Center on Social Dynamics and Policy at BROOKINGS

JANUARY 2024

THE HUMAN FACTOR:

ANTICIPATING PITFALLS IN THE APPLICATION OF ARTIFICIAL INTELLIGENCE TO HEALTH CARE

Matt Kasman, Adam B. Sedlak, Nicole Strombom, Ross A. Hammond

DISCLOSURES

The Brookings Institution is financed through the support of a diverse array of foundations, corporations, governments, individuals, as well as an endowment. A list of donors can be found in our annual reports, published online. The findings, interpretations, and conclusions in this report are solely those of its authors and are not influenced by any donation.

ACKNOWLEDGEMENTS

The authors would like to thank Sanjay Patnaik for his careful review and helpful comments, Stephanie Aaronson for her feedback, and Chris Miller for design assistance.

Introduction

Artificial intelligence (AI) shows tremendous promise for applications in health care. Tools such as machine learning algorithms, artificial neural networks, and generative AI (e.g., Large Language Models) have the potential to aid with tasks such as diagnosis, treatment planning, and resource management.[1–7] However, their ultimate impact on health outcomes will be shaped not only by the sophistication of the algorithms that they employ but by external "human factors" as well.

Foremost among these are the data given to AI tools; inaccurate, inappropriate, or incomplete data can result in poor performance--often in ways that are not anticipated by designers or transparent to users. Broadly, the datasets used by developers to train AI tools can have key gaps which can cause responses from AI tools that are lower quality for some users or situations.[2,8–16] These pitfalls are well-known among experts, but the extent to which they will hinder specific applications of AI in health care remains unclear.

Other ways in which applying AI to health care might unexpectedly produce diluted or even negative outcomes are less widely discussed. For AI tools to reach their potential, they must be trusted—potential users must believe that recommendations provided by an AI tool will result in positive outcomes that outweigh any possible costs (e.g., loss of privacy). Social factors that determine who uses AI tools and what actions they take in response to interactions with them will shape AI's impact on health care. Underpinning both behaviors are trust and social influence. Evidence suggests an initial skepticism among many potential users (who may be particularly hesitant to trust AI for consequential decisions that affect their own or other's health).[17–23] In addition, attitudes may differ within the population based on attributes such as age and familiarity with technology.[24] Direct experience with AI tools (positive or negative) can alter users' opinions about their trustworthiness, although it is likely that early experiences will weigh heavily in the formation of lasting impressions.[25–27] Similarly, indirect experiences can affect perceptions, as people communicate their opinions to others and change their own views based on what they learn.[28–34]

In this report, we use analysis from a simple computational model to illustrate how, when, and to what degree such human factors may cause applications of AI to health care to fall short of expectations and to consider ways to mitigate such circumstances. Given the limited amount of relevant data on social uptake of AI, our work here is intended as a set of "thought experiments" conducted with the assistance of computational simulation. Given the high degree of heterogeneity (among individuals and contexts) and interdependence (between them) inherent in the intersection of AI, health care, and social dynamics, we use agent-based modeling (ABM) to gain insight into these potential pitfalls we might encounter. ABM is a computational simulation approach in which individual entities (e.g., patients) are explicitly represented as they interact with one another and their environment over time.[35–39] Our ABM identifies initial insights into the conditions under which AI tools might fail to improve overall population health, might introduce health disparities, or might leave existing inequities in place.

Model Summary

We summarize the design of our model here, and provide a complete description of our model, including functional forms and specific parameter values, in Supplementary Materials.

AGENTS

Our model represents a hypothetical community of individuals receiving health care that is based on an artificial intelligence (AI) tool designed to provide medical recommendations. Simulated individuals ("agents") in our model have one of two different group affiliations (either "A" or "B"). This categorization is intended to abstractly represent a meaningful social categorization or combination of categorizations (e.g., race, gender, income, education, age cohort, etc.) along which AI bias can occur, trust might initially differ, or a tendency to form within-group social connections might take place. That is, in line with our uncertainty about how social factors might affect the application of AI to health care, group affiliation allows for the potential that our agents are functionally differentiated, while the specific groups are left intentionally undefined.

Agents each have properties denoting their health and their current level of trust in the AI tool. As time progresses in the model, the health of agents slowly deteriorates, reflecting natural age-related processes, occurrence of illnesses, and injury. In addition to this entropy, all agents have ongoing opportunities over time to interact with the AI, doing so through a set of serial decisions, actions, and effects shown in Figure 1 and summarized as follows:

FIGURE 1

Conceptual Figure Depicting Agent-Based Model (ABM) Operation

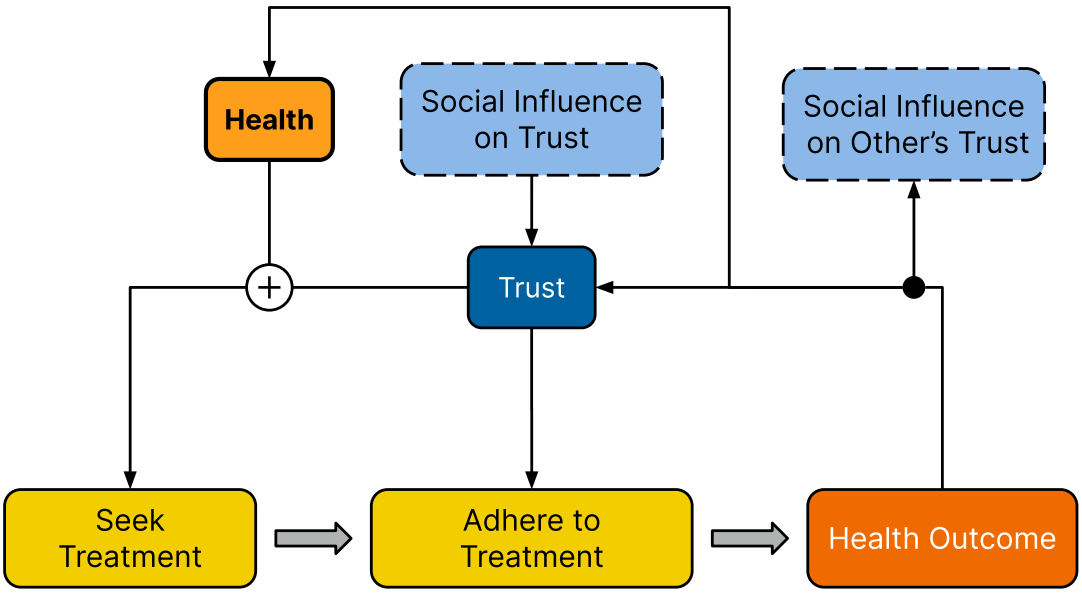


Figure 1: Flow chart between agent actions and properties. An agent's health, along with the extent to which the agent trusts AI-based healthcare, jointly influence their propensity to seek treatment. If an agent chooses to seek treatment, they then decide whether to adhere to healthcare recommendations, doing so based on their trust. If they do, then this will affect both health and trust. Agents can communicate the effects of treatment on health to one another, which results in social influence that can shift agents' trust values.

1. Seek Treatment. Agents choose whether to seek treatment from the AI health care tool. This decision is influenced by an agent's current health and current trust, with both lower health and greater trust increasing the chance of seeking treatment. If an agent chooses to seek treatment, they will consult the AI and receive a treatment recommendation.

2. Adhere to Treatment. For medical advice to have an impact on health, the patient must follow it ("adhere"). In the real world, there are many factors associated with adherence (e.g., treatment cost or unpleasantness).[40–42] In our model, we focus on a single one that is of immediate relevance: trust in the health care provider. An agent will choose to adhere to the recommendation they receive in proportion to their trust in the AI-based health care.

3. Health Outcome. Adherence to a treatment recommendation will affect an agent's health, with the effect dependent on the quality of the recommendation for the agent's health status. As discussed below, quality of recommendations is a continuous value that is probabilistically sampled each instance from a distribution that is based on AI characteristics. Thus, a highly accurate AI is likely to give high-quality recommendations that can improve health by a large amount, while a highly inaccurate one may frequently give low-quality advice that can have a negative effect on health. An agent's health outcome from a treatment recommendation will in turn influence their trust of the AI—where recommendations that increased health will increase an agent's trust of the AI and vice versa. Initial interactions with the AI will influence an agent's trust more than those later on, reflecting both a large body of social science literature on opinion formation and the common adage that "a first impression is a lasting one."[25–27]

4. Social Influence on Other's Trust. Every agent has social connections to a small number of other agents in the model. For every given simulation run, agents are randomly placed in a social network with a high degree of clustering (i.e., as in many real-world settings, there are many mutual friends in the social network). After experiencing a treatment outcome, an agent may share information about the effectiveness of the AI with their social contacts. If the treatment recommendation exceeds the agent's expectation of the AI (which is based on trust), the trust of the agents that they are socially connected to will increase. Similarly, if the AI underperforms, the agent will share this information, decreasing the trust of the agents to whom they are socially connected.

AI

The way in which each AI recommendation will affect an agent's health if followed is determined by randomly sampling from a predefined impact distribution. Our ABM allows us to control this distribution to conduct thought experiments connecting AI performance to health outcomes in three ways:

1. Average Accuracy: Allows for the adjustment of the quality of medical recommendations made by the model, which affects how much (and in what direction) average population health tends to change after adhering to a recommendation.

2. Variance: Allows for the adjustment of how variable the quality of AI recommendations are (i.e., the range of prediction accuracy). For example, do model recommendations consistently have the same accuracy or do they span a large range of accuracies?

3. Skew: We can adjust whether and to what extent the distribution is left-skewed. The presence of skew results in the AI occasionally experiencing "catastrophic failure," cases in which recommendations, if followed, result in very bad health outcomes.

SIMULATION RUNS

Each run consists of 450 "ticks" (simulated timesteps). During each of these ticks, agents' health can change through natural processes (i.e., health tends to deteriorate over time) as well as treatment-seeking and adherence. Similarly, during each tick, agents' trust values can change through experience with AI or social influence. As shown below, we can explore trends in health or trust within runs or focus on comparisons of end-of-run health and trust in the populations of agents across different runs.

Experimental Design

We use our model to conduct a large—but far from exhaustive—set of experiments to explore the potential impact of AI-based health care under a range of hypothetical scenarios. Our primary goal is to gain intuition into how different AI tools might shape health outcomes across a range of settings. Therefore, we first define sets of tools and settings and then experimentally apply each tool to each setting.

AI TOOLS

We first define a baseline AI with a distribution of recommendation quality such that average expected impact will be an increase in agents' health, but positive health effects are not guaranteed. We then vary from this baseline to explore potential impacts of AI bias. Specifically, we use our model to explore all combinations of the presence or absence of three types of potential bias that might conceivably occur due to limited or inconsistent training of the AI training with data that are more relevant for some users than others[43]:

- Differential mean accuracy: Recommendations for group A are sampled from a distribution with higher mean accuracy than group B (i.e., group A tends to receive recommendations that would result in better health outcomes if followed)
- Differential variance: Recommendations for group A are sampled from a distribution with lower variance than group B (i.e., there is less "noise" in the provision of recommendations to group A)
- Differential skew: Recommendations for group A are sampled from a distribution with no skew while recommendations for group B are sampled from a distribution with a high degree of left skew (i.e., members of group B have a greater chance of encountering rare instances of "catastrophic failure" in which they receive recommendations that would have negative health consequences if followed).

SOCIAL SETTINGS

We define social settings according to the presence or absence of two characteristics. The first is social segregation, which increases the frequency of same-group social connections and results in most communication comprising social influence on agents' trust in AI occurring between individuals who belong to the same group. Second, we allow for levels of trust to potentially differ between groups, with the possibility that trust begins substantially lower for group B than A. Such systematic differences might plausibly arise for a variety of reasons, including familiarity with technology in general or AI in particular; previous experiences within health care settings, especially with discrimination, stereotyping, stigma, or exclusion; and whether and how the use of a specific AI tool has been communicated.[24,43]

ADDITIONAL VARIATION

Many of the behaviors related to AI that we include in our model cannot be well-quantified with existing observational data. To understand the impact of potential variation in such behaviors, we explore a wide range of possible values that determine:

- When agents' experiences with AI are sufficiently satisfactory or disappointing for them to convey to others.
- How important early experiences with AI-based health care are in the formation of trust in AI-based health care relative to later experiences and social influence.
- The importance of trust in AI-based health care relative to agent health in treatment-seeking decisions.
- How much any given treatment can affect an agent's health.

Because many of the functions within our model are probabilistic, we ran the model 32 times for each of the experimental and behavioral conditions outlined above, building confidence that our results are not being driven by rare, random occurrences.

Results

After conducting simulation runs with our model, we identified several high-level findings that shed light on potential pitfalls of applying AI to health care. We summarize the most interesting of these here. Interested readers can find full details on all simulation runs in Supplementary Materials along with additional findings.

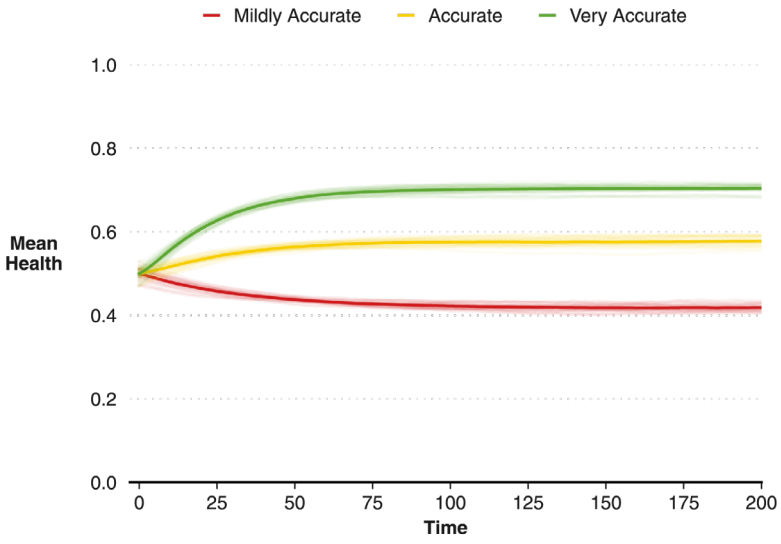
Result 1: In order to have positive impact on population health, an AI tool might have to be highly accurate; failing to reach this standard, an objectively accurate health care AI tool can potentially produce negative health outcomes if its performance is perceived as underwhelming by users.

Figure 2 shows that it is possible that human factors—how experiences are filtered through expectations, communicated to others, and then shift others' expectations—can result in surprising effects when applying AI-based health care. Here, we have a population with moderate initial trust who are willing to reshape their views over time based on their own experiences and what they learn from others. Under these plausible starting conditions, an accurate or very accurate AI tends to produce population-level health increases (shown in yellow and green in Figure 2, respectively). However, a mildly accurate AI that, on average, provides treatment recommendations with positive effects for users nonetheless can have an overall negative impact on population health (the red line in Figure 2). The key to this dynamic is in user expectations and social influence: If the quality of treatment received is lower than their moderate initial trust levels, then agents will adjust their own trust (and expectations) downward. If treatment quality is sufficiently lower than expectations ("is worth mentioning"), they will let their social network peers know, who in turn will adjust their own trust (and expectations) downward. With an AI that tends to give only slightly salubrious guidance—that is, in expectation it won't give recommendations that are harmful or completely ineffective, but also won't produce major improvements in health—moderate initial levels of trust erode as modest expectations are frequently disappointed. As this sentiment spreads through the community, agents are less likely to seek treatment or, if they do (e.g., because they are in particularly poor health) are less likely to adhere to recommendations given. As a result, we see a gradual decline in population health.

FIGURE 2

AI Accuracy Comparison

Average population health trajectories for agents interacting with AI models of various accuracies across multiple simulation runs.



Note: Darker lines depict the average model behavior while lighter colored lines show individual model runs.

Figure 2: Trends in health outcomes over time across multiple runs conducted under each of three scenarios. In each, the AI tool tends to provide guidance that, if followed, would result in positive health outcomes. In one, a "very accurate" AI provides recommended courses of action that would produce large increases in health outcomes if followed. However, this figure shows a plausible set of behavioral conditions under which an objectively effective AI tool with lower levels of accuracy results in only minor gains or even a population-level decrease in overall health over time. This occurs because of how agents' form trust in the AI-based health care tool and communicate their experiences to others (in turn affecting their social network peers' trust). The ways in which such factors affect trust, along with how trust affects treatment seeking, can make it such that an AI has a fairly high accuracy bar to clear before it effectively increases population health.

Result 2: Even an objectively accurate health care AI tool can worsen health disparities if its performance differs across users and is perceived as underwhelming by some.

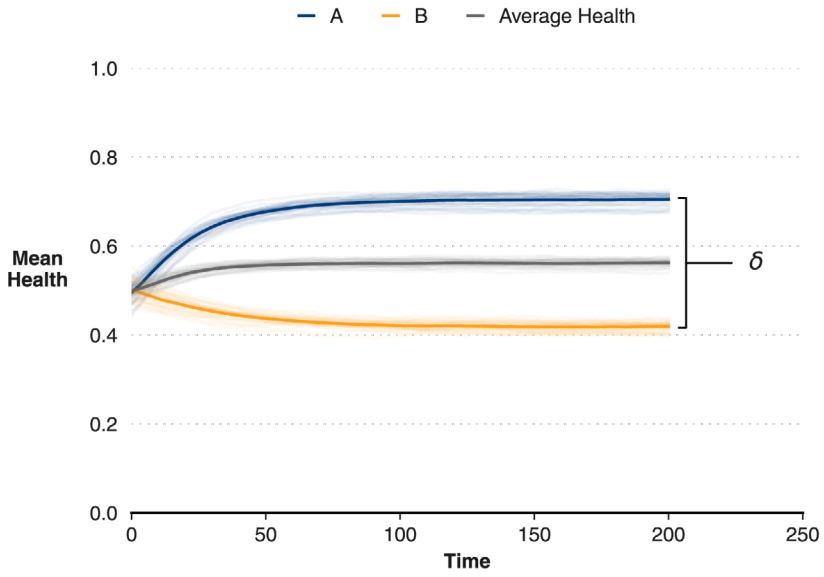
Figure 3 shows a dynamic similar to the one observed in the top line of Figure 2 (a highly accurate AI) but with one important difference: The AI is differentially accurate across groups within the population. Thus, although AI's recommendations result in treatment quality that tends to satisfy the expectations of one group, it does not for the other. Here, social influence is not a factor; agents will receive mixed messages about the AI that generally cancel one another out. However, the differences in direct experience with AI will drive down trust levels in one group, which in turn will make them less likely to seek out or adhere to treatment. Given the differences in the experiences and subsequent behaviors across groups here, we see an overall population-level health increase and the introduction of underlying health disparities.

Of course, providing different mean qualities of recommendations across groups is the clearest but not only form of bias that AI might exhibit. What if distributions of recommendation quality provided to groups had different levels of variance or skew? In the case of greater variance for one group, this would imply that recommendations provided by the AI contain less signal and more noise for members of one group (e.g., as a result of not being sufficiently trained on data relevant to that group), with agents from that group

FIGURE 3

Accurate but Biased AI

Average health of groups A and B after interacting with an accurate but biased AI.



Note: The symbol delta is the resulting health disparity between group A and B. Darker lines depict the average model behavior while lighter colored lines show individual model runs.

Figure 3: Health trends from repeated runs of a simulated scenario in which the AI tool is differentiallyaccurate across two groups. This figure shows both overall mean trends and mean trends disaggregated by group for a plausible set of behavioral conditions. The tool tends to provide guidance that would result in positive health outcomes for members of either group if followed, but provides higher quality guidance to members of group A than group B. This occurs because of how agents' form trust in the AI-based health care tool and communicate their experiences to others (in turn affecting their social network peers' trust), along with how trust is used in treatment-seeking decisions. Group A has positive experiences and Group B underwhelming ones; these cancel each other out in terms of how trust is affected by social influence. However, Group B has lower levels of trust through their direct experiences, thus seeking out treatment less frequently and suffering from worse health as a result.

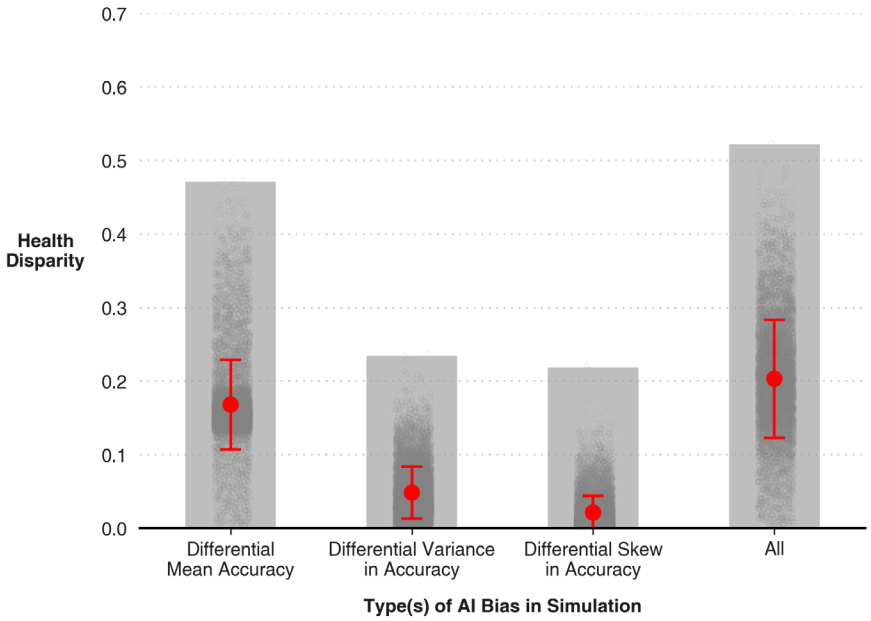
more likely to receive guidance that differs substantially from the tool's mean accuracy value. In the case of greater skew for one group, members of that group are more likely to encounter rare instances of "catastrophic failure" in which the AI provides recommendations that, if followed, would have strongly negative health consequences.

As we see in figure 4, on their own, variance and skew biases are less of a cause for concern; within the large set of behavioral conditions that we explore, we see that these forms of bias can produce at most limited health disparities. However, in combination with mean differential bias, they can potentially exacerbate the magnitude of disparities that emerge under many behavioral conditions. Although a full exploration of the intersection of bias and behavioral conditions is beyond the scope of this report, we do provide some additional information on this in Supplementary Materials.

FIGURE 4

Potential Impact of Differential AI Behavior Across Groups on Health Disparities

Resulting magnitudes of health disparities between groups A and B for different forms of AI bias across multiple simulation runs that represent a large sweep of plausible behavioral condition sets.



Note: Red dots denote the average model disparity and whiskers indicate and standard deviation radius. Gray bars show the minimum and maximum values and gray markers show individual model runs.

Figure 4: Health disparity magnitudes at the end of simulation runs given the presence of different forms of AI bias. On its own, differential mean accuracy (with one group tending to receive higher-quality treatment recommendations than the other) can result in meaningful health disparities under many behavioral conditions. Differential variance (a wider distribution of recommendation quality for one group) and skew (with one group more likely to receive rare, exceptionally low-quality recommendations) both introduce small health disparities at most. When all three forms of bias are present at the same time, substantial health disparities emerge under many behavioral conditions, with slightly greater variance in disparity magnitude than when mean AI accuracy is the only form of bias present.

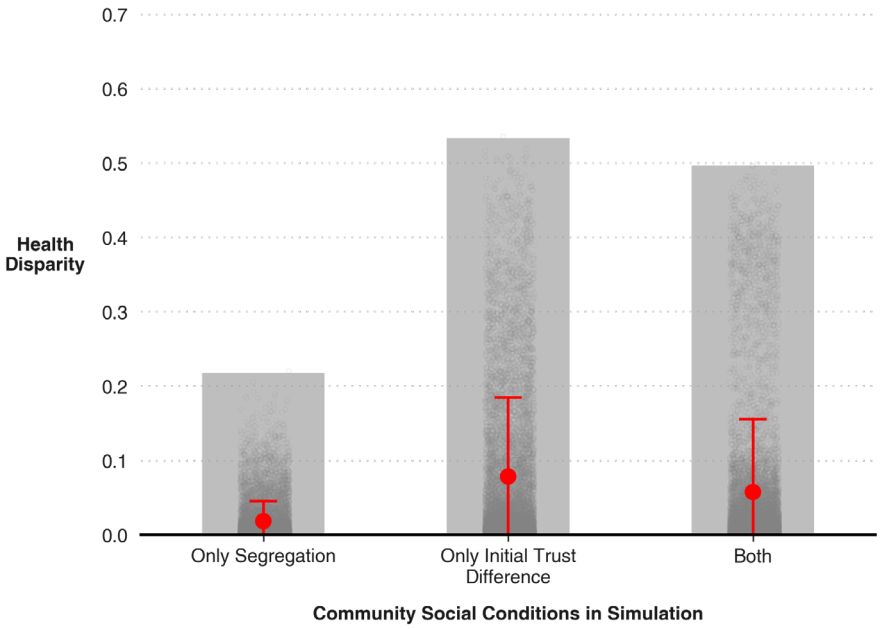
Result 3: If differences in initial trust in AI-based health care are present in a community, these may be hard to overcome and, as they drive different use patterns, may contribute to widening health disparities as AI is introduced.

Next, we explore whether and in what ways conditions that are external to AI itself but instead are socially situated within communities in which AI-based health care tools are applied might affect health outcomes. In figure 5, we have a completely unbiased AI tool; in expectation, it tends to provide all users with treatment recommendations that would result in positive health outcomes, and there is no difference in its performance related to a user's group membership. However, we see that when there are initial differences between groups in trust in AI-based health care (e.g., as a result of one group's systematically negative previous experiences with health care or due to use of the AI tool being poorly communicated to one group), there are many scenarios in which this translates into substantial health disparities between groups. This is driven by behavioral conditions that we consider plausible but are neither proven nor inevitable: when agents hold onto perceptions held early on, regardless of later direct experience or hearing others' opinions to the contrary. Importantly, these disparities driven by differential initial trust emerge regardless of whether groups are socially segregated. In other words, while fostering greater levels of communication between groups within the community would undoubtedly confer many benefits, it is not sufficient in our model to offset health disparities introduced by differences in trust in the AI tool.

FIGURE 5

Potential Impact of Community Conditions Related to AI Use on Health Disparities

Resulting magnitudes of health disparities between groups A and B under different community conditions across multiple simulation runs that represent a large sweep of plausible behavioral condition sets.



Note: Red dots denote the average model disparity and whiskers indicate and standard deviation radius. Gray bars show the minimum and maximum values and gray markers show individual model runs.

Figure 5: Health disparity magnitudes at the end of simulation runs given the presence of different community conditions and the absence of any AI bias. On its own, social segregation (the tendency for members of the same group to share their perceptions of AI more frequently with one another) does not introduce appreciable health disparities under most conditions. An initial difference in trust in AI (with trust being substantially lower for one group) can result in the introduction of at least small health disparities under many conditions, though, either by itself or in combination with social segregation. This is largely driven by conditions in which trust is less easily influenced by direct experiences or social influences, leaving one group seeking treatment at lower rates and suffering from worse health as a result.

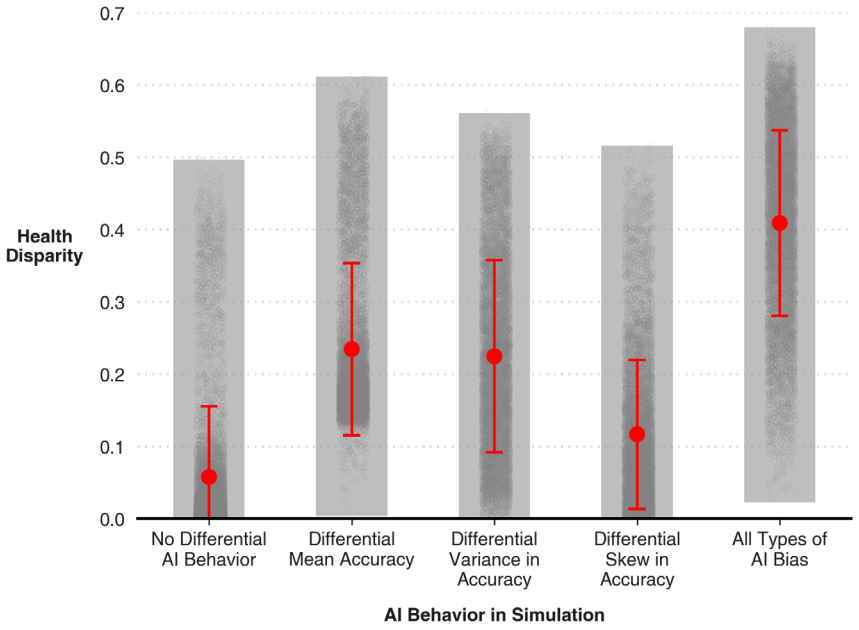
Result 4: Social conditions within communities can exacerbate the impact of biased AI tools on health disparities.

In figure 6, we see how the combination of differences in initial trust and social segregation affect health outcomes across scenarios in which the AI performs identically across groups (i.e., what was shown in Figure 5) as well as ones in which AI is biased. Here, we explore the same three types of bias that we did before: differences in mean treatment recommendation quality across groups, difference in variance, and difference in skew. Unlike before, when there were no social differences between groups within communities, here we see that there are many scenarios across those that we explore in which bias in variance and bias in skew can result in substantial health disparities. This is because with lower initial trust in one group, the greater likelihood of early experiences that vastly underperform already low expectations can solidify that lack of trust for many group members, with deleterious impact on health as a result; this is magnified by social segregation, which makes it likely that members of the lower initial trust group will not hear many opinions to the contrary. As before, a complete exploration of which combinations of behavioral conditions under which these disparities occur is beyond the scope of this report, but more information on these is provided in Supplementary Materials.

FIGURE 6

Potential Impact of Differential AI Behavior Combined with Community Conditions Related to AI Use on Health Disparities

Resulting magnitudes of health disparities between groups A and B under different forms of AI bias given the simultaneous presence of initial trust differences and social segregation. Results are taken from multiple simulation runs that represent a large sweep of plausible behavioral condition sets.



Note: Red dots denote the average model disparity and whiskers indicate and standard deviation radius. Gray bars show the minimum and maximum values and gray markers show individual model runs.

Figure 6: Health disparity magnitudes at the end of simulation runs given the presence of different community conditions along with different forms of AI bias. The first bar, in which there is no bias in AI effectiveness across groups, is shown for comparison; it is identical to the final bar in Figure 5 and shows that initial differences in trust along with social segregation can produce health disparities. When mean accuracy bias (with one group tending to receive higher-quality treatment recommendations than the other) is added to this, at least moderate health disparities emerge under most conditions. Similarly, differential variance bias (a wider distribution of recommendation quality for one group) combined with these community conditions produces meaningful health disparities under many conditions. Combining these community conditions with skew bias (with one group more likely to receive rare, exceptionally low-quality recommendations) results in at least small health disparities under most conditions. Finally, all forms of bias combined with initial differences in health as well as social segregation results in observable health disparities in all conditions that we explored, with substantially large disparities appearing under a majority of these conditions.

Conclusions

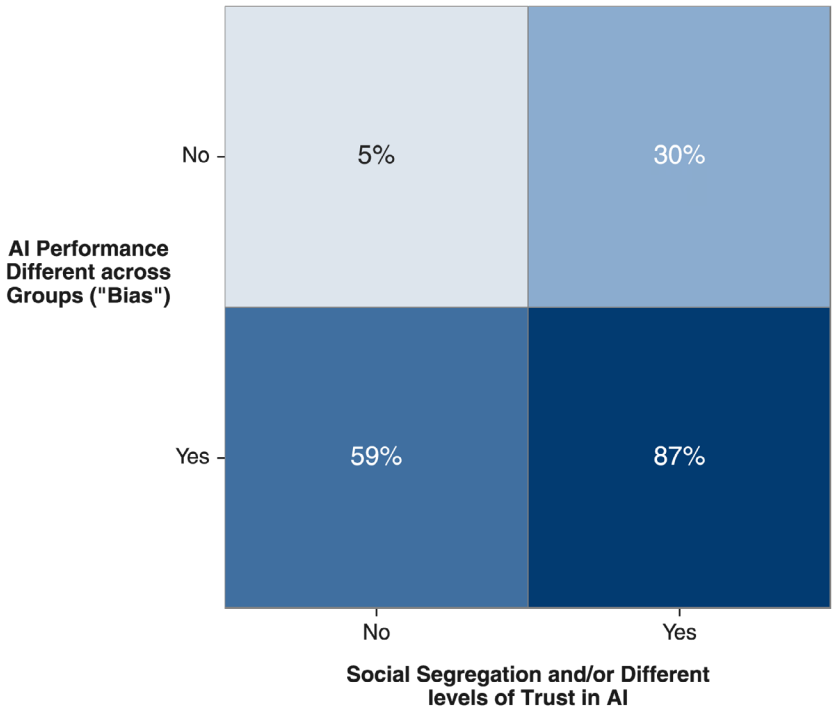
We used our ABM to conduct "computer-assisted thought experiments" to consider the possible ways in which applications of AI tools to health care might go awry. We focused our attention on two categories of "human factors" that can result in ineffective or inequitable health outcomes arising after the introduction of AI tools to health care settings: different types of AI bias and social differences between groups (i.e., systematic patterns of skepticism toward AI tools and social segregation). In addition, we grapple with whether to what extent these unintended consequences arise is dependent on a set of behavioral "known unknowns:" how treatment impacts health, trust affects treatment-seeking behavior, people form opinions about AI tools, and when people convey positive or negative perceptions of AI tools based on their experiences to others. Figure 7 summarizes the large number of simulation runs that we conducted across a wide range of experimental conditions and plausible behavioral circumstances, categorizing each simulation run by the presence or absence of any sort of AI bias or social difference between groups within a community. As anticipated, the absence of both is associated with health disparities between groups that appear only through random chance (i.e., "false positives" detected by the statistical test that we employ). As discussed above, either biased AI or social differences between groups on their own can create health disparities; here, we see that the former is much more likely to do so than the latter. However, the largest cause for concern is the simultaneous presence of both—when any type of bias is combined with either type of social difference between groups that we experimented with, health disparities are highly likely to emerge.

Although our experiments give us a first look at whether health disparities might emerge given specific AI or community conditions, they are predicated on hypo-

FIGURE 7

When Do Health Disparities Appear With AI-Based Health Care?

Proportions of simulation runs across condition categories in which we observed a statistically significant difference in health outcomes across two population groups.



Note: Statistical significance is determined using Walsh's t-test with a threshold of 95%

Figure 7: Proportion of simulation runs that ended with statistically significant health disparities between two groups among the large (but not exhaustive) range of conditions that we explored in which we varied behaviors related to AI (i.e., how people determine the extent to which they trust AI and communicate their experiences with AI to others). When AI is unbiased, there are not differences in initial trust in AI between groups, and no social segregation (tendency for people to communicate more with others in the same group), disparities occur infrequently, only emerging through random chance. However, the presence of some or all of these factors can result in health disparities, with this occurring most frequently when AI is biased in some fashion and community factors such as differential initial trust or social segregation are present.

thetical scenarios. That is, we have little "upstream" data to work with telling us how frequently AI biases might occur when applying such tools to health care and, when they do, what shape they might take. Similarly, we have little guidance on what social conditions related to the application of AI-based health care we might expect to find in real-world settings. Based on what we have seen here and what we know that we don't know, we have three sets of recommendations for policymakers, AI developers, health care workers, and researchers that can help proactively prevent unintended negative consequences of applying AI-based health care tools.

1. Consider potential pitfalls early in the process. The positive health impacts of AI tools can be constrained if the data samples used to train them don't sufficiently mirror diversity in the populations in which they will be deployed. Even tools that are effective in the aggregate can end up introducing and worsening disparities. Therefore, it is incumbent upon those creating and preparing these tools for use to carefully consider whether data used are representative across a wide variety of groups that might experience disparate impact. Categories to consider include (but are certainly not limited to) age, gender, culture, ethnicity, socioeconomic status, education, and language fluency. Given that the consequences of deleterious impact are so costly, it is also worth investing a substantial amount of time and resources to conduct pilot tests to assess performance across groups before large-scale application. Having done so, the resulting data can be used as input into an ABM like the one that we employ here to prospectively gauge the possible effects of applying the AI tool across a wide range of possible real-world settings and user behavior patterns to determine when and in what ways it might have suboptimal impact.

2. Proactively explore social contexts and behavioral factors related to AI. As our simulations show, the impact of AI tools on health outcomes can be highly dependent on social factors that are, at present, poorly understood. It is worth engaging in prospective data collection and subsequent qualitative and quantitative research to learn more about how people might think about AI tools, shape one another's perceptions, choose to engage with AI tools, and respond to guidance from these tools.

3. Continue thinking about long-term effects. The simulation model that we developed to explore how the impact of applying AI tools to health care might be moderated by social factors is a good first step on a journey that we believe should be ongoing. More advanced models that consider additional factors or respond to new data can provide further insights. For example, our ABM considers group affiliation in a very abstract way, with two groups that are differentiated but intentionally left undefined. A more realistic treatment of important social and demographic heterogeneity in future applications is almost certainly warranted. Second, our model considers activity within relatively short time horizons within which the effectiveness of AI tools does not change. However, the longer-term co-evolution of AI and human behavior is worth exploring in future work. Similarly, this initial model treats AI-based health care as a homogenous presence. In future work, it will be important to develop more sophisticated models that can provide insight into how AI-based health care might differ based on how health care professionals deploy it, how this is shaped by patient responses (including choices to embrace or avoid health care providers that utilize AI in different ways), and consider the implications of multiple competing or coordinating AI tools.

References

1. Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models. arXiv preprint arXiv:230509617. Published online 2023.
2. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. Published online 2023:1-9.
3. Benke K, Benke G. Artificial Intelligence and Big Data in Public Health. Int J Environ Res Public Health. 2018;15(12):E2796. doi:10.3390/ijerph15122796
4. Lorkowski J, Grzegorowska O, Pokorski M. Artificial Intelligence in the Health care System: An Overview. In: Pokorski M, ed. Best Practice in Health Care. Advances in Experimental Medicine and Biology. Springer International Publishing; 2021:1-10. doi:10.1007/5584_2021_620
5. Keel S, Lee PY, Scheetz J, et al. Feasibility and patient acceptability of a novel artificial intelligence-based screening model for diabetic retinopathy at endocrinology outpatient services: a pilot study. Scientific Reports (Nature Publisher Group). 2018;8:1-6. doi:10.1038/s41598-018-22612-2
6. Beede E, Baylor E, Hersch F, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ; 2020:1-12.
7. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? BMJ global health. 2018;3(4):e000798.
8. Liang W, Tadesse GA, Ho D, et al. Advances, challenges and opportunities in creating data for trustworthy AI. Nature Machine Intelligence. 2022;4(8):669-677.
9. Slota SC, Fleischmann KR, Greenberg S, et al. Good systems, bad data?: Interpretations of AI hype and failures. Proceedings of the Association for Information Science and Technology. 2020;57(1):e275.
10. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA internal medicine. 2018;178(11):1544-1547.
11. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in health care. Annual review of biomedical data science. 2021;4:123-144.
12. Obermeyer Z, Topol EJ. Artificial intelligence, bias, and patients' perspectives. Lancet. 2021;397(10289):2038. doi:10.1016/S0140-6736(21)01152-1
13. Noor P. Can we trust AI not to further embed racial bias and prejudice? BMJ. 2020;368:m363. doi:10.1136/bmj.m363
14. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-453.
15. Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. J Glob Health. 9(2):020318. doi:10.7189/jogh.09.020318
16. Panch T, Pearson-Stuttard J, Greaves F, Atun R. Artificial intelligence: opportunities and risks for public health. The Lancet Digital Health. 2019;1(1):e13-e14. doi:10.1016/S2589-7500(19)30002-0
17. DeCamp M, Tilburt JC. Why we cannot trust artificial intelligence in medicine. The Lancet Digital Health. 2019;1(8):e390. doi:10.1016/S2589-7500(19)30197-9
18. Bedué P, Fritzsche A. Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption. Journal of Enterprise Information Management. 2021;35(2):530-549. doi:10.1108/JEIM-06-2020-0233
19. von Eschenbach WJ. Transparency and the Black Box Problem: Why We Do Not Trust AI. Philos Technol. 2021;34(4):1607-1622. doi:10.1007/s13347-021-00477-0
20. Feldman RC, Aldana E, Stein K. Artificial Intelligence in the Health Care Space: How We Can

Trust What We Cannot Know. Stan L & Pol'y Rev. 2019;30:399.
21. Antes AL, Burrous S, Sisk BA, Schuelke MJ, Keune JD, DuBois JM. Exploring perceptions of health care technologies enabled by artificial intelligence: an online, scenario-based survey. BMC Medical Informatics and Decision Making. 2021;21(1):1-15. doi:10.1186/s12911-021-01586-8
22. Liyanage H, Liaw ST, Jonnagaddala J, et al. Artificial Intelligence in Primary Health Care: Perceptions, Issues, and Challenges. Yearb Med Inform. 2019;28(1):41-46. doi:10.1055/s-0039-1677901
23. Yokoi R, Eguchi Y, Fujita T, Nakayachi K. Artificial Intelligence Is Trusted Less than a Doctor in Medical Treatment Decisions: Influence of Perceived Care and Value Similarity. International Journal of Human-Computer Interaction. 2021;37(10):981-990. doi:10.1080/10447318.2020.1861763
24. Fritsch SJ, Blankenheim A, Wahl A, et al. Attitudes and perception of artificial intelligence in health care: A cross-sectional survey among patients. DIGITAL HEALTH. 2022;8:20552076221116772. doi:10.1177/20552076221116772
25. Hogarth RM, Einhorn HJ. Order effects in belief updating: The belief-adjustment model. Cognitive psychology. 1992;24(1):1-55.
26. Anderson NH. Primacy effects in personality impression formation using a generalized order effect paradigm. Journal of personality and social psychology. 1965;2(1):1.
27. Uleman JS, Kressel LM. A brief history of theory and research on impression formation. Oxford handbook of social cognition. Published online 2013:53-73.
28. Ramos M, Shao J, Reis SDS, et al. How does public opinion become extreme? Sci Rep. 2015;5(1):10032. doi:10.1038/srep10032
29. Friedkin NE, Johnsen EC. Social influence and opinions. The Journal of Mathematical Sociology. 1990;15(3-4):193-206. doi:10.1080/0022250X.1990.9990069
30. Del Vicario M, Scala A, Caldarelli G, Stanley HE, Quattrociocchi W. Modeling confirmation bias and polarization. Sci Rep. 2017;7(1):40391. doi:10.1038/srep40391
31. Deffuant GA. How can extremism prevail? A study based on the relative agreement interaction model. Published October 31, 2002. Accessed August 17, 2023. https://web-archive.southampton.ac.uk/cogprints.org/4355/1/1.html
32. Deffuant G, Neau D, Amblard F, Weisbuch G. Mixing beliefs among interacting agents. Advances in Complex Systems (ACS). 2001;(3):11. doi:10.1142/S0219525900000078
33. Kasman M, Hammond RA, Heuberger B, et al. Activating a Community: An Agent-Based Model of Romp & Chomp, a Whole-of-Community Childhood Obesity Intervention. Obesity. 2019;27(9):1494-1502. doi:https://doi.org/10.1002/oby.22553
34. Kasman M, Hammond RA, Mack-Crane A, et al. Using Agent-Based Modeling to Extrapolate Community-Wide Impact from a Stakeholder-Driven Childhood Obesity Prevention Intervention: Shape Up Under 5. Childhood Obesity. Published online May 24, 2022. doi:10.1089/chi.2021.0264
35. Kasman M, Kreuger LK. Best Practices for Systems Science Research. National Institutes of Health Office of Behavioral and Social Sciences Research; 2022. https://obssr.od.nih.gov/sites/obssr/files/inline-files/best-practices-consensus-statement-FINAL-508.pdf
36. Epstein JM. Agent-based computational models and generative social science. Complexity. 1999;4(5):41-60.
37. Hammond RA, Axelrod R. The Evolution of Ethnocentrism. Journal of Conflict Resolution. 2006;50(6):926-936. doi:10.1177/0022002706293470
38. Hammond RA. Considerations and best practices in agent-based modeling to inform policy. In: Assessing the Use of Agent-Based Models for Tobacco Regulation. National Academies Press (US); 2015.
39. Schelling TC. Dynamic models of segregation. Journal of mathematical sociology. 1971;1(2):143-186.
40. Dunbar-Jacob J, Houze MP, Kramer C, Luyster F, McCall M. Adherence to Medical Advice: Processes and Measurement. In: Steptoe A, ed. Handbook of Behavioral Medicine. Springer New York; 2010:83-95. doi:10.1007/978-0-387-09488-5_7

41. Dunbar J. Adhering to Medical Advice: A Review. International Journal of Mental Health. 1980;9(1-2):70-87. doi:10.1080/00207411.1980.11448852
42. Chambers JA, O'Carroll RE. Adherence to medical advice. Assessment in Health Psychology. Published online 2017:86.
43. Weidinger L, Mellor J, Rauh M, et al. Ethical and social risks of harm from Language Models. Published online December 8, 2021. doi:10.48550/arXiv.2112.04359

BROOKINGS

1775 Massachusetts Ave NW,
Washington, DC 20036
(202) 797-6000
www.brookings.edu