چینی یعنی Qwen [372,373] و بعد از آن DeepSeek-V3 [374] و R1 [375] منتشر شود. در حال حاضر، کل بسیاری از مناطق و کشورها نیز به طور فعال در توسعه هوش مصنوعی با وزن باز مشارکت دارند: Mistral در اروپا [376]، Falcon در امارات متحده عربی [377]، GigaChat [378] و YandexGPT [379] در فدراسیون روسیه، در کنار پروژه‌هایی در هند [380]، ژاپن [381]، جمهوری کره [382] و سایر کشورها.

چالش‌های کنترل و مدیریت ریسک

کنترل مدل‌های متن‌باز با قابلیت بالا دشوار است [383]. دسترسی محدود یا محدود شده پس از انتشار یک مدل در دامنه باز، امکان‌پذیر نیست و احتمال کاربرد مخرب (به عنوان مثال، تسهیل آسیب‌های سایبری مداوم، مانند تولید خودکار بدافزارهای پیچیده) را باقی می‌گذارد [21,384]. برای اطمینان از اینکه جوامع محلی می‌توانند با خیال راحت سیستم‌های هوش مصنوعی را تطبیق داده و ارزیابی کنند، به ظرفیت‌سازی جهانی قابل توجهی نیاز است. محققان به صورتی پیوسته از پروتکل‌های اندازه‌گیری و ارزیابی دقیق برای ارزیابی ایمنی یک مدل قبل از انتشار حمایت می‌کنند [385]. نهادهای بین‌المللی مانند سازمان ملل متحد می‌توانند نقش مهمی در هماهنگی استانداردهای جهانی ایفا کنند و اطمینان حاصل نمایند که انگیزه برای نوآوری باز در برابر نیاز به معیارهای ایمنی هماهنگ و اندازه‌گیری قوی خطرات تغییرناپذیر متعادل باقی می ماند.

۴. شکاف‌ها و مراحل بعدی

۱.۴ شکاف‌های شواهد

شواهدی که پنل در ارتباط با برخی از جنبه های هوش مصنوعی در اختیار دارد، نامتوازن یا ناکافی هستند. موارد زیر نمونه‌هایی از حوزه‌هایی می باشند که پنل هنوز نمی‌تواند در آنها نتیجه‌گیری‌های علمی مطمئنی داشته باشد.

• اقتصاد کلان و بهره‌وری

علم هنوز نمی‌تواند با اطمینان بگوید که آیا دستاوردهای بهره‌وری هوش مصنوعی می تواند به دستاوردهای کل اقتصاد منجر شود یا خیر. پیش‌بینی‌ها به دلیل فرضیات مختلف در مورد پذیرش و ایجاد مشاغل و وظایف جدید، به طور قابل توجهی متفاوت هستند. شواهد فعلی قادر هستند به راحتی کاهش هزینه ها در وظایف و مشاغل موجود را اندازه گیری نمایند ولی تعیین اینکه چه میزان از این کاهش در کالاها، خدمات و بازارهای جدید سهم هوش مصنوعی بوده است، دشوار می باشد.

• اثرات بر بازار کار

تحقیقات فعلی امکان نتیجه‌گیری روشنی در مورد شکل تاثیرات هوش مصنوعی را در بازار کار فراهم نمی‌کند. شواهد تاریخی نشان می‌دهند که اقتصادها می‌توانند کارهای جدید ایجاد کنند، اما این امر به طور گسترده ای رخ نداده است و در عین حال ممکن است مشاغل جدید کیفیت بالایی هم نداشته باشند.

• استفاده مخرب از فناوری‌های شیمیایی و بیولوژیکی توسط بازیگران غیردولتی

مطالعات نشان می‌دهند که هوش مصنوعی به تدریج آستانه تخصص برای توسعه و استقرار عوامل مهندسی زیستی را کاهش می‌دهد. با این حال، میزان واقعی این خطر و شرایطی که تحت آن می‌تواند منجر به ایجاد بیماری‌های همه‌گیر عمدی شود، هنوز به خوبی درک نشده است.

• محیط زیست و منابع

گسترش سریع هوش مصنوعی، تقاضا برای زیرساخت‌های دیجیتال، افزایش مصرف انرژی و آب، انتشار گازهای گلخانه‌ای، فشار بر زنجیره‌های تأمین مواد معدنی حیاتی و زباله‌های الکترونیکی [386] را افزایش می‌دهد. با این حال، اندازه‌گیری ها و گزارش های استاندارد که بتوانند موارد فوق را در کل چرخه عمر هوش مصنوعی روشن نمایند کماکان وجود ندارند.

• زنجیره تأمین جهانی هوش مصنوعی

این زنجیره شامل استخراج مواد اولیه، تولید تراشه، جمع‌آوری و پردازش داده‌ها، آموزش مدل، زیرساخت، و ایجاد سیستم مدیریت ضایعات سخت‌افزاری در سراسر کشورها است. برای درک بهتر اثرات کامل آن نیازمند تحقیقات بیشتری هستیم.

• اثربخشی ابزارهای حاکمیتی

این پنل، ابزارهای حاکمیتی را در لایه‌های شرکتی، ملی و بین‌المللی فهرست‌بندی کرده است ولی شواهد اثربخشی آنها در دنیای واقعی کماکان اندک می باشند.

• اثرات هوش مصنوعی در سطوح فردی و جمعی

گردآوری و مطالعه شواهد تأثیر هوش مصنوعی بر شکوفایی و استقلال فرهنگی و انسانی هنوز در مراحل ابتدایی خود می باشند. مسیر تعاملات هوش مصنوعی در رسیدن از سطح فردی تا ظهور پیامدهای آن در سطوح اجتماعی مانند فرسایش معرفتی، مشارکت مدنی و انسجام اجتماعی همچنان به خوبی درک نشده است. شواهد موجود، تصاویری لحظه‌ای - معیارهای تعامل، آسیب‌های مستند شده، مطالعات موردی - را ثبت می‌کنند، اما مسیر و فرآیند این تبدیل را نشان نمی‌دهند.

۲.۴ محدوده اختیارات

کاربردهای نظامی هوش مصنوعی و سیستم‌های تسلیحاتی خودکار کشنده در گزارش حاضر توسط پنل بررسی نشده است. قطعنامه 325/79 مجمع عمومی به صراحت فعالیت‌های این پنل را به حوزه غیرنظامی محدود می‌کند: «فعالیت‌های پنل و گفتگو به حوزه غیرنظامی محدود می‌شود و به هوش مصنوعی برای اهداف نظامی اشاره نمی‌کند». به همین دلیل، خطرات مربوط به استفاده مخرب از فناوری‌های شیمیایی و بیولوژیکی تا حدی که به حوزه نظامی مربوط می‌شوند، نیز خارج از اختیارات این پنل در نظر گرفته می‌شوند.

۳.۴ مراحل بعدی

این گزارش اولیه، آغاز کار پنل است، نه پایان آن. پنل از طریق مشورت‌های ساختاریافته و تعامل نزدیک با جامعه علمی، به تعمیق پایگاه شواهد علمی خود ادامه خواهد داد. گام‌های مشخص بعدی عبارتند از:

1. خلاصه‌های موضوعی. قطعنامه 325/79 مجمع عمومی صراحتاً صدور خلاصه‌های موضوعی را علاوه بر گزارش سالانه پیش‌بینی می‌کند.

این پنل قصد دارد خلاصه‌های تخصصی در مورد مسائل فوری که مطرح می‌شوند، از جمله (اما نه محدود به) موارد زیر، منتشر کند: هوش مصنوعی و محیط زیست؛ هوش مصنوعی و ایمنی کودکان؛ ابزارهای حاکمیت هوش مصنوعی و ارزیابی اثربخشی آنها، یا خلاصه‌های بخشی گسترده‌تر در مورد کاربردهای هوش مصنوعی در فضا، سیستم‌های کوانتومی، حقوقی و قضایی و همچنین بازارهای مالی.

2. ورودی از گفتگوی جهانی.

پنل نتایج گفتگوی جهانی در مورد حاکمیت هوش مصنوعی را در نظر خواهد گرفت و آماده است تا وظایف جدیدی را که توسط کشورهای عضو تعیین شده است، بر عهده بگیرد.

درباره پنل علمی بین‌المللی مستقل در مورد هوش مصنوعی

ترکیب و ماموریت

پنل علمی بین‌المللی مستقل در مورد هوش مصنوعی در سازمان ملل متحد از طریق قطعنامه 325/79 مجمع عمومی، همانطور که از طریق پیمان جهانی دیجیتال و پیمان برای آینده متعهد شده است، تأسیس شد.

این پنل متشکل از 40 متخصص مستقل است که بر اساس تخصص خود در هوش مصنوعی و زمینه‌های مرتبط، توسط مجمع عمومی برای مدت سه سال منصوب شده‌اند. این پنل از نظر جنسیتی متعادل است و شامل اعضایی از هر پنج گروه منطقه‌ای کشورهای عضو در رشته‌های مختلف از جمله هسته فنی هوش مصنوعی، کاربرد های هوش مصنوعی، ایمنی و زیرساخت، و سیاست، اخلاق و تأثیرگذاری هوش مصنوعی است.

این پنل موظف به انتشار ارزیابی‌های علمی مبتنی بر شواهد می باشد. در عین حال بایستی تحقیقات موجود مربوط به فرصت‌ها، خطرات و تأثیرات هوش مصنوعی را از طریق یک گزارش خلاصه سالانه مرتبط با سیاست گذاری (اما غیر تجویزی)، که شامل خلاصه‌های موضوعی می گردند، در صورت لزوم ترکیب و تجزیه و تحلیل کند. کار علمی آن باید با اصول استقلال، اعتبار و دقت علمی، مشارکت چند رشته‌ای و فراگیر هدایت شود. این پنل همچنین موظف است گزارش خلاصه سالانه خود را در گفتگوی جهانی سازمان ملل متحد در مورد حاکمیت هوش مصنوعی ارائه نماید. این پنل با آگاه‌سازی گفتگوی جهانی و فرآیندهای بین‌المللی گسترده‌تر، جامعه جهانی را قادر می‌سازد تا چالش‌های نوظهور را پیش‌بینی کند، تصمیمات حاکمیتی آگاهانه‌تری اتخاذ نماید و زمینه گسترده اطلاعات را برای سیاست‌گذاران در سراسر جهان هموار سازد. این پنل توسط دبیرخانه آن که توسط دفتر فناوری‌های دیجیتال و نوظهور سازمان ملل متحد هماهنگ می‌شود، پشتیبانی می‌شود.

روند رسیدن به گزارش حاضر

در سه ماه پس از اولین جلسه پنل در مارس 2026، پس از انتصاب آنها در فوریه 2026، پنل به طور فشرده برای ارائه گزارش حاضر تلاش کرد. این فرآیند فشرده تبادل علمی و تحلیل جمعی شامل یک جلسه عمومی حضوری سه روزه و بیش از 60 جلسه مجازی پنل بود که توسط روسای مشترک منتخب آن تسهیل شد. این گزارش اولیه، پایه و اساسی برای مشاوره گسترده‌تر با کارشناسان خارجی و خلاصه‌های موضوعی بعدی و گزارش‌های خلاصه سالانه را ایجاد می‌کند.

استقلال علمی پنل

اعضای پنل به عنوان کارشناسان علمی مستقل، به صورت شخصی در خدمت آن هستند. هر یک از آنها با انتصاب خود به عنوان کارشناسان سازمان ملل متحد در ماموریت، اعلام و قول داده‌اند که در رابطه با انجام وظایف خود از هیچ دولت یا منبع دیگری دستورالعملی را جستجو یا قبول نکنند. مقررات حاکم بر جایگاه، حقوق اساسی و وظایف کارشناسان در ماموریت (ST/SGB/2002/9) همچنین شامل مقرراتی در مورد رفتار و پاسخگویی آنها می‌شود.

مراجع و ماخذ:

1. Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. Science, 358(6370), 1530–1534. https://doi.org/10.1126/science.aap8062

2. Schölkopf, B. (2022). Causality for machine learning. In H. Geffner, R. Dechter, & J. Y. Halpern (Eds.), Probabilistic and causal inference: The works of Judea Pearl (pp. 765–804). ACM.

3. Hu, K. (2023, February 2). ChatGPT sets record for fastest-growing user base—Analyst note. Reuters. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

4. AI Alliance Network. (2025). AI horizons: What will AI technologies look like in 10 years? A research project. AI Alliance Network.

5. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. arXiv preprint arXiv:2206.07682. https://arxiv.org/abs/2206.07682

6. METR. (2025, March 19). Measuring AI ability to complete long tasks. https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/

7. Crafts, N. (2021). Artificial intelligence as a general-purpose technology: An historical perspective. Oxford Review of Economic Policy, 37(3), 521–536. https://academic.oup.com/oxrep/article/37/3/521/6374675

8. Bottou, L., & Schölkopf, B. (2025). The fiction machine. SIAM News, 58(3).

9. Costa-Gomes, B., Tolmachev, P., Taysom, E., Sounderajah, V., Richardson, H., Schoenegger, P., & King, D. (2026). Public use of a generalist LLM chatbot for health queries. Nature Health, 1–8.

10. Bengio, Y., Clare, S., Prunkl, C., et al. (2026). International AI Safety Report 2026. International AI Safety Report. https://internationalaisafetyreport.org/

11. Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., ... & Chan, L. (2026). Measuring AI ability to complete long software tasks. Advances in Neural Information Processing Systems, 38, 92213-92266.

12. Anthropic. (2025). Responsible scaling policy. https://www.anthropic.com/rsp

13. OpenAI. (2025). Preparedness framework version 2. https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

14. Google DeepMind. (2025). Frontier safety framework. https://deepmind.google/blog/updating-the-frontier-safety-framework/

15. Dragan, A., et al. (2024). Introducing the frontier safety framework. Google DeepMind. https://deepmind.google/blog/introducing-the-frontier-safety-framework/

16. Anthropic. (2026, April 7). Claude Mythos Preview (Frontier Red Team technical report). https://red.anthropic.com/2026/mythos-preview/

17. Anthropic. (2026, April 7). Project Glasswing: Securing critical software for the AI era. https://www.anthropic.com/glasswing

18. Stanford Institute for Human-Centered AI. (2026). AI Index report 2026. Stanford University. https://hai.stanford.edu/ai-index/2026-ai-index-report

19. Scale AI. (2025). Humanity's last exam. https://labs.scale.com/leaderboard/humanitys_last_exam

20. Epoch AI. (2026). AI capabilities. https://epoch.ai/benchmarks

21. Bengio, Y., Clare, S., Prunkl, C., et al. (2026). International AI Safety Report 2026. arXiv. https://doi.org/10.48550/arXiv.2602.21012

22. Xu, C., Guan, S., Greene, D., & Kechadi, M.-T. (2024). Benchmark data contamination of large language models: A survey. arXiv. https://arxiv.org/abs/2406.04244

23. Akhtar, M., Reuel, A., Soni, P., Ahuja, S., Ammanamanchi, P. S., Rawal, R., et al. (2026). When AI benchmarks plateau: A systematic study of benchmark saturation. arXiv. https://arxiv.org/abs/2602.16763

24. Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. Patterns, 5(5), Article 100988. https://doi.org/10.1016/j.patter.2024.100988

25. Needham, J., Edkins, G., Pimpale, G., Bartsch, H., & Hobbhahn, M. (2025). Large language models often know when they are being evaluated. arXiv. https://arxiv.org/abs/2505.23836

26. Van Der Weij, T., Hofstätter, F., Jaffe, O., Brown, S., & Ward, F. (2025, May). Ai sandbagging: Language models can strategically underperform on evaluations. In International Conference on Learning Representations (Vol. 2025, pp. 73152-73189).

27. Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., ... & Maharaj, T. (2023, June). Harms from increasingly agentic algorithmic systems. In Proceedings of the 2023 ACM conference on fairness, accountability, and transparency (pp. 651-666).

28. Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., ... & Rahwan, I. (2025). Multi-agent risks from advanced ai. arXiv preprint arXiv:2502.14143.

29. Folkerts, L., Payne, W., Inman, S., Giavridis, P., Skinner, J., Deverett, S., et al. (2026). Measuring AI agents' progress on multi-step cyber attack scenarios. arXiv. https://arxiv.org/abs/2603.11214

30. Patwardhan, T., Dias, R., Proehl, E., Kim, G., Wang, M., Watkins, O., et al. (2025). GDPval: Evaluating AI model performance on real-world economically valuable tasks. arXiv. https://arxiv.org/abs/2510.04374

31. Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., et al. (2025). Chain of thought monitorability: A new and fragile opportunity for AI safety. arXiv. https://arxiv.org/abs/2507.11473

32. Azaria, A., & Mitchell, T. (2023). The internal state of an LLM knows when it's lying. arXiv. https://arxiv.org/abs/2304.13734

33. Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., et al. (2024). Black-box access is insufficient for rigorous AI audits. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3630106.3659037

34. Tamkin, A., McCain, M., Handa, K., Durmus, E., Lovitt, L., Rathi, A., et al. (2024). Clio: Privacy-preserving insights into real-world AI use. arXiv. https://arxiv.org/abs/2412.13678

35. European Commission. (2025). AI Act: Draft guidance and reporting template for serious AI incidents. https://digital-strategy.ec.europa.eu/en/consultations/ai-act-commission-issues-draft-guidance-and-reporting-template-serious-ai-incidents-and-seeks

36. Organisation for Economic Co-operation and Development. (n.d.). AI Incident Monitor (AIM). https://oecd.ai/en/catalogue/tools/ai-incident-database

37. MIT FutureTech. (n.d.). AI Risk Repository / Incident Tracker. https://airisk.mit.edu

38. Organisation for Economic Co-operation and Development. (2025). Competition in artificial intelligence infrastructure (OECD Roundtables on Competition Policy Papers No. 330). OECD Publishing.

39. Sajadieh, S., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Santarlasci, L., et al. (2026). AI Index report 2026. Stanford Institute for Human-Centered AI.

40. Pilz, K. F., Sanders, J., Rahman, R., & Heim, L. Trends in AI Supercomputers. In ICML Workshop on Technical AI Governance (TAIG).

41. Kalluri, P. R., Agnew, W., Cheng, M., et al. (2025). Computer-vision research powers surveillance technology. Nature, 643, 73–79. https://doi.org/10.1038/s41586-025-08972-6

42. Office of the United Nations High Commissioner for Human Rights. (2022). The right to privacy in the digital age (A/HRC/51/17).

43. Teklehaymanot, H. K., & Nejdl, W. (2025). Tokenization disparities as infrastructure bias: How subword systems create inequities in LLM access and efficiency. arXiv preprint arXiv:2510.12389.

44. Occhini, G., Tanaka-Ishii, K., Barford, A., Tikochinski, R., Hu, S., Reichart, R., ... & Korhonen, A. (2026). Artificial intelligence is creating a new global linguistic hierarchy. arXiv. https://arxiv.org/abs/2602.12018

45. Alhanai, T., Kasumovic, A., Ghassemi, M. M., Zitzelberger, A., Lundin, J. M., & Chabot-Couture, G. (2025, April). Bridging the gap: enhancing LLM performance for low-resource African languages with new benchmarks, fine-tuning, and cultural adjustments. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 27, pp. 27802-27812).

46. Bhutani, M., Robinson, K., Prabhakaran, V., Dave, S., & Dev, S. (2024). SeeGULL multilingual: A dataset of geo-culturally situated stereotypes. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Vol. 2, pp. 842–854).

47. Cazzaniga, M., Jaumotte, F., Li, L., Melina, G., Panton, A. J., Pizzinelli, C., & Tavares, M. M. (2024). Gen-AI: Artificial intelligence and the future of work (IMF Staff Discussion Note SDN/2024/001). International Monetary Fund.

48. McElheran, K., Yang, M.-J., Kroff, Z., & Brynjolfsson, E. (2024). The rise of industrial AI in America: Microfoundations of the productivity J-curve(s) (NBER Working Paper No. 32937). National Bureau of Economic Research.

49. Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 469–481). ACM. https://doi.org/10.1145/3351095.3372828

50. Skirzynski, J., Danks, D., & Ustun, B. (2025). Discrimination exposed? On the reliability of explanations for discrimination detection. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (pp. 2554–2569).

51. Zollo, T., Rajaneesh, N., Zemel, R., Gillis, T., & Black, E. (2025). Towards effective discrimination testing for generative AI. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (pp. 1028–1047).

52. Lazard, L., Capdevila, R., Turley, E. L., Gilfoyle, K., & Stavropoulou, N. (2025). Deepfake Technology and Gender-Based Violence: A Scoping Review. Trauma, Violence, & Abuse, 15248380251384271.

53. Posetti, J., Williams, K., Hellmueller, L., Renaud, P., Shabbir, N., & Aboulez, N. (2026). Tipping point: Online violence impacts, manifestations and redress in the AI age. UN Women.

54. Brynjolfsson, E., Chandar, B., & Chen, R. (2025). Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence. Stanford Digital Economy Lab.

55. Humlum, A., & Vestergaard, E. (2025). Large language models, small labor market effects (NBER Working Paper No. 33777). National Bureau of Economic Research.

56. Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381(6654), 187–192. https://doi.org/10.1126/science.adh2586

57. Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. Quarterly Journal of Economics.

58. International Monetary Fund. (2024). Broadening the gains from generative AI: The role of fiscal policies (IMF Staff Discussion Note).

59. Organisation for Economic Co-operation and Development. (2026). The OECD.AI index. OECD Publishing.

60. Attard-Frost, B., & Lyons, K. (2025). AI governance systems: A multi-scale analysis framework, empirical findings, and future directions. AI and Ethics, 5(3), 2557-2604.

61. UNESCO. (2023). Readiness assessment methodology. UNESCO.

62. Hawkins, Z. J., Lehdonvirta, V., & Wu, B. (2025). AI compute sovereignty: Infrastructure control across territories, cloud providers, and accelerators. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5312977

63. Markus, A., Carolus, A., & Wienrich, C. (2025). Objective measurement of AI literacy: Development and validation of the AI Competency Objective Scale (AICOS). Computers and Education: Artificial Intelligence.

64. Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion.

65. Math Matters AI. (n.d.). Math Matters AI. https://www.mathmatters.ai

66. Hinojosa, T., Rapaport, A., Jaciw, A., & Zacamy, J. (2016). Exploring the foundations of the future STEM workforce: K–12 indicators of postsecondary STEM success. Regional Educational Laboratory Southwest. https://ies.ed.gov/use-work/resource-library/report/systematic-literature-review/exploring-foundations-future-stem-workforce-k-12-indicators-postsecondary-stem-success

67. United Nations Conference on Trade and Development. (2025). Technology and innovation report 2025: Inclusive artificial intelligence for development. United Nations. https://doi.org/10.18356/9789211068016

68. Chauhan, P. (2025). AI and human rights: Global South perspectives. International Journal of Humanities Social Science and Management, 5(4), 563–568. https://ijhssm.org/issue_dcp/AI%20and%20Human%20Rights%20%20Global%20South%20Perspectives.pdf

69. Roberts, H., Hine, E., Taddeo, M., & Floridi, L. (2024). Global AI governance: Barriers and pathways forward. International Affairs, 100(3), 1275–1286. https://doi.org/10.1093/ia/iiae073

70. Khodabin, M., & Arsalani, A. (2025). Artificial intelligence literacy as national strategy: A systematic review of policy, equity, and capacity building across the Global South. World Studies in Policy Sciences, 9, 777–814. https://doi.org/10.22059/wsps.2025.396472.1530

71. Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., et al. (2024). On the societal impact of open foundation models. arXiv. https://arxiv.org/abs/2403.07918

72. Grinstead, B., Holler, C., & Braun, F. (2026, May). Behind the scenes hardening Firefox with Claude Mythos Preview. Mozilla Hacks. https://hacks.mozilla.org/2026/05/behind-the-scenes-hardening-firefox/

73. Wang, W., Tu, Z., Chen, C., Yuan, Y., Huang, J.-T., Jiao, W., & Lyu, M. R. (2024). All languages matter: On the multilingual safety of LLMs. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 5865–5877). https://doi.org/10.18653/v1/2024.findings-acl.349

74. Nigatu, H. H., Mehandru, N., Abadi, N. H., Gebremeskel, B., Alaa, A., & Choudhury, M. (2025). Viability of machine translation for healthcare in low-resourced languages. In Proceedings of EMNLP 2025 (pp. 10584–10598).

75. Cazzaniga, M., Jaumotte, F., Li, L., Melina, G., Panton, A. J., Pizzinelli, C., Rockall, E., & Tavares, M. M. (2024). The global impact of AI: Mind the gap (IMF Working Paper No. 24/136). International Monetary Fund.

76. World Bank. (2025). Digital progress and trends report 2025: Strengthening AI foundations. World Bank.

77. McElheran, K., Yang, M.-J., Kroff, Z., & Brynjolfsson, E. (2024). The rise of industrial AI in America: Microfoundations of the productivity J-curve(s) (NBER Working Paper No. 32937).

78. Brynjolfsson, E., Rock, D., & Syverson, C. (2017). Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. NBER Working Paper No. 24001

79. Calvino, F., & Fontanelli, L. (2023). A portrait of AI adopters across countries: Firm characteristics, assets' complementarities and productivity. OECD Science, Technology and Industry Working Papers, No. 2023/11.

80. McKinsey & Company. (2026, March 25). State of AI trust in 2026: Shifting to the agentic era. https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/tech-forward/state-of-ai-trust-in-2026-shifting-to-the-agentic-era

81. Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., & Vardoulakis, L. M. (2020, April). A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In Proceedings of the 2020 CHI conference on human factors in computing systems (pp. 1-12).

82. Brant, A., Singh, P., Yin, X., et al. (2025). Performance of a deep learning diabetic retinopathy algorithm in India. JAMA Network Open, 8(3), e250984. https://doi.org/10.1001/jamanetworkopen.2025.0984

83. UNESCO. (2025). AI and the future of education: disruptions, dilemmas and directions

84. Cristia, J. et al. (2017). Technology and child development: evidence from the One Laptop per Child program. *American Economic Journal: Applied Economics*, 9(3), 295–320.

85. Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. Societies, 15(1), Article 6. https://doi.org/10.3390/soc15010006

86. Ojija, F., Ogwu, M. C., Ally, J., John, J. P., Stephano, A., Felix, N., & Tekka, R. (2025). Artificial intelligence-driven solutions for mitigating human–wildlife conflict in biodiversity hotspots. Science Progress, 108(4), 00368504251394584.

87. Noy, S. & Zhang, W. (2023). "Experimental evidence on the productivity effects of generative artificial intelligence." Science, 381(6654), 187–192. https://doi.org/10.1126/science.adh2586.

88. Cui, Z., Demirer, M., Jaffe, S., Musolff, L., Peng, S. & Salz, T. (2026). "The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers." Management Science. https://doi.org/10.1287/mnsc.2025.00535

89. Dell'Acqua, F., McFowland, E. III, Mollick, E., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F. & Lakhani, K. R. (2023). "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." Harvard Business School Working Paper 24-013. SSRN: https://ssrn.com/abstract=4573321

90. Ali, Z., Muhammad, A., Lee, N., Waqar, M., & Lee, S. W. (2025). Artificial Intelligence for Sustainable Agriculture: A Comprehensive Review of AI-Driven Technologies in Crop Production. Sustainability, 17(5), 2281. https://doi.org/10.3390/su17052281

91. Dhal, S. B., & Kar, D. (2024). Transforming Agricultural Productivity with AI-Driven Forecasting: Innovations in Food Security and Supply Chain Optimization. Forecasting, 6(4), 925–951. https://doi.org/10.3390/forecast6040046

92. Pearlman, K., Wan, W., Shah, S., & Laiteerapong, N. (2025). Use of an AI scribe and electronic health record efficiency. JAMA Network Open, 8(10), e2537000.

93. Afshar, M., Ryan Baumann, M., Resnik, F., Hintzke, J., Gravel Sullivan, A., Wills, G., ... & Gordon, J. (2025). A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. NEJM AI, 2(12), Aloa2500945.

94. Tierney, A. A., Gayre, G., Hoberman, B., Mattern, B., Ballesca, M., Wilson Hannay, S. B., ... & Lee, K. (2025). Ambient artificial intelligence scribes: learnings after 1 year and over 2.5 million uses. NEJM Catalyst Innovations in Care Delivery, 6(5), CAT-25.

95. Paul A. David (1990, May). The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox. The American Economic ReviewVol. 80, No. 2, Papers and Proceedings of the Hundred and Second Annual Meeting of the American Economic Association, pp. 355-361

96. Brynjolfsson, E., & Hitt, L. M. (2000). Beyond computation: Information technology, organizational transformation and business performance. Journal of Economic Perspectives, 14(4), 23–48. https://doi.org/10.1257/jep.14.4.23

97. Shaw, S. D., & Nave, G. (2026). Thinking—fast, slow, and artificial: How AI is reshaping human reasoning and the rise of cognitive surrender. SSRN. https://doi.org/10.2139/ssrn.6097646

98. Bauer, E., Greiff, S., Graesser, A. C., Scheiter, K., & Sailer, M. (2025). Looking beyond the hype: Understanding the effects of AI on learning. Educational Psychology Review, 37(2), 45.

99. Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., Van Smeden, M., & Boulesteix, A.-L. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078451. https://doi.org/10.1136/bmj-2024-078451

100. Rotz, S., Duncan, E., Small, M., Botschner, J., Dara, R., Mosby, I., Reed, M., & Fraser, E. D. G. (2019). The politics of digital agricultural technologies: A preliminary review. Sociologia Ruralis, 59(2), 203–229. https://doi.org/10.1111/soru.12233

101. United Nations General Assembly. (2024). Global Digital Compact (Annex II to the Pact for the Future, A/RES/79/1). United Nations. https://docs.un.org/en/A/RES/79/1

102. UNESCO. (2022). K-12 AI curricula: A mapping of government-endorsed AI curricula. https://www.unesco.org/en/articles/k-12-ai-curricula-mapping-government-endorsed-ai-curricula

103. Almatrafi, O., Johri, A., & Lee, H. (2024, 2024/06/01/). A systematic review of AI literacy conceptualization, constructs, and implementation and assessment efforts (2019–2023). Computers and Education Open, 6, 100173. https://doi.org/https://doi.org/10.1016/j.caeo.2024.100173

104. Ma, M., Ng, D. T. K., Liu, Z., & Wong, G. K. W. (2025, 2025/06/01/). Fostering responsible AI literacy: A systematic review of K-12 AI ethics education. Computers and Education: Artificial Intelligence, 8, 100422. https://doi.org/https://doi.org/10.1016/j.caeai.2025.100422

105. Atias, O., & Mawasi, A. (2025, 2025/12/01/). Conceptualizing AI literacies for children and youth: A systematic review on the design of AI literacy educational programs. Computers and Education: Artificial Intelligence, 9, 100491. https://doi.org/https://doi.org/10.1016/j.caeai.2025.100491

106. Kasirzadeh, A., & Gabriel, I. (2025, April). Characterizing AI Agents for Alignment and Governance. https://arxiv.org/abs/2504.21848

107. Wijk, H., Lin, T. R., Becker, J., Jawhar, S., Parikh, N., Broadley, T., ... & Barnes, E. (2025, October). RE-Bench: Evaluating Frontier AI R&D Capabilities of Language Model Agents against Human Experts. In International Conference on Machine Learning (pp. 66772-66832). PMLR.

108. Chan, J. S., Chowdhury, N., Jaffe, O., Aung, J., Sherburn, D., Mays, E., ... & Weng, L. (2025, May). Mle-bench: Evaluating machine learning agents on machine learning engineering. In International Conference on Learning Representations (Vol. 2025, pp. 50466-50494).

109. Pichai, S. (2026, April 22). Cloud Next '26: Momentum and innovation at Google scale. Google Blog. https://blog.google/innovation-and-ai/infrastructure-and-cloud/google-cloud/cloud-next-2026-sundar-pichai/

110. Cybersecurity and Infrastructure Security Agency. (2025, January 14). CISA, JCDC, government and industry partners publish AI cybersecurity collaboration playbook. U.S. Department of Homeland Security. https://www.cisa.gov/news-events/news/cisa-jcdc-government-and-industry-partners-publish-ai-cybersecurity-collaboration-playbook

111. Liu, Y., Zhao, Y., Lyu, Y., Zhang, T., Wang, H., & Lo, D. (2025). "Your AI, my shell": Demystifying prompt injection attacks on agentic AI coding editors. https://doi.org/10.48550/arXiv.2509.22040

112. Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107(6), 1753–1820. https://doi.org/10.15779/Z38RV0D

113. Schiff, K. J., Schiff, D. S., & Bueno, N. S. (2025). The liar's dividend: Can politicians claim misinformation to evade accountability?. American Political Science Review, 119(1), 71-90.

114. Chan, Y.-T., et al. (2024). Assessing the article screening efficiency of artificial intelligence for systematic reviews. Journal of Dentistry, 149, 105259. https://doi.org/10.1016/j.jdent.2024.105259

115. Delgado-Licona, F., Alsaiari, A., Dickerson, H., Klem, P., Ghorai, A., Canty, R. B., Bennett, J. A., Jha, P., Mukhin, N., Li, J., López-Guajardo, E. A., Sadeghi, S., Bateni, F., & Abolhasani, M. (2025). Flow-driven data intensification to accelerate autonomous inorganic materials discovery. Nature Chemical Engineering, 2, 436–446. https://doi.org/10.1038/s44286-025-00249-z

116. Chan, A., Wei, K., Huang, S., Rajkumar, N., Perrier, E., Lazar, S., Hadfield, G. K., & Anderljung, M. (2025). Infrastructure for AI Agents. http://arxiv.org/abs/2501.10114

117. Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., & Narayanan, A. AI Agents That Matter. Transactions on Machine Learning Research.

118. Zhu, L., Lu, Q., Ding, M. et al. Designing meaningful human oversight in AI. AI Ethics 6, 286 (2026). https://doi.org/10.1007/s43681-026-01147-7

119. Ferrara, E. (2024). GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. Journal of Computational Social Science, 7, 549–569. https://doi.org/10.1007/s42001-024-00250-1

120. Djiré, A. E., Kaboré, A. K., Samhi, J., et al. (2026). Learned or memorized? Quantifying memorization advantage in code LLMs. In Proceedings of the International Conference on Software Engineering. https://arxiv.org/abs/2604.13997

121. Gowal, S., Bunel, R., Stimberg, F., Stutz, D., Ortiz-Jimenez, G., Kouridi, C., ... & Kohli, P. (2025). SynthID-Image: Image watermarking at internet scale. arXiv preprint arXiv:2510.09263.

122. United Nations Human Rights Council. Expert Mechanism on the Right to Development. (2024). AI, cultural rights and the right to development (A/HRC/EMRTD/11/CRP.2). United Nations. https://undocs.org/A/HRC/EMRTD/11/CRP.2

123. Brennan Center for Justice. (2025). Gauging AI threat to free and fair elections. https://www.brennancenter.org/our-work/analysis-opinion/gauging-ai-threat-free-and-fair-elections

124. Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). The levers of political persuasion with conversational AI. Science, 390(6777), eaea3884. https://doi.org/10.1126/science.aea3884

125. Santos, F. P., Lelkes, Y., & Levin, S. A. (2021). Link recommendation algorithms and dynamics of polarization in online social networks. Proceedings of the National Academy of Sciences, 118(50), e2102141118.

126. Cho, J., Ahmed, S., Hilbert, M., Liu, B., & Luu, J. (2020). Do search algorithms endanger democracy? An experimental investigation of algorithm effects on political polarization. Journal of broadcasting & Electronic media, 64(2), 150-172.

127. Feezell, J. T., Wagner, J. K., & Conroy, M. (2021). Exploring the effects of algorithm-driven news sources on political behavior and polarization. Computers in human behavior, 116, 106626.

128. Argyle, L. P. (2025). Political persuasion by artificial intelligence. Science, 390(6777), 983-984.

129. Lin, H., Czarnek, G., Lewis, B., White, J. P., Berinsky, A. J., Costello, T., ... & Rand, D. G. (2025). Persuading voters using human–artificial intelligence dialogues. Nature, 1-8.

130. Myra Cheng et al. (2026). ,Sycophantic AI decreases prosocial intentions and promotes dependence.Science391,eaec8352(2026).DOI:10.1126/science.aec8352

131. Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. Science, 391(6792), eaec8352. https://doi.org/10.1126/science.aec8352

132. Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., ... & Pollak, T. A. (2026). Artificial intelligence-associated delusions and large language models: risks, mechanisms of delusion co-creation, and safeguarding strategies. The Lancet Psychiatry, 13(6), 522-530.

133. Balan, R., & Gumpel, T. P. (2025). ChatGPT Clinical Use in Mental Health Care: Scoping Review of Empirical Evidence. JMIR Mental Health, 12, e81204.

134. Hudon, A., & Stip, E. (2025). Delusional experiences emerging from AI chatbot interactions or "AI Psychosis". JMIR Mental Health, 12(1), e85799.

135. Osler, L. (2026). Hallucinating with AI: distributed delusions and "AI psychosis". Philosophy & Technology, 39(1), 30.

136. Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., ... Kaplan, J. (2021). A general language assistant as a laboratory for alignment (arXiv:2112.00861). arXiv. https://doi.org/10.48550/arXiv.2112.00861

137. Organisation for Economic Co-operation and Development. (2024). Facts not fakes: Tackling disinformation, strengthening information integrity. OECD Publishing. https://doi.org/10.1787/d909ff7a-en

138. Waight, H., Yang, E., Yuan, Y., et al. (2026). State media control influences large language models. Nature. https://doi.org/10.1038/s41586-026-10506-7

139. Kalina Bontcheva (2024). Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities. https://edmo.eu/wp-content/uploads/2023/12/Generative-AI-and-Disinformation_-White-Paper-v8.pdf

140. Laudrain, A. (2026, April 29). When AI governs (dis)information: Five lessons for democracy. Center for Security Studies (CSS), ETH Zurich. https://css.ethz.ch/en/center/CSS-news/2026/04/when-ai-governs-disinformation-five-lessons-for-democracy.html

141. Boine, C. (2023). Emotional attachment to AI companions and European law. MIT Case Studies in Social and Ethical Responsibilities of Computing (Winter 2023). https://doi.org/10.21428/2c646de5.db67ec7f

142. Department of Enterprise, Tourism and Employment. (2026, February 4). General scheme of the Regulation of Artificial Intelligence Bill 2026. Government of Ireland. https://www.gov.ie/en/department-of-enterprise-tourism-and-employment/publications/general-scheme-of-the-regulation-of-artificial-intelligence-bill-2026

143. United States Congress. Senate. (2025). S. 3062—GUARD Act: Guidelines for User Age-verification and Responsible Dialogue Act of 2025 (119th Congress). Congress.gov. https://www.congress.gov/bill/119th-congress/senate-bill/3062/text

144. European Commission. (2026, April 29). Blueprint for an age verification solution to help protect minors online. Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/factpages/blueprint-age-verification-solution-help-protect-minors-online

145. Reuters. (2026, April 24). From Australia to Europe, countries move to curb children's social media access. https://www.reuters.com/legal/government/australia-europe-countries-move-curb-childrens-social-media-access-2026-04-24/

146. Morrin H, Nicholls L, Levin M et al. (2026). Artificial intelligence-associated delusions and large language models: risks, mechanisms of delusion co-creation, and safeguarding strategies. The Lancet Psychiatry, 2026; 13, 522-530

147. Examining the harm of AI chatbots, Hearing before the Subcomm. on Crime and Counterterrorism of the S. Comm. on the Judiciary, 119th Cong. (2025) (testimony of Megan Garcia). https://www.judiciary.senate.gov/download/2025-09-16-pm_-testimony_-garciapdf

148. Solove, D. J. (2025). Artificial intelligence and privacy. Florida Law Review, 77. https://scholarship.law.ufl.edu/flr/vol77/iss1/1

149. Office of the United Nations High Commissioner for Human Rights. (2025). The right to privacy in the digital age (UN Doc. A/HRC/60/45). https://www.ohchr.org/en/documents/thematic-reports/ahrc6045-right-privacy-digital-age-reportoffice-united-nations-high

150. Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. Journal of Financial Economics, 143(1), 30–56. https://doi.org/10.1016/j.jfineco.2021.05.047

151. UNESCO, IRCAI, & University College London. (2024). Challenging systematic prejudices: An investigation into bias against women and girls in large language models. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000388971

152. Shah, S. S. (2024). Gender bias in artificial intelligence: Empowering women through digital literacy. Premier Journal of Artificial Intelligence, 1, 1000088. https://doi.org/10.70389/PJAI.1000088

153. Haider, S. A., Borna, S., Gomez-Cabello, C. A., et al. (2026). The algorithmic divide: A systematic review on AI-driven racial disparities in healthcare. Journal of Racial and Ethnic Health Disparities, 13, 188–217. https://doi.org/10.1007/s40615-024-02237-0

154. International Telecommunication Union. (2025). Joint statement on AI and the rights of the child. International Telecommunication Union. https://www.itu.int/hub/publication/d-str-cyb_joint-2025

155. Johnson, A. K., Winther, D. K., & Bhargava, A. (2026). Artificial intelligence and child sexual exploitation and abuse: Emerging risks and implications for children's rights (Issue brief). United Nations Children's Fund (UNICEF). https://www.unicef.org/media/178571/file/UNICEF%20AI%20CSEA%20Brief_FINAL3.pdf

156. Thiel, D. (2023). Identifying and Eliminating CSAM in Generative ML Training Data and Models. Stanford Digital Repository. Available at https://purl.stanford.edu/kh752sm9123

157. Internet Watch Foundation. (2026). Harm without limits: AI child sexual abuse material through the eyes of our Analysts. https://www.iwf.org.uk/media/hl1nvdti/iwf-ai-csam-report-2026.pdf

158. Goodacre, E.J. and Gibson, J.L. (2026) AI in the early years: Examining the implications of GenAI toys for young children. Cambridge: University of Cambridge (unpublished report, available via Apollo repository).

159. Kurian, N. (2025). Developmentally aligned AI: a framework for translating the science of child development into AI design. AI, Brain and Child, 1(1). https://doi.org/10.1007/s44436-025-00009-z

160. Livingstone, S., & Sylwander, K. R. (2025). Conceptualizing age-appropriate social media to support children's digital futures. British Journal of Developmental Psychology. https://doi.org/10.1111/bjdp.70006

161. Xiao, W., & Gonçalves, A. (2025). Intelligent toys, complex questions: A literature review of artificial intelligence in children's toys and devices. Big Data & Society, 12(4), 20539517251389860.

162. Grimmelikhuijsen, S. (2022). Explaining why the computer says no: Algorithmic transparency affects the perceived trustworthiness of automated decision-making. Public Administration Review, 82(4), 706–718. https://doi.org/10.1111/puar.13483

163. Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. New York University Law Review Online, 94, 15–55. https://ssrn.com/abstract=3333423

164. Mantelero, A. (2022). Human rights impact assessment and AI. In Beyond data: Human rights, ethical and social impact assessment in AI (pp. 45-91). The Hague: TMC Asser Press.

165. Mantelero, A., & Esposito, M. S. (2021). An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems. Computer Law and Security Review, 41. https://doi.org/10.1016/j.clsr.2021.105561

166. United Nations Educational, Scientific and Cultural Organization. (2025). How should children's rights be integrated into AI governance? https://www.unesco.org/en/articles/how-should-childrens-rights-be-integrated-ai-governance

167. Livingstone, S. & Pothong, K. (2025). Child Rights Impact Assessment: A Policy Tool for aRights-Respecting Digital Environment. Policy & Internet.

168. Denain, J.-S., & Barry, A. (2026, April 16). Have AI capabilities accelerated? Epoch AI. https://epoch.ai/blog/have-ai-capabilities-accelerated

169. Model Evaluation & Threat Research (METR). (2026, May 8). Task-completion time horizons of frontier AI models. https://metr.org/time-horizons/

170. Juniewicz, I. (2026, February 26). Hyperscaler capex has quadrupled since GPT-4's release. Epoch AI. https://epoch.ai/data-insights/hyperscaler-capex-trend

171. Epoch AI. (2026, May 28). Data on AI companies. https://epoch.ai/data/ai-companies?view=graph&tab=revenue&showTotal=true

172. Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. Proceedings of the IEEE, 109(5), 612-634.

173. AI Alliance Network. (2025). AI Horizons: What will AI technologies look like in 10 years? A Research Project. November 2025, p. 83. Based on 21 foresight sessions and 32 in-depth interviews with over 270 AI researchers from 36 countries.

174. Gommers, J., Hernström, V., Josefsson, V., Sartor, H., Schmidt, D., Hjelmgren, A., ... & Lång, K. (2026). Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. The Lancet, 407(10527), 505-514.

175. Reichstein, M., et al. (2025). Early warning of complex climate risk with integrated artificial intelligence. Nature Communications, 16, 2564. https://www.nature.com/articles/s41467-025-57640-w

176. Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. Scientific Reports, 15(1), 17458.

177. Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., & Léger, P. M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education. npj Science of Learning, 10(1), 29.

178. Delgado-Licona, F., Alsaiari, A., Dickerson, H., Klem, P., Ghorai, A., Canty, R. B., ... & Abolhasani, M. (2025). Flow-driven data intensification to accelerate autonomous inorganic materials discovery. Nature Chemical Engineering, 2(7), 436-446.

179. Rotz, S., Duncan, E., Small, M., Botschner, J., Dara, R., Mosby, I., ... & Fraser, E. D. (2019). The politics of digital agricultural technologies: a preliminary review. Sociologia ruralis, 59(2), 203-229.

180. Afshar, M., Ryan Baumann, M., Resnik, F., Hintzke, J., Gravel Sullivan, A., Wills, G., ... & Gordon, J. (2025). A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. NEJM AI, 2(12), AIoa2500945.

181. Chen, M., Wu, Y., Ma, J., Jia, X., Gao, C., Zhao, F., & Qiao, Y. (2026). Independent and collaborative performance of large language models and healthcare professionals in diagnosis and triage. npj Digital Medicine.

182. Tao, X., Zhou, S., Ding, K., Li, S., Li, Y., Wu, B., ... & Han, S. (2026). An LLM chatbot to facilitate primary-to-specialist care transitions: a randomized controlled trial. Nature Medicine, 1-9.

183. Costa-Gomes, B., Tolmachev, P., Taysom, E., Sounderajah, V., Richardson, H., Schoenegger, P., ... & King, D. (2026). Public use of a generalist LLM chatbot for health queries. Nature Health, 1-8.

184. Scientific Computing World. (2025, August 5). AI health assistant displays high diagnostic accuracy in tests. https://www.scientific-computing.com/article/sber-healths-gigachat-powered-ai-health-assistant-displays-high-diagnostic-accuracy-tests

185. MASTEL, P. M., de Dieu Nyandwi, J., Rutunda, S., & Kabanda, K. (2025). Mbaza RBC: Deploying and evaluation of an LLM powered Chatbot for Community Health Workers in Rwanda. In Workshop on Large Language Models and Generative AI for Health at AAAI 2025.

186. Mateen, B. A., Williams, G., Korom, R., Mwaniki, P., Emmanuel-Fabula, M., & Agweyu, A. (2026). Learning Effects from A GenAI-based Clinical Decision Support System in Primary Healthcare. medRxiv, 2026-05.

187. OECD (2025), Results from TALIS 2024: The State of Teaching, TALIS, OECD Publishing, Paris, https://doi.org/10.1787/90df6235-en.

188. OECD (2023), PISA 2022 Results (Volume II): Learning During – and From – Disruption, PISA, OECD Publishing, Paris, https://doi.org/10.1787/a97db61c-en .

189. Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., & Léger, P. M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K–12 education. npj Science of Learning, 10(1), 29. https://www.nature.com/articles/s41539-025-00320-7

190. OECD (2023), PISA 2022 Results (Volume II): Learning During – and From – Disruption, PISA, OECD Publishing, Paris, https://doi.org/10.1787/a97db61c-en.

191. OECD (2023), PISA 2022 Results (Volume II): Learning During – and From – Disruption, PISA, OECD Publishing, Paris, https://doi.org/10.1787/a97db61c-en.

192. Vodafone Foundation.(2025) AI in European Schools: A European Report --- comparing seven countries

193. World Food Programme. (2024). HungerMap LIVE: Global hunger monitoring. https://hungermap.wfp.org/

194. Yakov and Partners. (2024). Artificial intelligence in Russia's agricultural sector: Hype or real money? https://yakovpartners.com/publications/ai-in-agriculture/

195. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences, 122(26), e2422633122. https://doi.org/10.1073/pnas.2422633122

196. Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. Scientific Reports, 15, 17458. https://doi.org/10.1038/s41598-025-97652-6

197. Shneiderman, B. (2022). Human-Centered AI. Oxford University Press.

198. Sparling, T. M., Offner, C., Deeney, M., Denton, P., Bash, K., Juel, R., ... & Kadiyala, S. (2024). Intersections of climate change with food systems, nutrition, and health: an overview and evidence map. Advances in Nutrition, 15(9), 100274.

199. Reichstein, M., et al. (2025). Early warning of complex climate risk with integrated artificial intelligence. Nature Communications, 16, 2564. https://doi.org/10.1038/s41467-025-57640-w

200. Becker-Reshef, I., Justice, C., Barker, B., Humber, M., Rembold, F., Bonifacio, R., Zappacosta, M., Budde, M., Magadzire, T., Shitote, C., Pound, J., Constantino, A., Nakalembe, C., Mwangi, K., Sobue, S., Newby, T., Whitcraft, A., Jarvis, I., & Verdin, J. (2020). Strengthening agricultural decisions in countries at risk of food insecurity: The GEOGLAM Crop Monitor for Early Warning. Remote Sensing of Environment, 237, 111553. https://doi.org/10.1016/j.rse.2019.111553

201. Food and Agriculture Organization of the United Nations. (2025). Evidence in action: How anticipatory cash transfers reduce humanitarian needs and strengthen resilience in Somalia. https://www.fao.org/agrifood-economics/publications/detail/en/c/1756210/

202. World Food Programme. (2025). WFP's evidence base on anticipatory action 2015–2024. https://www.wfp.org/publications/wfps-evidence-base-anticipatory-action-2015-2024

203. Nakalembe, C., Kerner, H. R., Zvonkov, I., Humber, M., Galvez, A. S., Venturini, S., & Becker Reshef, I. (2025). A framework for EO based National Agricultural Monitoring for the African context. npj Sustainable Agriculture, 3, 45. https://www.nature.com/articles/s44264-025-00083-z

204. Nowak, A. C., et al. (2024). Opportunities to strengthen Africa's efforts to track national level climate adaptation. Nature Climate Change, 14, 876–882.https://www.nature.com/articles/s41558-024-02054-7

205. Karger, E., Kuusela, O., Abaluck, J., Bryan, K. A., Halperin, B., Jones, T. R., et al. (2026). Forecasting the economic effects of AI (NBER Working Paper No. w35046). National Bureau of Economic Research. https://doi.org/10.3386/w35046

206. Goldman Sachs (2023). Generative AI Could Raise Global GDP by 7 Percent. https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent

207. Anantrasirichai, N., & Bull, D. (2022). Artificial intelligence in the creative industries: a review. Artificial intelligence review, 55(1), 589-656.

208. United Nations News. (2026, February 18). Artists face steep income decline due to AI, UNESCO finds. United Nations. https://news.un.org/en/story/2026/02/1166989

209. Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The productivity J-curve: How intangibles complement general purpose technologies. American Economic Journal: Macroeconomics, 13(1), 333-372.

210. Gmyrek, P., Winkler, H., & Garganta, S. (2024). Buffer or Bottleneck? Employment Exposure to Generative AI and the Digital Divide in Latin America (ILO Working Paper 121 / World Bank Policy Research Working Paper 10863). International Labour Organization & World Bank.

211. Autor, D., Chin, C., Salomons, A., & Seegmiller, B. (2024). New frontiers: The origins and content of new work, 1940–2018. Quarterly Journal of Economics, 139(3), 1399–1465.

212. The Conversation. (2026, March 18). Tech companies are blaming massive layoffs on AI. What's really going on? https://theconversation.com/tech-companies-are-blaming-massive-layoffs-on-ai-whats-really-going-on-278314

213. Society for Human Resource Management. (2026, May). The AI layoffs narrative: Real transformation, or scapegoat? https://www.shrm.org/topics-tools/news/technology/ai-layoffs-transformation-scapegoat

214. OpenAI. (2026, February 27). Company communication accompanying a US$110 billion private funding round [Company communication]

215. Rodrik, D., & Sabel, C. (2020). Building a Good Jobs Economy (HKS Faculty Research Working Paper RWP20-001). Harvard Kennedy School.

216. Acemoglu, D. (2025). The simple macroeconomics of AI. Economic Policy, 40(121), 13–58. Acemoglu's headline figure is a TFP gain of no more than 0.71% over ten years.

217. Korinek, A., & Suh, D. (2024). Scenarios for the Transition to AGI (NBER Working Paper No. 32255). In their full-automation scenario, output rises sharply while wages collapse once automation crosses a critical threshold.

218. Karger, E., Kuusela, O., Abaluck, J., Bryan, K., Halperin, B., Jones, T., et al. (2026). Forecasting the Economic Effects of AI. Federal Reserve Bank of Chicago and Forecasting Research Institute, March 2026.

219. Jones, C. I., & Tonetti, C. (2026). Past Automation and Future AI: How Weak Links Tame the Growth Explosion. Working paper presented at the Bendheim Center for Finance, Princeton University, March 2026.

220. Trammell, P., & Korinek, A. (2025). Economic Growth under Transformative AI (NBER Working Paper No. 31815, revised September 2025).

221. OECD (2024). Artificial Intelligence, Data and Competition (OECD Artificial Intelligence Papers No. 18). OECD Publishing, Paris. https://doi.org/10.1787/e7e88884-en

222. Acemoglu, D., & Restrepo, P. (2026). Automation and rent dissipation: Implications for wages, inequality, and productivity. Quarterly Journal of Economics, 141(2), 1521–1579.

223. Liu, Y., Zhao, Y., Lyu, Y., Zhang, T., Wang, H., & Lo, D. (2025). "Your AI, my shell": Demystifying prompt injection attacks on agentic AI coding editors. arXiv. https://doi.org/10.48550/arXiv.2509.22040

224. Regilme, S. S. F. (2024). Artificial Intelligence Colonialism: Environmental Damage, Labor Exploitation, and Human Rights Crises in the Global South. SAIS Review of International Affairs, 44(2), 75–92. https://muse.jhu.edu/pub/1/article/950958

225. Barnett-Itzhaki, Z. (2026). The water footprint of artificial intelligence: Emerging solutions and governance imperatives. Water Research, 299, 125866. https://doi.org/10.1016/j.watres.2026.125866

226. International Energy Agency. (2025). Global Critical Minerals Outlook 2025 – Analysis (p. 312). https://iea.blob.core.windows.net/assets/ef5e9b70-3374-4caa-ba9d-19c72253bfc4/GlobalCriticalMineralsOutlook2025.pdf

227. Zhu, L., & Lu, Q. (2026). Verifiability-First AI Engineering in the Era of AIware: A Conceptual Framework, Design Principles, and Architectural Patterns for Scalable Verification. Design Principles, and Architectural Patterns for Scalable Verification (January 07, 2026).

228. UN Scientific Advisory Board. (2026, March 19). AI deception: Brief of the Scientific Advisory Board. United Nations. https://www.un.org/scientific-advisory-board/en/ai-deception

229. Bengio, Y., Clare, S., Prunkl, C., Rismani, S., Andriushchenko, M., Bucknall, B., ... & Zhu, L. (2025). International AI Safety Report 2025: First Key Update: Capabilities and Risk Implications. arXiv preprint arXiv:2510.13653.

230. United Nations Secretary-General. (2024). Human rights due diligence guidance on digital technology use. https://www.ohchr.org/sites/default/files/2024-08/digital-technology-use-guidance-sg-1-en.pdf

231. AI Equality Toolbox. (2025). Human rights impact assessment methodology. https://aiequalitytoolbox.com/tools/hria-workbook/

232. Baldé, C. P., Kuehr, R., Yamamoto, T., McDonald, R., D'Angelo, E., Althaf, S., Bel, G., Deubzer, O., Fernandez-Cubillo, E., Forti, V., Gray, V., Herat, S., Honda, S., Iattoni, G., Khetriwal, D. S., Luda di Cortemiglia, V., Lobuntsova, Y., Nnorom, I., Pralat, N., & Wagner, M. (2024). The global e-waste monitor 2024: Electronic waste rising five times faster than documented e-waste recycling. United Nations Institute for Training and Research & International Telecommunication Union. https://ewastemonitor.info/the-global-e-waste-monitor-2024/

233. International Panel on the Information Environment. (2024). Expert survey on the global information environment 2024: Searching for solutions (SFP2024-1). https://www.ipie.info/research/sfp2024-1

234. Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: A review. Social Network Analysis and Mining, 13(1), 30. https://doi.org/10.1007/s13278-023-01028-5

235. James, K., Perveen, S., & Jacobson, B. (2025). Drained by data: The cumulative impact of data centers on regional water stress. Ceres. https://www.ceres.org/resources/reports/drained-by-data-the-cumulative-impact-of-data-centers-on-regional-water-stres

236. Office of the United Nations High Commissioner for Human Rights. (2024). Taxonomy of generative AI-related human rights harms. https://www.ohchr.org/sites/default/files/documents/issues/expression/statements/2025-10-24-joint-declaration-artificial-intelligence.pdf

237. Patel, A. (2025, May 19). Freedom of expression, artificial intelligence and elections. United Nations Educational, Scientific and Cultural Organization (UNESCO) & United Nations Development Programme (UNDP). https://www.undp.org/publications/freedom-expression-artificial-intelligence-and-elections

238. Scharfbillig, M., Lewandowsky, S., Altay, S., Van Alstyne, M., Kozyreva, A., Hertwig, R., Lorenz-Spreen, P., DiResta, R., Valenzuela, S., Egidy, S., Quattrociocchi, W., & Orben, A. (2026). Fractured reality: How democracy can win the global struggle over the information space (JRC144603). Publications Office of the European Union. https://doi.org/10.2760/9358883

239. Observatory on Information and Democracy. (2024). Information ecosystems and troubled democracy: A global synthesis of the state of knowledge on news, media, AI, and data governance. https://observatory.informationdemocracy.org/report/information-ecosystem-and-troubled-democracy/

240. Solove, D. J. (2025). Artificial intelligence and privacy. Florida Law Review, 77. https://scholarship.law.ufl.edu/flr/vol77/iss1/1

241. Office of the United Nations High Commissioner for Human Rights (2025). The right to privacy in the digital age. URL https://www.ohchr.org/en/documents/thematic-reports/ahrc6045-right-privacy-digital-age-reportoffice-united-nations-high. UN Doc. A/HRC/60/45.

242. Council of Europe. Steering Committee on Media and Information Society. (2025). Guidance note on the implications of generative artificial intelligence for freedom of expression (CDMSI(2025)15rev). https://rm.coe.int/cdmsi-2025-15rev-guidance-note-on-the-implications-of-generative-artif/488029df80

243. European Union. (2024, June 13). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

244. Expert Mechanism on the Right to Development. (2024). Artificial intelligence, cultural rights and the right to development (A/HRC/EMRTD/13/CRP.1). United Nations Human Rights Council. https://www.ohchr.org/sites/default/files/documents/issues/development/emd/session13/a-hrc-emrtd-13-crp-1.pdf

245. Council of Europe, Steering Committee on Media and Information Society. (2025). Guidance note on the implications of generative artificial intelligence for freedom of expression (CDMSI(2025)15rev). https://rm.coe.int/cdmsi-2025-15rev

246. Waight, H., Yang, E., Yuan, Y., et al. (2026). State media control influences large language models. Nature. https://doi.org/10.1038/s41586-026-10506-7

247. Li, P., Yang, J., Islam, M. A., & Ren, S. (2025). Making AI less "thirsty": Uncovering and addressing the secret water footprint of AI models. Communications of the ACM, 68(7), 54–61. https://doi.org/10.1145/3724499

248. Elsworth, C., Huang, K., Patterson, D., Schneider, I., Sedivy, R., Goodman, S., ... & Manyika, J. (2025). Measuring the environmental impact of delivering AI at Google Scale. arXiv preprint arXiv:2508.15734.

249. Arsenault, A. C., & Kreps, S. (2026). Whose voice counts? The role of large language models in public commenting. Big Data & Society, 13(1). https://doi.org/10.1177/20539517261419341

250. Alslaity, A., Chan, G., & Orji, R. (2023). A panoramic view of personalization based on individual differences in persuasive and behavior change interventions. Frontiers in Artificial Intelligence, 6, 1125191. https://doi.org/10.3389/frai.2023.1125191

251. Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. Science, 385, eadq1814. https://doi.org/10.1126/science.adq1814

252. Boissin, E., Costello, T. H., Spinoza-Martín, D., Rand, D. G., & Pennycook, G. (2025). Dialogues with large language models reduce conspiracy beliefs even when the AI is perceived as human. PNAS Nexus, 4(11), pgaf325. https://doi.org/10.1093/pnasnexus/pgaf325

253. Schroeder, D. T., Cha, M., Baronchelli, A., Bostrom, N., Christakis, N. A., Garcia, D., ... & Kunst, J. R. (2026). How malicious AI swarms can threaten democracy. Science, 391(6783), 354-357.

254. Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). The levers of political persuasion with conversational AI. Science, 390(6777), eaea3884. https://doi.org/10.1126/science.aea3884

255. Germano, F., Gómez, V., & Sobbrio, F. (2025). Ranking for engagement: How social media algorithms fuel misinformation and polarization. Barcelona School of Economics, Working Paper No. 1501. https://bw.bse.eu/wp-content/uploads/2025/07/1501.pdf

256. Council of Europe CDMSI. (2025). Guidance Note on Generative AI and Freedom of Expression. CDMSI(2025) https://rm.coe.int/cdmsi-2025-15rev-guidance-note-on-the-implications-of-generative-artif/488029df80

257. Buyl, M., Rogiers, A., Noels, S., Bied, G., Dominguez-Catena, I., Heiter, E., ... & De Bie, T. (2026). Large language models reflect the ideology of their creators. npj Artificial Intelligence, 2(1), 7.

258. Wang, P., Zhang, L.-Y., Tzachor, A., & Chen, W.-Q. (2024). E-waste challenges of generative artificial intelligence. Nature Computational Science, 4, 818–823. https://doi.org/10.1038/s43588-024-00700-3

259. Schroeder, D. T., Cha, M., Baronchelli, A., Bostrom, N., Christakis, N. A., Garcia, D., ... & Kunst, J. R. (2026). How malicious AI swarms can threaten democracy. Science, 391(6783), 354-357.

260. Council of Europe. (2024). Council of Europe framework convention on artificial intelligence and human rights, democracy and the rule of law (Council of Europe Treaty Series No. 225). https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

261. Observatory on Information and Democracy. (2024). Information ecosystems and troubled democracy: A global synthesis of the state of knowledge on news, media, AI, and data governance. https://observatory.informationdemocracy.org/report/information-ecosystem-and-troubled-democracy/

262. Council of Europe, Steering Committee on Media and Information Society. (2025). Guidance note on the implications of generative artificial intelligence for freedom of expression (CDMSI(2025)15rev). https://rm.coe.int/cdmsi-2025-15rev

263. OECD. (2024). Facts not fakes: Tackling disinformation, strengthening information integrity. OECD Publishing. https://doi.org/10.1787/d909ff7a-en

264. United Nations General Assembly. (2017). Promotion and protection of human rights: Human rights questions, including alternative approaches for improving the effective enjoyment of human rights and fundamental freedoms: Report of the Third Committee, 72nd session (A/72/...). United Nations Digital Library. http://digitallibrary.un.org/record/1326669

265. Penney, J. W. (2025). Chilling effects: Repression, conformity, and power in the digital age. Cambridge University Press. https://doi.org/10.1017/9781108918022

266. UN Women. (2022). Tipping point: The chilling escalation of online violence against women in the public sphere. UN Women. https://www.unwomen.org/sites/default/files/2025-12/tipping-point-the-chilling-escalation-of-violence-against-women-in-the-public-sphere-in-the-age-of-ai-en.pdf

267. United Nations Educational, Scientific and Cultural Organization. (2024). Challenging systematic prejudices: An investigation into bias against women and girls in large language models. https://unesdoc.unesco.org/ark:/48223/pf0000388971

268. Chowdhury, R., & Lakshmi, D. (2023). "Your opinion doesn't matter, anyway": Exposing technology-facilitated gender-based violence in an era of generative AI (2nd ed.). UNESCO.

269. United Nations Educational, Scientific and Cultural Organization. (2024, March 7). Generative AI: UNESCO study reveals alarming evidence of regressive gender stereotypes. https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes

270. Fung, P. (2019, June 30). This is why AI has a gender problem. World Economic Forum. https://www.weforum.org/stories/2019/06/this-is-why-ai-has-a-gender-problem/

271. United Nations Conference on Trade and Development. (2025). Technology and innovation report 2025: The AI divide. https://unctad.org/publication/technology-and-innovation-report-2025

272. Ahmed, N., & Wahed, M. (2020). The De-democratization of AI: Deep learning and the compute divide in artificial intelligence research. arXiv preprint arXiv:2010.15581.

273. Coeckelbergh, M. (2026) Technofascism: AI, Big Tech, and the New Authoritarianism. AI & Society https://doi.org/10.1007/s00146-026-02862-9

274. Varoufakis, Y. (2024). Technofeudalism: What killed capitalism. Melville House.

275. Kalluri, P. R., Agnew, W., Cheng, M., Owens, K., Soldaini, L., & Birhane, A. (2025). Computer-vision research powers surveillance technology. Nature, 643(8070), 73-79. https://www.nature.com/articles/s41586-025-08972-6

276. Fola-Rose, A., Solomon, E., Bryant, K., & Woubie, A. (2024, August). A systematic review of facial recognition methods: Advancements, applications, and ethical dilemmas. In Proceedings of the 2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI 2024) (pp. 314–319). IEEE. https://doi.org/10.1109/IRI62200.2024.00070

277. Fussey, P., & Murray, D. (2025). Facial Recognition Surveillance: Policing and Human Rights in the Age of Artificial Intelligence. Oxford University Press.

278. Office of the United Nations High Commissioner for Human Rights. (2024). Mapping report: Human rights and new and emerging digital technologies (A/HRC/56/45). https://www.ohchr.org/en/documents/reports/mapping-report-human-rights-and-new-and-emerging-digital-technologies

279. Secretary-General. (2024). Human rights in the administration of justice (A/79/296). United Nations. https://digitallibrary.un.org

280. American Civil Liberties Union, More than a Dozen Wrongful Arrests Due to Police Reliance on Facial Recognition Technology (2025), https://www.aclu.org/news/privacy-technology/more-than-a-dozen-wrongful-arrests-due-to-police-reliance-on-facial-recognition-technology: Documentation of at least 14 publicly known wrongful arrests in the United States attributed to police use of facial recognition; in nearly all cases the persons wrongfully arrested were Black.

281. Saxena, D., & Guha, S. (2024). Algorithmic harms in child welfare: Uncertainties in practice, organization, and street-level decision-making. ACM Journal on Responsible Computing, 1(1), Article 2, 1–32. https://doi.org/10.1145/3616473

282. German Marshall Fund. (2024). Spitting Images: Tracking Deepfakes and Generative AI in Elections.

283. International Institute for Democracy and Electoral Assistance. (2024). The 2024 global elections super-cycle. https://www.idea.int/initiatives/the-2024-global-elections-supercycle

284. Recorded Future. (2024). 2024 Deepfakes and Election Disinformation Report: Key Findings and Mitigation Strategies.

285. Associated Press. (2025, June 13). New Hampshire jury acquits consultant behind AI robocalls mimicking Biden on all charges.

286. Federal Communications Commission. (2024, September). $6 million fine against Steven Kramer for AI-generated robocalls.

287. Global Witness. (2024). What Happened on TikTok Around the Romanian Elections?

288. IFES. (2024). The Romanian 2024 Election Annulment: Addressing Emerging Threats to Electoral Integrity.

289. Associated Press. (2024). Election disinformation takes a big leap with AI being used to deceive worldwide.

290. United Nations. (1966). International Covenant on Civil and Political Rights. Articles 17, 18, 19 and 25.

291. Council of Europe. (1950). Convention for the Protection of Human Rights and Fundamental Freedoms. Articles 8, 9 and 10; Protocol No. 1, Article 3.

292. EQUATE Language AI Readiness Index (https://equate.vercel.app/en)

293. Han, W., Zhang, Y., Chen, Z., Liu, B., Lin, H., Zhang, B., ... & Zheng, Y. (2025). MuBench: Assessment of Multilingual Capabilities of Large Language Models Across 61 Languages. arXiv preprint arXiv:2506.19468.

294. Lissak, S., Calderon, N., Shenkman, G., Ophir, Y., Fruchter, E., Klomek, A. B., & Reichart, R. (2024, June). The colorful future of llms: Evaluating and improving llms as emotional supporters for queer youth. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 2040-2079).

295. Gamboa, L. C. L., Feng, Y., & Lee, M. (2025, November). Social Bias in Multilingual Language Models: A Survey. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 27845-27868).

296. Choudhury, M., Sitaram, S., Vashistha, A., et al. (2026). AI for the Global South: 12 critical research questions for the next decade. AI for the Global South (AI4GS), Mohamed bin Zayed University of Artificial Intelligence. https://ai4gs.github.io/

297. Bartl, M., Mandal, A., Leavy, S., & Little, S. (2025). Gender bias in natural language processing and computer vision: A comparative survey. ACM Computing Surveys, 57(6), 1-36. https://dl.acm.org/doi/pdf/10.1145/3700438

298. Robb, M. B., & Mann, S. (2025). Talk, trust, and trade-offs: How and why teens use AI companions. Common Sense Media. https://www.commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf

299. U.S. PIRG Education Fund. (2025). AI comes to playtime: Artificial companions, real risks. U.S. PIRG Education Fund. https://pirg.org/edfund/wp-content/uploads/2025/12/AI-Comes-to-Playtime-Artifical-companions-real-risks.pdf

300. Staksrud, E., Mascheroni, G., Milosevic, T., Ní Bhroin, N., Ólafsson, K., Şengül-İnal, G., & Stoilova, M. (2026). European children's use and understanding of generative AI. EU Kids Online V.

301. Committee on the Rights of the Child. (2021). General comment No. 25 (2021) on children's rights in relation to the digital environment (CRC/C/GC/25). United Nations. https://digitallibrary.un.org/record/3906061

302. UNICEF (2025). Guidance on AI and Children: Updated guidance for governments and businesses to create AI policies and systems that uphold children's rights. https://www.unicef.org/innocenti/media/11991/file/UNICEF-Innocenti-Guidance-on-AI-and-Children-3-2025.pdf

303. UNESCO (2025). How should children's rights be integrated into AI governance? https://www.unesco.org/en/articles/how-should-childrens-rights-be-integrated-ai-governance

304. Grossman, S., Pfefferkorn, R., & Liu, S. (2025). AI-Generated Child Sexual Abuse Material: Insights from Educators, Platforms, Law Enforcement, Legislators, and Victims. Version 1. Stanford Digital Repository. Available at https://purl.stanford.edu/mn692xc5736/version/1. https://doi.org/10.25740/mn692xc5736.

305. Livingstone, S., Atabey, A., Stoilova, M., & Sylwander, K. R. (2025). How does, and how could, generative AI respect and enable children's rights? In N. Ní Loideain (Ed.), AI and power: Regulation and rights. University of London Press.

306. European Parliamentary Research Service. (2024). Children and deepfakes. European Parliament. https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2025)775855

307. United Nations Children's Fund (UNICEF). (2026). Artificial intelligence and child sexual abuse and exploitation [Issue brief]. https://www.unicef.org/reports/artificial-intelligence-and-child-sexual-abuse-and-exploitation

308. Ozcan, B., Sperati, V., Giocondo, F., Schembri, M., & Baldassarre, G. (2022, June). Interactive soft toys to support social engagement through sensory-motor plays in early intervention of kids with special needs. In Proceedings of the 21st Annual ACM Interaction Design and Children Conference (pp. 625-628).

309. Goodacre, E., & Gibson, J. (2026). AI in the Early Years: Examining the implications of GenAI toys for young children. https://www.cam.ac.uk/stories/ai-toys-study-play

310. British Standards Institution. (2026, May). Half of children have AI toys despite safety concerns. https://www.bsigroup.com/en-GB/insights-and-media/media-centre/press-releases/2026/may/half-of-children-have-ai-toys-as-parents-allow-widespread-use-despite-safety-concerns-and-gaps-in-guidance-parents/

311. Goodacre, E., & Gibson, J. (2026). AI in the Early Years: Examining the implications of GenAI toys for young children. https://doi.org/10.17863/CAM.126270

312. Chou, C. Y., Chan, T. W., Chen, Z. H., Liao, C. Y., Shih, J. L., Wu, Y. T., & Hung, H. C. (2025). Defining AI companions: a research agenda--from artificial companions for learning to general artificial companions for Global Harwell. Research & Practice in Technology Enhanced Learning, 20.

313. Ho, J. Q., Hu, M., Chen, T. X., & Hartanto, A. (2025). Potential and pitfalls of romantic artificial intelligence companions: A systematic review. Computers in Human Behavior Reports, 19, 100715.

314. Hollanek, T., & Sobey, A. (2025). AI companions for health and mental wellbeing: opportunities, risks and policy implications. Leverhulme Centre for the Future of Intelligence

315. De Freitas, J., Oguz-Uguralp, Z., & Kaan-Uguralp, A. (2025). Emotional manipulation by AI companions. arXiv preprint arXiv:2508.19258.

316. Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R., & Yang, D. (2025). The rise of AI companions: how human-chatbot relationships influence well-being. arXiv preprint arXiv:2506.12605.

317. Dewitte, P. (2024). Better alone than in bad company: Addressing the risks of companion chatbots through data protection by design. Computer Law & Security Review, 54, 106019.

318. Zhang, R., Li, H., Meng, H., Zhan, J., Gan, H., & Lee, Y. C. (2025, April). The dark side of ai companionship: A taxonomy of harmful algorithmic behaviors in human-ai relationships. In Proceedings of the 2025 CHI conference on human factors in computing systems (pp. 1-17).

319. De Freitas, J., & Cohen, I. G. (2024). The health risks of generative AI-based wellness apps. Nature medicine, 30(5), 1269-1275.

320. Radesky, J., Bragg, M. A., & Hiniker, A. (2026). Risks and Consequences of Children's Use of Social AI—A Framework. JAMA Pediatrics. https://doi.org/10.1001/jamapediatrics.2026.1349

321. Muldoon, J., & Parke, J. J. (2025). Cruel companionship: How AI companions exploit loneliness and commodify intimacy. new media & society, 14614448251395192.

322. Jacobs, K. A. (2024). Digital loneliness—changes of social recognition through AI companions. Frontiers in Digital Health, 6, 1281037.

323. Ho, J. Q., Hu, M., Chen, T. X., & Hartanto, A. (2025). Potential and pitfalls of romantic Artificial Intelligence (AI) companions: A systematic review. Computers in Human Behavior Reports, 19, 100715.

324. Hollanek, T., & Sobey, A. (2025). AI companions for health and mental wellbeing: opportunities, risks and policy implications.

325. Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R., & Yang, D. (2025). The rise of AI companions: how human-chatbot relationships influence well-being. arXiv preprint arXiv:2506.12605.

326. Dewitte, P. (2024). Better alone than in bad company: Addressing the risks of companion chatbots through data protection by design. Computer Law & Security Review, 54, 106019.

327. Robb, M. B., & Mann, S. (2025). Talk, trust, and trade-offs: How and why teens use AI companions. Common Sense Media. https://www.commonsensemedia.org/research/talk-trust-and-trade-offs-how-and-why-teens-use-ai-companions

328. Rousmaniere, T., Zhang, Y., Li, X., & Shah, S. (2025). Large language models as mental health resources: Patterns of use in the United States. Practice Innovations. https://doi.org/10.1037/pri0000292

329. Callahan, C., Tanner, L., Coe, C., Davis, M., Glover, J., Bernstein, E., ... & Kunkle, S. (2026). Real-World Use of a Mental Health AI Companion: Multiple Methods Study. JMIR Formative Research, 10, e86904.

330. Associated Press. (2026, June 1). Chatbot AI lawsuit alleges links to teen suicide. https://apnews.com/article/chatbot-ai-lawsuit-suicide-teen-artificial-intelligence-9d48adc572100822fdbc3c90d1456bd0

331. Hudon, A., & Stip, E. (2025). Delusional experiences emerging from AI chatbot interactions or "AI Psychosis". JMIR Mental Health, 12(1), e85799.

332. Green, H. H. (2026, March 14). New study raises concerns about AI chatbots fueling delusional thinking. The Guardian. https://www.theguardian.com/technology/2026/mar/14/ai-chatbots-psychosis

333. Ministère du Travail, de la Santé, des Solidarités et des Familles. (2025, March 24). La santé mentale, grande cause nationale 2025. Gouvernement de la France. https://solidarites.gouv.fr/la-sante-mentale-grande-cause-nationale-2025

334. American Psychiatric Association. (n.d.). Applications of artificial intelligence in mental health care. https://www.psychiatry.org/psychiatrists/practice/artificial-intelligence/applications

335. U.S. Food and Drug Administration. (2025). Executive summary for the Digital Health Advisory Committee meeting: Generative artificial intelligence–enabled digital mental health medical devices. https://www.fda.gov/media/189391/download

336. Hollis, A., & McKeown, G. (2024, September). Empathic AI for autism: Potential and pitfalls of empathic social chatbots in addressing loneliness. In 24th ACM International Conference on Intelligent Virtual Agents: CONNECT, A Workshop on Connecting Interdisciplinary Research on Connections With and Through Technology: IVA 2024.

337. Sharma, D., Meshkat, S., Perivolaris, A., Kamaleddin, M. A., Teferra, B. G., Rueda, A., ... & Bhat, V. (2026). Reimagining psychiatric care with agentic AI: promise, challenges, and a roadmap forward. npj Digital Medicine.

338. Straw, I., & Callison-Burch, C. (2020). Artificial Intelligence in mental health and the biases of language based models. PloS one, 15(12), e0240376.

339. Garg, M. (2024). Towards mental health analysis in social media for low-resourced languages. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(3), 1-22.

340. Wang, W., Tu, Z., Chen, C., Yuan, Y., Huang, J.-T., Jiao, W., & Lyu, M. R. (2024). All languages matter: On the multilingual safety of LLMs. Findings of the Association for Computational Linguistics: ACL 2024, 5865–5877. https://doi.org/10.18653/v1/2024.findings-acl.349

341. Nigatu, H. H., Mehandru, N., Abadi, N. H., Gebremeskel, B., Alaa, A., & Choudhury, M. (2025). Viability of machine translation for healthcare in low-resourced languages. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP), 10584–10598. https://aclanthology.org/2025.emnlp-main.535/

342. Fu, Y. V., Ramachandran, G. K., Park, N., Lybarger, K., Xia, F., Uzuner, Ö., & Yetisgen, M. (2025). BioMistral-NLU: Towards more generalizable medical language understanding through instruction tuning. AMIA Joint Summits on Translational Science Proceedings, 2025, 149–158. https://pubmed.ncbi.nlm.nih.gov/40502228/

343. Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., & Dufour, R. (2024). BioMistral: A collection of open-source pretrained large language models for medical domains. Findings of the Association for Computational Linguistics: ACL 2024, 5848–5864. https://aclanthology.org/2024.findings-acl.348/

344. Qiu, P., Wu, C., Zhang, X., Lin, W., Wang, H., Zhang, Y., Wang, Y., & Xie, W. (2024). Towards building multilingual language models for medicine. Nature Communications, 15, 8384. https://doi.org/10.1038/s41467-024-52417-z

345. Nwabufo, J., Ogueji, K., Adelani, D. I., Alabi, J., et al. (2025). Healthcare NLP for African Languages: Current State and Challenges. Proceedings of the AfricaNLP Workshop (AfricaNLP 2025). https://aclanthology.org/2025.africanlp-1.32/

346. Okafor, U. (2025). Multilingual NLP for African Healthcare: Bias, Translation, and Explainability Challenges. Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), 221–229. https://aclanthology.org/2025.africanlp-1.32/

347. Skianis, K., Doğruöz, A. S., & Pavlopoulos, J. (2024). Leveraging LLMs for translating and classifying mental health data. In J. Sälevä & A. Owodunni (Eds.), Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024) (pp. 236–241). https://doi.org/10.18653/v1/2024.mrl-1.20

348. Cronin, A., Kelly, A., Wrona, M., O'Donnell, P., Hassan, A., Myles, T., Fallon, T., & MacFarlane, A. (2025). The patient-safety implications of AI-based communication with migrants in general practice: a scoping review. BJGP Open, 9(4), BJGPO.2025.0107. https://bjgpopen.org/content/9/4/BJGPO.2025.0107

349. House of Lords Public Services Committee. (2025). Lost in translation? Interpreting services in the courts (2nd Report of Session 2024–26, HL Paper 87). https://committees.parliament.uk/publications/44602/documents/221328/default/

350. Nigatu, H. H., Mehandru, N., Abadi, N. H., Gebremeskel, B., Alaa, A., & Choudhury, M. (2025). Viability of machine translation for healthcare in low-resourced languages. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 10584–10598. https://aclanthology.org/2025.emnlp-main.535/

351. MIT AI Risk Initiative. (2025). AI risk mitigation database and draft taxonomy. https://airisk.mit.edu/ai-riskmitigations

352. Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., ... & Manyika, J. (2024). The ethics of advanced AI assistants. arXiv preprint arXiv:2404.16244.

353. Staufer, L., Feng, K., Wei, K., et al. (2026). The 2025 AI agent index: Documenting technical and safety features of deployed agentic AI systems. https://doi.org/10.48550/arXiv.2602.17753

354. Weidinger, L., Raji, I. D., Wallach, H., Mitchell, M., Wang, A., Salaudeen, O., ... & Isaac, W. (2025). Toward an evaluation science for generative ai systems. arXiv preprint arXiv:2503.05336.

355. Rabanser, S., Kapoor, S., Kirgis, P., et al. (2026). Towards a science of AI agent reliability. https://doi.org/10.48550/arXiv.2602.16666

356. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2024). Generative AI can harm learning. The Wharton School Research Paper.

357. Shen, J. H., & Tamkin, A. (2026). How AI impacts skill formation. arXiv preprint arXiv:2601.20245.

358. Budzyń, K., Romańczyk, M., Kitala, D., Kołodziej, P., Bugajski, M., Adami, H. O., ... & Mori, Y. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. The Lancet Gastroenterology & Hepatology, 10(10), 896-903.

359. Epoch AI. (2026). Data on AI models. https://epoch.ai/data/ai-models

360. Feng, K. J., McDonald, D. W., & Zhang, A. X. (2025). Levels of autonomy for ai agents. arXiv preprint arXiv:2506.12469.

361. Fink, M. (2025). Operationalizing meaningful human oversight under Article 14 of the EU AI Act. In AI Act commentary: A thematic analysis (forthcoming). Hart-Bloomsbury.

362. Shapira, N., Wendler, C., Yen, A., Sarti, G., Pal, K., Floody, O., ... & Bau, D. (2026). Agents of chaos. arXiv preprint arXiv:2602.20021.

363. Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., ... & Rahwan, I. (2025). Multi-agent risks from advanced ai. arXiv preprint arXiv:2502.14143.

364. Soares, N., Fallenstein, B., Armstrong, S., & Yudkowsky, E. (2015). Corrigibility. Machine Intelligence Research Institute. https://intelligence.org/files/Corrigibility.pdf.

365. Zeng, Y., Lu, E., Guo, X., Huangfu, C., Xie, J., Chen, Y., ... & Younas, A. (2025). AI Governance InternationaL Evaluation Index (AGILE Index) 2025. arXiv preprint arXiv:2507.11546.

366. Bommasani, R., Kapoor, S., Klyman, K., Longpre, S., Ramaswami, A., Zhang, D., Schaake, M., Ho, D. E., Narayanan, A., & Liang, P. (2023, December 13). Considerations for governing open foundation models. Stanford Institute for Human-Centered Artificial Intelligence. https://hai.stanford.edu/policy/issue-brief-considerations-governing-open-foundation-mode

367. Tzachor, A., Devare, M., Richards, C., Pypers, P., Ghosh, A., Koo, J., ... & King, B. (2023). Large language models and agricultural extension services. Nature food, 4(11), 941-948.

368. Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature, 616, 259–265. https://doi.org/10.1038/s41586-023-05881-4

369. Bommasani, R., Klyman, K., Kapoor, S., Longpre, S., Ramaswami, A., Zhang, D., Schaake, M., Ho, D. E., Narayanan, A., & Liang, P. (2024). The foundation model transparency index v1.1. arXiv:2407.12929. https://arxiv.org/abs/2407.12929

370. BigScience Workshop, Le Scao, T., Fan, A., et al. (2022). BLOOM: A 176B-parameter open-access multilingual language model. arXiv. https://doi.org/10.48550/arXiv.2211.05100

371. Touvron, H., Lavril, T., Izacard, G., et al. (2023). LLaMA: Open and efficient foundation language models. arXiv. https://arxiv.org/abs/2302.13971

372. Qwen Team. (2024). QwenLM. GitHub. https://github.com/QwenLM/Qwen

373. Qwen Team. (2024). Qwen2 technical report. arXiv. https://arxiv.org/abs/2407.10671

374. DeepSeek-AI. (2024). DeepSeek-V3 technical report. arXiv. https://doi.org/10.48550/arXiv.2412.19437

375. DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv. https://arxiv.org/abs/2501.12948

376. Jiang, A. Q., et al. (2023). Mistral 7B. arXiv. https://arxiv.org/abs/2310.06825. Mistral AI.

377. Technology Innovation Institute. (2023). The Falcon series of open language models. arXiv. https://arxiv.org/abs/2311.16867

378. SberDevices. (2026, March). GigaChat-3.1: Большое обновление больших моделей. Habr. https://habr.com/ru/companies/sberbank/articles/1014146/

379. Yandex. (2025, February 25). YandexGPT 5 — в Алисе, облаке и опенсорсе. Habr. https://habr.com/ru/companies/yandex/articles/885218/

380. Sarvam AI. (2025). Sarvam AI models on Hugging Face. https://huggingface.co/sarvamai

381. SB Intuitions. (2025). Sarashina models on Hugging Face. https://huggingface.co/sbintuitions

382. Naver Cloud. (2024). HyperCLOVA X technical report. arXiv. https://arxiv.org/abs/2404.01954

383. Seger, E., Dreksler, N., Moulange, R., et al. (2023). Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. Centre for the Governance of AI. https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models

384. Anthropic. (2025). Disrupting the first reported AI-orchestrated cyber espionage campaign. https://www.anthropic.com/news/disrupting-AI-espionage

385. Partnership on AI. (2023). PAI's guidance for safe foundation model deployment. https://partnershiponai.org/modeldeployment/

386. International Energy Agency. (2025). Energy and AI. International Energy Agency. https://www.iea.org/reports/energy-and-ai