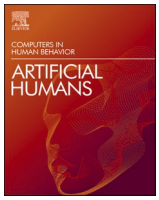
Computers in Human Behavior: Artificial Humans 2 (2024) 100066

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans



Confirmation bias in AI-assisted decision-making: AI triage recommendations congruent with expert judgments increase psychologist trust and recommendation acceptance

Anna Bashkirova **, Dario Krpan *

Department of Psychological and Behavioral Science, London School of Economics and Political Science, London WC2A 2AE, UK

ARTICLE INFO

Keywords:
Mental health service
Confirmation bias
Artificial intelligence
Healthcare technology
Diagnostic decision-making
Practitioner-AI interaction

ABSTRACT

The surging global demand for mental healthcare (MH) services has amplified the interest in utilizing AI-assisted technologies in critical MH components, including assessment and triage. However, while reducing practitioner burden through decision support is a priority in MH-AI integration, the impact of AI systems on practitioner decisions remains under-researched. This study is the first to investigate the interplay between practitioner judgments and AI recommendations in MH diagnostic decision-making. Using a between-subjects vignette design, the study deployed a mock AI system to provide information about patient triage and assessments to a sample of MH professionals and psychology students with a strong understanding of assessments and triage procedures. Findings showed that participants were more inclined to trust and accept AI recommendations when they aligned with their initial diagnoses and professional intuition. Moreover, those claiming higher expertise demonstrated increased skepticism when AI's suggestions deviated from their professional judgment. The study underscores that MH practitioners neither show unwavering trust in, nor complete adherence to AI, but rather exhibit confirmation bias, predominantly favoring suggestions mirroring their pre-existing beliefs. These insights suggest that while practitioners can potentially correct faulty AI recommendations, the utility of implementing debiased AI to counteract practitioner biases warrants additional investigation.

1. Introduction

Mental healthcare (MH) services worldwide are grappling with rising demand. In the UK, diagnoses of depression and anxiety disorders have surged by 13% in the past decade, exacerbated by the pandemic (McManus et al., 2009; WHO, 2022). This demand contrasts with staff shortages, resulting in increased staff workload and poorer treatment outcomes (Viswanathan et al., 2022). Patient triage and initial assessment stages of care are particularly staff intensive, with practitioners being required to evaluate the patient's MH status, and quickly and accurately assess case severity (Moss, 2016).

MH services are increasingly deploying AI assistive technologies to support patients and psychologists, particularly in the triage stage due to its structured procedure (Koutsouleris et al., 2022; Rollwage et al., 2022). MH AI chatbots, which support wait-list patients and collect the necessary patient information (e.g., symptoms, medical history), are being increasingly implemented by MH services such as the NHS 'Talking Therapies'. Research has shown that these chatbots can be effective in improving patient outcomes (Demner-Fushman et al., 2009; D'Alfonso et al., 2017). However, how practitioners interact with and trust the AI diagnostic recommendations—an understanding that is vital for the successful integration of AI systems into MH—remains under-investigated (Koutsouleris et al., 2022; Viswanathan et al., 2022).

Keeping the practitioner 'in the loop' during the AI integration into MH means that expert intuition will continue to play an integral role in clinical decision-making. While intuition allows practitioners to make decisions efficiently, especially under uncertainty, clinical judgement has been shown to possess a range of cognitive biases (Whelehan et al., 2020). No research to date has explored the interplay between practitioner intuition and AI recommendations in MH decision-making, including potential biases. Therefore, the present research draws on studies in other domains that investigated decision-making with AI technologies (i.e., street-level bureaucrats, SLBs; Snow, 2021; Selten et al., 2022; Meijer et al., 2021) and is grounded in psychological

* Corresponding author. .
** Corresponding author .
E-mail addresses: anna.bashkirova00@gmail.com (A. Bashkirova), d.krpan@lse.ac.uk (D. Krpan).

https://doi.org/10.1016/j.chbah.2024.100066
Received 24 November 2023; Received in revised form 18 March 2024; Accepted 19 March 2024
Available online 26 March 2024

theorizing on confirmation bias (Jonas et al., 2001; Selten et al., 2022).

We aimed to answer the following research question: *What is the effect of AI recommendations on the diagnostic decision-making of psychologists?* More specifically, we examined how the congruence of the AI recommendation with the psychologists' preliminary diagnosis would affect their likelihood of accepting and incorporating the recommendation into their final decision. The second aim of the study was to investigate whether this congruence would affect the perceived trustworthiness of the AI tool. In this context, we also examined the relationship between the perceived trustworthiness and the likelihood of acceptance of the AI tool, given that trust has been found to influence expert decision-making and AI technology acceptance (Viswanathan et al., 2022; Yu et al., 2019). Finally, we explored whether self-reported expertise in assessment and triage would impact the relationship between diagnosis congruency and AI recommendation acceptance or trust.

To accomplish our research objectives, we conducted a between-subjects, two-armed randomized control trial. The participants were either mental health practitioners or psychology/medical students with knowledge of assessments and triage procedures, all referred to as 'psychologists' for their role in the study. Participants were allocated into either the 'congruent' or 'incongruent' condition, which determined whether the AI recommendations they saw aligned with their preliminary diagnosis. The design of the study involved viewing three clinical case vignettes where participants first established a preliminary patient diagnosis and were then provided with an interface displaying triage information collected by a mock AI chatbot assistant called 'MindAssist'.

Overall, the results of this study indicate that psychologists were significantly more likely to accept and incorporate AI recommendations as part of their decision-making if they confirmed their preliminary diagnosis. Congruent AI recommendations were also perceived as significantly more trustworthy. The findings also showed that higher perceived trust was related to a higher likelihood of following the diagnosis. Finally, self-reported expertise moderated the relationships between congruence and perceived trustworthiness, and congruence and acceptance likelihood. More specifically, the higher the participants reported their expertise level in assessment and triage, the less likely they were to trust and accept the chatbot recommendations if they did not match their preliminary diagnoses. In the next sections, we first overview the literature that underpins the hypotheses that generated the described findings, and then report the research we conducted to test them.

2. Literature review

2.1. Automation technologies in mental healthcare (MH)

The healthcare sector is starting to embrace digital and automated technologies across various domains, aiding professionals in navigating increasingly complex administrative, diagnostic, and organizational challenges (Davenport & Kalakota, 2019). The MH domain has been more tentative in adopting automation, with the hesitancy often attributed to the importance of soft skills such as direct behavioral observation and rapport building (Gabbard & Crisp-Han, 2017; Graham et al., 2019). Nonetheless, AI tools have the potential to bring transformative benefits to the MH field by enhancing diagnostic capabilities, foreseeing the onset risks of mental disorders, and freeing up practitioner capacity (Bzdok & Meyer-Lindenberg, 2018; Luxton, 2014).

The MH stages of assessment and triage can particularly benefit from automation due to the standardized nature of the self-report questionnaires and assessments, which lends itself well to AI integration (Rollwage et al., 2022; Viswanathan et al., 2022). AI chatbots that have so far been integrated into MH systems demonstrated the ability to conduct psychotherapeutic conversations with patients and collect necessary triage data (Demner-Fushman et al., 2009; D'Alfonso et al., 2017). Despite the increasing integration of AI chatbots into MH services such as NHS's 'Talking Therapies', literature addressing the effectiveness of this technology, particularly in patient triage, remains highly limited (Car et al., 2020; NHS, 2023; Wilson et al., 2023). More precisely, studies have so far only looked at the effectiveness of AI tools in MH in isolation. For example, it was found that chatbots led to significant improvements in depression and anxiety outcomes and significantly reduced MH costs (Inkster et al., 2018; Rollwage et al., 2022). However, no research so far has examined how MH AI recommendations interact with and influence MH practitioner decision-making. As the AI tools are currently introduced to assist practitioners, a deeper exploration of their relationship with expert intuition is needed.

2.2. Practitioner intuition and decision-making in mental health

Practitioner psychologists implicitly use their professional training and lived experience when in contact with patients (Boomsma-van Holten et al., 2023). 'Professional intuition' among MH professionals denotes the automatic responses rooted in extensive, explicit learning from academic sources and hands-on clinical experience (Witteman et al., 2012). Despite being more error-prone than empirically based methods, practitioners argue for the importance of intuitive decision-making as clinical scenarios seldom present a singular 'optimal' solution and require them to understand the unique nature of each patient (Dawes et al., 1989; Grove et al., 2000). Clinicians often operate in conditions of high complexity and uncertainty and typically report combining an empirical approach with their professional intuitions in their decision-making (Witteman et al., 2012).

The naturalistic decision making (NDM) school of thought considers professional intuitions as 'expert intuitions'—valuable pattern-recognition tools honed over their careers (Kahneman & Klein, 2009). Conversely, the heuristics and biases approach casts a more critical eye, suggesting that human intuition is often laden with biases and heuristics, potentially leading to consistent errors (Kahneman & Klein, 2009; Moynihan & Lavertu, 2012). Although evidence in the MH field backs both perspectives, recent studies signal that MH professionals exhibit various cognitive biases (Bowes et al., 2020; Whelehan et al., 2020). Such biases are systematic but flawed reactions to judgment and decision-making challenges (Bowes et al., 2020; Wilke & Mata, 2012). Notably, biases play a substantial role in practitioner decision-making (Bowes et al., 2020; Featherston et al., 2020). Research estimates that cognitive biases contribute to between 36.5% and 77% of diagnostic errors in examined medical cases (Saposnik et al., 2016).

2.3. Confirmation bias

Clinical simulation research indicates that confirmation bias is one of the main contributors to premature case closures and incorrect diagnoses in healthcare (Prakash et al., 2017). Confirmation bias represents the tendency to search for information that confirms one's initial beliefs and ignore or distort data which contradicts them (DeWall et al., 2015). In the MH setting, confirmation bias may manifest through practitioners giving greater weight to data supporting their preliminary diagnosis and failing to seek out contradictory evidence supporting an alternative diagnosis (Mendel et al., 2011). In a study by Mendel et al. (2011) 13% of psychiatrists displayed confirmation bias, failing to seek evidence to contradict their preliminary diagnosis, resulting in diagnostic error. Diagnostic decisions are some of the most frequent error-prone decisions made by physicians (Newman-Toker & Pronovost, 2009). Therefore, factors such as cognitive biases that contribute to diagnostic inaccuracies significantly impact the quality of care (Mendel et al., 2011).

2.4. Practitioner-AI interaction

There has been extensive theoretical deliberation in healthcare on

the effect that substituting human decision-makers with algorithms would have on diagnostic error rates and care quality (Berner et al., 1999; Mendel et al., 2011; Sunstein, 2022). Research has shown that algorithms can help mitigate practitioner biases in MH (Brown et al., 2023; Ramnarayan et al., 2007). Authors such as Sunstein (2022) have proposed that algorithms are able to increase diagnostic accuracy through more appropriate symptom weighting and reduce biases such as 'current symptom' and 'availability bias' that commonly impact practitioners (Li et al., 2020; Sunstein, 2022). It is important to note, however, that algorithms have also been found to contain a range of errors, including biases, in clinical decision-making (Igoe, 2021; Parikh et al., 2019). AI decision-making errors across disciplines can generally be divided into the ones that are systematic (i.e., consistent and predictable, such as biases) and random (i.e., without a discernible pattern; Vicente & Matute, 2023). There are currently historical, sample, and knowledge-based biases in AI systems, stemming from training on outdated data that mirrors previous human errors. Such biases have led to clinical inaccuracies and disproportionate adverse impacts on marginalized groups (Minerva & Giubilini, 2023, pp. 1–9; Timmons et al., 2022). However, given that biases are systematic errors, there is a possibility for humans to investigate under what circumstances they occur to predict them (Hemmer et al., 2021).

To best mitigate both human and algorithmic bias, experts recommend integrating AI as decision-support systems for practitioners within MH (Busuioc, 2021; Koutsouleris et al., 2022). In this context, to achieve complementary team performance (CTP) that exceeds either AI or human performance individually (Hemmer et al., 2022), humans need to display appropriate reliance (Schemmer et al., 2023), which refers to individuals feeling empowered to differentiate when to rely on the AI advice versus their own choices (Wang & Yin, 2021; Yang et al., 2020).

The human 'in the loop' approach allows practitioners to still partly rely on their professional expertise and 'uncommon' sense to make certain decisions, retaining the critical human factors element of MH. The goal is for de-biased AI systems to address human biases while practitioners rely partly on their expertise to rectify potential algorithmic errors (Bullock et al., 2020; Miller et al., 2018; Veale & Brass, 2019).

2.5. Current research on the challenges of practitioner-AI interactions

In the human 'in-the-loop' approach, the degree of success for AI implementation will ultimately depend on how the practitioners will choose to use these tools (Snow, 2020). Practitioner intuition will continue to play an important role in clinical decision-making even after AI implementation (Snow, 2020).

Current research on AI in the MH field primarily contrasts the accuracy of purely algorithmic versus practitioner-based decisions (Vallejos et al., 2021; Wilson et al., 2023). To the authors' knowledge, no research has examined the interaction of practitioner judgement with algorithmic tools in MH or the potential biases present in such interactions.

The intersection of expert intuition and AI tools in professional decision-making has so far been mainly investigated in SLBs (Meijer et al., 2021; Selten et al., 2022; Snow, 2020). Intuition plays a significant role in SLBs as professionals rarely work with complete evidence and often need to 'satisfice' and make decisions under time and informational constraints (Kirkman & Melrose, 2014). Therefore, to inform the hypotheses of the current study, we relied on previous research regarding how SLBs use AI tools in their professional decision-making, and the biases that play a role in this context.

2.6. Two opposing theories regarding professional decision-making and AI recommendations—confirmation and automation bias

Automation and confirmation bias are two competing psychological theories that explore the interplay between professional knowledge and AI recommendations (Selten et al., 2022). Automation bias denotes the propensity to favor and over-rely on the recommendations made by automated systems and aids (e.g., computers, AI chatbots and assistants, recommendation algorithms, decision-making software) when making decisions (Goddard et al., 2012; Mosier & Skitka, 1996; Skitka et al., 1999). In healthcare, evidence suggests that practitioners display automation bias in highly complex tasks requiring multitasking (Lyell & Coiera, 2017). Broader research, however, suggests that automation bias might not be the primary way experts engage with AI tools (Meijer et al., 2021; Snow, 2021).

Indeed, outside of the healthcare field, Grgić-Hlača et al. (2022) have demonstrated that individuals do not always accept recommendations by AI systems. Instead, they are prone to accept the recommendations when the AI systems make judgment errors similar to their own errors, but are significantly less likely to in other circumstances. In line with this finding, research has consistently shown that professionals are likely to exhibit confirmation bias rather than automation bias when it comes to recommendations provided either by human colleagues (Elston, 2020; Mendel et al., 2011) or AI systems (Selten et al., 2022; Snow, 2021). In other words, professionals tend to accept information that confirms their prior beliefs rather than generally accepting human or automated recommendations irrespective of their views. For example, in the context of SLBs, child service professionals frequently engaged in "biased artificing' (Snow, 2021), as explained by the 'Heuristics and Biases' approach (Kahneman & Tversky, 2009), which means that they rejected AI recommendations that did not align with their pre-existing beliefs.

In a separate study, Selten et al. (2022) explored how police officers interacted with algorithmic decision tool recommendations after a crime. The findings indicated that officers displayed confirmation bias, showing a significantly higher propensity to trust and accept AI recommendations that aligned with their intuitive professional judgments. This pattern was also observed in teachers, who were hesitant to accept AI suggestions that contradicted their prior knowledge of students (Nazertsky et al., 2021). Alon-Barkat and Busuioc (2021) further discovered that teachers demonstrated a selective preference for AI recommendations that reinforced their pre-existing stereotypes. In a human- AI interaction study on ethical decision-making for allocating kidney transplants, Narayanan et al. (2023) found that participant reliance on AI advice did not only depend on the ethical values exhibited by humans and AI algorithms, but also the similarities between them. Individuals were much more likely to accept a recommendation given by an AI that displayed ethical values which were similar to the ones they themselves possessed (Narayanan et al., 2023).

To test whether these insights informed by research on confirmation and automation bias would apply in the MH context, our study measured the likelihood of accepting and incorporating the AI recommendation as the main outcome variable. We proposed the following hypothesis.

H1. Psychologists will be more likely to accept and incorporate congruent AI recommendations as part of their diagnosis than incongruent AI recommendations.

2.7. Perceived trustworthiness of AI tools

One of the major challenges for adopting AI tools for triage and decision support in MH is practitioner trust in the AI solutions (Aktan et al., 2022; Viswanathan et al., 2022). Trust critically impacts human-machine interaction, thus influencing decision-making, performance, and overall experience in the context of this interaction (Yu et al., 2019). Aktan et al. (2022) found that individuals who worked in professions related to psychology were particularly reluctant to accept AI tools in psychotherapeutic interventions. Recent AI attitude studies revealed that over a third of European practitioners harbor mistrust towards AI tools, often due to concerns about dehumanizing healthcare and undermining the therapist-patient bond (Darau, 2022; Minerva & Giubilini, 2023, pp. 1–9). Although presenting AI as an assistant rather

than a primary decision-maker partially mitigates these concerns, many practitioners still feel compelled to continuously validate the algorithm's decisions (Viswanathan et al., 2022). Since algorithms are often expected to be perfect decision-makers, AI judgement error disproportionately decreases trust in a way that would not occur for a human practitioner (Leichtmann et al., 2023; Nazaretsky et al., 2021; Yu et al., 2019).

The predictability of a system plays a fundamental role in trust; however, due to the increased complexities of automated systems, practitioners are no longer likely to grasp all the technicalities behind an algorithmic decision (Lee & Moray, 1992; Yu et al., 2019). Explainable AI, where the 'why' behind the AI tools decision is provided to the individual, has been argued to be an important way to increase perceived trustworthiness (Ahmad et al., 2018; Leichtmann et al., 2023; Miller, 2018). However, research on expert AI dynamics in SLB professions found that such explanations did not significantly affect trustworthiness (Selten et al., 2022). Selten et al. (2022) highlighted that a professional's prior knowledge and alignment of AI explanations with existing beliefs played a more crucial role in trust perception. Moreover, trust in technology was found to be more significantly influenced by confirmation bias, with individuals being more skeptical of technology that challenges their pre-held beliefs (Nazaretsky et al., 2021; Selten et al., 2022).

In line with these insights, we proposed the following hypothesis.

H2. Psychologists will perceive AI recommendations congruent with their preliminary diagnosis as more trustworthy than AI recommendations incongruent with this diagnosis.

Research has also shown a positive relationship between perceived trustworthiness and the likelihood of accepting an AI recommendation, with an increase in trustworthiness being associated with an increased acceptance likelihood (Selten et al., 2022). This is related to evidence that trust largely shapes the behaviors surrounding AI tools (Aktan et al., 2022; Nazaretsky et al., 2021). To investigate the relationship between acceptance likelihood and perceived trustworthiness of the AI recommendation in the context of MH, we proposed the following hypothesis.

H3. Higher perceived trustworthiness of the AI chatbot triage recommendation will be related to an increased likelihood of accepting the recommendation.

2.8. Self-reported expertise in assessment and triage

Another factor that needs to be explored is the effect of domain expertise on the individual propensity to have confirmation bias, which has so far yielded mixed findings. (Bowes et al., 2011; Krems & Zierer, 1994; Mendel et al., 2011). Krems & Zierer, 1994 found that expertise in terms of years worked reduced confirmation bias for medical diagnoses, with high-domain knowledge experts being more likely to modify their assumptions when faced with contradictory evidence. However, Mendel (2011) found that students were only slightly more susceptible to bias-driven errors than expert psychiatrists, regardless of the latter being substantially more experienced. On a broader scale, some research suggests that expertise does not offer significant protection against cognitive biases or markedly enhance accuracy (Bowes et al., 2011; Spengler et al., 2009). Mizrahi (2018) contends that expert judgments are as prone to cognitive biases as novices, especially under uncertainty. Goldberg et al. (2016) even argue that increased experience could lead psychiatrists to make less accurate diagnoses. Additionally, when individuals rate their own expertise in a field, they are more strongly inclined to defend their initial point of view and construct explanations regarding why the contradiction is incorrect (Atir et al., 2015).

In line with these assumptions, we proposed the following hypothesis.

H4. The link between AI chatbot recommendation congruency and the acceptance likelihood or perceived trustworthiness will be moderated by self-reported diagnosis expertise, with higher expertise increasing the positive impact of congruent (vs. incongruent) recommendations on the trustworthiness and acceptance likelihood.

3. Materials and methods

3.1. Design

The study was an online, between-subjects, two-armed (congruent vs. incongruent) randomized control trial. In the study, participants read three vignettes describing imaginary clinical cases of individuals who might suffer from a mental health disorder and were asked to make a preliminary diagnosis. Participants allocated to the congruent condition then saw an AI e-triage chatbot recommendation that aligned with their preliminary diagnosis. In the incongruent condition, the diagnosis recommendation from the chatbot did not align with the participants' preliminary diagnosis. Even though the participants completed three separate vignettes, the study was not a mixed design as the vignette responses were collated into one average, and only between-group differences were analyzed.

3.2. Participants

3.2.1. Sample size and demographics

Participants had to be either psychology/medical students with a strong understanding of assessments and triage procedures or clinical practitioners, trainees, or MH counsellors. The strong understanding of assessment and triage was self-determined by the students; they were asked to judge their own knowledge base on the subject. Participants also had to be over 18 and fluent in English. In total, 161 participants completed the study, and 114 were found eligible. An a priori sample size calculation was conducted using G* Power (setting: power = 0.80, Cohen's f effect size = 0.25 [i.e., medium effect], Type I error = 0.5; Faul et al., 2007). The sample size required was N = 128. A sensitivity power analysis further showed that, with the sample of 114 participant who were eventually used in statistical analyses, the study had a power of 0.80 (Type I error = 0.5) to detect Cohen's f of 0.26, which is close to the medium effect size originally intended (Faul et al., 2007).

Participant demographics can be seen in Table 1. For reasons of privacy and lack of necessity for other data, the only demographic information collected involved gender and profession. The sample was largely female (67%), partly representative of the UK's mental health practitioner and student demographics (NHS 75, 2018; MSC, 2018). Mental health professionals comprised 20% of the sample, which was expected due to known recruitment issues (Asch et al., 2000). The

Table 1
Frequency table for demographic variables.

Variable	Category	Frequency (%)
Gender	Male	35 (30%)
	Female	76 (67%)
	Other	2 (2%)
	Prefer not to say	1 (1%)
Profession	Practicing clinical psychologist	2 (2%)
	Trainee clinical psychologist	13 (11%)
	Licensed mental health counsellor	4 (4%)
	Undergraduate medical/psychology student	59 (51%)
	Postgraduate medical/psychology student	33 (29%)
	Other	3 (3%)
Recruitment Channel	Volunteer	31 (27%)
	Prolific	83 (73%)

Note. In the present research, we use the term *Gender* in relation to participants' gender identity—that is, whether they identify themselves as males, females, a non-binary gender (i.e., Other), or prefer not to say. For Profession, all participants who selected Other were mental health nurses.

participants who selected 'other' were all mental health nurses.

3.2.2. Sampling and recruitment

There were two recruitment channels: volunteer and a participant recruitment platform Prolific.com. Data collection started on June 26, 2023 and ended on July 16, 2023. Two recruitment strategies were employed within the volunteer sample: snowball and gatekeeper sampling. The snowball sampling strategy was majorly applied to psychology and medical students. The questionnaire was distributed to a group of students who were asked to complete and distribute the study. The study was also advertised on the researcher's social media channels (WhatsApp, Instagram, LinkedIn).

Gatekeeper sampling was applied to recruit clinical and trainee psychologists. A psychology professor from a university in London, UK was asked to advertise the research. The professor distributed the link to the questionnaire to the psychologists via email.

Participants were also recruited through the Prolific online platform and compensated at Prolific's standard rate of £9/hour. Prolific recruitment was conducted in two waves. The first aimed to accumulate numerous participants fluent in English, over 18, and with an educational background in psychology or medicine, regardless of professional status. The second wave focused solely on practitioner psychologists and mental health professionals. Overall, this recruitment strategy was necessary so we could recruit participants with different levels of expertise to be able to investigate how expertise impacts the relationship between diagnosis congruency and AI recommendation acceptance or trust.

3.3. Procedure

3.3.1. Data-collection and randomization

On average, participants took 13.6 min to complete the study. The questionnaire was distributed online using Qualtrics XM. The participants were randomized into either the congruent or incongruent condition using the Qualtrics randomizer and the 'embedded data' feature.

3.3.2. Ethics and consent

The study was rated low-risk and received ethical approval from the Research Ethics Committee of the authors' university on May 24, 2023, reference: 225,044. Participants gave written informed consent to take part in the study. In the information sheet, participants were made aware of the nature of the study, their involvement, researcher contact details, and their rights as participants (see Supplementary Information, pp.2-3). The participants were also informed that there were no risks associated with the study and that they could benefit from a better understanding of their attitudes towards E-triage recommendations. No sensitive information was collected, and participants were not from vulnerable groups (e.g., under 18s). The data was completely anonymized; Qualtrics was set not to record identifiable information such as participant IP address.

3.3.3. Experimental procedure

As can be seen in Fig. 1, the Prolific and volunteer participants completed the same questionnaire. The only difference was the unique participant ID and completion code required for Prolific participants to prove completion.

The steps of study completion can be seen in Fig. 1. The participants started the experiment with an introductory page, proceeded to check the inclusion criteria, and were then taken to the consent form. After consenting, participants completed the demographic questions, self-reported expertise, and the general attitudes towards artificial intelligence scale (GAAIS; Schepman & Rodway, 2020) presented in a matrix format. For the main task, participants reviewed three hypothetical vignettes and established a preliminary diagnosis. The preliminary diagnosis selection had six options: Depression, GAD, Social Anxiety, Panic Disorder, PTSD, and OCD. After the diagnosis, participants were provided with information about the MindAssist chatbot. They were asked to imagine that the hypothetical patients underwent the initial triage and assessment stage with the AI chatbot. To increase explainability, they were told about the information the chatbot collects, the psychological assessments it uses, and that it employs natural language processing (NLP). They were also informed that it was designed in collaboration with clinicians. Then, they saw the congruent or incongruent chatbot interface with a diagnosis recommendation and asked how likely they were to accept and incorporate the recommendation into their diagnosis. This process was repeated for all three vignettes.

The final section of the study included the perceived trustworthiness and human vs. AI practitioner preference measures presented in matrix tables. The section also included an open qualitative question which asked participants to state what made them perceive MindAssist's recommendations as trustworthy/untrustworthy. The qualitative data were collected for exploratory purposes and were not used in analyses. After completion, participants were thanked for their time and informed that their response had been recorded.

3.4. Materials

3.4.1. Main measures used in hypothesis testing

3.4.1.1. Likelihood of acceptance. Likelihood of acceptance was measured on a 6-item Likert scale framed as a statement 'How likely are you to accept the AI recommendation and incorporate it as part of the diagnosis?' The study response options ranged from 1 (very likely) to 6 (very unlikely). This item was reverse-coded for the purpose of analyses to facilitate interpretability. The participants rated their likelihood of accepting the recommendation for each of the three vignettes, and the average score of the reverse-coded ratings was used as the *likelihood of acceptance* variable. This measure was modelled on Selten et al.'s (2022) acceptance measure.

3.4.1.2. Perceived trustworthiness. Perceived trustworthiness was measured using a scale developed by Grimmelikhuijsen (2023) to assess trust in AI systems. The content of the questions was adapted for the chatbot context. The scale therefore included the following five items 1) I trust that the AI chatbot was able to collect participant information required for assessment and triage, 2) I trust that the information used was correct, 3) I trust the AI chatbot in giving the correct diagnosis recommendation, 4) I trust the AI chatbot assessed the situation honestly, 5) I believe that all information provided was relevant. The items have been measured on a 1 (no trust at all) to 6 (complete trust) scale. This altered scale was highly reliable (Cronbach's alpha = 0.88), similar to the original scale (Cronbach's alpha = 0.92).

3.4.1.3. Self-reported expertise. Self-reported expertise was measured with the self-perceived knowledge scale (Atir et al., 2015). The 2-item scale was originally designed to measure self-reported financial knowledge but was adapted to test MH expertise and used in previous studies (Browne et al., 2007). The two items were "In general, how knowledgeable would you say you are about mental health triage and assessment?" and "Compared to an average person living in the UK, how knowledgeable are you about triage and assessment?". The items were measured on a 5-item Likert scale from (1) Not knowledgeable at all to (5) extremely knowledgeable.

3.4.2. Secondary measures used as covariates

3.4.2.1. General attitudes towards artificial intelligence. The GAAIS questionnaire (Schepman & Rodway, 2020) was used to measure participants' general attitudes towards AI technologies. The original scale had 20 items (Schepman & Rodway, 2020) and was reduced to 10 items due to timing concerns. The top five positive and five negative items

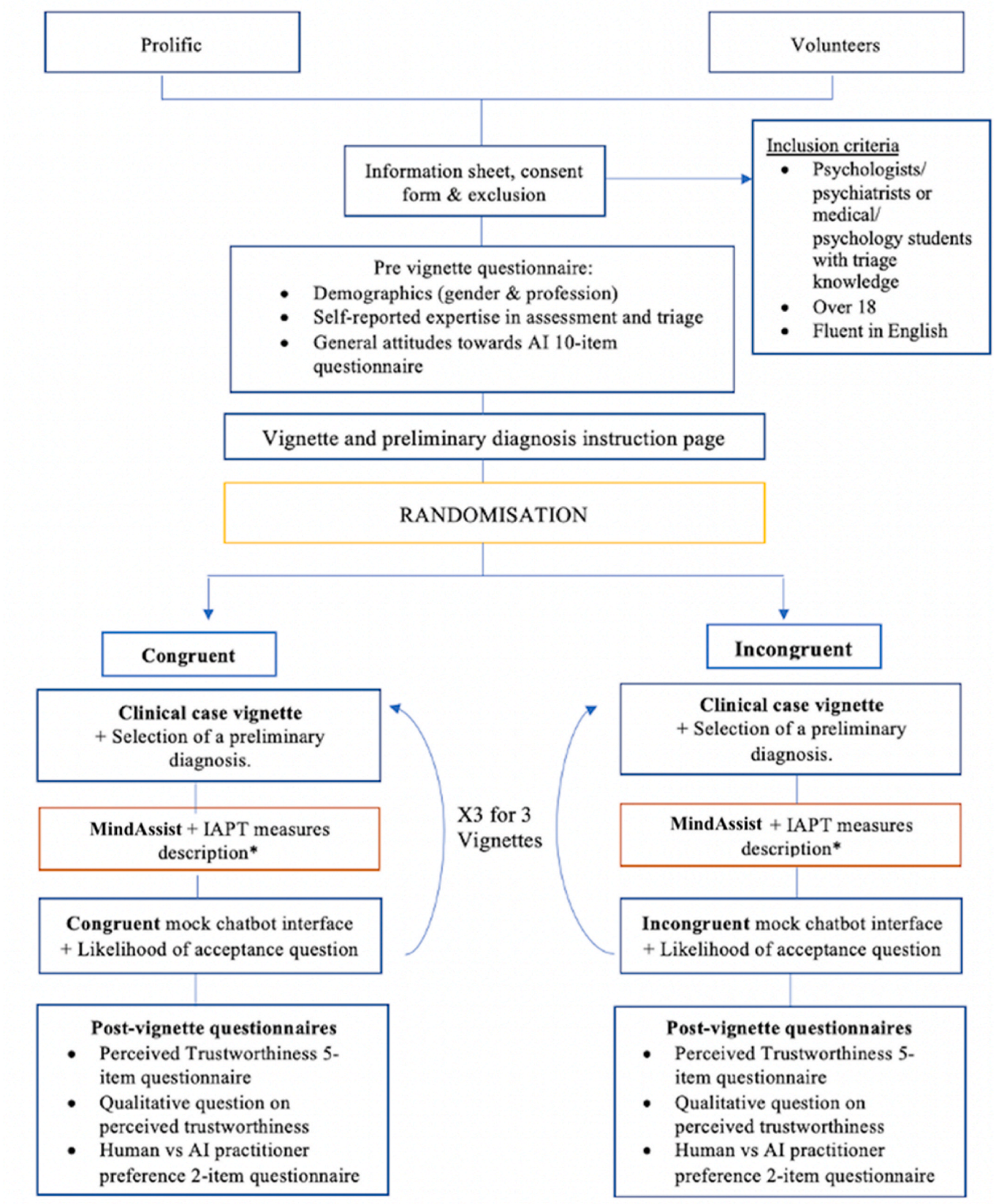


Fig. 1. Flowchart of the experimental procedure.

with the highest factor loadings were chosen. The items were measured on a 5-item Likert scale from (1) Strongly Disagree to (5) Strongly Agree. Some items were: 'I am interested in using artificially intelligent systems in my everyday life' and 'Artificial intelligence is dangerous' (for all items, see Supplementary Information, p.4). The reduced scale had good reliability (Cronbach's alpha = 0.79) that was almost identical to the original scale (Cronbach's alpha = 0.80).

3.4.2.2. Human vs. AI practitioner preference. To measure participants' general preference between AI and human practitioner triage recommendations, the following two statements assessed on a 5-item Likert scale from (1) strongly disagree to (5) strongly agree were used: 'I would trust the diagnosis recommendations more if they were provided by a human' and 'I would be more likely to accept the diagnosis recommendation if it was provided by a human'. The questions were modelled after qualitative research on human-AI interaction in MH (Viswanathan et al., 2022).

3.4.3. Vignette design and diagnosis selection

The three vignettes were designed to purposefully have an ambiguous diagnosis, with no one correct disorder. There were six preliminary diagnosis options for the most common anxiety and depression disorders: Depression, GAD, Social Anxiety, Panic Disorder, Post-Traumatic Stress Disorder and OCD (McManus, Bebbington, Jenkins, & Brugha, 2016). For each vignette, certain disorder choices were more obvious than others. Vignette 1 - OCD/GAD/Panic Disorder, Vignette 2 - PTSD/Depression, and Vignette 3 - Social Anxiety/GAD (the vignettes are available in Supplementary Information, p.5). To avoid the chance that the participants would recognize the vignettes from clinical training or university, the vignettes were generated by ChatGPT 4 (for the prompts used, see Supplementary Information, p.6). To ensure that the vignettes had clinical validity, they were validated by two clinical

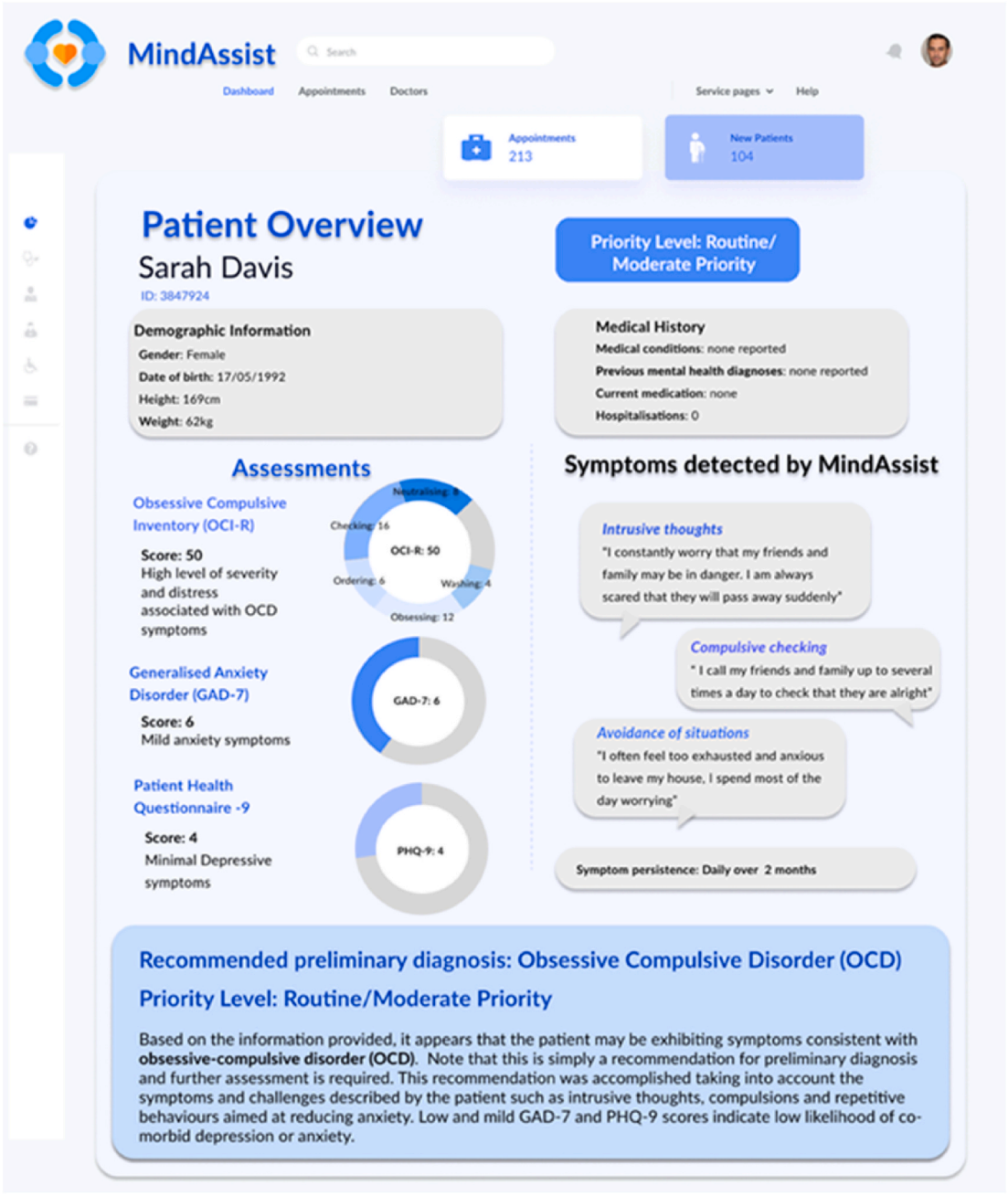


Fig. 2. MindAssist chatbot interface for Vignette 1 OCD diagnosis recommendation.

psychologists from the UK and India.

3.4.4. Mock AI chatbot interface

The mock AI chatbot interfaces were designed using Figma. A total of 18 interfaces were designed, 6 for each vignette. One interface was designed for each of the six diagnostic options for the congruent condition, allowing participants to get the diagnosis they selected confirmed by the chatbot. The interfaces for the incongruent chatbot were the same as the congruent ones, but participants were presented with an interface that did not match their selected diagnosis.

The interfaces was designed based on the Improving Access to Psychological Therapies (IAPT) manual guide for the assessment and triage of mental disorders(England, 2018). The report was used for guidance on the standard information collected during triage and what type of mental health questionnaires are utilized for each diagnosis. To further inform design, interface content and visual aspects from chatbots such as Wysa and Limbic AI were used as guidance. Additional evidence to inform interface design was gathered through academic research on e-triage and chatbot company websites (Inkster et al., 2018; Rollwage et al., 2023; Limbic Assess, 2023). To make sure that the interfaces are similar to the triage information usually provided to clinical psychologists, two clinical master's graduates validated the vignette designs.

The mock AI chatbot was called 'MindAssist'. It provided the following information to the participants: demographic information, medical history (previous mental health/medical conditions etc.), priority level, symptoms detected by MindAssist, psychological assessments administered (e.g., PHQ-9), and preliminary diagnosis recommendation. The participants were repeatedly told that MindAssist is there to provide assistance and guidance, and their diagnosis recommendations are preliminary. Fig. 2 shows the interface design vignette 1 for an OCD diagnosis recommendation.

4. Results

4.1. Data cleaning

In total, 47 participants were excluded after the study completion. Nine did not meet the inclusion criteria (six were not the right profession, one not fluent in English and two under 18). 20 participants did not move past the consent page, two failed at least one of the attention checks, and 10 provided only incomplete information and therefore could not be used in analyses. Four participants had some incomplete information but were used in analyses because they responded to several of the variables that were used in hypothesis testing.

Stata automatically omitted the empty variables from the analysis for scales with no responses from these participants. Supplementary Information (p.7) contain comprehensive information on how all variables used in analyses were coded and computed in Stata.

4.2. Descriptive statistics

Table 2 describes the number of participants and the means and standard deviations for the three main measures—trustworthiness, acceptance likelihood, and perceived expertise. The numbers of participants are different between outcome measures due to certain responses having incomplete data.

Table 2
Means and standard deviations (SDs) of the main study measures divided by diagnosis congruence.

Variable	Congruent			Incongruent		
	N	*Mean*	*SD*	*N*	*Mean*	*SD*
Acceptability	58	4.966	0.821	53	4.101	0.955
Trustworthiness	58	4.559	0.763	52	3.719	0.960
Expertise	59	3.678	0.662	55	3.509	0.767

Note. Acceptability ranges from 1 (very unlikely) to 6 (very likely), with higher scores corresponding to higher acceptability; Trustworthiness ranges from 1 (no trust at all) to 6 (complete trust); and Expertise ranges from (1) Not knowledgeable at all to (5) extremely knowledgeable.

4.3. Testing H1

To test the first hypothesis that psychologists will be more likely to accept the AI recommendations that are congruent with their preliminary diagnosis, we first conducted an independent samples *t*-test. In line with the prediction, the test showed that the mean likelihood of acceptance was significantly higher for participants who received congruent ($M = 4.966$, $SD = 0.821$) versus incongruent ($M = 4.101$, $SD = 0.955$) chatbot recommendations, $t\ (109) = 5.129$, $p = <0.001$, $d = 0.971$. To ensure that this effect was robust regardless of potential confounds, we conducted a multiple linear regression with congruence as the independent variable and gender, recruitment channel, profession, average GAAIS score, and human vs. AI preference added as covariates. As can be seen from Table 3, the effect of congruence versus incongruence on acceptance likelihood remained highly significant.

4.4. Testing H2

To test the second hypothesis that psychologists will perceive AI recommendations that are congruent with their preliminary diagnosis as more trustworthy, we first conducted an independent samples *t*-test. In line with the prediction, the test showed that the mean perceived trustworthiness score was significantly higher for participants who received congruent ($M = 4.559$, $SD = 0.763$) versus incongruent ($M = 3.719$, $SD = 0.960$) chatbot recommendations, $t\ (108) = 5.102$, $p = <0.001$, $d = 0.969$. To ensure that this effect was robust regardless of potential confounds, we conducted a multiple linear regression with congruence as the independent variable and gender, recruitment channel, profession, average GAAIS score, and human vs. AI preference added as covariates. As can be seen from Table 4, the effect of congruence versus incongruence on perceived trustworthiness remained highly

Table 3
Results of a Multiple Linear Regression Analysis Examining the Influence of Diagnosis Congruence on the Likelihood of Acceptance of E-Triage Chatbot Recommendations, while Controlling for Gender, Recruitment Channel, Profession, GAAIS, and Human vs. AI Preference.

Variable	*B*	*SE*	*t*	*p*	95% CI	
					LL	*UL*
Constant	3.251	0.840	3.871	<0.001	1.585	4.916
Congruence	0.821	0.171	4.807	<0.001	0.482	1.160
Gender Female	0.380	0.198	1.922	0.057	−0.012	0.772
Gender Other	−0.267	0.537	−0.498	0.620	−1.332	0.798
Recruitment Channel	0.152	0.203	0.748	0.456	−0.251	0.556
Profession	0.177	0.223	0.796	0.428	−0.264	0.619
GAAIS	0.272	0.152	1.786	0.077	−0.030	0.573
Human vs. AI Preference	−0.178	0.116	−1.532	0.129	−0.410	0.053

Note: Model $R^2 = 0.263$, $F(7, 102) = 5.206$, $p < 0.001$. *B* refers to raw regression coefficients. For Congruence, 0 = Incongruent and 1 = Congruent; for Gender, only 3 participants in total responded with "other" and "prefer not to say" (Table 1), so for this regression analysis they were collapsed into one category—Gender Other; for both Gender Female and Gender Other, Gender Male is the comparison category; for Recruitment Channel, 0 = Volunteer and 1 = Prolific.com; for Profession, categories that comprise practicing and trainee clinical psychologists, licensed mental health counsellors, and others (i.e., mental health nurses; Table 1) are coded as 1 = Practitioners, and categories that comprise postgraduate and undergraduate medical/psychology students (Table 1) are coded as 2 = Students. Finally, the scores for GAAIS and Human vs. AI Preference can range from (1) Strongly Disagree to (5) Strongly Agree. None of the variables in the analysis were standardized.

Table 4
Results of a Multiple Linear Regression Analysis Examining the Influence of Diagnosis Congruence on Perceived Trustworthiness of E-Triage Chatbot Recommendations, while Controlling for Gender, Recruitment Channel, Profession, GAAIS, and Human vs. AI Preference.

Variable	*B*	*SE*	*t*	*p*	95% CI	
					LL	*UL*
Constant	1.673	0.773	2.165	0.033	0.140	3.206
Congruence	0.777	0.157	4.938	<0.001	0.465	1.089
Gender Female	0.182	0.182	0.998	0.320	−0.179	0.543
Gender Other	−0.318	0.494	−0.644	0.521	−1.299	0.662
Recruitment Channel	0.049	0.187	0.262	0.794	−0.322	0.421
Profession	0.241	0.205	1.176	0.243	−0.166	0.647
GAAIS	0.540	0.140	3.858	<0.001	0.262	0.818
Human vs. AI Preference	−0.086	0.107	−0.806	0.422	−0.299	0.126

Note: Model $R^2 = 0.331$, $F(7, 102) = 7.226$, $p < 0.001$. *B* refers to raw regression coefficients. For Congruence, 0 = Incongruent and 1 = Congruent; for Gender, only 3 participants in total responded with "other" and "prefer not to say" (Table 1), so for this regression analysis they were collapsed into one category—Gender Other; for both Gender Female and Gender Other, Gender Male is the comparison category; for Recruitment Channel, 0 = Volunteer and 1 = Prolific.com; for Profession, categories that comprise practicing and trainee clinical psychologists, licensed mental health counsellors, and others (i.e., mental health nurses; Table 1) are coded as 1 = Practitioners, and categories that comprise postgraduate and undergraduate medical/psychology students (Table 1) are coded as 2 = Students. Finally, the scores for GAAIS and Human vs. AI Preference can range from (1) Strongly Disagree to (5) Strongly Agree. None of the variables in the analysis were standardized.

significant.

4.5. Testing *H3*

The next step was to investigate whether the perceived trustworthiness of the AI recommendations was related to the psychologists' reported likelihood of accepting the diagnosis recommendation. This was examined using a Pearson correlation analysis, which showed that higher acceptance likelihood was strongly associated with higher perceived trustworthiness, $r(110) = 0.751$, $p < 0.001$. The scatter plot in Fig. 3 visually demonstrates the positive relationship between perceived trustworthiness and acceptance likelihood.

4.6. Testing *H4*

Finally, to test *H4* that the relationship between recommendation congruence and both the acceptance likelihood and perceived trustworthiness of chatbot would be moderated by the self-reported diagnosis expertise, we computed the interactions between congruence and the expertise using linear regressions. As can be seen in Table 5, the interaction between the two variables concerning acceptance likelihood was significant. The analyses of simple slopes further showed that, at higher levels of expertise (+1 *SD*), congruence significantly increased acceptance likelihood compared to incongruence ($M_{diff} = 1.244$, $p < 0.001$), whereas at lower levels of expertise (−1 *SD*) the effect was smaller and non-significant ($M_{diff} = 0.475$, $p = 0.054$). To ensure that the interaction effect remained significant regardless of potential confounds, we also computed it while adding gender, recruitment channel, profession, average GAAIS score, and human vs. AI preference as covariates. As can be seen in Table 5, the interaction effect remained significant. Moreover, the analyses of simple slopes showed that, at higher levels of expertise (+1 *SD*), congruence significantly increased acceptance likelihood compared to incongruence ($M_{diff} = 1.197$, $p < 0.001$), whereas at lower levels of expertise (−1 *SD*) the effect was smaller and non-significant ($M_{diff} = 0.439$, $p = 0.078$). Therefore, the analyses remained robust despite covariates.

Concerning acceptance likelihood as the dependent variable, the interaction between congruence and the self-reported diagnosis expertise was also significant (Table 6). The analyses of simple slopes further showed that, at higher levels of expertise (+1 *SD*), congruence significantly increased perceived trustworthiness compared to incongruence ($M_{diff} = 1.235$, $p < 0.001$), whereas at lower levels of expertise (−1 *SD*) the effect was smaller and non-significant ($M_{diff} = 0.424$, $p = 0.077$). To ensure that the interaction effect remained significant regardless of potential confounds, we also computed it while adding gender, recruitment channel, profession, average GAAIS score, and human vs. AI preference as covariates. As can be seen in Table 6, the interaction effect remained significant. Moreover, the analyses of simple slopes showed that, at higher levels of expertise (+1 *SD*), congruence significantly increased acceptance likelihood compared to incongruence ($M_{diff} = 1.126$, $p < 0.001$), whereas at lower levels of expertise (−1 *SD*) the effect

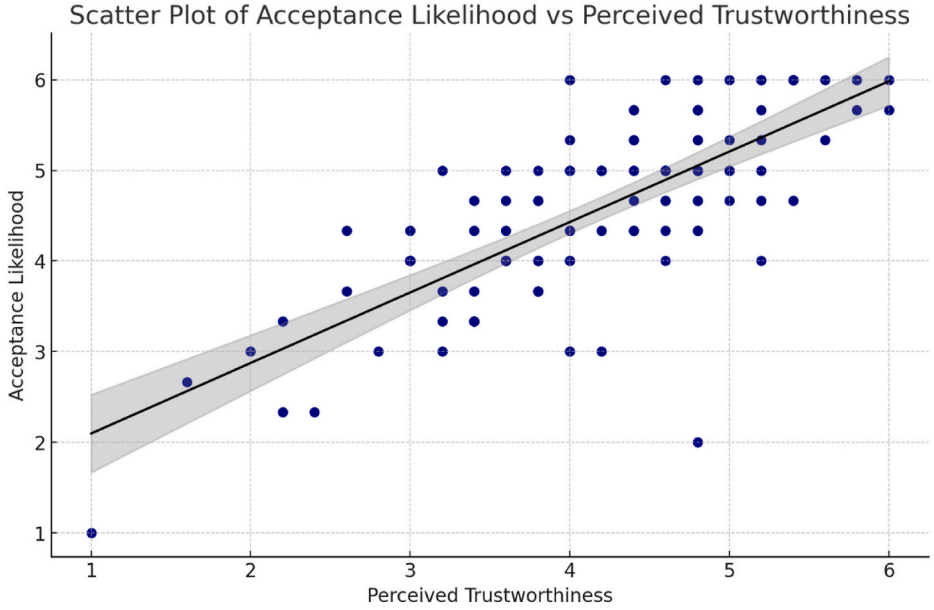


Fig. 3. A scatter plot graph demonstrating the relationship between Acceptance Likelihood and Perceived Trustworthiness. The grey region in the figure indicates 95% confidence intervals.

Table 5
Results of a multiple linear regression analysis examining the effect of diagnosis congruence on likelihood of acceptance of E-triage chatbot recommendations with expertise as a moderator, without and with controlling for covariates.

Variable	*B*	*SE*	*t*	*P*	95% CI	
					LL	*UL*
Model 1: Interaction Between Congruence and Expertise						
Constant	4.946	0.564	8.768	<0.001	3.828	6.065
Congruence	−1.143	0.920	−1.243	0.217	−2.967	0.680
Expertise	−0.242	0.158	−1.535	0.128	−0.555	0.071
Congruence × Expertise Interaction	0.555	0.249	2.226	0.028	0.061	1.050
Model 2: Interaction Between Congruence and Expertise with Covariates						
Constant	4.295	1.047	4.101	<0.001	2.217	6.373
Congruence	−1.149	0.929	−1.237	0.219	−2.991	0.694
Gender Female	0.385	0.195	1.972	0.051	−0.002	0.772
Gender Other	−0.173	0.539	−0.320	0.749	−1.242	0.896
Recruitment Channel	0.107	0.203	0.527	0.599	−0.295	0.509
Profession	0.146	0.230	0.635	0.527	−0.310	0.602
GAAIS	0.243	0.152	1.601	0.113	−0.058	0.544
Human vs. AI Preference	−0.209	0.118	−1.778	0.078	−0.443	0.024
Expertise	−0.211	0.164	−1.288	0.201	−0.536	0.114
Congruence × Expertise Interaction	0.545	0.252	2.159	0.033	0.044	1.046

Note: Model 1: $R^2 = 0.230$, $F(3, 107) = 10.668$, $p < 0.001$.; Model 2: $R^2 = 0.296$, $F(9, 100) = 4.673$, $p < 0.001$. *B* refers to raw regression coefficients. For Congruence, 0 = Incongruent and 1 = Congruent; Expertise ranges from (1) Not knowledgeable at all to (5) extremely knowledgeable; for Gender, only 3 participants in total responded with "other" and "prefer not to say" (Table 1), so for this regression analysis they were collapsed into one category—Gender Other; for both Gender Female and Gender Other, Gender Male is the comparison category; for Recruitment Channel, 0 = Volunteer and 1 = Prolific.com; for Profession, categories that comprise practicing and trainee clinical psychologists, licensed mental health counsellors, and others (i.e., mental health nurses; Table 1) are coded as 1 = Practitioners, and categories that comprise postgraduate and undergraduate medical/psychology students (Table 1) are coded as 2 = Students. Finally, the scores for GAAIS and Human vs. AI Preference can range from (1) Strongly Disagree to (5) Strongly Agree. None of the variables in the analysis were standardized.

Table 6
Results of a multiple linear regression analysis examining the effect of diagnosis congruence on perceived trustworthiness of E-triage chatbot recommendations with expertise as a moderator, without and with controlling for covariates.

Variable	*B*	*SE*	*t*	*P*	95% CI	
					LL	*UL*
Model 1: Interaction Between Congruence and Expertise						
Constant	4.530	0.546	8.301	<0.001	3.448	5.612
Congruence	−1.276	0.890	−1.434	0.155	−3.040	0.488
Expertise	−0.232	0.153	−1.521	0.131	−0.535	0.070
Congruence × Expertise Interaction	0.583	0.241	2.419	0.017	0.105	1.062
Model 2: Interaction Between Congruence and Expertise with Covariates						
Constant	2.594	0.963	2.693	0.008	0.683	4.505
Congruence	−1.083	0.854	−1.268	0.208	−2.777	0.612
Gender Female	0.186	0.180	1.034	0.304	−0.171	0.542
Gender Other	−0.216	0.495	−0.437	0.663	−1.199	0.767
Recruitment Channel	0.004	0.186	0.020	0.984	−0.366	0.374
Profession	0.222	0.211	1.048	0.297	−0.198	0.641
GAAIS	0.510	0.140	3.658	<0.001	0.233	0.787
Human vs. AI Preference	−0.119	0.108	−1.096	0.276	−0.333	0.096
Expertise	−0.179	0.151	−1.188	0.238	−0.478	0.120
Congruence × Expertise Interaction	0.513	0.232	2.211	0.029	0.053	0.974

Note: Model 1: $R^2 = 0.236$, $F(3, 106) = 10.935$, $p < 0.001$; Model 2: $R^2 = 0.363$, $F(9, 100) = 6.326$, $p < 0.001$. *B* refers to raw regression coefficients. For Congruence, 0 = Incongruent and 1 = Congruent; Expertise ranges from (1) Not knowledgeable at all to (5) extremely knowledgeable; for Gender, only 3 participants in total responded with "other" and "prefer not to say" (Table 1), so for this regression analysis they were collapsed into one category—Gender Other; for both Gender Female and Gender Other, Gender Male is the comparison category; for Recruitment Channel, 0 = Volunteer and 1 = Prolific.com; for Profession, categories that comprise practicing and trainee clinical psychologists, licensed mental health counsellors, and others (i.e., mental health nurses; Table 1) are coded as 1 = Practitioners, and categories that comprise postgraduate and undergraduate medical/psychology students (Table 1) are coded as 2 = Students. Finally, the scores for GAAIS and Human vs. AI Preference can range from (1) Strongly Disagree to (5) Strongly Agree. None of the variables in the analysis were standardized.

was smaller and non-significant ($M_{diff} = 0.413$, $p = 0.071$). Therefore, the analyses remained robust despite covariates.

5. Discussion

The findings indicate that psychologists were more likely to accept and incorporate the AI recommendation in their decision-making if it confirmed their preliminary diagnosis (*H1*). The same effect was observed for perceived trustworthiness, where psychologists perceived the AI chatbot as significantly more trustworthy when its recommendations confirmed their prior beliefs about the diagnosis (*H2*). Additionally, an increase in perceived trustworthiness was related to a higher likelihood of accepting and following the AI recommendation (*H3*). Finally, the relationship between recommendation congruence and acceptance likelihood, as well as congruence and perceived trustworthiness, was moderated by self-reported expertise. The higher the participants reported their expertise level in assessment and triage, the more they were likely to accept and trust the chatbot recommendations that did versus did not match their preliminary diagnoses (*H4*). These findings spawn four core conclusions that warrant discussion in relation to previous research.

The first conclusion is that the risk of automation bias appears to be less prominent in the psychologists' diagnostic decision-making; instead, they tend to be prone to confirmation bias. Both students and practitioners were significantly more likely to accept and incorporate AI recommendations into their decision-making when they aligned with their preliminary diagnoses. This finding aligns with the literature on confirmation bias in MH, where practitioners tend to give more weight to data supporting their preliminary diagnosis (Elston, 2020; Mendel et al., 2011). The finding is also in line with qualitative findings on SLBs' interactions with AI tools, indicating that decision-makers weigh the information provided by AI with their intuitive professional knowledge (e.g., Meijer et al., 2021; Selten et al., 2022). Research on the interactions between SLBs and AI has similarly demonstrated that confirmation bias plays an important role in the likelihood of acceptance of AI recommendations (Selten et al., 2022; Snow, 2021). Importantly, the integration of conversational AI technology, and NLP more generally, is very new to the MH field (Car et al., 2020; Graham et al., 2019), and automation bias primarily occurs in fields where automation is highly established (Peeters, 2020). Future research should therefore investigate whether the relative importance of the confirmation and automation biases in the context of MH AI recommendations changes when AI tools are used repeatedly for a prolonged period.

The second conclusion that can be drawn is that confirmation bias affects perceived trustworthiness towards chatbot technologies. Participants who saw AI recommendations that confirmed their preliminary diagnoses later rated the AI tool as more trustworthy than those who saw the congruent recommendations. This phenomenon also largely aligns with the literature regarding how experts perceive trustworthiness of AI tools (Nazaretsky et al., 2021; Selten et al., 2022; Yu et al., 2019). In addition, research on police officers, teachers and general corporate employees also found that individuals would be more likely to perceive

an AI tool as more trustworthy if it aligns with their beliefs and preconceived judgements (Nazaretsky et al., 2021; Meijer et al., 2021; Selten et al., 2022).

Third, our findings indicate that an increase in perceived trustworthiness is strongly related to an increase in the likelihood of accepting and incorporating the recommendation as part of the practitioners' diagnoses. This finding aligns with previous research showing that individuals' behavior towards AI tools is largely shaped by trust in these tools (Nazaretsky et al., 2021). This is also in line with previous research on SLB's where individuals who perceived AI recommendations as more trustworthy were more likely to follow them (Selten et al., 2022).

Finally, our research indicates that confirmation bias is significantly exacerbated by self-reported expertise. Indeed, psychologists who reported having higher triage expertise were more likely to both accept the congruent (vs. incongruent) recommendations and trust them. Whereas several research studies have examined expert interactions with AI in specific fields (Meijer et al., 2021; Selten et al., 2022; Snow et al., 2021), to our knowledge the present research is the first to directly investigate how expertise shapes the propensity to confirmation bias in the context of AI recommendations. Our findings partly align with other research on self-reported expertise and decision accuracy that did not directly focus on confirmation bias, but which showed that individuals with higher reported expertise are typically more motivated to construct explanations dismissing the contradictory point of view (Atir et al., 2015). It is important, however, to consider alternative explanations for this finding. Namely, the possibility stands that self-reported experts in MH diagnosis have an enhanced ability to draw conclusions from limited information provided by the vignettes, based on having enhanced experience dealing with self-report patient data (Reverberi et al., 2022). While research primarily shows that both expert and non-expert individuals are susceptible to bias when interacting with AI decision-making in healthcare (Adam et al., 2022), further research needs to be done with more objective measures of expertise, mainly years of experience instead of self-report.

5.1. Research implications

Collectively these conclusions indicate that psychologists do not blindly trust or incorporate AI recommendations that go against their professional intuition for diagnosis and triage. As AI gets used for increasingly complex tasks, the number and severity of the errors AI "advisors" make may increase (Frey & Osborne, 2017). Therefore, as mentioned previously, it is vital to ensure appropriate reliance where the humans feel empowered to critically and objectively evaluate the advice that they receive (Hemmer et al., 2021; Lai et al., 2021; Schemmer et al., 2023). The current findings demonstrate that practitioners are at least partly capable of mitigating any unfair or biased outcomes in AI decision-making and are not prone to automation bias (Veale & Binns, 2017). This insight supports the importance of keeping practitioners 'in the loop' and maintaining the right to overturn AI decisions (Selten et al., 2022). However, the choice to maintain partial reliance on human intuition may introduce new challenges. As shown by our research, practitioners may be unlikely to correct biased AI systems if they align with their initial hypotheses, and solely relying on individual expert decision-making to judge the quality of AI recommendations may be unreliable.

Considering that reliance on AI healthcare technologies is based on open communication and enforceable systems of responsibility (Kerasidou et al., 2022), explainable AI has emerged as a promising approach to address the challenge of reliance behaviors (Amann et al., 2022). Schemmer et al. (2023) demonstrated that explaining the AI decision-making process was effective in increasing appropriate reliance when an individual had to change their initially incorrect decision to a correct one following AI advice (Schemmer et al., 2023). Therefore, the impact AI decisions can have on patients' lives makes explainable AI crucial in the field of healthcare. However, the heterogenous and unstructured data format and the dynamic nature of clinical knowledge significantly complicate explainability of AI in healthcare (Alam et al., 2023; Anton et al., 2022). Since explaining AI decision making allows individuals to acknowledge how the information provided by the technology compliments their own understanding of the diagnosis, it is crucial that explainable AI systems are integrated into clinical workflows (e.g., clinical support systems) and are built collaboratively with clinicians, data scientists, and ethicists alike (Alam et al., 2023; Amann et al., 2022).

In addition to creating more explainable AI systems, the provision of coordinated training and education about the use of AI and the mitigation of professionals' biases would need to be provided to MH practitioners (Viswanathan et al., 2022). Hospitals and mental health clinics that are implementing AI systems as diagnostic assistants need to consider cognitive bias mitigation systems such as computer-based systems, simulation, workshops, seminars, and comprehensive curricula (Doherty & Carroll, 2020). Computer based diagnostic reasoning interfaces used to detect and measure bias as well as simulation exercises, particularly for more expert practitioners have shown notable promise in raising awareness of cognitive strategies to mitigate cognitive error traps (Bond et al., 2004; Crowley et al., 2013; Doherty & Carroll, 2020).

5.2. Limitations and future research

This study had several limitations that need to be addressed. Firstly, the sampling strategies used for both Prolific and volunteer participants did not use random selection, as they largely relied on the existing professional networks of the researcher and the gatekeeper. For the volunteers, both practitioners and students were sampled from specific university networks, which may limit the representativeness and generalizability of the study. Because most participants who received the study link were based in either UK or US, they were also more likely to be familiar with the IAPT diagnostic tools and more likely to trust AI chatbots that used them. Because individuals across different geographic locations and nationalities have been found to have varying levels of trust and likelihood of acceptance of AI tools (Dang & Liu, 2022; Meijer et al., 2021), our findings need to be replicated in countries that employ distinct diagnosis tools and may have alternative approaches to patient triage to improve generalizability.

Secondly, as stimuli in the present research, we used 3 vignettes which provided information necessary for a MH diagnosis (e.g., chief patient complaints, past medical history; for an example, see Fig. 2). Although these vignettes were validated by clinical psychologists and cover a broad range of MH diagnoses used by the IAPT program (England, 2018)(National Collaborating Centre for Mental Health, 2018a), which positively contributes to their ecological validity, the limited number of vignettes raises potential generalizability issues. In this context, it is important to emphasize that our research is a first step toward understanding the interaction between human-AI recommendations in the field of mental health diagnostics and is therefore a building block for future research to investigate more ample domains and types of stimuli that could lead to similar patterns in expert decision-making in the presence of AI. Generalizability is often a joint effort of various researchers who investigate a phenomenon of interest in different domains and circumstances, and we hope that our findings will inspire other researchers to jointly create a more comprehensive picture of the phenomenon the present article tackles.

Thirdly, it was unclear whether participants were on average more likely to incorporate AI recommendations when they were accurate or inaccurate. This is because our research focused specifically on mental health diagnostics, a field that is historically ambiguous, in part due to comorbidities with other mental health conditions, issues with accuracy due to overlapping symptoms, and reliance on patient's subjective recollection of behaviors (Goldman et al., 1999). Therefore, in this domain, investigating how accuracy or inaccuracy of AI advice shapes

expert decisions would be challenging and potentially methodologically flawed, since for many diagnoses the ground truth is difficult to establish, and giving the type of participants we examined (i.e., individuals with advanced psychology knowledge) an obviously wrong diagnosis recommendation would make it unlikely that anyone would accept it. For that reason, all recommendations provided in our vignettes were to some degree plausible, and examining how diagnosis accuracy shapes AI recommendation acceptance in addition to people's initial judgment may be more appropriate for fields where clear ground truth is easily established.

Finally, self-reported expertise was subjective in this study, where some undergraduate students rated themselves as highly knowledgeable, and some practitioners had a moderate rating. Some evidence indicates that actual expertise, measured in years of experience, may mitigate the effects of confirmation bias (Krems & Zierer, 1994). Future research should therefore examine the role actual expertise plays in practitioner-AI interactions. Nevertheless, our findings indicate that how psychologists perceive themselves in terms of expertise is highly important for their susceptibility to confirmation bias, and an intervention aimed at helping them to realistically appraise their expertise may be important for tackling this bias when it comes to AI recommendations.

6. Conclusion

The increasing integration of AI assistive technologies in MH services offers the potential to mitigate current challenges in the field, such as long-waiting times and over-stretched MH practitioners. This research illuminates the intricate relationship between practitioners' intuition and AI recommendations, shedding light on the psychological dynamics of clinical decision-making in the presence of AI. Our findings underscore that practitioners are discerning in adopting and trusting AI recommendations, especially when such suggestions deviate from their initial diagnoses. However, the highly significant presence of confirmation bias in their decision-making implies that a practitioner-AI collaboration in MH is unlikely to eliminate diagnostic bias. To harness the full potential of AI in MH services, institutions must prioritize organization-wide policies, education, and training focused on better understanding of AI technologies and bias mitigation. It is pivotal to strike a balance between the advantages of AI-driven recommendations and the invaluable human touch in MH practice, ensuring optimal patient outcomes and trust in emerging technologies.

CRediT authorship contribution statement

Anna Bashkirova: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Dario Krpan:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Formal analysis.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT and QuillBot in order to check the appropriateness of certain idiomatic expressions and sentence structures. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The researchers would like to thank Juliet Foster, Liam Delaney and Joanna Jiang for their active advisory role throughout this research and valuable feedback throughout the process.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chbah.2024.100066.

References

Adam, H., Balagopalan, A., Alsentzer, E., Christia, F., & Ghassemi, M. (2022). Mitigating the impact of biased artificial intelligence in emergency decision-making. *Communication and Medicine, 2*(1), 149.

Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics* (pp. 559–560).

Aktan, M. E., Turhan, Z., & Dolu, I. (2022). Attitudes and perspectives towards the preferences for artificial intelligence in psychotherapy. *Computers in Human Behavior, 133*, Article 107273.

Alam, M. N., Kaur, M., & Kabir, M. S. (2023). Explainable AI in Healthcare: Enhancing transparency and trust upon legal and ethical consideration. *Int Res J Eng Technol, 10*(6), 1–9.

Alon-Barkat, S., & Busuioc, M. (2023). Human–AI interactions in public sector decision making: "automation bias" and "selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory, 33*(1), 153–169.

Amann, J., Vetter, D., Blomberg, S. N., Christensen, H. C., Coffee, M., Gerke, S., & Z-Inspection Initiative. (2022). To explain or not to explain?—artificial intelligence explainability in clinical decision support systems. *PLOS Digital Health, 1*(2), Article e0000016.

Anton, N., Doroftei, B., Curteanu, S., Catălin, L., Ilie, O. D., Târcoveanu, F., & Bogdănici, C. M. (2022). Comprehensive review on the use of artificial intelligence in ophthalmology and future research directions. *Diagnostics, 13*(1), 100.

Asch, S., Connor, S. E., Hamilton, E. G., & Fox, S. A. (2000). Problems in recruiting community-based physicians for health services research. *Journal of General Internal Medicine, 15*(8), 591–599.

Atir, S., Rosenzweig, E., & Dunning, D. (2015). When knowledge knows no bounds: Self-perceived expertise predicts claims of impossible knowledge. *Psychological Science, 26*(8), 1295–1303.

Berner, E. S., Maisiak, R. S., Cobbs, C. G., & Taunton, O. D. (1999). Effects of a decision support system on physicians' diagnostic performance. *Journal of the American Medical Informatics Association, 6*(5), 420–427.

Bond, W. F., Deitrick, L. M., Arnold, D. C., Kostenbader, M., Barr, G. C., Kimmel, S. R., & Worrilow, C. C. (2004). Using simulation to instruct emergency medicine residents in cognitive forcing strategies. *Academic Medicine, 79*(5), 438–446.

Boomsma-van Holten, M., Weerman, A., Karbouniaris, S., & Van Os, J. (2023). The use of experiential knowledge in the role of a psychiatrist. *Frontiers in Psychiatry, 14*, Article 1163804.

Bowes, S. M., Ammirati, R. J., Costello, T. H., Basterfield, C., & Lilienfeld, S. O. (2020). Cognitive biases, heuristics, and logical fallacies in clinical practice: A brief field guide for practicing clinicians and supervisors. *Professional Psychology: Research and Practice, 51*(5), 435.

Brown, C., Story, G. W., Mourão-Miranda, J., & Baker, J. T. (2021). Will artificial intelligence eventually replace psychiatrists? *The British Journal of Psychiatry, 218*(3), 131–134.

Browne, M. O., Lee, A., & Prabhu, R. (2007). Self-reported confidence and skills of general practitioners in management of mental health disorders. *Australian Journal of Rural Health, 15*(5), 321–326.

Bullock, J., Young, M. M., & Wang, Y. F. (2020). Artificial intelligence, bureaucratic form, and discretion in public service. *Information Polity, 25*(4), 491–506.

Busuioc, M. (2021). Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review, 81*(5), 825–836.

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3*(3), 223–230.

Car, L., Dhinagaran, D. A., Kyaw, B. M., Kowatsch, T., Joty, S., Theng, Y. L., & Atun, R. (2020). Conversational agents in health care: Scoping review and conceptual analysis. *Journal of Medical Internet Research, 22*(8), Article e17158.

Crowley, R. S., Legowski, E., Medvedeva, O., Reitmeyer, K., Tseytlin, E., Castine, M., & Mello-Thoms, C. (2013). Automated detection of heuristics and biases among pathologists in a computer-based system. *Advances in Health Sciences Education, 18*, 343–363.

D'alfonso, S., Santesteban-Echarri, O., Rice, S., Wadley, G., Lederman, R., Miles, C., & Alvarez-Jimenez, M. (2017). Artificial intelligence-assisted online social therapy for youth mental health. *Frontiers in Psychology, 8*, 796.

Dang, J., & Liu, L. (2022). Implicit theories of the human mind predict competitive and cooperative responses to AI robots. *Computers in Human Behavior, 134*, Article 107300.

Darau, G. (2022, December 23). *Trust of doctors in adopting AI-powered tools into their daily practice: Complicated Relationship? AEGIS IT Research.* https://aegisresearch.eu/tr

ust-of-doctors-in-adopting-ai-powered-tools-into-their-daily-practice-complicated-relationship/#:~:text=More%20than%20half%20of%20the,support%20for%20their%20daily%20decisions.
Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal, 6*(2), 94–98.
Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668–1674.
Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics, 42*(5), 760–772.
DeWall, C. N., Myers, D. G., May, C., & Einstein, G. (2015). Teaching current directions in psychological science. *APS Observer, 28*.
Doherty, T. S., & Carroll, A. E. (2020). Believing in overcoming cognitive biases. *AMA journal of ethics, 22*(9), 773–778.
Elston, D. M. (2020). Confirmation bias in medical decision-making. *Journal of the American Academy of Dermatology, 82*(3), 572.
England, NHS. (2018). *NHS Talking Therapies for anxiety and depression Manual* (pp. 12–14).
Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.
Featherston, R., Downie, L. E., Vogel, A. P., & Galvin, K. L. (2020). Decision making biases in the allied health professions: A systematic scoping review. *PLoS One, 15*(10), Article e0240716.
Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change, 114*, 254–280.
Gabbard, G. O., & Crisp-Han, H. (2017). The early career psychiatrist and the psychotherapeutic identity. *Academic Psychiatry, 41*, 30–34.
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.
Goldberg, S. B., Rousmaniere, T., Miller, S. D., Whipple, J., Nielsen, S. L., Hoyt, W. T., & Wampold, B. E. (2016). Do psychotherapists improve with time and experience? A longitudinal analysis of outcomes in a clinical setting. *Journal of Counseling Psychology, 63*, 1–11.
Goldman, L. S., Nielsen, N. H., Champion, H. C., & Council on Scientific Affairs, American Medical Association. (1999). Awareness, diagnosis, and treatment of depression. *Journal of General Internal Medicine, 14*(9), 569–580.
Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H. C., & Jeste, D. V. (2019). Artificial intelligence for mental health and mental illnesses: An overview. *Current Psychiatry Reports, 21*, 1–18.
Grgić-Hlača, N., Castelluccia, C., & Gummadi, K. P. (2022). Taking advice from (dis) similar machines: The impact of human-machine similarity on machine-assisted decision-making. October *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 10*, 74–88.
Grimmelikhuijsen, S., & Knies, E. (2017). Validating a scale for citizen trust in government organisations. *International Review of Administrative Sciences, 83*(3), 583–601.
Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19–30.
Hemmer, P., Schellhammer, S., Vössing, M., Jakubik, J., & Satzger, G. (2022). *Forming effective human-AI teams: Building machine learning models that complement the capabilities of multiple experts. arXiv preprint arXiv:2206.07948*.
Hemmer, P., Schemmer, M., Vössing, M., & Kühl, N. (2021). *Human-AI complementarity in hybrid intelligence systems: A structured literature review* (Vol. 78). PACIS.
Igoe, K. J. (2021, March 12)). *Algorithmic bias in health care exacerbates social inequities—how to prevent it. Harvard T.H. Chan School of Public Health.* https://www.hsph.harvard.edu/ecpe/how-to-prevent-algorithmic-bias-in-health-care/.
Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth, 6*(11), Article e12106.
Jonas, E., Schulz-Hardt, S., Frey, D., & Thelen, N. (2001). Confirmation bias in sequential information search after preliminary decisions: An expansion of dissonance theoretical research on selective exposure to information. *Journal of Personality and Social Psychology, 80*(4), 557.
Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*(6), 515.
Kerasidou, C. X., Kerasidou, A., Buscher, M., & Wilkinson, S. (2022). Before and beyond trust: Reliance in medical AI. *Journal of Medical Ethics, 48*(11), 852–856.
Kirkman, E., & Melrose, K. (2014). *Clinical judgement and decision-making in children's social work: An analysis of the 'front door' system. Research report.* TSO.
Koutsouleris, N., Hauser, T. U., Skvortsova, V., & De Choudhury, M. (2022). From promise to practice: Towards the realisation of AI-informed mental health care. *The Lancet Digital Health, 4*(11), e829–e840.
Krems, J. F., & Zierer, C. (1994). Sind experten gegen kognitive täuschungen gefeit? Zur abhängigkeit des confirmation bias von Fachwissen. *Zeitschrift für experimentelle und angewandte Psychologie, 41*(1), 98–115.
Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., & Tan, C. (2021). *Towards a science of human-ai decision making: A survey of empirical studies. arXiv preprint arXiv: 2112.11471*.
Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*(10), 1243–1270.
Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior, 139*, Article 107539.
Li, P., yan Cheng, Z., & lin Liu, G. (2020). Availability bias causes misdiagnoses by physicians: Direct evidence from a randomised controlled trial. *Internal Medicine, 59*(24), 3141–3146.
Luxton, D. D. (2014). Artificial intelligence in psychological practice: Current and future applications and implications. *Professional Psychology: Research and Practice, 45*(5), 332.
Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: A systematic review. *Journal of the American Medical Informatics Association, 24*(2), 423–431.
McManus, S., Bebbington, P. E., Jenkins, R., & Brugha, T. (2016). *Mental health and wellbeing in england: The adult psychiatric morbidity survey 2014*. NHS digital.
McManus, S., Bebbington, P. E., Jenkins, R., & Brugha, T. (2016). *Mental health and wellbeing in England: the adult psychiatric morbidity survey 2014. NHS digital*.
Meijer, A., Lorenz, L., & Wessels, M. (2021). Algorithmisation of bureaucratic organisations: Using a practice lens to study how context shapes predictive policing systems. *Public Administration Review, 81*(5), 837–846.
Mendel, R., Traut-Mattausch, E., Jonas, E., Leucht, S., Kane, J. M., Maino, K., & Hamann, J. (2011). Confirmation bias: Why psychiatrists stick to wrong preliminary diagnoses. *Psychological Medicine, 41*(12), 2651–2659.
Miller, D. D., & Brown, E. W. (2018). Artificial intelligence in medical practice: The question to the answer? *The American Journal of Medicine, 131*(2), 129–133.
Minerva, F., & Giubilini, A. (2023). *Is AI the future of mental healthcare? Topoi*.
Mizrahi, M. (2018). Arguments from expert opinion and persistent bias. *Argumentation, 32*(2), 175–195.
Mosier, K. L., & Skitka, L. J. (2018). Human decision makers and automated decision aids: Made for each other?. In *Automation and human performance* (pp. 201–220). CRC Press.
Moss, W. (2016). *Manual for the health care of children in humanitarian emergencies; triage and emergency assessment*. World Health Organization. https://www.ncbi.nlm.nih.gov/books/NBK143755/.
Moynihan, D. P., & Lavertu, S. (2012). Cognitive biases in governing: Technology preferences in election administration. *Public Administration Review, 72*(1), 68–77.
Narayanan, S., Yu, G., Ho, C. J., & Yin, M. (2023). How does value similarity affect human reliance in AI-assisted ethical decision making?. August. In *Proceedings of the 2023 AAAI/ACM conference on AI* (pp. 49–57). Ethics, and Society.
National Collaborating Centre for Mental Health. (2018). *The improving access to psychological therapies manual*. UK: NCCMH. https://www.england.nhs.uk/publication/the-improving-access-to-psychological-therapies-manual/.
Nazaretsky, T., Cukurova, M., Ariely, M., & Alexandron, G. (2021). Confirmation bias and trust: Human factors that influence teachers' attitudes towards AI-based educational technology. September. In *, Vol. 3042. CEUR workshop proceedings*.
Newman-Toker, D. E., & Pronovost, P. J. (2009). Diagnostic errors—the next frontier for patient safety. *JAMA, 301*(10), 1060–1062.
NHS 75 Digital. (2020). Psychological Therapies. In *Annual report on the use of IAPT services 2019-20*. NHS Digital. digital.nhs.uk/data-and information/publications/statistical/psychological-therapies-annual-reports-on-the-use-of-iapt-services/annual-report-2019-20#.
Parikh, R. B., Teeple, S., & Navathe, A. S. (2019). Addressing bias in artificial intelligence in health care. *JAMA, 322*(24), 2377–2378.
Peeters, R. (2020). The agency of algorithms: Understanding human-algorithm interaction in administrative decision-making. *Information Polity, 25*(4), 507–522.
Prakash, S., Bihari, S., Need, P., Sprick, C., & Schuwirth, L. (2017). Immersive high fidelity simulation of critically ill patients to study cognitive errors: A pilot study. *BMC Medical Education, 17*(1), 1–12.
Ramnarayan, P., Cronje, N., Brown, R., Negus, R., Coode, B., Moss, P., Hassan, T., Hamer, W., & Britto, J. (2007). Validation of a diagnostic reminder system in emergency medicine: A multi-centre study. *Emergency Medicine Journal, 24*, 619–624.
Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P., & Cherubini, A. (2022). Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific Reports, 12*(1), Article 14952.
Rollwage, M., Habicht, J., Juchems, K., Carrington, B., Stylianou, M., Hauser, T., & Harper, R. (2023). Using conversational AI to facilitate mental health assessment and improve clinical efficiencies in psychotherapy services in large real-world dataset. https://doi.org/10.2196/preprints.44358.
Rollwage, M., Juchems, K., Habicht, J., Carrington, B., Hauser, T., & Harper, R. (2022). Conversational AI facilitates mental health assessments and is associated with improved recovery rates. *medRxiv, 11*. https://doi.org/10.1101/2022.11.03.22281887
Saposnik, G., Redelmeier, D., Ruff, C. C., & Tobler, P. N. (2016). Cognitive biases associated with medical decisions: A systematic review. *BMC Medical Informatics and Decision Making, 16*, 138.
Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. March. In *Proceedings of the 28th international conference on intelligent user interfaces* (pp. 410–422).
Schepman, A., & Rodway, P. (2020). Initial validation of the general attitudes towards artificial intelligence scale. *Computers in Human Behavior Reports, 1*, Article 100014.
Selten, F., Robeer, M., & Grimmelikhuijsen, S. (2023). 'Just like I thought': Street-level bureaucrats trust AI recommendations if they confirm their professional judgment. *Public Administration Review, 83*(2), 263–278.
Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies, 51*(5), 991–1006.
Snow, T. (2021). From satisficing to artificing: The evolution of administrative decision-making in the age of the algorithm. *Data & Policy, 3*, e3.
Spengler, P. M., Miller, D. J., & Spengler, E. S. (2016). Psychological masquerade embedded in a cluster of related clinical errors: Real prac- tice, real solutions, and

their scientific underpinnings. *Psychotherapy, 53*, 336–341. https://doi.org/10.1037/pst0000076

Sunstein, C. R. (2023). The use of algorithms in society. *The Review of Austrian Economics*, 1–22.

Timmons, A. C., Duong, J. B., Simo Fiallo, N., Lee, T., Vo, H. P. Q., Ahle, M. W., & Chaspari, T. (2022). A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. *Perspectives on Psychological Science*, Article 17456916221134490.

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society, 4*(2), Article 2053951717743530.

Veale, M., & Brass, I. (2019). Administration by Algorithm? Public Management Meets Public Sector Machine Learning. In K. Yeung, & M. Lodge (Eds.), *Algorithmic Regulation* (pp. 121–149). Oxford University Press.

Vicente, L., & Matute, H. (2023). Humans inherit artificial intelligence biases. *Scientific Reports, 13*(1), Article 15737. https://doi.org/10.1038/s41598-023-42384-8

Viswanathan, S., Rollwage, M., & Harper, R. (2022). Promises and challenges of AI-enabled mental healthcare: A foundational study. October. In *Empowering Communities: A participatory approach to AI for mental health*.

Wang, X., & Yin, M. (2021). Are explanations helpful? A comparative study of the effects of explanations in ai-assisted decision-making. April. In *26th international conference on intelligent user interfaces* (pp. 318–328).

Whelehan, D. F., Conlon, K. C., & Ridgway, P. F. (2020). Medicine and heuristics: Cognitive biases and medical decision-making. *Irish Journal of Medical Science, 189*, 1477–1484.

Wilke, A., & Mata, R. (2012). Cognitive bias. In *Encyclopedia of human behavior* (pp. 531–535). Academic Press.

Wilson, R. L., Higgins, O., Atem, J., Donaldson, A. E., Gildberg, F. A., Hooper, M., & Welsh, B. (2023). Artificial intelligence: An eye cast towards the mental health nursing horizon. *International Journal of Mental Health Nursing, 32*(3), 938–944.

Witteman, C. L., Spaanjaars, N. L., & Aarts, A. A. (2012). Clinical intuition in mental health care: A discussion and focus groups. *Counselling Psychology Quarterly, 25*(1), 19–29.

World Health Organization. (2022). *World mental health report: Transforming mental health for all*.

Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020). How do visual explanations foster end users' appropriate trust in machine learning?. March. In *Proceedings of the 25th international conference on intelligent user interfaces* (pp. 189–201).

Yu, K., Berkovsky, S., Taib, R., Zhou, J., & Chen, F. (2019). Do I trust my machine teammate? An investigation from perception to decision. March. In *Proceedings of the 24th international conference on intelligent user interfaces* (pp. 460–468).