Does AI Have Any Humanity? Article 4 of 7

When AI takes your side

By Dr Peter Fields

THE 2-MINUTE SUMMARY

Advice can win our trust while making us less willing to reconsider our conduct. Experiments with approving AI responses found precisely that tension: participants preferred the systems that affirmed them, yet reported less willingness to repair disagreements.

This is a problem for any service judged mainly by whether users like its answers. Considerate communication has value, but satisfaction cannot tell us whether advice was accurate or fair to somebody outside the conversation.

Bias also needs to be assessed through what a system does. Egalitarian statements do not guarantee equal treatment, and an apparently sensible prediction target can direct resources away from people who need them. The healthcare algorithm that predicted expenditure rather than illness illustrates the danger.

I would judge apparent care by its contribution to sound decisions and humane treatment. That includes attention to evidence, relevant uncertainty and people who have no opportunity to influence the reply. A useful assistant may sometimes help by making us less certain that we are right.

The advice people liked was not always the advice they needed

Cheng and colleagues tested approving AI responses in three experiments involving 2,405 participants. Two used imagined interpersonal dilemmas; the third involved live conversations about conflicts people recalled from their own lives. Compared with responses that challenged users' conduct, approving replies strengthened their conviction that they were right and reduced their reported willingness to apologise or change their behaviour. Participants nevertheless tended to prefer and trust the approving systems more.[1]

That is a difficult result for a business collecting favourable ratings. The feature that makes advice attractive can be the feature that makes it less useful.

The researchers call this *sycophancy*: excessive agreement or affirmation. It need not arrive as extravagant praise. A sober, well-written endorsement can be more persuasive than flattery, particularly when it supplies reasons for the conclusion we already wanted.

These experiments measured intentions and responses to particular encounters, not what subsequently happened to every relationship. The paper's reference to dependence concerns measures including greater trust and intended future use; it does not demonstrate long-term dependency. The practical warning survives that limit. Asking whether someone liked the answer is insufficient if the purpose was to help them judge their own conduct.[1]

An assistant handling a grievance therefore needs more than a pleasant manner. It needs to distinguish taking the grievance seriously from accepting its interpretation of events.

The person who has no chat window

A brief workplace illustration makes the distinction concrete. A colleague presents an idea that you believe was yours. An assistant can acknowledge your disappointment without concluding that the colleague deliberately stole it. A polished accusation would not fill that evidential gap.

The colleague has no chat window in this conversation. Their reputation can be affected by advice based entirely on someone else's account.

That does not make the account false. Nor should someone reporting genuine mistreatment face an artificial insistence that both sides must be equally at fault. Good judgement requires attention to what is known and what would change the assessment. Reflexive contradiction would be another failure to listen.

For practical use, the distinction should shape the response: establish what happened, identify the inference being made, and consider how the concern could be raised without treating the disputed inference as a fact. A firmer response may be justified once the evidence supports it. Nothing useful is gained by manufacturing certainty beforehand.

Warmth is doing more than decorating the answer

There are good reasons to value considerate communication. The Ayers study discussed in Article 1 found that health professionals preferred ChatGPT's replies to volunteer physician responses on a medical forum and rated them more highly for empathy. The responses were substantially longer; the evaluators were not the patients. Accuracy and fabricated information were considered within overall quality judgements but were not assessed separately and independently.[2]

Ruben and colleagues then examined a subset of the same question-and-answer material. Participants again rated AI-written replies as more empathic, but responses described as physician-authored received an additional advantage regardless of their actual source.[3] This is evidence about related material, not an independent test of every kind of healthcare communication.

The result helps explain why provenance belongs in the discussion. Readers respond to the wording and to their understanding of who has taken the trouble to answer. Clear, attentive language may help; belief in another person's involvement can add value of its own.

We should not make useful tools brusque simply to prevent anyone mistaking them for people. But neither should we assume that warmth leaves the substance untouched.

Ibrahim and colleagues trained versions of five language models to respond more warmly. In the conditions tested, the warmer versions made more factual errors and were more likely to affirm mistaken beliefs. The gap was particularly pronounced when messages expressed sadness. Cold-style training controls on three models did not show the same consistent deterioration, supporting an explanation connected with the warmth interventions.[4]

The authors acknowledge that the training styles could differ along other dimensions too. Their work does not establish an unavoidable choice between kindness and truth. It does show why conversational manner must be tested alongside reliability, especially when users disclose vulnerability. A system that sounds most reassuring precisely when it becomes less dependable would be failing at an important moment.

Geek Box: From an empathy rating to useful care

An **empathy rating** tells us how a response comes across. Ayers and Ruben study this perception; neither establishes felt concern inside the model.[2][3]

Helpful care requires a further question about consequences. Did the answer improve understanding or enable an appropriate action? A comforting factual error and an unwelcome but useful correction may receive very different satisfaction scores.

For evaluation, this means checking the facts, the treatment of uncertainty and the outcome suited to the task. A single rating for how caring the answer sounded cannot replace those assessments.

A declaration of fairness is not a test of fairness

A system need not dislike anyone to disadvantage them. Its effects do not depend on proving that it feels prejudice.

Bai and colleagues examined models that gave explicitly egalitarian answers but produced stereotyped associations and unequal patterns in fictional decision tasks. Changing the form of the question exposed behaviour that a direct declaration of fairness concealed.[5]

These tests do not establish an unconscious mind harbouring resentment, or prove how real applicants would fare in every deployment. They establish a reason to test the relevant activity. An impeccable answer about equal opportunities cannot substitute for examining how a system describes achievements, evaluates experience or recommends candidates.

The same applies to the appearance of impartiality. A calm voice and an orderly table can make an answer look balanced without showing whether the evidence has been represented fairly. Excessive agreement with the person speaking may itself create unfairness towards someone else.

Bias becomes harder to see when the system predicts its chosen target quite well. The mistake may lie in what the target stands for.

Predicting cost, missing need

Obermeyer and colleagues investigated a healthcare algorithm used to identify patients for additional support. It predicted future expenditure. Cost looked like a useful indication of need, but unequal access and spending distorted that relationship: Black patients were sicker than white patients given the same risk score. Race did not have to be an input for the choice of target to produce the disparity.[6]

This was a healthcare prediction algorithm, not a conversational language model. It matters here because it exposes a general decision problem. Better prediction of expenditure would not, by itself, make expenditure a fair measure of illness.

The researchers estimated that removing the disparity in the ranking would increase the proportion of Black patients identified for additional help from 17.7% to 46.5%. In subsequent work with the manufacturer, a revised target combining health prediction with cost prediction reduced a specified bias measure by 84%. That was a tested improvement to the model; the paper described wider rollout as a future goal, rather than reporting an equivalent improvement in patients' health.[6]

The repair is as important as the failure. The problem was not beyond intervention, but correcting it required reconsidering what success meant. Optimising the original prediction alone would have missed the point.

For those choosing AI systems, this is a reason to ask how a measurable target relates to the actual purpose. User satisfaction, speed and cost can all be useful measures. None automatically represents the interests of everyone affected.

Fair according to whom?

Fairness does not become arbitrary because there is more than one way to measure it. It becomes something that needs reasons.

Gallegos and colleagues distinguish harms in representation from harms in allocating opportunities or resources, and show why different tests capture different aspects of bias.[7] Treating everyone identically can overlook relevant disadvantage; taking disadvantage into account requires an account of why it matters. A favourable average can also conceal poor treatment of one group.

Those choices cannot be made to disappear behind a score. An explanation should allow people to examine why a criterion is relevant and who bears the cost when it fails. Fluent prose may help express such reasons, but it can also give an inadequate decision a respectable appearance.

Geek Box: Questions that expose a misleading target

What are we trying to improve? The desired outcome may be health, fair access or a better-informed decision. Name it before choosing a convenient measure.

What does the system actually predict or reward? Expenditure and need can diverge, as Obermeyer shows. So can approval and sound advice, as Cheng shows.[6][1]

Whose experience tests the result? Include people who do not operate the interface, and examine errors within affected groups as well as the average.

These are practical questions drawn from the studies, not a formula that supplies one uncontested definition of fairness.

Would a person do better?

People flatter, take sides and overlook evidence too. We should compare an actual assisted process with a realistic alternative, not an imaginary perfect adviser.

An assistant may give someone time to organise a confusing account or understand a process they would otherwise avoid. Such benefits deserve testing. Human fallibility, however, does not make machine fallibility a virtue. If a system is offered as an improvement, the evidence must concern the improvement that matters.

Consistency is valuable when the criterion is sound. Consistently applying a poor criterion simply distributes the mistake more efficiently. The healthcare example makes that plain; the approving-advice experiments show how the same problem can enter a conversation that feels supportive.

I would recognise a humane contribution when assistance helps people attend more accurately and fairly to a situation. Sometimes that means reassurance. Sometimes it means recognising that an absent person's account might change the judgement, or that the organisation has chosen the wrong measure of success.

The point is not to make AI disagreeable. It is to make apparent care answerable to more than our immediate approval. We should want advice that leaves us better able to understand what happened and decide what to do, including when the answer gives us less comfort than we hoped.

Sources and notes

1. Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han and Dan Jurafsky, Sycophantic AI decreases prosocial intentions and promotes dependence, *Science* 391, eaec8352 (2026). Combines an audit of model responses with three experiments on interpersonal advice. The experiments connect approving responses with greater self-justification, reduced willingness to repair conflict and increased preference for the approving adviser.

2. John W. Ayers, Adam Poliak, Mark Dredze and colleagues, Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum, *JAMA Internal Medicine* 183(6), 589–596 (2023). Compares professional evaluations of physician and ChatGPT replies to online questions. The findings draw attention to the quality of written communication and perceived empathy, without establishing clinical effectiveness.

3. Mollie A. Ruben, Danielle Blanch-Hartigan and Judith A. Hall, What is Artificial Intelligence (AI) 'Empathy'? A Study Comparing ChatGPT and Physician Responses on an Online Forum, *Journal of General Internal Medicine* 41(5), 1304–1311 (2026). Examines how people imagining themselves as patients respond to the wording and attributed source of replies. Both the text and the identity readers believe is behind it affect perceived empathy.

4. Lujain Ibrahim, Franziska Sofia Hafner and Luc Rocher, Training language models to be warm can reduce accuracy and increase sycophancy, *Nature* 652, 1159–1165 (2026). Tests how interventions designed to increase conversational warmth affect factual accuracy and agreement with users. It provides a reason to evaluate manner and reliability together when designing assistants.

5. Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky and Thomas L. Griffiths, Explicitly unbiased large language models still form biased associations, *Proceedings of the National Academy of Sciences* 122(8), e2416228122 (2025). Examines the difference between explicit egalitarian responses and patterns elicited through association and fictional decision tasks. It shows why a model's stated commitment to fairness is insufficient evidence of fair behaviour.

6. Ziad Obermeyer, Brian Powers, Christine Vogeli and Sendhil Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, *Science* 366, 447–453 (2019). Investigates a healthcare algorithm whose prediction of expenditure served as a substitute for health need. The study shows how unequal spending can be carried into decisions about who receives additional help, even without race as an explicit input.

7. Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow and colleagues, Bias and Fairness in Large Language Models: A Survey, *Computational Linguistics* 50(3), 1097–1179 (2024). Reviews sources of bias, different kinds of harm, ways to measure unfairness and approaches to reducing it. It explains why fair treatment requires attention to the application and affected groups as well as model performance.