scientific reports

OPEN



A paper-and-pencil questionnaire outperforms GPT for measuring cognitive vulnerability to depression and predicting depressive symptoms

Jane K. Stallman & Gerald J. Haeffel[⊠]

We tested if GPT could predict changes in depressive symptoms using participants' (*n* = 930) causal explanations for negative life events. Results showed that 2 of 30 GPT prompts yielded output that could reliably predict changes in future depressive symptoms; but this output was not a better predictor than the traditional paper-and-pencil measure of cognitive risk for depression (Cognitive Style Questionnaire). These findings highlight potential limitations of large language models like GPT. Human thought is complex, and language may not accurately represent people's internal cognitive processes. In this case, participants' written explanations for negative life events did not contain meaningful information that could be used for differentiation (or was indicative of some latent construct). We found that people could generate equally negative causal explanations for negative events yet hold different beliefs about the changeability of those causes. Our results support the hypothesis that it is the perceived changeability, not the overall negativity, of causal beliefs that determines risk for depressive symptoms. GPT cannot yet discern this changeability as well as a paper-and-pencil questionnaire.

Keywords GPT, Depression, Cognitive vulnerability, AI, Natural language processing, Hopelessness theory

Artificial intelligence (AI) has changed how students learn, people work, and scientists do research^{1–3}. AI can process large amounts of data fast and effectively making it useful for detecting patterns and improving measurement. The purpose of this study was to determine if AI could improve the prediction of depressive symptoms compared to the standard paper-and-pencil questionnaire typically used.

Depression is one of the most common forms of mental illness with nearly 300 million people struggling worldwide^{4,5}. It is the leading cause of disability for people ages 15–44 and is among the strongest predictors of suicide⁶. Therefore, it is critical to develop reliable and valid methods to identify those at risk for depression to try to prevent its occurrence.

There are several theories for why some people are at greater risk for depression than others, including cognitive, biochemical, neuroanatomical, and behavioral explanations. In this study, we focus on cognition given the strong empirical support for its role in both the etiology and treatment of depression^{7–9}. According to the cognitive theories of depression^{10,11}, some people are at a heightened risk for developing depression because they tend to generate overly negative interpretations of life stress (i.e., they have a "cognitive vulnerability"). Specifically, people who attribute negative life events to factors that are stable (consistent over time) and global (affecting many areas of their life) and infer negative consequences and self-worth implications are more likely to develop depression than people who do not generate these types of inferences.

There is strong empirical support for this "cognitive vulnerability" hypothesis^{8,12}. Longitudinal studies show that those with high levels of cognitive vulnerability are at greater risk for developing depressive symptoms and depressive disorders than those who exhibit low levels of cognitive vulnerability (e.g.,^{13–15}). This research establishes temporal precedence and suggests that cognitive vulnerability may be a causal contributor to at least some types of depression (assuming the existence of multiple pathways and multiple depressive subtypes).

Cognitive vulnerability, as conceptualized in the hopelessness theory of depression, is measured with the cognitive style questionnaire (CSQ)⁸. The CSQ is a self-report questionnaire in which participants write down

Department of Psychology, University of Notre Dame, Notre Dame, IN 46656, USA. [™]email: ghaeffel@nd.edu

their causal explanations for hypothetical negative life events. Participants then complete Likert-scale ratings about the causes and implications of hypothetical negative life events on dimensions of stability/globality, consequences, and self-worth. Our primary question is—can OpenAI's large language model (LLM) use the written responses from the CSQ to improve the measurement of cognitive vulnerability and, in turn, better predict risk for future depressive symptoms compared to the traditional CSQ scoring method?

LLMs use natural language processing, machine learning, and reinforcement learning with human feedback to process input text and respond probabilistically with output text that replicates human semantics and syntax. Among the most widely used LLM is OpenAI's GPT. The first version, GPT-1, was released in 2018, and new versions that increase the intelligence of the software are continuously released. For example, GPT-3.5 failed to outperform human experts on questions rated to be difficult, but GPT-4 showed consistently higher accuracy than humans on similar questions¹⁶.

There is emerging evidence that LLMs can advance measurement of psychological constructs and improve prediction relative to existing measures. For example, a machine learning model using OpenAI's text embedding model achieved 85% sensitivity and 75% specificity in predicting future childbirth-related post-traumatic stress disorder cases based on narrative language¹⁷. Similarly, natural language processing models like GPT were shown to have adequate degrees of accuracy and precision when identifying participants with depression from structured clinical interviews¹⁸. Further, natural language processing of psychotherapy notes demonstrated incremental validity in measuring suicide risk over time¹⁹. These are just a few examples from a growing body of research showing the potential of LLMs to advance psychological research.

In this study, we tested if GPT could improve the measurement of cognitive vulnerability to depression compared to the traditional paper-and-pencil CSQ measure. Specifically, we tested if GPT could use participants' causal explanations for hypothetical events to identify those at greatest risk for future depressive symptoms. Additionally, we examined whether this LLM assessment could contribute unique predictive validity accounting for the CSQ. The implications of the results of this work are twofold. First, the findings will determine the degree to which LLMs such as GPT can use causal explanations to make predictions. Second, the results have potential to improve the identification of individuals who may be at high risk for depression, which informs both theory (e.g., determining the kinds of cognitions put people at most risk), and prevention intervention research.

Method

Participants

Participants were 930 undergraduates (60% female, 40% male) aged 17–22 (M = 18.75, SD = 1.15) aggregated from 6 previous longitudinal studies that used measures of cognitive vulnerability and depressive symptoms^{20–25}. The studies were aggregated to create a large sample size that would provide strong power to test the study hypotheses. Participants were recruited via a medium-sized private midwestern university's psychology department's online extra credit participant pool, except for one study²³ in which participants were recruited from the general first-year class at the university. All participants from each study were included in the aggregated data set (no exclusions). Racial and ethnic demographic data of the samples was collected for 4 of the 6 studies (excluding^{24,25}). The demographic breakdown for the studies was 74% Caucasian, 8% Asian, 3% Black, and 3% other; 12% identified as Hispanic. The use of human subjects in all prior studies was approved by the University of Notre Dame Institutional Review Board (IRB) and all research was conducted in alignment with the approved IRB protocol (17-08-4039) and institutional guidelines and regulations. Informed consent was obtained from all participants in each of the studies before they started the studies. The aggregated data set used in this study was completely anonymous (with no human subject interactions).

Measures

Cognitive vulnerability to depression

The Cognitive Style Questionnaire $(CSQ)^8$ measures the cognitive vulnerability factor featured in the hopelessness theory of depression¹⁰. Participants are instructed to read 12 negative hypothetical events and imagine themselves in each situation (see Fig. 1 for example scenario). Participants then write down what they believe to be the cause of the event; this written narrative is referenced for answering questions about the globality/specificity and stability/instability of the cause of the event using a 7-point Likert scale. Participants then rate the meaning of the hypothetical event regarding future consequences and dimensions of self-worth. The CSQ score is calculated by taking the average of the Likert scale responses for the four components (stability/globality/future consequences/self-worth) for the negative hypothetical events. Composite scores range from 1 to 7, with higher scores representing greater degrees of cognitive vulnerability. The CSQ has demonstrated excellent internal consistency (coefficient alpha typically >.90)⁸, strong test–retest reliability over months and even years (e.g., 1-year test–retest is .80)²⁶, and predictive validity⁸.

Depressive symptoms

The Beck Depression Inventory $(BDI)^{27}$ is a commonly used measure of depressive symptoms. It is a 21-item self-report questionnaire in which participants rate symptoms on a 4-point Likert scale. Scores range from 0–63 with higher scores indicating greater levels of depressive symptomatology. The BDI has shown strong test-retest reliability (>.60) and construct validity²⁸. Internal consistency of the measure in this study was strong, alpha = .87.

Procedure

Participants completed assessments at 2 time points, a baseline assessment and follow-up ranging from 4 weeks to 6 months later depending on the study. At baseline, participants completed both the BDI and CSQ. At the second time point, participants completed the BDI. GPT was used to evaluate the written causal explanations

Imagine that the following situation actually happens to you:

SITUATION: <u>An important romantic relationship you are involved in breaks up because the other person no</u> longer wants a relationship with you.

Questions A-C ask about the cause of the person not wanting a romantic relationship with you.

A) One the line below, write down the <u>one</u> major cause of the person not wanting a romantic relationship with you.

CAUSE:

B) Think about the <u>cause</u> (i.e., what you wrote down on the line above) of the person not wanting a romantic relationship with you. Is this <u>cause</u> something that leads to problems just in your romantic relationship in that instance, or does this <u>cause</u> also lead to problems in <u>other</u> areas of your life? (Circle one number.)

This cause leads to								This cause leads to
problems <u>just</u> in my	1	2	3	4	5	6	7	problems in <u>all</u> areas of
romantic relationship								my life
in that instance								

C) Think about the <u>cause</u> (i.e., what you wrote down on the line above) of the person not wanting a romantic relationship with you. Now assume that in the future, you approach the same person on other occasions to find out how the person feels about having a romantic relationship with you. Will the <u>cause</u> of the person not wanting a romantic relationship with you now as described above again cause that person to not want a romantic relationship with you in the future? (Circle one number.)

Will never again cause								Will always cause that
that person to not want	1	2	3	4	5	6	7	person to not want a
a romantic relationship								romantic relationship
with me								with me

Questions D-E ask for your views about the <u>meaning</u> of the situation of the other person no longer wanting a romantic relationship with you rather than about the cause of this situation.

D) How likely is it that the other person no longer wanting a romantic relationship with you will lead to other negative things happening to you? (Circle one number.)

Not at all likely to								Extremely likely to lead
lead to other negative	1	2	3	4	5	6	7	to other negative things
things happening to								happening to me
me								

E) To what degree does the other person no longer wanting a romantic relationship with you mean to you that you are flawed in some way? (Circle one number.)

Definitely does not								Definitely does mean I
mean I am flawed in	1	2	3	4	5	6	7	am flawed in some way
some way								

Fig. 1. Example scenario from the Cognitive Style Questionnaire.

.....

from the CSQ. GPT was accessed using OpenAI's API via R Studio. The primary GPT model used for this project was GPT-4. The code used for this study was adapted from an instructional video explaining how to access GPT for psychological research that was developed and posted by the Social Identity & Morality Lab at New York University (https://www.youtube.com/watch?v=Mm3uoK4Fogc). In total, 30 different prompts were used to create GPT-generated cognitive vulnerability scores. All prompts and results can be found in the Supplemental Materials (https://osf.io/c3yf6/files/osfstorage). Prompts were created sequentially based on results (or lack thereof), and were driven by three questions:

1. Could GPT use written causal explanations for hypothetical negative events to create reliable vulnerability scores that predict depressive symptoms? Twenty-three of the 30 prompts were focused on this question (see Supplemental Materials for exact wording of all prompts). These prompts requested that GPT return cognitive

vulnerability scores for participants based on their causal explanations for hypothetical negative life events. Most of the prompts asked GPT to generate risk scores on a 1–7 scale (like the CSQ), with 1 being the lowest risk for future depression and 7 being the highest, whereas others asked GPT to return 'high' or 'low' risk scores, as either a 0 (low) or 1 (high). The prompts also provided varying degrees of context, starting with only participants' causal explanations and no other information and then increasing to explanations of hopelessness theory and information about the CSQ and BDI scores of the sample (e.g. range, percentiles, means, standard deviations). The ability of the GPT-derived risk scores to predict changes in depressive symptoms were tested using hierarchical linear regression.

Note that for 15 of these 23 prompts, the CSQ written response data (i.e., the causal explanations) for the hypothetical scenarios was combined into a single (composite) narrative text for each participant. The other 8 prompts used written response data (i.e., causal explanations) from a single hypothetical negative event from the CSQ rather than the composite of responses to all the hypothetical events. These prompts used 3 scenarios, respectively—"you are unhappy," "you take an exam and receive a low grade on it," and "an important romantic relationship you are involved in breaks up because the other person no longer wants a relationship with you."

2. Could GPT use cognitive vulnerability scores (CSQ) and current depression scores (BDI) to create reliable vulnerability scores that predict depressive symptoms? Given the lack of significant results for question 1, we decided to provide "hard" data to GPT. Specifically, we provided baseline data that we knew was predictive of depressive symptoms at follow up: baseline CSQ and BDI scores. Seven prompts were used to test this question (see Supplemental Materials for the exact wording for all prompts used).

3. Could GPT use written causal explanations to differentiate (qualitatively) those who scored high versus low on cognitive vulnerability, and to differentiate (qualitatively) those who reported increases versus decreases in depressive symptoms? Given the lack of significant findings for questions 1 and 2, we stopped asking GPT to generate numerical values for cognitive risk. Instead, we prompted GPT to use its natural language abilities to qualitatively describe the differences in the written causal explanations for participants who scored high vs. low on the CSQ and for participants who experienced an increase or decrease in depressive symptoms, respectively. GPT was also asked to provide examples of participants and their text that best demonstrated these differences. Six prompts were used to test this question (see Supplemental Materials for the exact wording for all prompts used).

Results

Descriptive statistics for the study variables are shown in Table 1. The data generated and/or analyzed is available in the OSF repository: https://osf.io/c3yf6/files/osfstorage/.

Hierarchical linear regression (jamovi, ver. 2.6). was used to test predictions. In all regression models, level of depressive symptoms (BDI) at Time 2 was the dependent variable. CSQ and GPT derived risk scores (alone and combined) were the independent variables. Baseline levels of depressive symptoms were used as a covariate to control for individual differences in initial depressive symptoms. We report total model metrics (R^2), unstandardized coefficients (b), t-values, *p* values, ΔR^2 , and 95% confidence intervals.

The CSQ was normally distributed (Fig. 2) with little evidence of skewness (-.21; see Fig. 2). The BDI was positively skewed (skewness = 1.80; see Fig. 3), as expected, given the low rates of depressive symptoms in non-clinical populations.

Replication of Prior Work—Traditional Paper-and-Pencil Questionnaire (CSQ) predicting depressive symptoms

Hierarchical multiple regression was used to test the effect of baseline cognitive vulnerability on depressive symptoms at time 2 controlling for depressive symptoms at baseline. Replicating previous research, cognitive vulnerability was a significant predictor of depressive symptoms at time 2 when controlling for depressive symptoms at baseline (R^2 = 0.377, b = .62, t = 3.03, p = .003, 95% CI [0.22, 1.02]).

GPT predicting depressive symptoms using participants' causal explanations

GPT used participants' written causal explanations to create cognitive vulnerability risk scores. The risk scores were then entered into a hierarchical multiple regression equation to determine if the GPT output could predict depressive symptoms at time 2 controlling for depression symptoms at baseline (as demonstrated by the CSQ; see analysis above). We started by giving GPT this prompt: "*Participants were administered a survey in which they read about hypothetical negative events. They were instructed to picture each event as clearly as they could, and as if the events were happening to them right now. Then, they were instructed to write down what they feel caused each event to happen to them. Using the text they wrote, assign each participant a score from 1 to 7 to indicate their risk for depression. 1 indicates the lowest risk for depression and 7 indicates the highest risk for depression. Only give the*

	Mean	Median	Standard deviation
Cognitive Vulnerability	4.00	4.04	0.86
Depressive Symptoms—Baseline	6.20	5.00	5.70
Depressive Symptoms—Time 2	5.20	3.00	6.04

Table 1. Means, medians, and standard deviations for study variables at baseline and time 2. Cognitivevulnerability was measured with the Cognitive Style Questionnaire; Depressive symptoms were measured withthe Beck Depression Inventory.



Fig. 2. Distribution of scores for cognitive vulnerability to depression.



Fig. 3. Distribution of Baseline scores for depressive symptomatology.

score as a double without explanation, and only use the participant's text to determine their score. Here is the text:." Note that the term 'double' in the prompt refers to floating-point numbers with double precision, meaning that the returned score will have a decimal point.

This prompt generated cognitive vulnerability risk scores that did not significantly predict depressive symptoms at time 2 ($R^2 = 0.370$, b = 0.20, t = 0.73, p = 0.467, 95% CI [-0.31, 0.74]). Thus, we gradually increased the amount of information that was provided to GPT in the prompt, including explanations of cognitive theories of depression, and providing examples of 3 participants' text and corresponding CSQ scores. In addition, we fine-tuned 2 new models using OpenAI's fine-tuning capabilities in the API interface. Fine-tuning allows users to create a new model that is trained on a given data set, and we trained the 2 models using the first 75 participants' composite text and either the corresponding CSQ scores or BDI scores. Finally, we provided GPT with a subset of the data set that was split into the top and bottom 15% of BDI difference scores. BDI difference scores were calculated by subtracting baseline depressive symptoms from time 2 depressive symptoms. Positive values represented increases in depressive symptoms over the interval whereas negative values indicated decreases in symptoms. This data set was used to test if removing the middle 70% of participants (i.e., providing the extremes) would help GPT pick up more clearly on linguistic differences between those experiencing high increases and decreases in depressive symptoms.

In all, 15 prompts were used to test if GPT could generate cognitive vulnerability risk scores when given participants' composite causal explanations. (see Supplemental Materials). Two of the prompts resulted in GPT outputs that could predict changes in depressive symptoms. The first prompt that yielded significant results provided GPT-4 with an explanation of hopelessness theory and then asked GPT to return a cognitive

vulnerability score based on the aggregated text from each participant. The prompt read: "you are a psychologist who is an expert at studying and measuring depression. You have completed many thousands of hours scoring measures of depression and determining patients' risk for future depression. Specifically, you are an expert in the hopelessness theory of depression. According to the hopelessness theory of depression, some people are at high risk for developing depression because they have a 'negative cognitive style' for interpreting negative life events. These people generate overly negative inferences about the cause, consequences, and self-worth implications of negative life events. Specifically, when faced with a negative life event, an individual with a negative cognitive style is likely to: (a) attribute the event to stable (enduring over time) and global (affects many areas of their life) causes; (b) view the event as likely to lead to other negative consequences; and (c) construe the event as implying that he or she is unworthy or deficient. Individuals who generate these negative inferences are hypothesized to be at risk for depression. You will now be asked to predict participants' risk for depression based on a specific study. In this study, participants were administered a survey in which they read about hypothetical negative situations. They were instructed to picture each situation as clearly as they could, and as if the situation was happening to them right now. Then, they were instructed to write down what they feel caused each situation to happen to them. Using the text they wrote to explain the cause of these negative events and your knowledge of hopelessness theory, assign each participant a score of either 0 or 1 indicating their risk for depression. 1 indicates very high risk for increased depression, and 0 indicates lower risk for increased depression. All of the events participants' were asked to consider were negative, so the participants' explanations will use negative language. Therefore, do not just return '1' due to this negative language, be very discriminative only assign a 1 if the participant is clearly at very high risk for depression according to the hopelessness theory of depression. Only give the score as an integer without explanation, and do not return any other words or characters. Here is the text:.'

The output from this prompt was a statistically significant predictor of depressive symptoms, controlling for baseline levels of depressive symptoms (R^2 =0.373, b=0.88, t=2.20, p=0.028, 95% CI [0.09, 1.67]). Next, we added the CSQ to the model to compare the predictive power of the two constructs. In the combined model (predictor 1=CSQ, predictor 2=GPT), GPT output was no longer a significant predictor of depressive symptoms (ΔR^2 =0.002, b=0.73, t=1.82, p=0.069, 95% CI [-0.06, 1.52]). In contrast, CSQ score remained a significant predictor of depressive symptoms in this model (b=0.57, t=2.76, p=0.006, 95% CI [0.16, 0.97]). In other words, adding the GPT output did not significantly increase the predictive power of the model above and beyond variance already explained by the CSQ.

To test the reliability of the significant GPT finding for this prompt, we re-ran the same prompt. The output generated the second time failed to significantly predict depressive symptoms (R^2 =0.372, b=0.81, t=1.76, p=0.079, 95% CI [-0.09, 1.70]) and was only moderately correlated with the data generated the first time we ran the prompt (r=.44).

The second prompt that yielded a significant result was a model fine-tuned in OpenAI's API platform. In this case, the base GPT-3.5 model was trained using the first 75 participants' BDI scores and composite causal explanations uploaded via JSONI file to the fine-tuning interface. Then, the new model created from this training was prompted to generate cognitive vulnerability scores for the full sample of 930 participants. The prompt read: "We would like you to determine a person's risk for future depression based on their responses to a survey. Participants were instructed to picture hypothetical negative life events as clearly as they could, and as if the events were happening to them right now. Then, they were instructed to write down what they feel caused each event to happen to them. Participants also took a survey to assess their current depression levels. The current depression measure is scored on a scale from 0 to 63, with a score of 0 indicating no depression and a score of 63 indicating the highest level of depression. Using the text the participants wrote, assign each participant a score from 1 to 7 to indicate their risk for future depression. 1 indicates the lowest risk for depression and 7 indicates the highest risk for depression. Only return the score as a double without explanation. Here is the text:"

The output from this prompt was a statistically significant predictor of depressive symptoms, controlling for baseline levels of depressive symptoms (R^2 =0.382, b=2.18, t=3.98, p<.001, 95% CI [1.10, 3.26]). Next, we added the CSQ to the model to compare the predictive power of the two constructs. In the combined model (predictor 1=CSQ, predictor 2=GPT), GPT output remained a significant predictor of depressive symptoms (ΔR^2 =0.01, b=2.04, t=3.71, p<.001, 95% CI [0.96, 3.11]). CSQ score also remained a significant predictor of depressive symptoms in this model (b=0.544, t=2.68, p=.008, 95% CI [0.15, 0.94]). This means that GPT output significantly increased the predictive power of the model above and beyond variance already explained by the CSQ (by 1%).

To test the reliability of the significant GPT finding for this prompt, we re-ran the same prompt. The output generated the second time replicated the results from the first running (output was correlated with the first running of the prompt, r = .93). GPT output was a significant predictor of depressive symptoms, controlling for baseline depressive symptoms ($R^2 = 0.381$, b = 2.11, t = 3.94, p < .001, 95% CI [1.06, 3.16]). Next, we added the CSQ to the model to compare the predictive power of the two constructs. In the combined model (predictor 1 = CSQ, predictor 2 = GPT), GPT output remained a significant predictor of depressive symptoms ($\Delta R^2 = 0.01$, b = 1.97, t = 3.69, p < .001, 95% CI [0.93, 3.02]). CSQ score also remained a significant predictor of depressive symptoms in this model (b = 0.55, t = 2.70, p = 0.007, 95% CI [0.15, 0.95]). This means that GPT output significantly increased the predictive power of the model above and beyond variance already explained by the CSQ (by 1%). We also attempted to replicate this finding in an independent sample of participants. In this case, the GPT prompt did not yield output that predicted depressive symptoms ($R^2 = 0.476$, b = 1.643, t = 1.111, p = .270, 95% CI [-1.31, 4.59]). In contrast, the CSQ did predict depressive symptoms in the independent sample ($R^2 = 0.508$, b = 1.56, t = 2.35, p = .022, 95% CI [0.23, 2.88]; see Supplemental Materials for details).

Given the overall lack of significant GPT predictors using the aggregated causal explanations (only 2 of 15 prompts), we decided to try using written explanations for a single negative life event, rather than the aggregate of all the negative events. We used 3 of the 12 CSQ scenarios—"you are unhappy," "you take an exam and

receive a low grade on it," and "an important romantic relationship you are involved in breaks up because the other person no longer wants a relationship with you," respectively. Eight more prompts were evaluated (see Supplemental Materials).

One of the 8 prompts yielded significant results. The prompt provided GPT with participants' causal explanations of the CSQ scenario "you take an exam and receive a low grade on it," an explanation of hopelessness theory, and the 15th, 50th, and 80th percentile of CSQ scores in the sample population. Specifically, the prompt read: "You are a psychologist who is an expert at studying and measuring depression. You have completed many thousands of hours scoring measures of depression and determining patients' risk for future depression. Specifically, you are an expert in the hopelessness theory of depression. According to the hopelessness theory of depression, some people are at high risk for developing depression because they have a "negative cognitive style" for interpreting negative life events. These people generate overly negative inferences about the cause, consequences, and self-worth implications of negative life events. Specifically, when faced with a negative life event, an individual with a negative cognitive style is likely to: (a) attribute the event to stable (enduring over time) and global (affects many areas of their life) causes; (b) view the event as likely to lead to other negative consequences; and (c) construe the event as implying that he or she is unworthy or deficient. Individuals who generate these negative inferences are hypothesized to be at risk for depression. You will now be asked to predict participants' risk for depression based on a specific study. In this study, participants were administered a survey in which they read about hypothetical negative situations. They were instructed to picture each situation as clearly as they could, and as if the situation was happening to them right now. Then, they were instructed to write down what they feel caused each situation to happen to them. One specific negative situation included in this survey was this: You take an exam and receive a low grade on it. Here are a few examples of text from participants explaining what caused them to receive a low grade on the exam and the corresponding depression risk score that they received: "the test was difficult" received a score of 3.13 and was the 15th percentile of risk for depression. "I didn't spend enough time studying" received a score of 4.19 and was the 50th percentile of risk for depression. "I blew off the exam because I didn't think it was going to be hard" received a score of 4.85 and was the 85th percentile of risk for depression. Using the text they wrote to explain what caused them to receive a low grade on the exam, assign each participant a score from 1 to 7 to indicate their risk for depression. 1 indicates the lowest risk for depression and 7 indicates the highest risk for depression. Only give the score as a double without explanation, and only use the participant's text to determine their score. Here is the text:."

The output from this prompt was a statistically significant predictor of depressive symptoms, controlling for baseline levels of depressive symptoms (R^2 =0.378, b=1.69, t=3.15, p=.002, 95% CI [0.64, 2.74]). Next, we added the CSQ to the model to compare the predictive power of the two constructs. In the combined model (predictor 1=CSQ, predictor 2=GPT), GPT output remained a significant predictor of depressive symptoms (ΔR^2 =0.007, b=1.67, t=3.13, p=.002, 95% CI [0.62, 2.71). CSQ score also remained a significant predictor of depressive symptoms in this model (b=0.603, t=2.97, p=.003, 95% CI [0.20, 1.00]). This means that GPT output significantly increased the predictive power of the model above and beyond variance already explained by the CSQ (by .7%).

To test the reliability of the significant GPT finding for this prompt, we re-ran the same prompt. The output generated the second time replicated the results from the first running (output was correlated with the first running of the prompt, r=.93). GPT output was a significant predictor of depressive symptoms, controlling for baseline depressive symptoms ($R^2=0.375$, b=1.31, t=2.52, p=.012, 95% CI [0.29, 2.33]). Next, we added the CSQ to the model to compare the predictive power of the two constructs. In the combined model (predictor 1 = CSQ, predictor 2 = GPT), GPT output remained a significant predictor of depressive symptoms ($\Delta R^2 = 0.007$, b=1.30, t=2.52, p=.012, 95% CI [0.288. 2.315]). CSQ score also remained a significant predictor of depressive symptoms in this model (b=0.61, t=23.00, p=0.003, 95% CI [0.21, 1.01]). This means that GPT output significantly increased the predictive power of the model above and beyond variance already explained by the CSQ (by .7%). We also attempted to replicate this finding in an independent sample of participants. In this case, the GPT prompt did not yield output that predicted depressive symptoms ($R^2=0.472$, b=-0.119, t=-0.05, p=.960, 95% CI [-4.852, 4.614]). In contrast, the CSQ did predict depressive symptoms in the independent sample (0.508, b=1.555, t=2.35, p=0.022, 95% CI [0.234, 2.875]; see Supplemental Materials for details).

GPT predicting depressive symptoms using CSQ and BDI scores

Given the overall lack of significant findings for GPT using participants' written causal explanations, we decided to provide GPT with 'hard' data. We wanted to know if GPT could use information known to be predictive of participants' future depressive symptoms: baseline BDI and CSQ scores. We used 7 prompts that provided GPT with only the numerical scores from these measures. The initial prompts provided either the CSQ or BDI score, but more information was added in later prompts, including providing both CSQ and BDI scores and adding information about mean, median, standard deviation, and range of CSQ and BDI scores in the sample population.

One of the 7 prompts resulted in GPT output that significantly predicted depressive symptoms controlling for depression symptoms at baseline. This prompt provided GPT-4 with participants' baseline BDI score and CSQ score, as well as information about baseline BDI and CSQ score ranges, means, and standard deviations within the participant data set. The prompt read: "You are a psychologist who is an expert at studying and measuring depression. You have completed many thousands of hours scoring measures of depression and determining patients' risk for future depression. You are provided with two scores that will help you determine these participants' risk for future depression. The first score, which is listed first and before the comma, is the participant's Cognitive Style Questionnaire score, or CSQ score, which is a score that indicates a participant's risk for future depression based on the hopelessness theory of depression. The CSQ is scored on a scale from 1–7, with 1 indicating low risk for rises in depression and 7 indicating high risk for rises in depression. For this population, the mean CSQ score was 4.0, and the standard deviation was 1.0. The second score, which is listed after the comma, is the participant's Beck

Prompt number	GPT predictor alone	Controlling for cognitive vulnerability (CSQ)	Prompt repeated (replication)	Controlling for cognitive vulnerability (CSQ)	Second data set result
1	p=.467				
2	p=.637				
3	p=.370				
4	N/A				
5	p=.130				
6	p=.283				
7	p=.071				
8	<i>p</i> =.028*	p=.069	p=.079		
9	p=.417				
10	p=.668				
11	p=.884				
12	p=.163				
13	<i>p</i> <.001***	<i>p</i> < .001***	<i>p</i> <.001***	<i>p</i> <.001***	p=.270
14	p=.587				
15	p=.970				
16	p=.867				
17	p=.590				
18	p=.641				
19	p=.157				
20	<i>p</i> = .086				
21	p=.871				
22	<i>p</i> =.002**	<i>p</i> =.002**	<i>p</i> =.012*	p=.012*	p=.960
23	p=.551				
24	p=.091				
25	p=.391				
26	<i>p</i> =.061				
27	p=0.162				
28	p=0.120				
29	p=0.065				
30	<i>p</i> =.044*	<i>p</i> = .642	p=.123		

Table 2. Summary of GPT results predicting future depressive symptoms (controlling for baseline depressive symptoms). Prompt 4 did not produce usable data; see more information in the supplemental materials. (*) p = .05, (**) p = .01, (***) p < .001. Significant values are in bold.

Depression Inventory score, or BDI score, which indicates the participant's current level of depressive symptoms. The BDI scores in this sample ranged from 0–41, with 0 indicating the lowest current level of depressive symptoms and 41 indicating the highest current level of depressive symptoms. For this population, the mean BDI score was 6.2, and the standard deviation was 5.7. CSQ score and BDI score are both strong predictors of future depression. Given these two scores, assign each participant a score of either 0 or 1 indicating their risk for depression. 0 indicates low risk for future rises in depression and 1 indicates high risk for future rises in depression. Only give the score as an integer without explanation, and only use the participant's text to determine their score. Here are the scores:"

The output from this prompt was a statistically significant predictor of depressive symptoms, controlling for baseline levels of depressive symptoms (R^2 =0.373, b=0.82, t=2.02, p=0.044, 95% CI [0.02, 1.63]). Next, we added the CSQ to the model to compare the predictive power of the two constructs. In the combined model (predictor 1=CSQ, predictor 2=GPT), GPT output was not a significant predictor of depressive symptoms (ΔR^2 =0.000, b=0.23, t=0.47, p=0.642, 95% CI [-0.72, 1.17]). In contrast, the CSQ score remained a significant predictor of depressive symptoms in this model (b=0.56, t=2.30, p=0.022, 95% CI [0.08, 1.03]). In other words, adding the GPT output did not significantly increase the predictive power of the model above and beyond variance already explained by the CSQ.

To test the reliability of the significant GPT finding for this prompt, we re-ran the same prompt. The output generated the second time did not replicate the results from the first running (output was correlated with the first running of the prompt, r=.77). GPT output was not a significant predictor of depressive symptoms, controlling for baseline depressive symptoms (R^2 =0.372, b=0.63, t=1.54, p=.123, 95% CI [-0.17, 1.43]). For summary of all results see Table 2.

GPT natural language descriptions of high versus low CSQ and BDI scorers, respectively

GPT had difficulty using written causal explanations and hard data to create vulnerability scores to predict depressive symptoms. This may be because GPT is trained on a corpus of natural language and may be better at language-based tasks than those that involve processing and/or returning numerical information²⁹. Thus, we decided to ask ChatGPT (using the GPT-40 model) to provide language-based qualitative output rather than quantitative output (i.e. numerical risk scores). We used 6 prompts and provided ChatGPT with either the full 930 participant data set or a subset of the data split into the top and bottom 15% of BDI difference scores. ChatGPT was prompted to describe the differences between those who reported the greatest increases vs. decreases in depressive symptoms and to include examples of participants that highlight these differences. Six prompts were used (see Supplemental Materials).

For example, one prompt read: "You are a psychologist who is an expert at studying and measuring depression. You have completed many thousands of hours scoring measures of depression and determining patients' risk for future depression. This data set gives you written text from participants in a survey (the column named "Combined Text") and a score that indicates the participants' rise or fall in depressive symptoms over the course of 1 month (the column named "BDI Difference Score"). An increase in depressive symptoms is noted by a positive score, and a decrease in depressive symptoms is given by a negative score. Given this information, determine the dimensions in which the text differs for participants that had increases in depressive symptoms vs. those that had decreases. Summarize in a paragraph the thematic differences between the text of those who saw increases in depression vs. those who saw decreases."

The ChatGPT response was: "Thematic analysis of the written text from participants reveals distinct differences between those who experienced increases and decreases in depressive symptoms over one month. Participants with increased depressive symptoms frequently mentioned feelings of inadequacy and failure, as highlighted by common words such as "enough," "didn't," "not," and "study." These individuals often expressed sentiments of not meeting expectations or falling short in their endeavors, which likely exacerbated their depressive symptoms. In contrast, those with decreased depressive symptoms also used terms like "enough" and "didn't," but their texts included more words that could imply self-reflection and acceptance, such as "was," "too," and "time." This group's text suggests a focus on understanding and possibly coming to terms with their experiences, which may have contributed to a reduction in their depressive symptoms. Overall, the increased depression group's text suggested more self-awareness and contextual understanding of their situations."

Discussion

The purpose of this study was to determine if AI could predict depressive symptoms using participants' causal explanations for negative events. Results showed that, with a few exceptions, GPT was not able to use this information to generate reliable and valid predictors of depressive symptoms. Only 2 of the 30 prompts yielded positive, replicable findings. And in both these instances, the GPT generated scores had similar or smaller effect sizes compared to the traditional paper-and-pencil CSQ questionnaire. However, GPT was able to use its natural language capabilities to provide qualitative output that described the differences between those who scored high and low on cognitive vulnerability and depressive symptom measures in a way that was consistent with vulnerability as described in the cognitive theories of depression.

Why was GPT inferior to an old-school paper-and-pencil questionnaire? The most parsimonious explanation is that GPT is not yet capable of evaluating a large amount of narrative information and translating it into numerical output. Indeed, GPT had difficulty creating a normal distribution of scores when directly prompted to do so. Because GPT is based in natural language, it has been known to struggle with numerical tasks. For example, when GPT-4 is given a string of 1 s and 0 s and is simply asked to count the number of 1 s, it often gets the answer wrong.

Our results are consistent with several studies in which LLMs did not improve the measurement of psychological constructs. For example, Englhardt and colleagues³⁰ found that when asked to assign participants a Patient Health Questionnaire-4 score (PHQ-4), both GPT-3.5 and GPT-4 selected the middle value every time, making it an unproductive metric. And, a review of natural language processing for mental illness detection by Zhang and colleagues³¹ found that despite promising results in several studies, there are still key challenges in using machine learning and AI for these tasks, including issues of performance instability and interpretation (see also³²).

A second explanation for the results is that GPT was given an impossible task. It is possible that people's written causal explanations are not predictive of depressive symptoms. In other words, they cannot be used to differentiate those at high risk vs. low risk. This is because risk for depression might not be determined by the negativity of one's thoughts per se, but rather the stability of the negative thoughts over time. For example, Haeffel¹⁵ theorized that all people initially generate negative thoughts in response to negative life events. This means that it is impossible to predict future depression from these initial spontaneous cognitive appraisals because of a lack of variability—nearly everyone has some negative cognitions at first. However, over time, most people reappraise the negative event and start to generate more adaptive thoughts. Those who do not make this "correction" are those who are at risk for future depression. This explanation is consistent with our results. The CSQ, which measures the stability and globality of causal explanations, not the absolute negativity of the explanations, predicted depressive symptoms whereas participants' written content did not (as evaluated by GPT).

Our results highlight an inherent problem for LLMs that use verbal reports. It may be impossible to measure or discern the underlying motivations, intent, and meaning of the narratives that people generate through verbal reports. Two people can generate the same explanation for an event yet attach different underlying meanings to those words. To illustrate this point, we looked through the data and found numerous examples in which people wrote down nearly identical negative causal explanations but rated those explanations on dimensions of stability and globality very differently. For example, one participant wrote explanations of events such as: "Feelings Changed. Not outgoing enough. Poor Genetics. Didn't work hard enough. Didn't work hard enough. Didn't spend enough time on it." Another participant wrote similarly negative content including: "Don't like my habits. I'm not being outgoing enough. Don't care how I look. Didn't understand the material. Didn't study enough. I didn't work hard enough on it." Despite the similarity in the negative explanations, the first participant scored a 5.04 on the CSQ whereas the second participant scored 2.46. This example illustrates how people attach different underlying meanings and values to the same basic narratives. In this case, the stability and globality of the causal explanations people generated could not be discerned based on the words alone.

These results suggest that the best use of AI and LLMs might be to combine quantitative data. Although GPT was not able to use participants' casual explanations to derive reliable and valid cognitive risk scores, it was able to use existing quantitative data to slightly improve (~1% explained variance) the prediction of depressive symptoms. In the two instances in which GPT remained a significant predictor when controlling for cognitive vulnerability, the prompt provided additional quantitative data—BDI scores in one model, and CSQ percentiles in the other. Thus, the best use of GPT may not be trying to find patterns in human thought, but rather, patterns in existing data. GPT might be able to determine the best ways to combine reliable and valid data to make predictions (similar to how Meehl argued for statistical as opposed to clinical prediction (e.g.,³³)).

The study had several strengths including a large sample size, use of a longitudinal design, empirically supported measures, and a novel data set with the written causal explanations of negative events for nearly 1000 people. This study is also among the first to test AI's ability to create cognitive risk scores for depression.

The study also had limitations. For example, we used a college sample and thus, results may not generalize to more diverse populations. Similarly, we examined non-clinical depressive symptoms, so results may not apply to those with clinically significant mental health issues. Further we used a single measure of depression. Although the BDI is a well-established measure of depressive symptoms in both clinical and non-clinical samples, it does not measure depression-specific symptoms (e.g., low positive affect/anhedonia). Thus, future work using a depression-specific measure may be informative. Finally, we used OpenAI's GPT as the LLM for all analyses in this project, thus future research should compare the abilities of other LLMs, such as Anthropic's Claude or Google's Gemini, to predict depressive symptoms. We look forward to replication attempts using data sets with more diverse participants, greater clinical severity of symptoms, and testing alternate LLM models.

In conclusion, GPT did not improve the prediction of depressive symptoms relative to a widely used questionnaire of cognitive vulnerability (CSQ). However, there were a few GPT prompts that yielded positive results, and thus, the potential for this new technology to improve assessment cannot be ruled out. The results also highlight potential limitations of LLMs. For example, people's cognitive explanations and other narratives may not always contain meaningful information. It is possible for two people to generate the same negative cognitive explanation but differ greatly in their beliefs about the stability and globality of that explanation. Our results support the hypothesis that cognitive risk for depression is better conceptualized as how changeable cognitions are rather than how negative they are. And right now, the best way to measure this cognitive risk is with a questionnaire, not AI.

Data availability

The data generated and/or analyzed is available in the OSF repository: https://osf.io/c3yf6/files/osfstorage/. All prompts and results can be found in the Supplemental Materials (https://osf.io/c3yf6/files/osfstorage).

Received: 1 December 2024; Accepted: 22 May 2025 Published online: 05 June 2025

References

- Chen, Z. & Yadollahpour, A. A new era in cognitive neuroscience: The tidal wave of artificial intelligence (AI). BMC Neurosci. 25(1), 23. https://doi.org/10.1186/s12868-024-00869-w (2024).
- Van Noorden, R. & Perkel, J. M. AI and science: What 1,600 researchers think. Nature 621(7980), 672–675. https://doi.org/10.103 8/d41586-023-02980-0 (2023).
- Williams, G. Y. & Lim, S. Psychology of AI: How AI impacts the way people feell, think, and behave. Curr. Opin. Psychol. 58, 101835. https://doi.org/10.1016/j.copsyc.2024.101835 (2024).
- 4. Kessler, R. C. et al. Prevalence and treatment of mental disorders, 1990 to 2003. N. Engl. J. Med. 352, 2515–2523 (2005).
- Whiteford, H. A. et al. Global burden of disease attributable to mental and substance use disorders: Findings from the Global Burden of Disease Study 2010. Lancet 382, 1575–1586 (2013).
 World Health Organization ununuuche int/neuro recom/fact sheats/datail/damassion (2017).
- 6. World Health Organization. www.who.int/news-room/fact-sheets/detail/depression. (2017).
- Beck, A. T. The current state of cognitive therapy: A 40-year retrospective. Arch. Gen. Psychiatry 62, 953–959. https://doi.org/10.1 001/archpsyc.62.9.953 (2005).
- Haeffel, G. J. et al. Measuring cognitive vulnerability to depression: Development and validation of the Cognitive Style Questionnaire. Clin. Psychol. Rev. 28(5), 824–836. https://doi.org/10.1016/j.cpr.2007.12.001 (2008).
- 9. Lorenzo-Luaces, L. The evidence for cognitive behavioral therapy. JAMA 319, 831-832. https://doi.org/10.1001/jama.2017.20826 (2018).
- Abramson, L. Y., Metalsky, G. I. & Alloy, L. B. Hopelessness depression: A theory-based subtype of depression. *Psychol. Rev.* 96(2), 358–372. https://doi.org/10.1037/0033-295x.96.2.358 (1989).
- 11. Beck, A. T. Depression: Clinical, Experimental, and Theoretical Aspects (Harper & Row, 1967).
- Liu, R. T., Kleiman, E. M., Nestor, B. A. & Cheek, S. M. The hopelessness theory of depression: A quarter century in review. Clin. Psychol. A Publ. Div. Clin. Psychol. Am. Psychol. Assoc. 22(4), 345–365. https://doi.org/10.1111/cpsp.12125 (2015).
- 13. Abramson, L. Y. et al. Cognitive vulnerability to depression: Theory and evidence. J. Cogn. Psychother. 13(1), 5–20 (1999).
- Alloy, L. B. et al. Prospective incidence of first onsets and recurrences of depression in individuals at high and low cognitive risk for depression. J. Abnorm. Psychol. 115(1), 145–156. https://doi.org/10.1037/0021-843X.115.1.145 (2006).

- Haeffel, G. J. After further deliberation: Cognitive vulnerability predicts changes in event-specific negative inferences for a poor midterm grade. Cogn. Ther. Res. 35(4), 285–292. https://doi.org/10.1007/s10608-010-9298-y (2011).
- Moshirfar, M., Altaf, A. W., Stoakes, I. M., Tuttle, J. J. & Hoopes, P. C. Artificial intelligence in ophthalmology: A comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. *Cureus* 15(6), e40822. https://doi.org/10.775 9/cureus.40822 (2023).
- Bartal, A., Jagodnik, K. M., Chan, S. J. & Dekel, S. AI and narrative embeddings detect PTSD following childbirth via birth stories. Sci. Rep. 14(1), 8336. https://doi.org/10.1038/s41598-024-54242-2 (2024).
- Hadžić, B., Ohse, J., Danner, M., Peperkorn, N., Mohammed, P., Shiban, Y., & Rätsch, M. (2024). AI-supported diagnostic of depression using clinical interviews: A pilot study. In Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 500-507. https://doi.org/10.5220/0012439700003660
- Levis, M., Leonard Westgate, C., Gui, J., Watts, B. V. & Shiner, B. Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. *Psychol. Med.* 51(8), 1382–1391. https://doi.org/10.1017/S0033291720000173 (2021).
- Doom, J. R. & Haeffel, G. J. Teasing apart the effects of cognition, stress, and depression on health. Am. J. Health Behav. 37(5), 610–619. https://doi.org/10.5993/AJHB.37.5.4 (2013).
- Gorman, E. L., Burke, H. H., Rubino, L. G., Vargas, I. & Haeffel, G. Teasing apart the effect of depression specific and anxiety specific symptoms on academic outcomes. *Clin. Psychol. Spec. Educ.* 9, 91–104. https://doi.org/10.17759/cpse.2020090307 (2020).
 Haeffel, C. L. Dan't algon an it Less algon reduces risk for depression symptome in aggritude with earble undergraduates. *L Para*
- Haeffel, G. J. Don't sleep on it: Less sleep reduces risk for depressive symptoms in cognitively vulnerable undergraduates. J. Pers. Soc. Psychol. 113(6), 925–938. https://doi.org/10.1037/pspp0000119 (2017).
 Haeffel, G. L. & Hamag, L. Cognitive superphility to depression can be capterious. Clin. Basehol Sci. 2(1), 75–85. https://doi.org/10.1037/psp10000119 (2017).
- Haeffel, G. J. & Hames, J. L. Cognitive vulnerability to depression can be contagious. Clin. Psychol. Sci. 2(1), 75–85. https://doi.org /10.1177/2167702613485075 (2014).
- Haeffel, G. J. & Mathew, A. R. Inside thoughts and outside influences: Cognitive vulnerability moderates the effect of decreases in perceived social support on depressive symptoms. J. Soc. Clin. Psychol. 29(3), 281–300. https://doi.org/10.1521/jscp.2010.29.3.281 (2010).
- Haeffel, G. J. & Vargas, I. Resilience to depressive symptoms: The buffering effects of enhancing cognitive style and positive life events. J. Behav. Ther. Exp. Psychiatry 42(1), 13–18. https://doi.org/10.1016/j.jbtep.2010.09.003 (2011).
- Alloy, L. B. et al. The Temple-Wisconsin Cognitive Vulnerability to Depression Project: Lifetime history of axis I psychopathology in individuals at high and low cognitive risk for depression. J. Abnorm. Psychol. 109(3), 403–418 (2000).
- 27. Beck, A. T., Rush, A., Shaw, B. & Emery, G. *Cognitive Therapy of Depression* (The Guilford Press, 1979).
- Beck, A. T., Steer, R. A. & Garbin, M. G. Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clin. Psychol. Rev.* 8(1), 77–100. https://doi.org/10.1016/0272-7358(88)90050-5 (1988).
- 29. Chen, B. X. The new ChatGPT offers a lesson in A.I. Hype. The New York Times (2024). https://www.nytimes.com/2024/05/31/te chnology/personaltech/chatgpt-4o-openai-review.html
- Englhardt, Z., Ma, C., Morris, M. E., Chang, C.-C., Xu, X. "Orson," Qin, L., McDuff, D., Liu, X., Patel, S. & Iyer, V. From classification to clinical insights: Towards analyzing and reasoning about mobile and behavioral health data with large language models. *Proc.* ACM Interact. Mobile Wearable Ubiquitous Technol. 8(2), 1–25 (2024). https://doi.org/10.1145/3659604
- Zhang, T., Schoene, A. M., Ji, S. & Ananiadou, S. Natural language processing applied to Mental Illness Detection: A Narrative Review. Npj Digital Med. 5(1), 46. https://doi.org/10.1038/s41746-022-00589-7 (2022).
- 32. Demszky, D. et al. Using large language models in psychology. Nat. Rev. Psychol. https://doi.org/10.1038/s44159-023-00241-5 (2023).
- 33. Grove, W. M. Clinical versus statistical prediction: The contribution of Paul E. Meehl. J. Clin. Psychol. 61(10), 1233–1243 (2005).

Acknowledgements

This research was made possible in part by support from the Summer Comprehensive Grant Program, Institute for Scholarship in the Liberal Arts, College of Arts and Letters, the University of Notre Dame.

Author contributions

JS co-conceptualized the study, ran the study using GPT, conducted analyses, and wrote the main manuscript. GH co-conceptualized the study, conducted analyses, wrote portions of the manuscript, edited the manuscript, and provided research supervision.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/1 0.1038/s41598-025-03740-y.

Correspondence and requests for materials should be addressed to G.J.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025