A framework for language technologies in behavioral research and clinical applications: Ethical challenges, implications and solutions.

Catherine Diaz-Asper [1], Mathias K Hauglid [2,3], Chelsea Chandler [4], Alex S. Cohen [5,6], Peter W. Foltz [4] & Brita Elvevåg [3,7]

1. Department of Psychology, Marymount University
2. Faculty of Law, University of Tromsø - the Arctic University of Norway, Tromsø, Norway
3. Norwegian Center for Clinical Artificial Intelligence, University Hospital of North Norway, Tromsø, Norway
4. Institute of Cognitive Science, University of Colorado Boulder
5. Department of Psychology, Louisiana State University
6. Center for Computation and Technology, Louisiana State University
7. Department of Clinical Medicine, University of Tromsø - the Arctic University of Norway, Tromsø, Norway

Author Note

We have no known conflict of interest to disclose.

Correspondence concerning this article should be addressed to Catherine Diaz-Asper, Ph.D., Marymount University, 2807 N. Glebe Road, Arlington VA 22207, United States. Email: cdiazasp@marymount.edu

Abstract

Technological advances in the assessment and understanding of speech and language within the domains of automatic speech recognition, natural language processing, and machine learning present a remarkable opportunity for psychologists to learn more about human thought and communication, evaluate a variety of clinical conditions, and predict cognitive and psychological states. These innovations can be leveraged to automate traditionally time-intensive assessment tasks (e.g., educational assessment), provide psychological information and care (e.g., chatbots), and when delivered remotely (e.g., by mobile phone or wearable sensors) promise underserved communities greater access to healthcare. Indeed, the automatic analysis of speech provides a wealth of information that can be used for patient care in a wide range of settings (e.g., mHealth applications) and for diverse purposes (e.g., behavioral and clinical research, medical tools that are implemented into practice) and patient types (e.g., numerous psychological disorders, and in psychiatry and neurology). However, automation of speech analysis is a complex task that requires integration of several different technologies within a largely distributed process with numerous stakeholders. Many organizations have raised awareness about the need for robust systems for ensuring transparency, oversight, and regulation of technologies utilizing artificial intelligence. Since there is limited knowledge about the ethical and legal implications of these applications in psychological science, we provide a balanced view of both the optimism that is widely published on but also the challenges and risks of use, including discrimination and exacerbation of structural inequalities.

KEYWORDS:

Natural Language Processing (NLP), artificial intelligence (AI), ethical challenges, psychological research

Public Significance Statement

Computational advances in the domains of automatic speech recognition, natural language processing, and machine learning allow for the rapid and accurate assessment of a person's speech for numerous purposes. The widespread adoption of these technologies permits psychologists an opportunity to learn more about psychological function, interact in new ways with research participants and patients, and aid in diagnosis and management of various cognitive and mental health conditions. However, we argue that the current scope of the APA Ethics Code is insufficient to address the ethical issues surrounding the application of artificial intelligence. Such a gap in guidance results in the onus falling directly on psychologists to educate themselves about the ethical and legal implications of these emerging technologies, potentially exacerbating the risk of their use in both research and practice.

Section 1: The importance of Artificial Intelligence and language analysis

Language data are increasingly acquired via a myriad of technological innovations, such as telehealth, mobile devices, and social media. These systems have been implemented to predict, assess, and monitor psychological and cognitive state, as well as provide artificial intelligence (AI)-driven empathetic conversational agents for self-management of anxiety and depression (Bedi et al., 2015; Chandler et al., 2020a; De Choudhury et al., 2013; Coppersmith et al., 2018; Elvevåg et al., 2007; Faurholt-Jepsen et al., 2019). Many voices have extolled the potential for greater efficiency, precision, and equity in healthcare if these digital solutions are implemented (e.g., Hirschtritt & Insel, 2018). Computational psychological assessments enable more frequent remote monitoring to facilitate detection and diagnosis and promote adherence to treatments (e.g., Nasland et al., 2019), given that the current alternative requires time-consuming in-person interactions with experts. Mobile applications further provide promise of rapid, personalized treatment through chat with intelligent agents (e.g., Miner et al., 2019). This is particularly important for vulnerable populations, whose access to traditional assessment may be challenged by distance, socioeconomic, cognitive, literacy, and sensory issues.

While there has been rapid growth and use of these innovations, the ethical principles of *how* and *when* they should be employed have not developed at the same rate. Indeed, leveraging natural language processing (NLP) methods for speech analysis in research and applied settings evokes a variety of legal and ethical issues (Hauglid, 2022). As the boundaries between research and practice are often permeable and iterative (especially due to the growth of research–practice partnerships in psychological science), the ethical issues we raise herein pertain to all aspects of research, development, and application. An ideal technology should have relatively high levels of human involvement in development and oversight of the system, users who understand the

benefits and limits of the technology, and relatively minor consequences if the system fails. Conversely, the highest risk occurs when fully autonomous systems with low user understanding are used to generate safety-critical outputs. Language-based AI systems are simply not (ethically) feasible without some degree of human involvement and oversight. The decision to employ NLP should be based upon evaluation of *why* it may be suitable in the specific context, *what* the intended purpose is, *who* specifically will be the user(s), *how* the appropriate level of human agency or oversight will be established, and *how* explainable or understandable the resulting technology is.

The focus of this paper is in keeping with an anticipatory ethics approach, recently adopted by Chiang et al. (2021) in neurological research, that encourages a pre-emptive examination of methodological and design choices to enable a careful evaluation of the ethical implications of decisions in the development, calibration, and implementation of algorithms. We examine widely recognized ethical AI principles and problematize their implementation in speech technologies in psychology, and thus do not evaluate the principles as such. Three examples are discussed that leverage current, or soon to be available, speech technologies.

The literature on medical AI applications and ethical challenges is considerable (e.g., Davenport & Kalakota, 2019; Rajpurkar et al., 2022). However, literature specifically addressing psychological applications or providing tailored AI ethics guidance for psychologists is limited. Here, we focus on *psychological applications* used by professionals conducting research and in clinical practice, and those available without a need for expert supervision. We evaluate the ethical considerations inherent to the use of such tools from the perspectives of various users. Lessons learned may also be relevant for applications across allied fields, some of which have been grappling with AI-related ethical issues for decades (Clancey & Shortliffe, 1984).

Section 2: Three examples of existing domains of AI in psychological science

Conversational agents

Conversational agents, also known as chatbots, use NLP and machine learning (ML) to simulate conversations with users from voice or text input. The projected worldwide healthcare chatbot market of over 900 million USD by 2032 (GlobeNewswire, 2023) is no doubt attributable to their impressive ability to improve standardization and efficiency and reduce costs. An increasing number of chatbots operate within psychological science, offering a variety of potential interventions (e.g., psychoeducation, skill building, self-care strategies), as well as standardized and interactional stimuli for affective and social sciences (Croes & Antheunis, 2021). To our knowledge, few commercially available chatbots have been developed in collaboration with psychologists, despite their "faux" psychological presentation and purpose (Ruane et al., 2019). Although not currently under the purview of the APA, psychologists should be concerned given the potential harm to both users and the field (McGreevey et al., 2020).

The capabilities of current language models to engage in conversations with users has recently been demonstrated by a series of generative AI-based conversational agent systems. Chatbots' natural, conversational language style promotes interaction, thus strengthening their potential as tools in psychological research and for clinical purposes. For instance, conversations may be recorded remotely, increasing the availability of relevant data in research projects or for clinical assessments. While this may be useful, the lack of psychologist involvement in each conversation raises ethical challenges.

Currently, chatbots are marketed as informational only, rather than as intended for clinical purposes. Consequently, the ethical and legal risks for their use, particularly as an adjunct for clinical services or to collect potentially sensitive information, are not systematically

considered by any agency. This increases the potential for user harm (Luxton, 2020) because users may not be aware of the systems' limits in mental health expertise and user protection. For example, a review of chatbots in healthcare (including mental health) found that 63% did not contain information about users' data privacy, and only 12% were HIPAA compliant (Parmar et al., 2022). Privacy safeguards, data access, and the "right to be forgotten" must be prioritized for all speech technologies. Further, "bot disclosure" requires that users be made aware at the outset that they are interacting with AI rather than a real person, to avoid the situation of "counterfeiting humanity" (Pasquale, 2020). An additional concern involves the potential for chatbots to emulate and perpetuate stereotypes about populations for which they are designed (Ruane et al., 2019).

Remote eHealth monitoring devices

Remote eHealth monitoring devices are AI-based systems that digitally transmit data from the patient to their clinician or healthcare center, potentially providing "real-time" monitoring. The last few decades have seen an explosion in both medical and consumer grade remote medical monitoring devices (Vegesna et al., 2017), new research directions (Ramesh et al., 2021), and a specific interest in utilizing these devices for psychological assessment. Language can be recorded in a variety of ways: when an individual is "actively" interacting with their device, through "passive" recording as someone navigates their daily routine, as part of natural phone and text use, and by scanning social media and online activity.

The use of AI, NLP, and ML built from complex and voluminous data in research and commercial applications leads to "black box" algorithms that are difficult to interpret by most stakeholders. There are additional ethical considerations with respect to agency, as decisions may necessitate resource-costly actions and even involuntary intervention (e.g., hospitalization).

Some devices can invade privacy by inadvertently recording ancillary activities (e.g., other speakers, social media posts). Patients have the right to know when their healthcare provider is receiving information or making decisions about them via an algorithm. Finally, remote assessment requires access to technology. Although smart devices are considered ubiquitous in most industrial nations, literacy/skill about their use and privacy, access to internet, and knowledge of relevant protection laws and resources vary considerably as a function of socio-economic status and culture. More insidious is the reality that language models have been shown to predict clinical phenomena differently based on age, language, gender, or ethnicity, potentially exacerbating existing inequities among disadvantaged groups.

Educational Assessment

Low-cost networked digital devices (e.g., mobile devices) are widely used to present educational material and to assess learning. The global digital education market is projected to be valued at 180 billion USD by 2033, growing over 11% annually from current levels (Factmr, 2023). In educational assessment, materials are developed to make valid inferences about constructs related to cognitive processes to target a student's knowledge, skills, and abilities (e.g., Williamson et al., 2012), rather than psychological well-being (although it may include assessment of affective, motivational, and behavioral attributes). Digital technologies can support the development of learning and assessment activities in ways that increase the inferential fidelity of assessments as well as allow automated assessments of writing and speaking that formerly was laboriously hand-scored by instructors (Behrens et al., 2019). For example, K-12 assessments, professional certifications, automated tutoring systems, and English language proficiency assessments are increasingly including more open-ended responses that are assessed by ML and NLP algorithms. Digital technologies are used both in assessment delivery

and interpretation (e.g., Yan et al., 2020), and are also widely used in research to understand the effectiveness of educational and psychological interventions and the underlying cognitive and social processes involved (D'Mello et al., 2022).

The American Educational Research Association, American Psychological Association, and National Council on Measurement joint standards (AERA, APA, & NCME, 2014) emphasize the need for appropriate considerations and procedures for validity, reliability, fairness, and testing methodologies in educational and psychological assessments. Although the standards do not focus on digital technologies, they recognize their impact on assessment and address ethical considerations around the development of AI-based assessments and the use of technology in scoring. For example, the standards address fairness of assessments that may be digitally delivered in different modalities (e.g., written, audio or American Sign Language) while maintaining valid interpretation of the scores for accommodations. Automated scoring may assess performance based on irrelevant features or introduce bias, hence the importance of using construct-relevant features that can be linked to targeted constructs. Similarly, scoring algorithms may have been developed from inherently biased human training data, hence the standards state that automated algorithms "need to be reviewed for potential sources of bias" (AERA, APA, & NCME, 2014, p. 66) and should be evaluated for their impact on marginalized groups.

Section 3: The risks and inadequate protections of AI and language technologies currently

AI-based language technologies have rapidly expanded in recent years and this growth is anticipated to continue as tools become more widely available and accepted, thus raising important ethical considerations. AI experts are generally pessimistic about the widespread adoption of ethical design principles, due to disagreements about ethics definitions and implementation and enforcement responsibility (Rainie et al., 2021). While the U.S. Food and

Drug Administration has released an action plan for the use of AI in medicine (U.S. Food and Drug Administration, 2020), it remains to be seen whether a regulatory framework for medical devices will address the full scope of ethical challenges (as regulation is primarily concerned with safety and effectiveness). The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research's *Belmont Report* (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979) and the APA's *Ethical Principles of Psychologists and Code of Conduct* (APA Ethics Code; American Psychological Association, 2017) set forth broad general principles and ethical standards for researchers and psychologists to follow and to guide decision making. The Belmont Report and APA Ethics Code are sufficiently broad to encompass many of the ethical dilemmas arising from these technologies, but they are arguably *insufficiently* specific at the time of writing to fully protect users (researchers, research participants, clinicians, patients, citizens) from harm.

Insufficient understanding of AI development and implementation raises risks of algorithmic bias, inaccurate predictions, and potential misuse of data, among other dangers. The "black box" nature of NLP/ML systems pose challenges to human agency, from the perspective of researchers, clinicians, and patients alike. Consider a real case: we developed a language-based app to predict participants' psychological state factors from daily speech samples provided on a smart device (Chandler et al., 2020a; Cohen et al., 2019; Holmlund et al., 2019). In such a setting, psychological researchers are tasked with providing complete and fully comprehensible information to obtain informed consent (Belmont Report, *Informed Consent;* APA section 8.02, *Informed Consent to Research;* APA section 8.03, *Informed Consent for Recording Voices and Images in Research*), understanding the psychometric properties of the assessment tools (APA section 9.02b, *Use of Assessments*), interpreting and explaining the results (APA section 9.06,

Interpreting Assessment Results; APA section 9.10, *Explaining Assessment Results*), and fully debriefing the participant upon completion of the study (APA section 8.08, *Debriefing*). However, if the technology lacks transparency and explainability, it is not possible for researchers to fulfill these ethical obligations (Chandler et al., 2020b). Similarly, participants cannot provide informed consent if researchers are unable to describe the technology, detail its limits, and explain how participant data may be used. While the general principles of the Belmont Report and the APA Ethics Code broadly encompass these concerns, more AI-specific considerations are warranted (e.g., UNESCO, 2022). Certainly, non-psychologists could argue that the ethics code does not apply to them, yet developers should consider the Belmont Report and APA Ethics Code principles as central when designing tools either with psychological content or to be used by psychologists for research and clinical purposes.

Section 4: Ethical AI guidance exists outside of psychology and should be applied

Guidelines for ethical and principled AI have proliferated over the last several years, developed within industry, governmental, and intergovernmental organizations (see Fjeld et al., 2020). Two prominent sets of international AI guidelines are the World Health Organization's (WHO) *Ethics and Governance of Artificial Intelligence for Health* **(**World Health Organization, 2021) and the European Union's (EU) *Ethics Guidelines for Trustworthy AI* (High-Level Expert Group on AI, 2019). The substantial overlap between the two indicates that there is an emerging consensus about key ethical principles for AI systems (see also Section 1.2 of the Organization for Economic Co-operation and Development's guidelines (OECD, 2019; Jobin et al., 2019)). Many of the EU and WHO principles are encompassed by the Belmont Report and APA Ethics Code, indicating that policy makers' core values regarding AI must have long been on the ethics agenda of psychologists. Most of the principles are familiar to psychology, although their

application to digital technologies creates new challenges. There is no APA principle that corresponds to the principle of "human agency and oversight" in the EU guidelines ("protection of autonomy" in the WHO guidance), nor is there an APA principle corresponding to "transparency" or "explainability." The most relevant APA principle for autonomy and human agency is *Principle E*, which requires "respect for people's rights and dignity" and promotes self-determination. However, the accompanying explanation shows that this guideline does not ensure overall human autonomy with digital technologies. Principle E protects vulnerable patients with reduced autonomy, and the Belmont Report's "respect for persons" principle acknowledges individual autonomy (being "capable of deliberation about personal goals and of acting under the direction of such deliberation") and the need to protect those with reduced autonomy. Neither code refers directly to the professional autonomy of the psychologist, nor to the need to ensure human involvement in psychological research and practice. In contrast, the WHO's principle on protection of autonomy suggests that "humans should remain in full control of health-care systems and medical decisions" (World Health Organization, 2021, p.25).

The absence of references to human agency/oversight and transparency/explainability in the Belmont Report and APA Ethics Code is not surprising. Although AI ethics have been discussed for decades, the need to develop ethical principles specifically for language-based technologies is invoked by recent and ongoing technological advancements, as demonstrated in Section 2. Traditional psychological methods and early versions of AI/NLP technologies did not give rise to these challenges in the past. Adoption of the emerging principles for ethical AI by the APA should be considered, given the growing use of digital technologies (including NLP) in psychology. We argue that the APA Ethics Code should be the focus of this effort because it applies to all psychologists, irrespective of setting (e.g., research vs. practice). The incorporation

of principles on human agency and transparency would arguably be a good start in terms of enhancing awareness of ethical challenges related to the development and use of digital technologies in the field. The next step should then be to develop more specific guidelines for the operationalization of the new ethical principles, as well as the existing APA Ethics Code principles, in relation to digital technologies [1]. Throughout the remainder of this paper, we explore three categories of ethical challenges related to language-based technologies in psychology and suggest components of an implementing framework for ethical principles in this context. This includes the ethical principles that are encompassed by the APA Ethics Code, and the more AI-specific principles on human agency/autonomy and transparency/explainability.

Human agency and oversight

As AI can increasingly undertake more tasks and responsibilities that have traditionally been allocated to the expert (e.g., the researcher or clinician), the distinct roles of the computer and the human must be delineated. A core element of human agency/autonomy is in preserving the human ability to control the extent to which decision-making is transferred from humans to AI. According to the EU guidelines, users of AI systems should be "given the knowledge and tools to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system" (High-Level Expert Group on AI, 2019, p. 16l). In healthcare, from a patient perspective, human agency/autonomy refers to individuals' level of control in their healthcare and decision-making. From a psychologist perspective, human agency is connected with professional autonomy, and is effectively reduced

[1] The EU's High-Level Expert Group on AI has translated general principles into an assessment list for trustworthy AI, but it is not specific enough to satisfy the need for guidance in psychology and/or NLP.

if decisions are based on AI predictions rather than professional (human) judgment (Zicari et al., 2021). The WHO guidance stresses that human autonomy should not be undermined by the extension of "machine autonomy" (World Health Organization, 2021, p.25), thus indicating that human autonomy and machine autonomy may be seen as competing notions and that the use of AI systems can be placed on a scale ranging from no human involvement (maximum machine autonomy) to maximum human control (minimal machine autonomy) (see also the 'graded autonomy' model for medical AI applications suggested by Bitterman et al., 2020).

Human oversight is described by the WHO as including effective, transparent monitoring of human values and moral considerations, emphasizing the ability of humans to override AI decisions. Moreover, the WHO sees risk of automation bias as a threat to human agency. The EU guidelines are somewhat more specific in terms of how human oversight may be achieved, by suggesting the use of approaches framed as "human-on-the-loop" (human intervention during design stages and monitoring during operation), "human-in-the-loop" (humans have full control over the decisions of the system and can approve or deny any action at all stages), and "human-in-command" (humans oversee the overall activity of the AI system and decide when and how to use the system, thus allocating the most control to humans). We refer to these terms as they provide helpful tools for categorizing the different types of human oversight which can be envisaged at all stages of development and use of AI systems. In Section 5, we discuss how these approaches may be implemented into psychological science involving NLP.

Human agency and oversight are complex issues to consider when developing guidelines for the use of NLP in psychological research and practice. While a benefit of automated systems is the promise of monitoring and reaching a greater proportion of the population, human oversight requirements may limit this. As such, the benefits of oversight must be weighed against

the benefits of automation. This trade-off, which has received limited attention in current AI ethics discourse, is crucial to certain AI applications of relevance to psychological research and practice. Consider the benefits of having participants engage with conversational agents remotely, generating more data than in-person conversations. To determine the appropriate type and degree of human oversight, one needs to consider the possible impacts on the research participant, as well as any implications of limited oversight during the conversations. Similarly, human oversight may be ethically required for AI systems used for important decision-making or monitoring. In the case of remote assessment for monitoring purposes, we have advocated for the employment of a human-in-the-loop oversight system (Chandler et al., 2022). Scenarios such as these allow the AI systems to generate predictions in cases that they are confident in, whilst deferring to a human in atypical cases with insufficient training data. After human input is collected, the system is strengthened and able to tackle such rarities in the future. The field of education encompasses many applications for language-based AI tools, from automated essay scoring to the development of curriculum and classroom assistants, each one with unique implementations of human agency and oversight. In essence, the level of human involvement should be such that systems with the highest risk incorporate the most human expertise to minimize or prevent spurious and dangerous predictions from being made and acted upon.

Transparency and explainability

Transparency and explainability are among the most widely discussed challenges of AI (e.g., Chandler et al., 2020b). Due to the complexity of certain AI technologies, there is concern that users and stakeholders may not be able to receive meaningful information, for instance about the basis, or logic of AI-based assessments (Casey et al., 2019; Selbst & Powles, 2017; Wachter et al., 2017). In the EU guidelines, the transparency principle is described in terms of

"traceability" (documentation of datasets and how they are processed), "explainability" (ability to explain technical processes and decisions) and "communication" (providing information about the use, capabilities, and limitations of the system). The WHO guidance similarly emphasizes the communication of information about AI systems, datasets, algorithms and processing methods. It follows that the system's strengths and limitations should be evident to not only developers, but also users of the technology (Ribera & Lapedriza, 2019). [2]

In terms of explainability, the APA Ethics Code requires psychologists to take reasonable steps to ensure that assessment results are explained to the individual (9.10 *Explaining Assessment Results*). However, when NLP-based systems are used in the assessment, it may not be obvious which steps it is reasonable for the psychologist to take. There may be limitations inherent in the technology, particularly if developers are not mindful of the explainability requirement that applies to psychologists. Determining what kind of explainability is reasonable may involve complex trade-offs. For instance, the EU and the WHO guidelines both recognize the potential trade-off between transparency and accuracy but provide little guidance as to how the appropriate balance should be struck. The WHO guidance suggests that AI systems should be "explainable to the extent possible and according to the capacity of those to whom the explanation is directed" (World Health Organization, 2021, p.27). This leaves a lot of room for domain-specific interpretation, which means that psychology and NLP experts must determine the appropriate level of explainability based on the state-of-the-art, the capacity of the users, and the values, interests, and risks at play in a specific application context.

As a rule, explainability and transparency are especially important for AI systems that may greatly impact the lives of many and those with dire consequences if incorrect predictions

[2] The EU's AI Act proposal further requires that AI systems be sufficiently transparent for users to interpret their output and use them appropriately, cf. Article 13(1) of the proposal.

are made. Transparency and explainability are challenging aspects in the implementation of language-based systems as many of the state-of-the-art models today are large, uninterpretable neural networks (i.e., "black boxes"). As the expanding body of literature on explainable AI shows, tools exist that allow researchers to probe these models and generate explanations (Gunning et al., 2019), and transparency is possible in terms of describing the model's purpose, training data, boundaries, potential sources of bias, and so on. The possible level of explainability and transparency will vary with the AI model type (e.g., neural networks are more difficult to explain than linear regression models) and with the application. Conversational agents may not require explainability with each output, but information on their development must be made available. In the case of remote monitoring where a clinician is involved and the models are generating predictions to aid in the understanding of patient state and decision making, each output must be able to be explained and understood by the clinician. The reasoning behind each individual prediction should be traceable to the raw patient inputs such that the clinician can decide whether the predictions are logical. Similar arguments hold for education, where transparency will always be important, and the level of explainability will depend on the use case.

To mitigate some of the inevitable negative effects, involvement of all stakeholders early in design decisions is essential, as are explicit explanations of the resulting model, similar to the Data Nutrition Label project (Holland et al. (2018) or the MINimum Information for Medical AI Reporting (MINIMAR; Hernandez-Boussard et al., 2020). The ideal scenario would thus provide the user with details about the training data (e.g., size, racial makeup, and so on, see Gebru et al., 2021), model development (e.g., specifics about algorithms), performance (e.g., accuracy and errors such as false positives and negatives), assessment evaluation (e.g., fairness, bias

attestations), validation (e.g., studies detailing safety and efficacy), purpose of the algorithm (e.g., verbal memory evaluation versus detection of suicide risk), and specifics of the last update (e.g., latest model version). Additionally, users may be interested in knowing why and how decisions about the model were made.

Equity, biases and non-discrimination

While AI and NLP technologies show great promise, concerns have been raised that digital health-care solutions may not address or may exacerbate inequity (e.g., Naslund et al., 2019). To accommodate these concerns, ethical guidance documents tend to promote ethical principles addressing the impact of AI systems on vulnerable or marginalized groups, particularly to mitigate the risk of unfair and/or discriminatory outcomes (e.g., Amnesty International, 2018). The WHO guidance lists as a main objective, to "ensure inclusiveness and equity," and emphasizes the need to identify and mitigate unintended biases in AI systems. The EU Guidelines similarly list "diversity, non-discrimination and fairness" as a key ethical principle for trustworthy AI. The guidelines elaborate that this principle concerns avoiding unfair biases, ensuring accessibility and universal design, and promoting stakeholder participation.

The ethical standards in the APA Ethics Code require that psychologists "use assessment instruments whose validity and reliability have been established for use with members of the population tested" (9.02 b, *Use of Assessments*). For NLP-based tools, we understand this as requiring testing procedures that consider the linguistic composition of the target population. At a more general level, the APA Ethics Code instructs psychologists to eliminate the impact of unacceptable biases on their work (*Principle E*) and to avoid unfair discrimination (3.01, *Unfair Discrimination*). Our view is that the APA Ethics Code is currently better prepared to address challenges related to equity, biases, and non-discrimination than the two other categories of

ethical challenges discussed here. However, to operationalize the principles of equity and non-discrimination in a meaningful way that protects stakeholders, these principles must be interpreted and elaborated for specific use cases. To inform the development of guidelines for digital technologies, the APA Ethics Code should be interpreted in conjunction with emerging ethical frameworks for AI and the available knowledge of how and why these tools may cause harm.

Language is inherently intertwined with sociocultural factors and its expression varies by culture, age, education and other demographic factors. There is increasing evidence that groups who are underrepresented in the datasets used to train algorithms are at risk of being assessed less accurately and that the models may reproduce undesirable stereotypes (Bolukbasi et al., 2016; Caliskan et al., 2017; Straw & Callison-Burch, 2020) pertaining to, for example, gender (e.g., Bailey et al., 2022) and race (e.g., Hitczenko et al., 2021). Moreover, contextual factors affect language expression, such as deference to a medical professional, willingness to self-disclose online, and level of wariness when interacting with authority figures. Unequal access to digital technologies, software and connectivity is a fundamental concern, particularly when access is mediated by cultural, language or legal issues. Notably, cultural mistrust of digitized personal data for private or governmental solutions is also common for many communities (Hsiao, 2003). This might reflect a barrier to utilization regardless of access to technologies.

The APA Ethics Code principles of beneficence and nonmaleficence, justice and respect for people's rights and dignity would dictate that psychologists pay particular attention to the source of a system’s training data (also shown in the principle of transparency). As noted above, models based on majority datasets may risk entirely missing vocabulary for illness or distress in minority populations. Given that algorithms are used at multiple stages of an NLP pipeline,

inadequate representation at any stage could systematically bias or invalidate the entire model for particular groups. This is why an understanding of the data and assumptions behind the creation of these models, as well as some information about how they work is required. Incorporating modern methods of bias mitigation is important for system developers to avoid potentially discriminatory predictions.

Section 5: Developing and implementing ethical guidelines for language technologies

Human-in-the-loop AI can be viewed as a specific implementation of an intelligent system where humans and computers work *together* towards a common goal. In these scenarios, the best of human and computer intelligence are harnessed alongside one another. Computers are well suited for efficient calculations and predictions on data within the scope of their training. Human input can supplement and fill gaps where outliers and unusual inputs are encountered to strengthen the algorithm and allow for the computer to become increasingly accurate and self-sufficient. The importance of a human-in-the-loop methodology is clear, however various levels of human oversight and intervention must be defined. Given the continuous improvement of NLP technologies, human oversight methods will continue to evolve. “Active learning,” where human labeling is used to steer the model towards improved learning in low confidence areas, is one way of enhancing human agency and oversight during research and development. In an implementation phase, humans may be used as safeguards to step in when the model is lacking confidence. We have demonstrated the effectiveness of this technique in the case of automated scoring of story recall in verbal memory assessment (Chandler et al. 2022). The story recall task measures a person’s verbal episodic memory by asking them to listen to a short story and recall it with as many details as possible, and is an important aspect of neuropsychological assessment. For the automated scoring of this task, human-computer collaboration was implemented as

follows: an initial regression model was trained to predict how close a given recall was to the original story. The initial training was performed on as diverse and representative a dataset as possible. The model was then employed on new, unseen data and either all, or simply low confidence predictions were verified by a human. Active learning was harnessed in order to retain the newly labeled cases, representing diverse data for updating the model and making it more robust in the future. We recommend beginning this process with a high level of human oversight (i.e., human-in-command) and as the model becomes more accurate, oversight can begin to be lifted (i.e., human-in-the-loop) and eventually only needed for extenuating circumstances (i.e., human-on-the-loop). (For an overview of human-in-the-loop systems, see Monarch, 2021). Incorporating human-in-the-loop methodologies significantly improved the model's awareness of low confidence or knowledge gaps and required less than half of the training data used in a traditional setting to achieve a sufficient accuracy level (Chandler et al. 2022). Moving forward, the extent to which this technique may be used in other NLP applications should be explored.

As the use cases in Section 2 illustrate, the feasibility of human involvement at all stages of psychological NLP will vary between applications, as will the demand for human involvement from an ethical perspective. The principle of human agency should not be understood as discouraging the use of fully autonomous systems, specifically in low stakes applications with high constraint and little room for error. For example, in simple remote monitoring applications where actions are not directly being taken and data are simply collected, processed, and transmitted, there is less need for constant human oversight. However, the involvement of a human is critical in the cases where remote monitoring detects scenarios of high risk or if the purpose of the monitoring is in high risk scenarios. With that said, in a scenario where AI

predictions are directly made as part of a clinical diagnosis, there must *always* be human verification so as to avoid spurious and/or life threatening decisions. Given the complexity of language, there may be some situations where complete autonomy is unrealistic. The use of chatbots, particularly for purposes involving interventions with at-risk populations, is one such example (Powell, 2019).

While the complexity of an AI algorithm can vary widely: from rule-based AI, to simple and well understood linear models, to deep neural networks with high dimensional inputs and parameters, a universal set of guiding principles should apply to any non-deterministic and (semi) autonomous systems. Items to consider when choosing the appropriate algorithm for modeling include (i) weighing positive impacts against potential risks (i.e., does impact outweigh risk or vice versa?) and (ii) deciding how to relate the output from tools built for explanation and interpretation of algorithms to end users (e.g., researchers, clinicians or patients). Special considerations must be made, however, for any "black box" AI system not understandable by humans. What constitutes a sufficient explanation should be defined with stakeholder co-design (i.e., a process that incorporates various points of view of the clinician, the researcher, and other relevant stakeholders; Foltz et al., 2022), whether the explanation entails linking features to medical biomarkers, denoting irregular areas of speech, or simply describing the decision in interpretable language (e.g., Tschandl et al., 2020). Importantly, the human expert must understand both how the model is making its decisions and its limitations to avoid blind trust of predictions. Such a problem is called "automation bias" (Goddard et al., 2012) and is a well-documented issue for human-computer interaction that must be avoided. For instance, reliance on chatbots is likely to increase as the technology improves and errors appear less frequently.

To promote ethical applications, Leidner and Plachouras (2017) advocate for "ethical review boards," functioning similarly to institutional review boards, to oversee the process of design, development and deployment, thus supporting an "ethical by design" approach to new AI-based research projects. All stakeholders should be involved early in the conceptualization of language technologies, to mitigate concerns surrounding barriers to use, and lack of transparency of the system and its components (Ribera & Lapedriza, 2019). Furthermore, as noted earlier, to protect patient and consumer dignity, automated tools should disclose at the outset that users are interacting with an AI system and not a human (Kretzschmar et al., 2019), and each implementation of an AI-based system should be accompanied with a plain language "user's guide" for non-technical users of the tool. Basic concepts and keyword definitions surrounding model architectures, the model training process, performance and fairness criteria, should be clearly defined.

INSERT FIGURE 1 HERE

Section 6: From recommendation to application: a case study in psychological research

Early signs of cognitive decline are observable in speech, often years prior to a formal diagnosis of dementia. We used NLP and ML to develop a prototype screening tool for detection of cognitive decline that could potentially be used at scale for dementia screening (Diaz-Asper et al., 2022; Chandler et al., submitted). As shown in Figure 1, our *Rationale for NLP use* was to both improve the accuracy of current dementia screening tools and enable remote assessments (see also *Role in the workflow of the clinician)*. To develop the tool, we used speech samples recorded over the telephone from older individuals who were cognitively healthy or diagnosed with mild cognitive impairment or mild Alzheimer's disease. *Transparency* was addressed by documenting explanations of algorithmic inferences, contents of the data, and model

performance in published materials. Possible biases in the training data (including participant demographics) and algorithm (such as avoiding selection bias in choosing to report only the best performing, and potentially idiosyncratic, algorithm) were discussed. Importantly, only NLP features that were fully explainable and aligned with clinical constructs were retained in the model and documented publicly. The principle of *Human Agency* was considered through requesting stakeholder (including participant and clinician) feedback regarding the acceptability and utility of the potential tool (Diaz-Asper et al., 2021), and consulting with IRBs regarding limits to privacy and confidentiality and access to data. *Human Oversight* was achieved via a human-in-the-loop approach to visualizing the model's decision-making process and prediction confidence, allowing clinicians the ability to step in in the case of a faulty prediction. We regarded the *Consequences of an error* in the technology to be of medium impact, given that an incorrect conclusion could either be a false alarm, leading to an unnecessary clinical follow-up, or miss the need for follow-up, potentially delaying timely diagnoses. Responsibility for the error would fall on us in our role as the developer of the technology, however this risk is minimized through human-in-the-loop methodologies enabling thoughtful collaboration between the clinician and computer.

Summary & Conclusions

Ethical guidelines for AI are currently at the very general principle stage. Reviews of digital technologies for mental health revealed that only 15% of the studies discussed ethical implications, with a focus on participant privacy (Fiske et al., 2019), and "a near-complete exclusion of service users in conceptualization or development of algorithmic and data-driven technologies and their application to mental health initiatives" (Gooding & Kariotis, 2021, p.10).

The role ethics guidelines can play in terms of realizing AI policy aims (including NLP) is debatable. Hitherto developed AI ethical guidelines are non-binding, meaning that the consequences of not abiding by them may be minimal in terms of enforcement. [3] Ethical principles are also vaguely articulated, leaving them vulnerable to opportunistic interpretation, such as by actors who may have a low threshold for declaring publicly that they are compliant with the relevant principles. However, in psychological and medical professions there is a strong tradition and expectation of adherence to ethical principles, both when it comes to research and practice. In these domains, ethical guidance could play an important role in shaping the digital technological future. Thus, we have emphasized the potential role of the APA Ethics Code, and suggested how it may be improved to better reflect generally accepted principles of AI ethics. The key challenge, we argue, is to translate the general ethical principles into application-specific guidance. Questions of exactly *when* and *how* and *to what extent* we implement the appropriate measures must necessarily depend on the specific use case. Through the analysis in this paper, we have arrived at certain considerations for developers and users of NLP-based systems, pertaining to commonly accepted AI ethics principles, and we have highlighted certain nuances between different applications of NLP in psychology and described their ethical implications as they are currently understood. While our aim is to enable and encourage psychologists to assess the trustworthiness of NLP-based psychological tools, particularly when those tools are developed by non-psychologists, our contribution is merely an early stepping stone in the development of ethical guidance for NLP in psychology.

References

[3] In the EU, however, a legal regulation (the AI Act) has been proposed, which will transpose the ethical principles from the EU-HILEG Guidelines for Trustworthy AI into binding law.

AERA, APA, & NCME (2014). Standards for Educational and Psychological Testing: National Council on Measurement in Education. Washington DC: American Educational Research Association.

Amnesty International (2018, May 18). The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems. https://www.amnesty.org/en/documents/pol30/8447/2018/en/

Bailey, A. H., Williams, A., & Cimpian, A. (2022). Based on billions of words on the internet, people=men. *Science Advances*, *8*(13), eabm2463.

Bedi, G., Carrillo, F., Cecchi, G.A., Slezak, D.F., Sigman, M., Mota, N.B., Ribeiro, S., Javitt, D.C., Copelli, M. & Corcoran, C.M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, *1*(1), 1-7.

Behrens, J. T., DiCerbo, K. E., & Foltz, P. W. (2019). Assessment of complex performances in digital environments. *The Annals of the American Academy of Political and Social Science*, 683(1), 217–232

Bitterman, D. S., Aerts, H. J. W. L., & Mak, R. H. (2020). Approaching autonomy in medical artificial intelligence. *The Lancet Digital Health, 2(9),* e447-e449.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in neural information processing systems*, *29*. https://doi.org/10.48550/arXiv.1607.06520

Caliskan, A., Bryson, J.J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356(6334).* doi: 10.1126/science.aal4230

Casey, B., Farhangi, A., & Vogl, R. (2019). Rethinking explainable machines: The GDPR's right to explanation debate and the rise of algorithmic audits in enterprise. *Berkeley Technology Law Journal*(1), 143-188.

Chandler, C., Diaz-Asper, C., Turner, R. S., Reynold, B., & Elvevåg, B. (submitted). An explainable machine learning model of cognitive decline derived from acoustic and linguistic speech features.

Chandler, C., Foltz, P.W., Cohen, A.S., Holmlund, T.B., Cheng, J., Bernstein, J.C., Rosenfeld, E.P., and Elvevåg, B. (2020a). Machine learning for ambulatory applications of neuropsychological testing. *Intelligence-Based Medicine, 1-2,* 100006. https://doi.org/10.1016/j.ibmed.2020.100006

Chandler, C., Foltz, P.W. & Elvevåg, B. (2020b). Using machine learning in psychiatry: The need to establish a framework that nurtures trustworthiness. *Schizophrenia Bulletin, 46,* 11-14. https://doi.org/10.1093/schbul/sbz105

Chandler, C., Foltz, P. W., & Elvevåg, B. (2022). Improving the applicability of AI for psychiatric applications through human-in-the-loop methodologies. *Schizophrenia Bulletin, 48(5),* 949-957. doi: 10.1093/schbul/sbac038.

Chiang, S., Picard, R. W., Chiong, W., Moss, R., Worrell, G. A., Rao, V. R., & Goldenholz, D. M. (2021). Guidelines for conducting ethical artificial intelligence research in neurology: a systematic approach for clinicians and researchers. *Neurology*, *97*(13), 632-640. doi: 10.1212/WNL.0000000000012570.

Clancey, W.J., & Shortliffe, E.H. (Eds.) (1984). *Readings in Medical Artificial Intelligence: The First Decade*. Addison Wesley, Reading, MA.

Cohen, A.S., Fedechko, T.L., Schwartz, E.K., Le, T.P., Foltz, P.W., Bernstein, J., Cheng, J., Holmlund, T.B**.** & Elvevåg, B. **(**2019). Ambulatory vocal acoustics, temporal dynamics and serious mental illness. *Journal of Abnormal Psychology*, 128, 97-105. doi: 10.1037/abn0000397

Cohen, A. S., Rodriguez, Z., Warren, K. K., Cowan, T., Masucci, M. D., Edvard Granrud, O., Holmlund, T.B., Chandler, C., Foltz, P.W., & Strauss, G. P. (2022). Natural language processing and psychosis: on the need for comprehensive psychometric evaluation. *Schizophrenia Bulletin*, 48(5), 939-948. doi: 10.1093/schbul/sbac051.

Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights, 10*, 1178222618792860. doi: 10.1177/1178222618792860

Croes, E. A., & Antheunis, M. L. (2021). Can we be friends with Mitsuku? A longitudinal study on the process of relationship formation between humans and a social chatbot. *Journal of Social and Personal Relationships, 38(1),* 279-300. https://doi.org/10.1177/0265407520959463

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future healthcare journal, 6(2),* 94-98. doi: 10.7861/futurehosp.6-2-94

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2021). Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media, 7*(1), 128-137. https://doi.org/10.1609/icwsm.v7i1.14432

D'Mello, S. K., Tay, L., & Southwell, R. (2022). Psychological measurement in the information age: machine-learned computational models. *Current Directions in Psychological Science, 31(1*), 76–87. https://doi.org/10.1177/09637214211056906

Diaz-Asper, C. Chandler, C. Turner, R.S. Reynolds, B. & Elvevåg, B. (2021). Acceptability of collecting speech samples from the elderly via the telephone. *Digital Health* **7**, 1-10. https://doi.org/10.1177/20552076211002103.

Diaz-Asper, C., Chandler, C., Turner, R. S., Reynolds, B., & Elvevåg, B. (2022). Increasing access to cognitive screening in the elderly: Applying natural language processing methods to speech collected over the telephone. *Cortex, 156,* 26-38. https://doi.org/10.1016/j.cortex.2022.08.005

Elvevåg, B., Foltz, P. W., Weinberger, D. R., & Goldberg, T. E. (2007). Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia Research, 93*(1-3), 304-316. DOI: 10.1016/j.schres.2007.03.001

Factmr (2023, January). *Digital Education Content Market Growth Outlook (2023 to 2033).* https://www.factmr.com/report/digital-education-content-market

Faurholt-Jepsen, M., Torri, E., Cobo, J., Yazdanyar, D., Palao, D., Cardoner, N., Andreatta, O., Mayora, O. & Kessing, L.V. (2019). Smartphone-based self-monitoring in bipolar disorder: evaluation of usability and feasibility of two systems. *International Journal of Bipolar Disorders, 7(1),* 1-11. doi: 10.1186/s40345-018-0134-8

Fiske, A., Henningsen, P., & Buyx, A. (2019). Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of Medical Internet Research, 21*(5), e13216. doi: 10.2196/13216

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication*, (2020-1).

Foltz, P.W., Chandler, C., Diaz-Asper, C., Cohen, A.S., Rodriguez, Z., Holmlund, T.B. & Elvevåg, B. (2022). Reflections on the nature of measurement in language-based automated assessments of patients' mental state and cognitive function. *Schizophrenia Research.* S0920-9964(22)00283-3. doi: 10.1016/j.schres.2022.07.011. Epub ahead of print.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64(12),* 86-92.

GlobeNewswire (2023, January). *Healthcare Chatbots Market Size to Surpass USD 944.65 BN by 2032*. https://www.globenewswire.com/news-release/2023/01/04/2582717/0/en/Healthcare-Chatbots-Market-Size-to-Surpass-USD-944-65-BN-by-2032.html

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association : JAMIA, 19(1),* 121–127. doi: 10.1136/amiajnl-2011-000089

Gooding, P., & Kariotis, T. (2021). Ethics and law in research on algorithmic and data-driven technology in mental health care: scoping review. *JMIR Mental Health*, *8*(6), e24668. doi: 10.2196/24668.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science robotics, 4(37),* eaay7120.

Hauglid, M.K. (2022). What's that noise? Interpreting algorithmic interpretation of human speech as a legal and ethical challenge. *Schizophrenia Bulletin*, *48,* 960-962. https://doi.org/10.1093/schbul/sbac008

Hernandez-Boussard, T., Bozkurt, S., Ioannidis, J. P., & Shah, N. H. (2020). MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. *Journal of the American Medical Informatics Association*, *27*(12), 2011-2015. https://doi.org/10.1093/jamia/ocaa088

High-Level Expert Group on Artificial Intelligence (2019). Ethics guidelines for trustworthy AI. European Commission.

Hirschtritt, M. E., & Insel, T. R. (2018). Digital technologies in psychiatry: present and future. *Focus, 16(3),* 251-258. doi: 10.1176/appi.focus.20180001.

Hitczenko, K., Cowan, H. R., Mittal, V. A., & Goldrick, M. (2021). Automated coherence measures fail to index thought disorder in individuals at risk for psychosis. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 129-150. Association for Computational Linguistics (ACL).

Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The dataset nutrition label: A framework to drive higher data quality standards. arXiv preprint arXiv:1805.03677

Holmlund, T.B., Foltz, P.W., Cohen, A.S., Johansen, H.D., Sigurdsen, R., Fugelli, P., Bergsager, D., Cheng, J., Bernstein, J., Rosenfeld, E. & Elvevåg, B. (2019). Moving psychological assessment out of the controlled laboratory setting: Practical challenges. *Psychological assessment, 31*(3), 292-303. doi: 10.1037/pas0000647.

Hsiao, R. L. (2003). Technology fears: distrust and cultural persistence in electronic marketplace adoption. *The Journal of Strategic Information Systems*, *12*(3), 169-199. https://doi.org/10.1016/S0963-8687(03)00034-9

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1(9*), 389-399. https://doi.org/10.1038/s42256-019-0088-2

Kretzschmar, K., Tyroll, H., Pavarini, G., Manzini, A., Singh, I., & NeurOx Young People's Advisory Group. (2019). Can your phone be your therapist? Young people's ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support. *Biomedical informatics insights*, *11*, 1178222619829083.

Leidner, J. L., & Plachouras, V. (2017). Ethical by design: ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40, Valencia, Spain. Association for Computational Linguistics.

Luxton, D. D. (2020). Ethical implications of conversational agents in global public health. *Bulletin of the World Health Organization*, *98*(4), 285. doi: 10.2471/BLT.19.237636

McGreevey, J. D., Hanson, C. W., & Koppel, R. (2020). Clinical, legal, and ethical aspects of artificial intelligence–assisted conversational agents in health care. *JAMA*, *324*(6), 552-553. doi: 10.1001/jama.2020.2724

Miner, A. S., Shah, N., Bullock, K. D., Arnow, B. A., Bailenson, J., & Hancock, J. (2019). Key considerations for incorporating conversational AI in psychotherapy. *Frontiers in psychiatry*, *10*, 746. https://doi.org/10.3389/fpsyt.2019.00746

Monarch, R.M. (2021). Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI. Shelter Island, NY: Manning Publications Co.

Naslund, J. A., Gonsalves, P. P., Gruebner, O., Pendse, S. R., Smith, S. L., Sharma, A., & Raviola, G. (2019). Digital innovations for global mental health: opportunities for data

science, task sharing, and early intervention. *Current treatment options in psychiatry*, *6*(4), 337-351. doi: 10.1007/s40501-019-00186-8.

OECD (2019). *Recommendation of the Council on Artificial Intelligence*. OECD/LEGAL/0449. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Parmar, P., Ryu, J., Pandya, S., Sedoc, J., & Agarwal, S. (2022). Health-focused conversational agents in person-centered care: a review of apps. *NPJ digital medicine*, *5*(1), 1-9. https://doi.org/10.1038/s41746-022-00560-6

Pasquale, F. (2020). *New laws of robotics : defending human expertise in the age of AI*. The Belknap Press of Harvard University Press.

Powell, J. (2019). Trust me, I'm a chatbot: how artificial intelligence in health care fails the Turing test. *Journal of Medical Internet Research, 21(10),* e16222. doi: 10.2196/16222

Rainie, L., Anderson, J., & Vogels, E.A. (2021, June 16). Worries about developments in AI. Pew Research Center. https://www.pewresearch.org/internet/2021/06/16/1-worries-about-developments-in-ai/

Rajpurkar, P., Chen, E., Banerjee, O. et al. (2022). AI in health and medicine. *Nat Med, 28*, 31-38. https://doi.org/10.1038/s41591-021-01614-0

Ramesh, J., Aburukba, R., & Sagahyroon, A. (2021). A remote healthcare monitoring framework for diabetes prediction using machine learning. *Healthcare Technology Letters, 8(3),* 45-57. https://doi.org/10.1049/htl2.12010

Ribera, M., & Lapedriza, A. (2019). Can we do better explanations? A proposal of user-centered explainable AI. In *IUI Workshops* (Vol. 2327, p. 38).

Ruane, E., Birhane, A., & Ventresque, A. (2019). Conversational AI: Social and Ethical Considerations. *Irish Conference on Artificial Intelligence and Cognitive Science*. (pp.104-115).

Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law, 7*(4), 233-242. https://doi.org/10.1093/idpl/ipx022

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

Tschandl, P., Rinner, C., Apalla, Z. *et al.* (2020). Human–computer collaboration for skin cancer recognition. *Nat Med,* 26, 1229–1234. https://doi.org/10.1038/s41591-020-0942-0

UNESCO (2022). Recommendation on the Ethics of Artificial Intelligence. Paris, UNESCO.

U.S. Food and Drug Administration. Proposed regulatory framework for modifications to artificial intelligence/machine learning-based software as a medical device. Accessed February 14, 2020. https://www.fda.gov/media/122535/download

Vegesna, A., Tran, M., Angelaccio, M., & Arcona, S. (2017). Remote patient monitoring via non-invasive digital technologies: a systematic review. *Telemedicine and e-Health, 23(1),* 3-17. doi: 10.1089/tmj.2016.0051.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*(2), 76-99. https://doi.org/10.1093/idpl/ipx005

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational measurement: issues and practice 31,* 1**,** 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

World Health Organization. (2021). Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization. ISBN: 9789240029200

Yan, D., Rupp., A. C. & Foltz, P. W. (Eds.) (2020). *Handbook of Automated Assessment: Theory into practice.* Taylor & Francis, CRC Press.

Zicari, R. V., Brusseau, J., Blomberg, S. N., Christensen, H. C., Coffee, M., Ganapini, M. B., . . . Hildt, E. (2021). On assessing trustworthy AI in healthcare. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Frontiers in Human Dynamics, 30*. https://doi.org/10.3389/fhumd.2021.673104.

Figure 1

Considerations before implementing language technologies in psychological research and practice. (Check marks in the figure refer to the case study in Section 6).