Viewpoint

Human-Centered Design to Address Biases in Artificial Intelligence

You Chen[1,2], PhD; Ellen Wright Clayton[3,4,5], MD, JD; Laurie Lovett Novak[1], PhD; Shilo Anders[1,2,6], PhD; Bradley Malin[1,2,4,7], PhD

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States
[2]Department of Computer Science, Vanderbilt University, Nashville, TN, United States
[3]Law School, Vanderbilt University, Nashville, TN, United States
[4]Center for Biomedical Ethics and Society, Vanderbilt University Medical Center, Nashville, TN, United States
[5]Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN, United States
[6]Department of Anesthesiology, Vanderbilt University Medical Center, Nashville, TN, United States
[7]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, United States

Corresponding Author:
You Chen, PhD
Department of Biomedical Informatics
Vanderbilt University Medical Center
2525 West End Ave
Nashville, TN, 37203
United States
Phone: 1 6153431939
Email: you.chen@vanderbilt.edu

Abstract

The potential of artificial intelligence (AI) to reduce health care disparities and inequities is recognized, but it can also exacerbate these issues if not implemented in an equitable manner. This perspective identifies potential biases in each stage of the AI life cycle, including data collection, annotation, machine learning model development, evaluation, deployment, operationalization, monitoring, and feedback integration. To mitigate these biases, we suggest involving a diverse group of stakeholders, using human-centered AI principles. Human-centered AI can help ensure that AI systems are designed and used in a way that benefits patients and society, which can reduce health disparities and inequities. By recognizing and addressing biases at each stage of the AI life cycle, AI can achieve its potential in health care.

(J Med Internet Res 2023;25:e43251) doi: 10.2196/43251

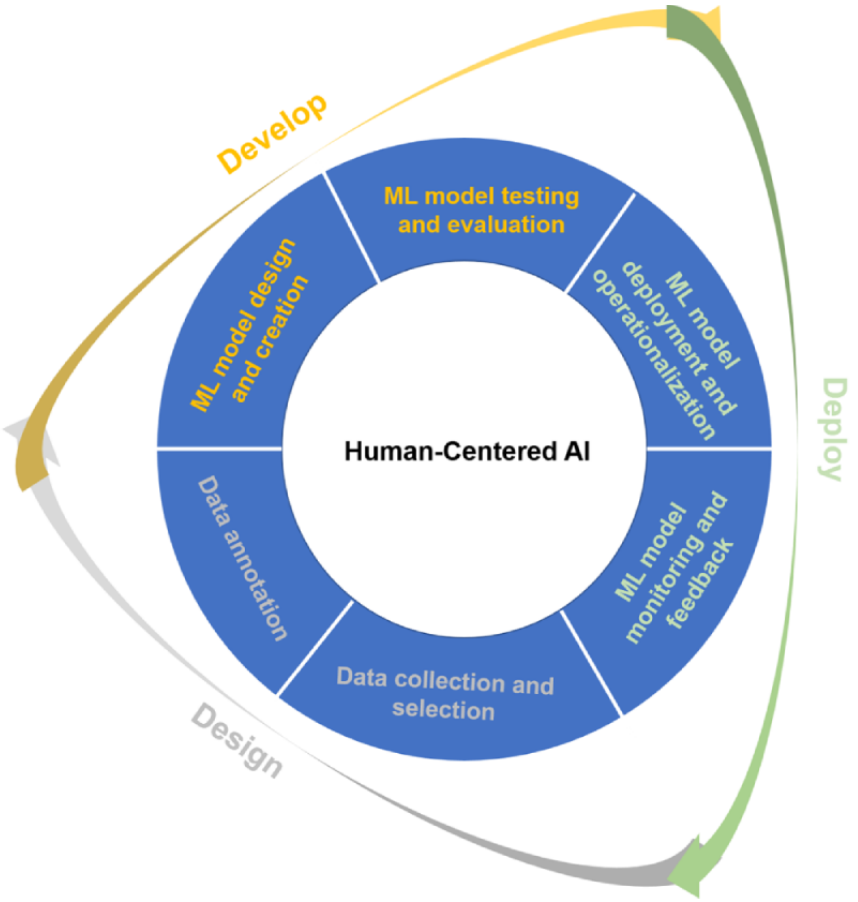
KEYWORDS

artificial intelligence; human-centered AI; biases; AI; care; biomedical; research; application; human-centered; development; design; patient; health; benefits

Introduction

Artificial intelligence (AI) promises to help health organizations deliver equitable care to their patients and optimize administrative processes [1,2]. However, the complex life cycle of AI can be biased in ways that exacerbate health disparities and inequities. As AI applications take on more central roles in biomedical research and health care [3], it is crucial to determine how best to maximize their benefits while minimizing their risks to patients and health care systems. One way to accomplish this is by involving a diverse group of stakeholders in the development and implementation of AI in health care. This perspective highlights the dual impact of AI on health disparities and inequalities; potential biases in each stage of AI design, development, and deployment life cycle; and tools for identifying and mitigating these biases. Finally, it illustrates how human-centered AI (HCAI) (Figure 1) can be applied to recognize and address the biases.

Figure 1. The artificial intelligence (AI) lifecycle has 3 primary phases: design, develop, and deploy. These phases are further partitioned into a series of stages, beginning with data collection and selection; data annotation; and proceeding through machine learning (ML) model design and creation, testing and evaluation, deployment and operationalization, and monitoring and integration of feedback loops for continuous improvement. Human-centered AI can help recognize and remediate the sources of bias that induce health disparities and inequities that can arise at each stage.


Develop
Deploy
Design
ML model testing and evaluation
ML model deployment and operationalization
ML model monitoring and feedback
Data collection and selection
Data annotation
ML model design and creation
Human-Centered AI

AI Is a Double-Edged Sword for Health Disparities and Inequities

Health disparities can stem from a variety of factors within and outside health care, such as differences in disease burden, access to health care, insurance coverage, and mortality rates. Population groups that are stratified by race and ethnicity, age, gender, socioeconomic status, geographic location, sexual orientation, gender identity, and disability can be affected differently [4]. Health disparities can result in unfair and unjust differences in health outcomes for certain groups of people, referred to as "health inequities" [4].

AI has the potential to reduce health disparities and inequities through various methods. One way is by using it to examine large amounts of data to identify patterns that may indicate a higher risk of certain health conditions in specific population groups. An example of this is using predictive modeling to identify patients with diabetes who are at risk of developing diabetic retinopathy, a serious complication that can cause blindness [5,6]. This may assist health care providers to allocate resources and develop interventions for the populations with the greatest need. Additionally, AI can be leveraged to analyze an individual patient's genetic and health data to identify the most effective treatment options [7-10]. This can help ensure that patients are getting the care that is most likely to work for them, rather than a one-size-fits-all approach.

On the contrary, AI also has the potential to amplify disparities and inequalities in health care [11-13]. This can occur because machine learning (ML) models are often trained on data from health care organizations that are already riddled with inequity, potentially creating bias in the data and the resulting recommendations. For example, ML algorithms that are designed to predict hospital mortality may be biased by the data used to train them [11-16]. In particular, these algorithms are often trained on data from electronic health records (EHRs). EHRs are designed to capture information related to patient care, such that they may include more information about patients who receive more intensive or prolonged care [14-16]. Relying on such data can create an imbalanced representation of the patient population. In addition, EHRs are typically generated by health care providers who may not always capture all relevant information about every patient [14-16]. This implies that the information recorded in EHRs is not always be complete nor is always accurate. These characteristics can create inaccurate predictions when using ML algorithms and induce negative patient outcomes.

In addition, health care providers of all types may lack the understanding, knowledge, and training required to use AI systems effectively [17], which can result in suboptimal care or unintended consequences. For instance, a provider may ignore the AI system's warning of high sepsis risk for a patient because the provider fails to comprehend how the system calculates risk, thinking that the patient is not showing any symptoms, when in reality, the patient is in the early stages of sepsis and the provider should act quickly [18]. Another example of this problem is that a provider might ignore the AI system's recommendations, thinking that the AI system is not as accurate as the provider's own judgment. This can lead to a scenario where a patient's condition is not treated appropriately, such that their health status deteriorates [19].

Finally, AI systems are often costly to develop and implement [20], such that they may not be accessible to all health care providers, particularly those serving low-income or underserved populations. This can exacerbate existing disparities and inequalities in health care access and outcomes.

AI Life Cycle and Biases

The AI life cycle refers to the process of designing, developing, and deploying AI systems, which typically includes data collection and selection; data annotation; model development and evaluation; and model deployment, monitoring, and maintenance [21,22]. It is important to note that these steps are not always linear and that the process of developing an AI system is often iterative, with feedback from one step being used to inform those that precede and follow.

Biases can exist in each step of the AI life cycle [23]. Moreover, they can be intertwined and induce a cascading effect on the final AI system performance and its potential biases. If the data used to train the model are not representative of the population, or if certain groups are underrepresented or excluded in the data, then biases are likely to exist in collection and preparation of the data. Algorithmic bias can be realized in model development if the model is not assessed for its ability to perform equally for different groups of people. Evaluation bias may transpire if evaluation metrics are not appropriate for the task or population or if the model is not tested on a diverse set of data. Additionally, bias can exist if the model is not sufficiently validated in a real-world setting or if the users of the model are not properly trained or supported. During the monitoring and maintenance phase, additional biases can arise when the model is not updated to reflect changes in the population it is being used for or if the monitoring process is not appropriate or fair.

Bias Auditing Tools

Notably, AI itself offers the potential to detect and mitigate biases in AI systems by involving open-source bias auditing tools [24]. Bias auditing tools typically involve a combination of techniques from statistics, computer science, social science, and organizational management. These tools are developed to audit the predictions of ML-based risk assessment models to understand different types of biases and make informed decisions about developing and deploying such systems. They further ensure that the ML models are appropriately trained from their inception to their completion and tested across the full diversity of patients. As illustrated in one recent study, it was shown that bias auditing tools can address inequities for race across risk models for breast cancer, renal disease, and cardiac disease [2].

Bias auditing tools typically rely on a combination of several methods to detect and analyze bias in AI systems. These methods can include fairness metrics, counterfactual analysis, sensitivity analysis, algorithmic transparency, and adversarial testing [25-27]. For example, a bias auditing tool may apply fairness metrics to spotlight potential biases in a model and then use counterfactual analysis to understand the underlying causes of the bias. After identifying the sources of bias, the tool may use sensitivity analysis to determine the factors that contribute to the bias and algorithmic transparency to understand how the model is making its predictions. Finally, the tool may use adversarial testing to identify potential weaknesses in the model. While useful in identifying potential biases in AI systems, bias auditing tools have certain limitations in detecting all forms of bias. These shortcomings may arise from various assumptions about bias and fairness, can be computationally demanding, may not provide solutions for removing the bias, may not be applicable for different forms of bias, may be difficult to interpret, and may be tested on a limited set of data. Furthermore, biases raised from AI practitioners using the AI system cannot be detected or addressed by the auditing tools. Therefore, auditing tools are clearly an incomplete solution for addressing biases in AI systems.

Human-Centered AI

As is true for all research, the first step should be an evaluation of the social and ethical merit of the project, considering the interests of the sponsors, the impact on social and organizational practices, and the impact on individuals and populations. HCAI places a strong emphasis on involving and collaborating with humans throughout the entire process of designing, developing, and implementing AI [28,29]. This emerging discipline is based on human-AI collaboration to ensure that AI operates transparently and delivers equitable outcomes. Meeting these goals requires a multidisciplinary team that includes people with a variety of expertise, including human-centered design (HCD) specialists, ethicists, social scientists, lawyers, frontline health care workers, health care managers, AI or ML practitioners, education or outreach specialists, communication scientists, and crucially patients and members of the community to ensure that AI systems are designed and used in ways that are beneficial for people and society. Below is a summary of how these different roles contribute to HCAI:

- HCD specialists play a key role in HCAI by designing and evaluating AI-based systems that are easy to use and understand by people [30-33]. They conduct foundational research on specifying the context of use and the information needs of individuals and groups, which can be translated into design requirements for AI-based systems. They also support participatory design sessions with potential users, design user interfaces, and evaluate the usability of AI systems. Finally, HCD specialists can help ensure that AI systems are accessible to people with disabilities and develop for universal access.
- Ethicists, social scientists, and lawyers advise on the ethical, social, and legal implications of AI [34]. They help organizations and governments develop responsible

practices for the use of AI and consider the impact of AI on society.

- Frontline health care workers, such as doctors and nurses, are the ones who will be using AI in the course of their work. They can provide valuable input on the design and usability of AI systems for health care and help ensure that AI systems are aligned with the needs of patients and health care professionals.
- Health care managers include executives and managers of specific clinical services. They are responsible for protecting patients and the organization by ensuring that AI tools being implemented are rigorously evaluated for appropriateness to the population served and that workflow changes are evaluated for risk and fairness.
- AI or ML practitioners are the ones who develop and apply AI systems. They need to be aware of the human-centered perspective and design AI systems with this in mind, to ensure that the systems are usable, fair, and safe for people.
- AI education or outreach specialists design and deliver education and outreach programs about AI for a variety of audiences [35-37]. They educate the public about AI, its capabilities and limitations, and how it can be used to improve people’s lives. By working with researchers, policy makers, and other stakeholders, AI education and outreach specialists can help to shape the development of AI in ways that align with the values of society and promote the responsible and ethical use of AI. This is important to ensure that AI can be used for the benefit of humanity and not to the detriment of it.
- Communication scientists play a vital role in ensuring that AI systems are developed, deployed, and communicated in ways that are aligned with the needs, values, and perspectives of the people who will be affected by them [38,39]. They can help to bridge the gap between the technical aspects of AI and the social and human aspects of its development and deployment to develop effective strategies for communicating about AI, its capabilities and limitations, and how it is being used in various domains, and to identify and mitigate potential ethical and societal implications of AI, such as issues related to privacy, bias, and fairness.
- Patients and their communities are the ultimate beneficiaries of the AI systems, and their feedback is critical to ensure that the AI systems are meeting their needs and are not causing any unintended consequences [40]. Their input, needs, and preferences should be taken into consideration throughout the development, testing, and deployment of the AI systems. Additionally, by gathering feedback from patients on the performance of the system, developers can make necessary adjustments and improvements to ensure that the system is meeting the needs of patients.

Collaboration among team members needs to be actively encouraged and supported at each stage of the AI life cycle. This can include having data scientists and AI or ML practitioners engage and work with the communities their work intends to affect to meet the distinct needs of the communities and having AI researchers or scientists collaborate with AI education or outreach specialists, policy makers, and other stakeholders to shape the development of AI in ways that align with the values of society and promote the responsible and ethical use of AI. The AI team should leverage the advantages of diversity and inclusion to create measurable and actionable debiasing strategies throughout the AI life cycle.

Mitigating Biases in Each Stage of the AI Life Cycle via HCAI

AI has the potential to magnify biases in health care due to the use of ML models that are trained on health care systems that are already unjust and unequal [41]. This raises concerns about biases in the data and the recommendations made by these models. Recognizing and mitigating biases needs to occur at each step in the AI life cycle to reduce health disparity and inequity.

Data Collection and Selection

Bias during data collection and selection refers to how the data used to train and test ML models may be unrepresentative or skewed in some way. It includes sampling, measurement, selection, confounding, and socioeconomic status bias. Sampling bias occurs when the data used to train and test an AI system are not representative of the population it will be used on. Measurement or selection bias occurs when the data are measured or selected in a way that is different for different patient groups. Confounding bias occurs when there are other factors that can impact the results but are not included. Socioeconomic bias occurs when the data are collected, measured, or selected in a way that is systematically different for patients with different socioeconomic status. The aforementioned biases can have serious consequences, such as producing AI systems that perform poorly for certain groups of patients or make decisions that are unfair or discriminatory. As a real illustration of the problem, in an ML model trained on x-rays, it was found that Black patients experienced higher levels of pain (disparities in pain) even with similar severity of osteoarthritis based on the radiographic measures of severity [42]. The ML model was trained without consideration for nonradiologic factors (eg, stress) that aggravated pain in Black patients with osteoarthritis [39]. Addressing the data bias in this study can enable the development of psychosocial interventions to address nonradiologic factors, while optimizing physical therapy, medications, and orthopedic procedures.

A multidisciplinary team can help address data collection biases by bringing an array of perspectives to creating representative training and test data and developing human-centered data evaluation strategies to ameliorate biases at the very beginning of the AI life cycle. For instance, AI or ML practitioners can apply resampling or oversampling methods to balance the data [43] and counterfactual analysis, which reveals how decisions are made by the model to ensure that the AI system does not discriminate against certain groups of patients [44,45]. Ethicists can evaluate the moral and ethical implications of the data bias and provide guidance on how to design AI systems in ways that align with societal values and ethical principles [46]. Communication scientists can develop effective strategies for communicating about the data bias, its capabilities, and limitations, and mitigate potential ethical and societal implications of data bias, such as issues related to privacy, bias, and fairness [47].

Data Annotation

Bias can be introduced during data annotation, which is typically overseen by humans who may let their prior knowledge and subjective perspectives affect their labeling processes [48,49]. Annotation bias occurs when the data used to train and test an AI system are labeled in a way that is unclear or systematically different for different patient groups. For example, the labeling of many of the data sets relied upon to train dermatological ML algorithms induced health disparities and inequities [48]. In 20 of 56 (36%) studies that developed ML algorithms for cutaneous malignant neoplasms, annotation did not satisfy gold-standard criteria for disease labeling and often failed to communicate critical information about the patients' skin tone or race [48].

When annotating data sets for AI system development, several types of biases can arise, such as cognitive, interannotator, and confirmation biases. Cognitive bias occurs when annotators' prior experiences or preconceptions influence their labeling decisions. For example, an annotator without a background in neurology may not accurately identify abnormalities in magnetic resonance imaging images due to lack of knowledge. Inter-annotator bias arises when different annotators have differing interpretations of the annotation task or expertise levels, leading to inconsistent labels. For instance, 2 annotators labeling ultrasound images of fetuses with varying levels of experience in obstetrics and gynecology may have different criteria for determining normality. Confirmation bias is similar to cognitive bias, as it occurs when annotators tend to seek out examples that confirm their preexisting beliefs, such as labeling records as compliant or noncompliant based on their personal beliefs rather than objective guidelines.

HCAI is an approach that aims to mitigate annotation biases by bringing together experts from various areas of expertise. This can be achieved by incorporating the context and background of both the annotators and the data being labeled, making the AI system more sensitive to potential biases [50-52]. To reduce bias, diversifying the annotator pool is key. By involving individuals with different backgrounds and perspectives, the data set is more likely to be labeled objectively. Additionally, implementing regular monitoring and checking of the annotation process, including interannotator agreement, can help identify and address any issues that may arise. Furthermore, providing clear guidelines for the annotation task is crucial to ensure that annotators understand the task and use consistent criteria.

ML Model Design, Creation, and Evaluation

During this phase of the life cycle, various decisions must be made by the AI team. These include how features should be engineered and selected for the data, what algorithms should be applied to train the machines, and what evaluation metrics or evaluation data should be developed. Psychologists have identified approximately 180 cognitive biases [53] that can lead to prejudiced hypotheses and inclusion biases when designing ML models. When the model architecture, selected features, algorithms, and evaluation metrics are not representative of the population, biases can manifest in various ways. For example, feature selection bias can occur when certain features are not collected for specific populations or when certain features are more prevalent in one group than another. Algorithmic bias refers to a term that transcends the technical definition of bias to encompass the broader societal meaning of prejudice and discrimination [54]. In technical terms, a large bias in an algorithm can cause the algorithm to neglect the relationship between input and output variables, whereas a large variance can lead to overfitting and poor generalization when a model is overly complex and learns almost all the data points in its training data. In societal terms, health disparities, cultural differences, and prevailing societal notions can all contribute to biased algorithms that perpetuate inequalities. For example, mental illnesses may be subject to stigmatization, leading to underreporting and underdiagnosis, particularly among marginalized groups [54]. Cultural differences in the perception and expression of symptoms, as well as language barriers, can also affect the diagnosis and treatment of mental health conditions. These factors can lead to biases in the algorithms used to diagnose and treat mental illnesses that may not accurately reflect the true prevalence and severity of these conditions. Additionally, evaluation bias can occur when the evaluation metrics or the data used to evaluate the model are not representative of the population. This can lead to inaccurate assessment of the model's performance and result in the selection of models that perform poorly for certain populations. To reduce the likelihood of these biases, it is important to ensure that the model architecture, features, algorithms, evaluation metrics, and data used are representative of the population that the model will be applied to.

HCAI has the potential to address biases in the model design, creation, and evaluation process [29,55,56]. The AI team responsible for the development of a fair and unbiased ML model should be process driven. First, they should review the features chosen for the model and ensure that they are relevant and representative of the patient population. They can also be involved in engineering new features that are relevant to the populations under study. Additionally, when designing the model architecture, the team should ensure that the model is capable of generalizing well to different groups of the population and not just performing well on the specific group of population the data used for training the model came from. They should also choose the algorithms and techniques that can provide fair and accurate diagnoses and treatments for all patients, regardless of demographic or cultural differences. To ensure that the model's performance is evaluated fairly, the team should use rigorous evaluation metrics that take into account potential biases and ensure that the model does not treat race and other aspects of social identity unfairly. Finally, the team should be involved in interpreting the model's performance and provide insights on whether the model's performance is adequate for the population under study. As an illustration of how this can be achieved, AI researchers have been partnering with people from health organizations around the world to validate and develop strategies to implement a risk assessment algorithm for breast cancer for diverse populations [57].

ML Model Deployment, Operationalization, Monitoring, and Maintenance

Integrating an AI system into a health care system is a multistep process that requires careful planning and execution. This process includes setting up the necessary infrastructure and technology to support the AI system, connecting it to other existing health technologies, training health care professionals on how to use and maintain the AI system, regularly monitoring its performance and making adjustments as needed, and ensuring compliance with all relevant regulations and guidelines for the use of AI in health care [58-62]. However, the implementation of these processes is not always straightforward. For instance, the setup of infrastructure may pose a challenge if certain hospitals or clinics lack the necessary hardware and software, which can limit access to AI-assisted care. Similarly, system integration can be a source of inefficiencies or inaccuracies if the AI system is not properly integrated with other health technologies. Local clinical personnel need to be involved in designing and implementing any workflow changes that accompany the AI tool, including training and communication. User implementation may be inconsistent if certain groups of health care professionals are not provided with adequate instruction or are not engaged in the design process. Furthermore, monitoring and maintenance may fail to detect inaccuracies or inefficiencies if certain groups of patients or health care professionals are not adequately monitored. Lastly, a body of research shows that compliance with regulations such as the Health Insurance Portability and Accountability Act of 1996 may not fully protect certain groups and their personal data when implementing AI in health care [63-65].

In addition to the aforementioned challenges, biases can arise during the integration phase, including overfitting, feedback loops, human bias and errors, and model interpretability. Overfitting occurs when a model is unable to generalize from training data to new data in the real world. Feedback loops occur when the ML model's predictions influence the data that are collected, leading to a self-fulfilling cycle of inaccurate predictions. Human bias and errors occur when decision-making and errors are introduced in the process of model deployment into busy clinical settings, operationalization, monitoring, and maintenance. Model interpretability is an issue when the model's decision-making process is not presented or accessible in a clear way to the user, which can lead to a lack of trust in the model and its predictions. Overall, current AI deployment faces significant challenges in terms of understanding how AI works in the real world.

The multidisciplinary AI team can ensure that ML models adapt to evolving data that manifest over time. It is important to note that these biases can be mitigated by careful model design, monitoring, and maintenance, and involving end users, a diverse team, and ethical considerations in the AI system's deployment, monitoring, and maintenance process. The AI team considers explanations for the predictions and the potential impact of decisions on different groups to ensure that the model's predictions used to make decisions are fair and equitable. In addition, the team can also assist in making sure that the model is able to handle new data, avoid biasing the data collection process, minimize human errors and decision-making, and make the model's decision-making process understandable and transparent. Various studies show that the HCAI framework can be used to develop explainable AI methods to make the model more interpretable and to involve domain experts in the process of interpreting the model's results and performances and providing feedback to improve the model [66,67].

Benefits of HCAI in Health Care

HCAI is a human-centered approach to designing, developing, and deploying AI systems that puts the needs and concerns of individuals at the forefront. This approach involves the participation of human stakeholders, such as patients, health care providers, health care institutions, government agencies, and insurance companies, throughout the entire process from design to deployment. By incorporating human perspectives and input throughout the AI life cycle, HCAI can help identify and mitigate biases, ensuring that the AI system is fair, ethical, and aligned with human values in health care. This is important for stakeholders as it can improve the quality of health care, increase efficiency, and reduce costs.

HCAI has multiple benefits, including the promotion of fair and unbiased care for patients, regardless of their demographics, particularly for marginalized populations who may be at a higher risk of experiencing bias in health care. HCAI can enable health care providers and patients to make decisions that are based on facts, and not on assumptions, biases, or stereotypes. This, in turn, can improve patient outcomes and reduce the risk of medical errors. In addition, it can ensure that the operations and policies of health care institutions are not discriminatory and that they provide fair and equitable care to all patients. Government agencies could further benefit from HCAI by ensuring that public health policies and interventions are not discriminatory and reach all members of the population. Insurance companies can use HCAI to ensure fair and unbiased coverage and claims processing, reducing the risk of discrimination against certain groups. In short, HCAI is important in addressing biases in AI systems because it can help ensure that AI systems are fair and equitable and that they do not perpetuate or exacerbate existing societal inequalities.

Limitations of HCAI

Still it should be recognized that HCAI has certain limitations with respect to addressing biases in AI systems. For example, it may not be possible to eliminate bias completely in data or models, and human perspectives and the input itself can be biased. Additionally, involving human stakeholders in the development process can be costly and time-consuming and requires openness to various perspectives and new collaborations, which may be difficult to achieve. The development and implementation of HCAI systems raises concerns about potential biases arising from various factors such as technical, ethical, industry, geographic, or socioeconomic. This can be observed in the different perspectives and understanding of AI among experts in different fields. For example, ML experts may have a strong understanding of the technical aspects of AI but lack an understanding of the broader societal implications or ethical considerations. On the other hand, ethicists may have a strong understanding of the ethical considerations surrounding AI, yet lack knowledge of the technical capabilities and limitations of the technology. Similarly, industry professionals may have a strong understanding of the practical applications and commercial potential of AI, yet lack knowledge of the ethical considerations and potential societal impacts of the technology. Additionally, people with different cultural backgrounds and different levels of access to resources, information, expertise, power and influence within an organization, and domain-specific knowledge may have different perspectives on the implications and potential impacts of AI. Furthermore, people from different socioeconomic backgrounds may also have different perspectives on the implications and potential impacts of AI. Although HCAI can be an effective approach to addressing biases in the AI life cycle, it is not a panacea and it is important to be aware of its limitations.

Although bias is a critical issue related to AI, it is not the only problem that needs to be addressed as AI advances and becomes increasingly prevalent in health care. Other ethical dimensions, such as data privacy and security, must also be taken into consideration. For instance, AI systems that are not developed with security in mind can be vulnerable to cyberattacks, which can compromise sensitive patient data and potentially harm patients. Moreover, ML algorithms require large amounts of data to learn and improve, and these data may contain sensitive information that needs to be safeguarded from unauthorized access and misuse.

In addition to privacy and security concerns, AI can raise ethical questions about the appropriateness of using it to make decisions that affect patients' lives, such as diagnosing illnesses or recommending treatments. This includes issues related to transparency, accountability, and the potential for unintended consequences. Furthermore, there are ongoing debates about the appropriateness of certain types of AI research, such as research involving human subjects or research that could lead to discriminatory or harmful outcomes.

Incorporating HCAI requires addressing all of these issues to ensure ethical and effective AI development in health care. This approach involves the participation of human stakeholders throughout the AI life cycle, from design to deployment, to help identify and mitigate biases, ensure transparency and accountability, and protect patient privacy and security. By prioritizing the needs and concerns of patients and health care providers, HCAI can help to ensure that AI is developed and deployed in a responsible and beneficial manner that advances the goals of health care.

Acknowledgments

This research was supported, in part, by the National Human Genome Research Institute of the National Institutes of Health under Award Number U54HG012510, the National Library of Medicine of the National Institutes of Health under Award Number R01LM012854, and the National Institutes of Health's Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity under Award Number OT2OD032581. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Authors' Contributions

The idea for this paper was initiated by YC, who also carried out the literature review and composed and drafted the manuscript. Feedback on the organization, content flow, and language was provided by EWC, LLN, SA, and BM, and the manuscript underwent critical revisions for important intellectual content with their input. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Braveman P. Health disparities and health equity: concepts and measurement. Annu Rev Public Health 2006;27:167-194. [doi: 10.1146/annurev.publhealth.27.021405.102103] [Medline: 16533114]
2. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight - reconsidering the use of race correction in clinical algorithms. N Engl J Med 2020;383(9):874-882. [doi: 10.1056/NEJMms2004740] [Medline: 32853499]
3. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2:719-731. [doi: 10.1038/s41551-018-0305-z]
4. Braveman PA, Kumanyika S, Fielding J, LaVeist T, Borrell LN, Manderscheid R, et al. Health disparities and health equity: the issue is justice. Am J Public Health 2011;101(suppl 1):S149-S155. [doi: 10.2105/ajph.2010.300062]
5. Nielsen KB, Lautrup ML, Andersen JK, Savarimuthu TR, Grauslund J. Deep learning-based algorithms in screening of diabetic retinopathy: a systematic review of diagnostic performance. Ophthalmol Retina 2019;3(4):294-304. [doi: 10.1016/j.oret.2018.10.014] [Medline: 31014679]
6. Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. Nat Med 2020;26(1):16-17. [doi: 10.1038/s41591-019-0649-2] [Medline: 31932779]
7. Dankwa-Mullan I, Bull J, Sy F. Precision medicine and health disparities: advancing the science of individualizing patient care. Am J Public Health 2015;105(suppl 3):S368. [doi: 10.2105/AJPH.2015.302755] [Medline: 26039545]

8. Griffith DM. Precision medicine approaches to health disparities research. Ethn Dis 2020;30(suppl 1):129-134 [FREE Full text] [doi: 10.18865/ed.30.S1.129] [Medline: 32269453]
9. Balogun OD, Olopade OI. Addressing health disparities in cancer with genomics. Nat Rev Genet 2021;22(10):621-622 [FREE Full text] [doi: 10.1038/s41576-021-00390-4] [Medline: 34244675]
10. Xu J, Yang P, Xue S, Sharma B, Sanchez-Martin M, Wang F, et al. Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. Hum Genet 2019;138(2):109-124 [FREE Full text] [doi: 10.1007/s00439-019-01970-5] [Medline: 30671672]
11. 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 [FREE Full text] [doi: 10.1126/science.aax2342] [Medline: 31649194]
12. Cho MK. Rising to the challenge of bias in health care AI. Nat Med 2021;27(12):2079-2081. [doi: 10.1038/s41591-021-01577-2] [Medline: 34893774]
13. O'Connor S, Booth RG. Algorithmic bias in health care: opportunities for nurses to improve equality in the age of artificial intelligence. Nurs Outlook 2022;70(6):780-782 [FREE Full text] [doi: 10.1016/j.outlook.2022.09.003] [Medline: 36396503]
14. Xie F, Chakraborty B, Ong MEH, Goldstein BA, Liu N. AutoScore: a machine learning-based automatic clinical score generator and its application to mortality prediction using electronic health records. JMIR Med Inform 2020;8(10):e21798 [FREE Full text] [doi: 10.2196/21798] [Medline: 33084589]
15. Jun E, Mulyadi AW, Choi J, Suk HI. Uncertainty-gated stochastic sequential model for EHR mortality prediction. IEEE Trans Neural Netw Learn Syst 2021;32(9):4052-4062. [doi: 10.1109/tnnls.2020.3016670]
16. Wanyan T, Honarvar H, Azad A, Ding Y, Glicksberg BS. Deep learning with heterogeneous graph embeddings for mortality prediction from electronic health records. Data Intelligence 2021;3(3):329-339. [doi: 10.1162/dint_a_00097]
17. Charow R, Jeyakumar T, Younus S, Dolatabadi E, Salhia M, Al-Mouaswas D, et al. Artificial intelligence education programs for health care professionals: scoping review. JMIR Med Educ 2021;7(4):e31043 [FREE Full text] [doi: 10.2196/31043] [Medline: 34898458]
18. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med 2018;46(4):547-553 [FREE Full text] [doi: 10.1097/CCM.0000000000002936] [Medline: 29286945]
19. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 2020;46(3):383-400 [FREE Full text] [doi: 10.1007/s00134-019-05872-y] [Medline: 31965266]
20. Desouza KC, Dawson GS, Chenok D. Designing, developing, and deploying artificial intelligence systems: lessons from and for the public sector. Bus Horiz 2020;63(2):205-213. [doi: 10.1016/j.bushor.2019.11.004]
21. Ng MY, Kapur S, Blizinsky KD, Hernandez-Boussard T. The AI life cycle: a holistic approach to creating ethical AI for health decisions. Nat Med 2022;28(11):2247-2249. [doi: 10.1038/s41591-022-01993-y] [Medline: 36163298]
22. Hwang TJ, Kesselheim AS, Vokinger KN. Lifecycle regulation of artificial intelligence- and machine learning-based software devices in medicine. JAMA 2019;322(23):2285-2286. [doi: 10.1001/jama.2019.16842] [Medline: 31755907]
23. Varona D, Suárez JL. Discrimination, bias, fairness, and trustworthy AI. Appl Sci 2022;12(12):5826. [doi: 10.3390/app12125826]
24. Courtland R. Bias detectives: the researchers striving to make algorithms fair. Nature 2018;558(7710):357-360.
25. Silberg J, Manyika J. Notes from the AI frontier: tackling bias in AI (and in humans). McKinsey Global Institute 2019 Jun:1-8.
26. Cornacchia G, Anelli VW, Biancofiore GM, Narducci F, Pomo C, Ragone A, et al. Auditing fairness under unawareness through counterfactual reasoning. Inf Process Manag 2023;60(2):103224. [doi: 10.1016/j.ipm.2022.103224]
27. Lee MSA, Singh J. The landscape and gaps in open source fairness toolkits. 2021 Presented at: CHI '21: CHI Conference on Human Factors in Computing Systems; May 8 - 13, 2021; Yokohama Japan p. 1-13. [doi: 10.1145/3411764.3445261]
28. Shneiderman B. Int J Hum Comput Interact 2020;36(6):495-504. [doi: 10.1080/10447318.2020.1741118]
29. Shneiderman B. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Trans Interact Intell Syst 2020;10(4):1-31. [doi: 10.1145/3419764]
30. Benda NC, Reale C, Ancker JS, Ribeiro J, Walsh CG, Novak LL. Purpose, process, performance: designing for appropriate trust of AI in healthcare. 2021 Presented at: Proceedings of the CHI Conference on Human Factors in Computing Systems; 2021; Yokohama, Japan p. 1-5.
31. Reale C, Novak LL, Robinson K, Simpson CL, Ribeiro JD, Franklin JC, et al. User-centered design of a machine learning intervention for suicide risk prediction in a military setting. AMIA Annu Symp Proc 2021;2020:1050-1058 [FREE Full text] [Medline: 33936481]
32. Salwei ME, Anders S, Novak L, Reale C, Slagle J, Harris J, et al. Preventing clinical deterioration in cancer outpatients: human centered design of a predictive model and response system. J Clin Oncol 2022;40(suppl 16):e13567-e13567. [doi: 10.1200/jco.2022.40.16_suppl.e13567]
33. Salwei ME, Novak LL, Vogus T, Tang LA, Anders S, Reale C, et al. Designing resilient cancer care. 2022 Presented at: Podium Abstract in AMIA; November 2022; Washington, DC.

34. Iphofen R, Kritikos M. Regulating artificial intelligence and robotics: ethics by design in a digital society. Contemp Soc Sci 2019;16(2):170-184. [doi: 10.1080/21582041.2018.1563803]
35. Russell RG, Lovett Novak L, Patel M, Garvey KV, Craig KJT, Jackson GP, et al. Competencies for the use of artificial intelligence-based tools by health care professionals. Acad Med 2023;98(3):348-356. [doi: 10.1097/ACM.0000000000004963] [Medline: 36731054]
36. Garvey KV, Thomas Craig KJ, Russell R, Novak LL, Moore D, Miller BM. Considering clinician competencies for the implementation of artificial intelligence-based tools in health care: findings from a scoping review. JMIR Med Inform 2022;10(11):e37478 [FREE Full text] [doi: 10.2196/37478] [Medline: 36318697]
37. Garvey KV, Craig KJT, Russell RG, Novak L, Moore D, Preininger AM, et al. The potential and the imperative: the gap in AI-related clinical competencies and the need to close it. Med Sci Educ 2021;31(6):2055-2060 [FREE Full text] [doi: 10.1007/s40670-021-01377-w] [Medline: 34956712]
38. Walsh CG, McKillop MM, Lee P, Harris JW, Simpson C, Novak LL. Risky business: a scoping review for communicating results of predictive models between providers and patients. JAMIA Open 2021;4(4):ooab092 [FREE Full text] [doi: 10.1093/jamiaopen/ooab092] [Medline: 34805776]
39. Bao L, Krause NM, Calice MN, Scheufele DA, Wirz CD, Brossard D, et al. Whose AI? How different publics think about AI and its social impacts. Comput Hum Behav 2022;130:107182. [doi: 10.1016/j.chb.2022.107182]
40. Nundy S, Montgomery T, Wachter RM. Promoting trust between patients and physicians in the era of artificial intelligence. JAMA 2019;322(6):497-498. [doi: 10.1001/jama.2018.20563] [Medline: 31305873]
41. Dickman SL, Himmelstein DU, Woolhandler S. Inequality and the health-care system in the USA. Lancet 2017;389(10077):1431-1441. [doi: 10.1016/S0140-6736(17)30398-7] [Medline: 28402825]
42. Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. Nat Med 2021;27:136-140. [doi: 10.1038/s41591-020-01192-7]
43. Nelson JC, Marsh T, Lumley T, Larson EB, Jackson LA, Jackson ML, Vaccine Safety Datalink Team. Validation sampling can reduce bias in health care database studies: an illustration using influenza vaccination effectiveness. J Clin Epidemiol 2013;66(suppl 8):S110-S121 [FREE Full text] [doi: 10.1016/j.jclinepi.2013.01.015] [Medline: 23849144]
44. Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nat Mach Intell 2020;2(7):369-375. [doi: 10.1038/s42256-020-0197-y]
45. Chandar P, Carterette B. Estimating clickthrough bias in the cascade model. 2018 Presented at: CIKM '18: The 27th ACM International Conference on Information and Knowledge Management; October 22 - 26, 2018; Torino, Italy p. 1587-1590. [doi: 10.1145/3269206.3269315]
46. Blumenthal-Barby J, Lang B, Dorfman N, Kaplan H, Hooper WB, Kostick-Quenet K. Research on the clinical translation of health care machine learning: ethicists experiences on lessons learned. Am J Bioeth 2022;22(5):1-3. [doi: 10.1080/15265161.2022.2059199] [Medline: 35475968]
47. Byrd N, Thompson M. Testing for implicit bias: values, psychometrics, and science communication. Wiley Interdiscip Rev Cogn Sci 2022;13(5):e1612. [doi: 10.31234/osf.io/y5nm9]
48. Daneshjou R, Smith MP, Sun MD, Rotemberg V, Zou J. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. JAMA Dermatol 2021;157(11):1362-1369 [FREE Full text] [doi: 10.1001/jamadermatol.2021.3129] [Medline: 34550305]
49. Kazimzade G, Miceli M. Biased priorities, biased outcomes: three recommendations for ethics-oriented data annotation practices. 2020 Presented at: AIES '20: AAAI/ACM Conference on AI, Ethics, and Society; February 7 - 9, 2020; New York, NY, USA p. 71. [doi: 10.1145/3375627.3375809]
50. Zhang Y, Bellamy R, Varshney K. Joint optimization of AI fairness and utility: a human-centered approach. 2020 Presented at: AIES '20: AAAI/ACM Conference on AI, Ethics, and Society; February 7 - 9, 2020; New York, NY, USA p. 400-406. [doi: 10.1145/3375627.3375862]
51. Bond RR, Mulvenna MD, Wan H, Finlay DD, Wong A, Koene A, et al. Human centered artificial intelligence: weaving UX into algorithmic decision making. 2019 Presented at: RoCHI 2019: International Conference on Human-Computer Interaction; October 17-18, 2019; Bucharest, Romania p. 2-9.
52. Monarch RM. Human-In-The-Loop Machine Learning : Active Learning and Annotation for Human-Centered AI. New York: Manning Publications Co. LLC; 2021.
53. Waldman AE. Cognitive biases, dark patterns, and the privacy paradox. Curr Opin Psychol 2020;31:105-109. [doi: 10.1016/j.copsyc.2019.08.025] [Medline: 31590106]
54. Walsh CG, Chaudhry B, Dua P, Goodman KW, Kaplan B, Kavuluru R, et al. Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence. JAMIA Open 2020;3(1):9-15 [FREE Full text] [doi: 10.1093/jamiaopen/ooz054] [Medline: 32607482]
55. Schmidt A, Giannotti F, Mackay W, Shneiderman B, Väänänen K. Artificial intelligence for humankind: a panel on how to create truly interactive and human-centered AI for the benefit of individuals and society. Cham: Springer International Publishing; 2021 Presented at: Human-Computer Interaction – INTERACT 2021: 18th IFIP TC 13 International Conference Part V; August 30 – September 3, 2021; Bari, Italy p. 335-339. [doi: 10.1007/978-3-030-85607-6_32]

56. van Stijn JJ, Neerincx MA, ten Teije A, Vethman S. Team design patterns for moral decisions in hybrid intelligent systems: a case study of bias mitigation. : CEUR-WS; 2021 Presented at: 2021 AAAI Spring Symposium on Combining Machine Learning and Knowledge Engineering, AAAI-MAKE 2021; March 22-24, 2021; Palo Alto, US p. 1-12.
57. Yala A, Mikhael PG, Strand F, Lin G, Smith K, Wan YL, et al. Toward robust mammography-based models for breast cancer risk. Sci Transl Med 2021;13(578):eaba4373. [doi: 10.1126/scitranslmed.aba4373] [Medline: 33504648]
58. Matheny M, Israni ST, Ahmed A, Whicher D. Artificial intelligence in health care. Washington, DC: National Academy of Sciences; 2019.
59. Braun M, Hummel P, Beck S, Dabrock P. Primer on an ethics of AI-based decision support systems in the clinic. J Med Ethics 2020;47(12):e3 [FREE Full text] [doi: 10.1136/medethics-2019-105860] [Medline: 32245804]
60. Juluru K, Shih HS, Keshava Murthy KN, Elnajjar P, El-Rowmeim A, Roth C, et al. Integrating Al algorithms into the clinical workflow. Radiol Artif Intell 2021;3(6):e210013 [FREE Full text] [doi: 10.1148/ryai.2021210013] [Medline: 34870216]
61. Dikici E, Bigelow M, Prevedello LM, White RD, Erdal BS. Integrating AI into radiology workflow: levels of research, production, and feedback maturity. J Med Imaging (Bellingham) 2020;7(1):016502 [FREE Full text] [doi: 10.1117/1.JMI.7.1.016502] [Medline: 32064302]
62. MacDonald S, Steven K, Trzaskowski M. Interpretable AI in healthcare: enhancing fairness, safety, and trust. In: Artificial Intelligence in Medicine. Singapore: Springer; 2022:241-258.
63. de Hond AAH, Leeuwenberg AM, Hooft L, Kant IMJ, Nijman SWJ, van Os HJA, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med 2022;5(1):2 [FREE Full text] [doi: 10.1038/s41746-021-00549-7] [Medline: 35013569]
64. Mazurek G, Małagocka K. Perception of privacy and data protection in the context of the development of artificial intelligence. J Manag Anal 2019;6(4):344-364. [doi: 10.1080/23270012.2019.1671243]
65. Osadchuk MA, Osadchuk AM, Kireeva NV, Trushin MV. Legal regulation in digital medicine. J Adv Res Law Econ 2020;11(1):148-155. [doi: 10.14505//jarle.v11.1(47).18]
66. Adadi A, Berrada M. Explainable AI for healthcare: from black box to interpretable models. In: Bhateja V, Satapathy S, Satori H, editors. Embedded Systems and Artificial Intelligence. Singapore: Springer; 2020:327-337.
67. Hanif AM, Beqiri S, Keane PA, Campbell JP. Applications of interpretability in deep learning models for ophthalmology. Curr Opin Ophthalmol 2021;32(5):452-458 [FREE Full text] [doi: 10.1097/ICU.0000000000000780] [Medline: 34231530]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
HCAI: human-centered artificial intelligence
HCD: human-centered design
ML: machine learning

Edited by T Leung; submitted 05.10.22; peer-reviewed by W Cao, B Chaudhry, JA Benítez-Andrades, A Arruda-Olson; comments to author 10.01.23; revised version received 30.01.23; accepted 28.02.23; published 24.03.23

Please cite as:
Chen Y, Clayton EW, Novak LL, Anders S, Malin B
Human-Centered Design to Address Biases in Artificial Intelligence
J Med Internet Res 2023;25:e43251
URL: https://www.jmir.org/2023/1/e43251
doi: 10.2196/43251
PMID: 36961506