No justice No peace
BLACK LIVES MATTER
</>
RESIST!
I CAN'T BREATHE
REGULATE!
BLACK LIVES MATTER
REIMAGINE!
I'M NOT A THREAT
REINFORCE!

Resist, Regulate, Reimagine, and Reinforce: How Social Workers can Advocate for Digital Inclusion

SARAH E. DILLARD

ABSTRACT

Algorithms have become more complex, creating artificial intelligence (AI) with the hope it will match human decision making. They are now being used behind-the-scenes in areas such as healthcare, housing and employment, and criminal justice. These computer formulas were created by a privileged set of individuals who often prioritized profit and growth over privacy and protection. This has led to gross injustices that have prevented marginalized communities from receiving care, finding jobs, or gaining freedom. Social workers must be able to digitally advocate for their clients. Resisting these technologies, regulating them through legislation, reimagining the role one can play, and reinforcing what is already experienced in day-to-day interactions with AI are all ways social workers can be involved in creating a world that is digitally inclusive.

Social workers are advocates for marginalized people. Clinicians, case workers, policy makers, and others all interact with populations that are considered “protected” or “sensitive” in the field of artificial intelligence (AI). These groups include, but are not limited to, women, immigrants, people of color (POC), LGBTQI+ people, those with disabilities, and those from low socioeconomic status (SES).

As AI becomes more integrated into our everyday lives, algorithms are influencing a variety of decisions previously handled by live humans: criminal risk assessment, health insurance costs, home loan eligibility, and employee resume review are just a few examples.

What can a social worker do when their client has been labeled high-risk for re-offense despite not having a history of violence? What can a social worker do when their client’s home care benefits have been cut in half? What can a social worker do if their client has high utility rates because their credit score factored in their social media activity?

This paper will cover the history of artificial intelligence, a partial overview of its current implementation and sources of bias, and four ways social workers can advocate for digital inclusion: through **resistance, regulation, reimagination, and reinforcement.** As technology is changing rapidly, the examples that follow may already have changed in the time that has passed between the writing of this article and its publication. The community organizations and movements highlighted can therefore be explored for the most recent updates.

HISTORY OF ARTIFICIAL INTELLIGENCE

An algorithm is a set of rules or calculations—like a recipe—that must be followed in order to solve a problem (Oxford Advanced Learner’s Dictionary, n.d.). In use for centuries, they were originally developed to aid in the construction of buildings, agriculture, and commerce in order to to streamline processes and create a uniform method for getting results (Ausiello, 2013). It was not until the mid-1950s that artificial intelligence became possible, a technology designed to reflect the problem-solving capabilities of humans (Anyoha, 2017). Using algorithms as the structure and data as the substance, technologists started to use AI as a substitute for human analysis and interpretation.

Once computers increased in storage and processing power, AI was able to flourish. More data could be stored on computers and in turn used to train machine learning (ML) algorithms (Anyoha, 2017). Computers became more adept at problem solving and interpreting language. As they became cheaper, more institutions became involved in research and development.

Today, we are seeing AI and big data—a term used to describe the vast amount of online information that companies are able to garner on an individual—come together (Bean, 2017). Digital footprints consisting of all the data a person has following them online are thus being used for AI machine learning in the fields of banking, healthcare, criminal justice, employment, and national security (Anyoha, 2017; Benjamin, 2019).

AREAS OF IMPACT

Technologists wrongfully assumed that a computer would eliminate bias by being based in formulas and mathematical calculations. These calculations were implemented to improve and optimize human decision-making, especially in areas such as criminal justice, in which judges were making subjective decisions (Eckhouse et al., 2019). However, the historic discrimination of marginalized groups is only reproduced and reinforced by these systems in what experts are calling “The New Jim Code” or “Coded Bias” (Benjamin, 2019; Buolamwini et al., 2018). One key example of this is an algorithm being used to determine a person’s risk of reoffending after arrest. Developers believed that by not including race as a data point, the machine would not produce racially biased results. They did, however, include people’s zip codes. Geographic data—such as the area someone lives in—is often a proxy for race due to residential segregation and redlining (Eckhouse et al., 2019). By not being familiar with this history, technologists created a system that reinforced existing biases. Accountability must be taken by companies instead of operating under the guise of expertise.

The following sections outline three areas where algorithms have greatly harmed marginalized communities and perpetuated systemic oppression.

HOUSING AND EMPLOYMENT

CREDIT SCORE AND HOMEOWNERSHIP

Algorithms have been used to determine credit score since the 1980s (Trainor, 2015). Before that, lenders would keep their own records of who they believed was "trustworthy" enough to receive a loan, often barring marginalized communities such as Black, Latinx, and Jewish identifying people from participating. Today, credit scores affect many facets of everyday life, including loan eligibility, home ownership, utility rates, and social standing.

The largest credit score companies—Experian, Equifax, and Transunion—use data such as bill payment history, employment information, and current debt to determine one's score. They also factor in child support payment history, arrest and incarceration history record, and app usage. The companies have not released information on what metrics are used to determine the weight of each category (Hao, 2020).

With the rise of big data, smaller credit score companies are beginning to use data outside the typical sources used by larger companies. This includes social media information (likes, friends, locations, and posts), the amount of time you spend on their website, and what percent of income is spent on rent given geographic location (Hurley et al., 2017). Any credit reporting agency (CRA) can request data from social media or data scraping companies (entities one can hire/pay to collect vast amounts of information from people online) in order to build their reports. Despite the Fair Credit Reporting Act of 1970 outlining what data can be used, federal legislation has not caught up with how quickly technology is changing and becoming more integrated in our everyday lives. Social media likes, for example, are not mentioned as an accepted or prohibited datum anywhere in the bill.

There is a large racial discrepancy between those with good vs. bad credit (Singletary, 2020). This directly correlates with the biased history of credit scoring and systemic oppression that is inherent in the rating. If no one is willing to provide a loan, building credit becomes significantly more difficult. Redlining of Black and Latinx neighborhoods made it impossible for families to qualify for mortgages by sanctioning these areas as "risky" for lending (Lerner, 2020). At the same time, these families pay an average of 13% more in taxes compared to white families living in homes of equivalent value.

In the US, only 45.1% of Black households own their home compared to 73.8% of white households, in part due to racist redlining (Campisi, 2021). Since credit score focuses on ownership through mortgages, the majority of Black Americans do not have this assurance to add into the algorithm. If rental payments, however, were taken into consideration, a greater swath of marginalized individuals would be able to build better credit.

The lack of homeownership in marginalized communities is perpetuated by AI in other ways. Without any privacy regulations or civil rights laws in place to regulate the use of electronic data, lenders are able to filter candidates using racial proxy data, resulting in digital discrimination and continued historic exclusion. For example, Black and Latinx individuals are charged more for home loans, amounting to an 11 to 17% additional profit for lenders (Counts, 2018). This adds up to $250-$500 million annually from Black and Latinx individuals. Despite Foggo et al. reporting that lending discrimination being on a "steady decline," the authors did not indicate how that was measured (2020). Most importantly, any lending discrimination directly impacts the potential to buy a home, one of the main ways a family can build generational wealth (DeMatteo, 2020). Owning a home means one can refinance or sell at a higher price than the home was purchased for, resulting in a profit that can be passed down to children or other dependents.

Achievements such as the Fair Housing Act of 1968—which disallowed discrimination in the buying, renting, or financing of a home—are unable to protect those they were meant to. Algorithms are often protected as trade secrets by technology companies. With no way to analyze or research the formulas being used, it is difficult to build a case that proves someone's civil rights were violated.

HIRING

AI is also being used by companies to accelerate the hiring process. Algorithms can go through thousands of applications quickly—choosing the candidates who best fit what is coded into the system as ideal (Heilweil, 2019). One of the main benefits is that more people can be considered for a position than, for example, when an individual in the HR department had to manually review resumes. However, AI is also being used for facial recognition to deduce applicants' personalities based on their expressions and appearance (Castelvecchi, 2020). Oftentimes, these photos are obtained through quick online searches of a candidate's social media platforms, such as LinkedIn or Facebook. The practice of discerning personality traits from face recognition algorithms has been proven generally inaccurate but some companies are still deploying this technology (Wells, 2020).

In addition, facial recognition technology is shown to be less accurate on dark skinned faces and women/femmes' faces (Buolamwini et al., 2018). This results in individuals from these historically marginalized, intersecting communities often registering as "non-human" to these computer systems. This will be discussed more in a later section.

Gender bias in hiring algorithms was most notably reported in 2018 when Amazon had to get rid of their system which penalized candidates who had "women's" in their application—that is, attended a women's college or were in a women's group (Vincent, 2018). According to sources at the company, this was because the algorithm was trained on existing employment information. Using the data that most of Amazon's employees are men, the algorithm decided that applications with the word "woman" or "women" should be rejected, reinforcing the pre-existing gender bias at the company. By learning from data based on existing inequities, the machine inherited a bias and thus perpetuated this unconscious preference in Silicon Valley.

HEALTHCARE

INSURANCE COSTS

Lifestyle data—the food you eat or how much you watch TV—is now readily available as industries collect information they hope to use to keep you as a customer. In addition, many insurance companies are also using this data to determine a patient’s risk of incurring high medical costs (Allen, 2019). Concerns are mounting over whether or not this data is impacting the cost a person is quoted for their monthly health insurance rate. In addition, the accuracy of the predictions is in question, as they reflect discriminatory assumptions about certain groups of people.

The Health Insurance Portability and Accountability Act (HIPAA) only covers medical information that was collected through a “covered entity,” which limits the bill's protective capabilities for health and mental health facilities. In recent years, health insurance companies such as Aetna and UnitedHealth have been collecting (either independently or through contracts) personal or lifestyle data such as social media activity, hours spent watching TV, education status, place of residence, and net worth (Allen, 2019).

By raising health insurance costs based on certain social demographics, especially static factors such as parents’ education level, marginalized communities become stuck in a cycle of poor health and poverty as the assessment is based on metrics they cannot change. In addition, by using data points that disproportionately impact POC, such as arrest records, health insurance companies perpetuate racist oppression. Thus, the algorithmic results are inherently biased.

AT-HOME CARE HOURS

The use of algorithms to make healthcare decisions is becoming more widespread as industries try to streamline processes in order to cut time and cost while also eliminating human bias. In Arkansas, a software was implemented to determine how many hours of at-home-care Medicaid patients needed (Lecher, 2018). Officials say that before this system, their assessments were done by individuals who would make decisions that favored some and were arbitrary with others.

After the algorithm, which was developed by a group of health researchers at InterAI, was implemented, many people had their hours cut—both patients receiving services and staff providing at-home assistance (Lecher, 2018). Legal Aid of Arkansas started receiving calls from individuals with complaints, some of whom were hospitalized due to lack of care.

When the president of InterAI was interviewed about transparency in the algorithm's metrics, he argued that one should trust that "a bunch of smart people determined this is the smart way to do it" (Lecher, 2018). However, during court proceedings it was revealed that the wrong calculation was being used for at least one case. This kind of error could have been caught if someone had overseen the deployment and checked all results.

POTENTIAL ILLNESSES

A risk-assessment tool used by large health systems in the United States was shown to give sick Black patients the same score it was giving to healthier white people (Obermeyer et al., 2019). Research showed that fixing this disparity would have caused an increase in Black patients who required extra care from 17.7% to 46.5%. This algorithm did not use race as one of its data points; it did, however, use insurance claims data over a certain year (information such as age and sex, insurance type, diagnosis, medications, and detailed costs). In the end, it predicted accurately what people would spend on healthcare the following year; it did not predict who was more in need of improved care due to adverse health conditions.

Proxies for race are often unknowingly used in developing algorithms, which then produce biased results. Ruha Benjamin refers to this as "coded inequality" and the entire system as "The New Jim Code" (Benjamin, 2019). Without proper knowledge of systemic racism, the individuals working for companies such as InterAI continue to perpetuate the oppression of marginalized groups while giving more power to the privileged. The notion that healthcare should be provided to an individual based on the amount of money they are able to spend furthers current racial disparities in life expectancy and benefits those with greater capital.

CRIMINAL JUSTICE

RISK ASSESSMENT

In the 1980s, lawmakers across the United States passed legislation for harsh, mandatory minimum sentencing in order to eliminate human bias in decision making (Forman, 2017). This meant an individual had to spend a certain amount of time in prison based on the crime they committed. With the crack-cocaine epidemic ravaging Black communities, substance use was further criminalized. Today, the prison industrial complex (PIC) in the U.S. has in part expanded because of this legislation as the number of people incarcerated rose from hundreds of thousands to millions over the following decades (The Sentencing Project, 2021). The need for improved criminal risk assessment therefore became present and private companies started creating algorithms in order to more accurately predict the probability of a defendant reoffending.

One of these tools, the Correctional Offender Management Profiling for Alternative Solutions (COMPAS), produces three categories of risk—low, medium, or high—and has been shown to reproduce racial disparities in its results (Angwin, 2016). Black people are twice as likely as white people to be labeled a higher risk but not reoffend. Overall, the software was shown to be accurate 61% of the time.

The biased results are not the only problem. The labels produced by the COMPAS algorithm do not correlate with a statistical chance of reoffending; they are generalizations or essentially randomly assigned numbers. Anyone can interpret the rating differently; a high score or high chance of reoffending does not correlate to a number of days, months, etc. In addition, these results are shown to judges without any explanation of the data that went into them or the formula used.

In 2016, one defendant challenged a Wisconsin court's ruling and the label produced by the risk-assessment. The judge decided that because the algorithm was not deterministic in the ruling, there was no way to prove it had such a grave influence on the decision (Eckhouse et al., 2019). However, the court failed to recognize that this decision goes against the purpose of using an algorithm—eliminating human bias—by adding the judge's input on top of the low, medium, or high result, and by not using the algorithm in a deterministic way, its objectivity (assuming they were objective, which they are not) is not being employed. At the end of the day, a judge—a human with bias—is making the decision and that decision is now being influenced by inaccurate algorithms.

In the Wisconsin case, the judge declared that since the defendant was able to see the results of the algorithm, there was nothing else that needed to be revealed (Eckhouse et al., 2019). However, the data, metrics, and formulation all impact the algorithm's output and can all be sources of bias (Miron, 2020). As stated previously, using static information (zip code at birth, last name, past criminal history) has been shown to correlate with the social factors of sensitive groups more so than dynamic information (current substance use, peer rejection, hostile behavior).

FACIAL RECOGNITION

Biometric identifications (fingerprints, voice, and iris scans) have been used by the criminal justice system for decades (Najibi, 2020). In addition, TSA's advanced imaging technology present at airport checkpoints requires agents to select one of two buttons when people enter the machine: man or woman. This means anyone who does not fit within this oppressive gender binary gets pulled aside and searched (Costanza-Chock, 2020).

Out of all the above biometric examples, facial recognition technology is being deployed across the widest variety of industries, including law enforcement, employers, manufacturers, and government housing authorities (Klosowski, 2020). In 2018, the Gender Shades study found that three different commercial algorithms were gravely inaccurate at identifying darker-skinned women, with error rates as high as 34.7% (Buolamwini et al., 2018). Compared to a 0.8% error rate for lighter-skinned males, the disparity is astonishing.

However, the impact of this bias is more frightening. In a test conducted by the ACLU of Amazon's facial recognition tool, which was available for anyone to use, the tool incorrectly identified 28 members of Congress as criminals (Snow, 2018). Black Congress members made up 40% of those matches despite only making up 20% of the House. These results reinforce the historic over-policing of the Black community and criminalization of individuals based on their skin tone.

Having more accurate facial recognition technology would not fix the problem of over-policing; in fact, it might exacerbate it. During slavery in the US, "lantern laws" were enacted in New York requiring enslaved people to carry a light by their faces in order to remain visible (Najibi, 2020). This same tracking of Black individuals could thus be done by high resolution cameras disproportionately located in certain neighborhoods which capture images and use them for databases.

DIGITAL INCLUSION: WHAT CAN SOCIAL WORKERS DO?

Even as technology expands and overtakes many human jobs, social workers are here to stay. According to a 2015 study done by NPR, mental health workers are the least likely profession to be automated by a machine (Bui, 2015). This means that for as long as AI affects our lives, there will be social workers ready and able to advocate.

According to the NASW Code of Ethics, social workers must challenge social injustice and address social problems (NASW, 2021). With technology companies often unknowingly perpetuating systemic oppression of marginalized groups through over-policing, inadequate healthcare, or discrimination, social workers have the responsibility to advocate for those targeted by these practices. The following outlines current models addressing algorithmic harm and ways social workers can be involved in mitigating the gap of algorithmic knowledge, digital inequality, and coded bias.

RESIST

There are many organizations working to ban the use of facial recognition software by police (Ozer et al., 2021). The website banfacialrecognition.com is supported by dozens of groups and they provide an interactive map marking places where facial recognition is used (Ban Facial Recognition, n.d.). This not only includes law enforcement agencies, but Amazon Ring devices as well.

It is nearly impossible today to avoid an online footprint. However, resisting the use of AI in one's everyday life is one of the main forms of not only advocacy but protection. Social workers can both inform their clients and resist these technologies in their own lives. Guidelines to follow include limiting the amount of information shared online, refusing to opt-in to monitoring services, and turning off smartphone features that group photos based on identified faces (Klosowski, 2020).

Oftentimes when signing up for an online account, websites will ask for personal identifying information (PII) such as full name, birthdate, and address. Unless absolutely needed, providing these sensitive facts about oneself can result in unwanted tracking and associations. Analytics such as cookies are another way websites use online history to filter ads and search results. They save certain types of data in order to track what individuals are clicking on, looking at, and engaging with. The social isolation this causes limits online content and can be dangerous for clients who find themselves locked into misinformation. Meanwhile, under the guise of social connection, facial recognition software—specifically in iPhones—allows users to tag their friends. However, this data is being shared beyond one's personal device. The setting must be turned off manually.

Resistance can come in many forms. Creative ways of avoiding the technology are prevalent especially in the past five years, most notably the Umbrella Movement in Hong Kong, in which protestors used open umbrellas to shield their faces from government surveillance cameras (BBC, 2019).

REGULATE

Currently, there are no federal laws in the US regulating AI. Governing bodies lack the expertise and knowledge to properly create legislation that protects privacy, limits surveillance, and bans discrimination (Pazzanese, 2020). These technologies, as outlined above, reflect the structural biases that have been present in society for centuries, and thus continue to harm marginalized communities. Privacy legislation from the 1960s-80s are now out of date. Data protection only covers government and medical databases while anti-discrimination in housing and employment does not extend to a computer formula (Bock, 2016). These policies need to be refreshed to reflect the vast growing implementation of AI.

Technology companies monitor their systems in-house and rarely provide the exact details of their algorithms for quality checks by outside researchers. They claim their system is protected by being a trade secret: intellectual property that cannot be released because it is integral to the financial well-being of the company and could put them out of business if copied (United States Patent and Trademark Office, n.d.). However, this claim prevents diverse and informed research entities from mitigating biased outputs or results which reflect historic discrimination. Due to a fear of losing profits if the company's reputation is harmed, many data-driven industries hide behind this trade secret policy, which intentionally obscures them from public review.

Social workers in policy can educate themselves on the uses of AI in a field they are experts in, healthcare, criminal justice, or another. They can write briefs on biased algorithms and the need for federal regulation as members of SAFElab at Columbia University did (Anguiano et al., 2021). Cities such as San Francisco and Boston have passed their own legislation disallowing facial recognition technology, ahead of federal changes (Associated Press, 2021).

Petitioning lawmakers to focus on AI and its potential for harm is another way social workers can get involved in advocating for digital inclusion. As stated before, with biometric systems such as facial recognition spreading surveillance, it is likely that a more accurate algorithm will be used to continue the over-policing of Black individuals. Social workers, who are educated in the historic and systemic harms done to marginalized communities, can inform those with political power the ways in which AI perpetuates this oppression.

Without regulation, technology companies will be unlikely to scrutinize their systems to the same degree as outside researchers. Limiting the uses of a product, whether by disallowing hate groups from posting on a platform or by ending data partnerships with other firms, means limiting business and therefore profit. There needs to be a monetary incentive in the form of a tax (ideally on data storage) that encourages these companies to delete digital footprints.

REIMAGINE

There are many other roles that social workers can take in advocating for digital inclusion. Technology companies are now creating jobs in fields such as research ethics and community relations and are attempting to diversify their hiring practices through apprenticeships for people with unconventional backgrounds. With an extensive understanding of systemic bias, social workers are well equipped to be a part of these discussions.

Ethical development and deployment of AI is one emerging field social workers must be a part of. Knowledge of criminal justice and healthcare is integral in decisions concerning what data should be used, whether that data is a proxy for race, and if the data results in biased outputs that harm marginalized communities. Applying this judgment and empathy will be a growing necessity as automation continues to expand (Johnson, 2021).

In research, teams improving machine learning algorithms need annotators from a wide range of backgrounds in order to capture the nuances of human expression (Johnson, 2021). By including stakeholders with varying sources of knowledge, discussions open up and opinions are provided which could not have been captured by people who mostly think the same. Time and diligence are also needed, something tech companies try to cut by paying annotators by the social media post. Working with a group means a consensus must be reached, rather than allowing one person to determine the meaning behind a post (Patton et al., 2020).

As technology companies seek to diversify their staff in order to improve the systems they create, social workers can be consultants for unbiased hiring practices. Firms such as Race Forward are employing people to look at structural oppression and find ways to eliminate it in different areas (Race Forward, n.d.). Big consulting firms are also taking on projects to create more inclusive employment searches and outreach, and socially responsible technology is an emerging field of research.

REINFORCE

Algorithmic knowledge gaps are another form of digital inequality impacting marginalized communities (Cotter et al., 2020). Understanding how personal data is used, where one may encounter bias due to AI, and ways to protect oneself are all crucial for agency in the digital world.

Socioeconomic status is viewed as the main determinant for algorithmic knowledge (Cotter et al., 2020). In the US, class often relates to one's race, as a disproportionate number of Black and Latinx individuals live below the poverty line (Creamer, 2020). However, 70% of Black people in the US use social media (Pew, 2021). This means a vast majority of Black users—given the disproportionate number of Black individuals who experience intersecting poverty—likely are not aware of the underlying algorithms, data scraping, or implications of their online presence in their physical lives.

Reinforcing base knowledge of technology—specifically AI and how it is used—is another way social workers can support digital inclusion efforts. The Algorithmic Justice League, for example, took a creative approach by creating a workshop called "Drag vs. AI" (AJL, 2020). Participants learn about facial recognition software and then learn from drag performers how to do their makeup in order to escape the machine's "coded gaze." It ends with a final runway show and additional information on how to resist, not only individually but as part of an oversight organization.

CONCLUSION

It is necessary for social workers to become advocates for digital inclusion. Technology is only progressing and becoming a greater part of our everyday lives. Currently, the AI systems being developed reflect the historic discrimination of marginalized individuals based on sensitive characteristics such as race, class, and gender. Well-versed in systemic oppression—its roots, causes, and manifestations—social workers must be involved in dismantling this latest iteration: coded inequality (Benjamin, 2019). Through resistance, regulation, reimagination, and reinforcement social workers in any position are able to advocate for those being harmed by an algorithm.

REFERENCES

Algorithmic Justice League. (2020). *Drag vs. AI*. https://www.ajl.org/drag-vs-ai

Allen, M. (2019). *Health insurers are vacuuming up details about you—and it could raise your rates.* ProPublica. https://www.propublica.org/article/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates

Anguiano, K., Darkwa, E., & Patton, D. (2021). *Recommendations to end 21st century online "stop and frisk" policing.* SAFELab. https://safelab.socialwork.columbia.edu/content/policy-0

Angwin, J. (2016). *When algorithms decide what you pay.* ProPublica. https://www.propublica.org/article/breaking-the-black-box-when-algorithms-decide-what-you-pay

Anyoha, R. (2017, August 28). *The history of artificial intelligence.* Science in the News. https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/

Associated Press. (2021). *States push back against use of facial recognition by police*. U.S. News. https://www.usnews.com/news/politics/articles/2021-05-05/states-push-back-against-use-of-facial-recognition-by-police

Ausiello, G., & Petreschi, R. (2013). *The power of algorithms: Inspiration and examples in everyday life*. Springer. https://doi.org10.1007/978-3-642-39652-6

BBC. (2019, September 28). Hong Kong protests: What is the 'umbrella movement'? *BBC*. https://www.bbc.co.uk/newsround/49862757

Bean, R. (2017). How big data is empowering AI and machine learning at scale. *MIT Sloan Management Review*. https://sloanreview.mit.edu/article/how-big-data-is-empowering-ai-and-machine-learning-at-scale/

Benjamin, R. (2019). Assessing risk, automating racism. *Science*, 366(6464), 421-422.

Bock, C. (2016). Preserve personal freedom in networked societies. *Nature*, 537(9).

Bui, Q. (2015, May 21). Will your job be done by a machine? *NPR*. https://www.npr.org/sections/money/2015/05/21/408234543/will-your-job-be-done-by-a-machine

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 1-15.

Campisi, N. (2021, June 18). *The black homeownership gap is worse. Here's what's being done*. Forbes. https://www.forbes.com/advisor/mortgages/black-homeownership-gap/

Campisi, N. (2021, February 26). *From inherent racial bias to incorrect data—the problems with current credit scoring models.* Forbes. https://www.forbes.com/advisor/credit-cards/from-inherent-racial-bias-to-incorrect-data-the-problems-with-current-credit-scoring-models/

Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need*. MIT Press.

Cotter, K., & Reisdorf, B. (2020). Algorithmic knowledge gaps: A new dimension of (digital) inequality. *International Journal of Communications*, 14, 745-765.

Council, J. (2021). Facial recognition tool used by police faces civil lawsuit in California. *Wall Street Journal*. https://www.wsj.com/articles/facial-recognition-tool-used-by-police-faces-civil-lawsuit-in-california-11615395179

Counts, L. (2018). Minority home buyers face widespread statistical lending discrimination, study finds. *Berkeley Haas School of Business*.

Creamer, J. (2020). Inequalities persist despite decline in poverty for all major race and Hispanic origin groups. *U.S. Census*. https://www.census.gov/library/stories/2020/09/poverty-rates-for-blacks-and-hispanics-reached-historic-lows-in-2019.html

DeMatteo, M. (2020). How this entrepreneur is working to help Black women build generational wealth through homeownership. *CNBC*. https://www.cnbc.com/select/what-is-generational-wealth/

Eckhouse, L., Lum, K., Conti-Cook, C., & Ciccolini, J. (2019). Layers of bias: A unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior,* 46(2), 185-209.

Foggo, V., & Villasenor, J. (2020). Algorithms, housing discrimination, and the new disparate impact rule. *Science and Technology Law Review*, 22(1), 1-62.

Forman, J. (2017). *Locking up our own: Crime and punishment in Black America.* Farrar, Straus, and Giroux.

Hao, K. (2020). The coming war on the hidden algorithms that trap people in poverty. *MIT Technology Review*. https://www.technologyreview.com/2020/12/04/1013068/algorithms-create-a-poverty-trap-lawyers-fight-back/

Heilweil, R. (2019, December 12). *Artificial intelligence will help determine if you get your next job.* Vox.vox.com/recode/2019/12/12/20993665/artificial-intelligence-ai-job-screen

Hurley, M., & Adebayo, J. (2016). Credit scoring in the era of big data. *Yale Journal of Law and Technology 18*(48).

Johnson, K. (2021, August 4). *This new way to train AI could curb online harassment.* Wired. https://www.wired.com/story/new-way-train-ai-curb-online-harassment/

Klowsowski, T. (2020). Facial recognition is everywhere. Here's what we can do about it. *The New York Times*. https://www.nytimes.com/wirecutter/blog/how-facial-recognition-works/

Lecher, C. (2018, March 21). *A healthcare algorithm started cutting care, and no one knew why.* The Verge. https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy

Lerner, M. (2020, July 23). One home, a lifetime of impact. *The Washington Post*. https://www.washingtonpost.com/business/2020/07/23/black-homeownership-gap/

Miron, M., Tolan, S., Gomez, E., & Castillo, C. (2021). Evaluating causes of algorithmic bias in juvenile criminal recidivism. *Artificial Intelligence and Law*, 29(2), 111-147.

Najibi, A. (2020). Racial discrimination in face recognition technology. *Science in the News*. https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

NASW. (n.d.) Code of ethics. *National Association of Social Workers*. https://www.socialworkers.org/About/Ethics/Code-of-Ethics/Code-of-Ethics-English

Obermeyer, Z., Powers, B., Vogell, C., & Mullainatthan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science 366*(6464), 447-453.

Oxford Advanced Learner's Dictionary. (n.d.) Definition of algorithm noun. *Oxford University Press*. https://www.oxfordlearnersdictionaries.com/us/definition/english/algorithm

Ozer, N., Ruane, K., & Cagle, M. (2021). Grassroots activists are leading the fight to stop face recognition. It's time for Congress to step up, too. *American Civil Liberties Union.* https://www.aclu.org/news/privacy-technology/grassroots-activists-are-leading-the-fight-to-stop-face-recognition-its-time-for-congress-to-step-up-too/

Patton, D., Frey, W. R., McGregor, K. A., Lee, F. T., Mckeown, K., & Moss, E. (2020). Contextual analysis of social media: The promise and challenge of eliciting context in social media posts with natural language processing. *Association for Computing Machinery*. 337-342.

Pazzanese, C. (2020). Ethical concerns mount as AI takes bigger decision-making role. *Havard Gazette*. https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/

Pew. (2021). *Social media fact sheet*. Pew Research Center. https://www.pewresearch.org/internet/fact-sheet/social-media/

Singletary, M. (2020). Credit scores are supposed to be race-neutral. But that's impossible. *Washington Post*. https://www.washingtonpost.com/business/2020/10/16/how-race-affects-your-credit-score/

Snow, J. (2018). Amazon's face recognition falsely matched 28 members of Congress with mugshots. *American Civil Liberties Union*. https://www.aclu.org/blog/privacy technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

Trainor, S. (2015, July 22). The long, twisted history of your credit score. *Time*. https://time.com/3961676/history-credit-scores/

Vincent, J. (2018, October 10). *Amazon reportedly scraps AI recruitment tool that was biased against women.* The Verge. https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report

Wells, S. (2020, May 22). *A.I. can now correctly predict something we think about privately.* Inverse. https://www.inverse.com/innovation/ai-photo-judgment

Cureus

Open Access
Editorial

DOI: 10.7759/cureus.44720

A Call to Address AI “Hallucinations” and How Healthcare Professionals Can Mitigate Their Risks

Rami Hatem [1] , Brianna Simmons [1] , Joseph E. Thornton [1]

[1]. Department of Psychiatry, University of Florida College of Medicine, Gainesville, USA

Corresponding author: Joseph E. Thornton, joethornton@ufl.edu

Review began 08/20/2023
Review ended 08/30/2023
Published 09/05/2023

Abstract

Artificial intelligence (AI) has transformed society in many ways. AI in medicine has the potential to improve medical care and reduce healthcare professional burnout but we must be cautious of a phenomenon termed "AI hallucinations" and how this term can lead to the stigmatization of AI systems and persons who experience hallucinations. We believe the term "AI misinformation" to be more appropriate and avoids contributing to stigmatization. Healthcare professionals can play an important role in AI’s integration into medicine, especially regarding mental health services, so it is important that we continue to critically evaluate AI systems as they emerge.

Categories: Medical Education, Public Health, Healthcare Technology
Keywords: ethics, misinformation, artificial intelligence, chatgpt, hallucinations, mental health, psychiatry, ai & robotics in healthcare

Editorial

Generative artificial intelligence (AI) has captivated the world by storm and revolutionized society in incomprehensible ways. AI chat boxes are designed to be easy to use, enhance our access to information, and improve our productivity as a society. While integrating AI services like ChatGPT (OpenAI, San Francisco, CA), Claude 2 (Anthropic, San Francisco), or Bard (Google, Mountain View, CA) into medicine can support mental health care and clinical decision-making, its ability to analyze large datasets and improve diagnostic accuracy and efficiency is not free of cost. Without careful consideration and monitoring by human healthcare professionals, AI algorithms, researchers, and practitioners can perpetuate existing biases, leading to unequal access to care, misdiagnoses, and inadequate treatment recommendations [1]. We call for each user of AI in healthcare to do their part in mitigating AI-generated misinformation.

What are AI hallucinations?

AI hallucinations, as defined by ChatGPT3.5 (August 16, 2023),

> "[...] refer to the generation of content that is not based on real or existing data but is instead produced by a machine learning model's extrapolation or creative interpretation of its training data. These hallucinations can manifest in various forms, such as images, text, sounds, or even video. AI hallucinations occur when a machine learning model, particularly deep learning models like generative models, tries to generate content that goes beyond what is has learned from its training data. These models learn patterns and correlations from the data they are trained on and attempt to produce new content based on those patterns. However, in some cases, they can generate content that seems plausible but is actually a blend of various learned elements, resulting in content that might not make sense or could even be surreal, dream-like, or fantastical."

This can have consequences in healthcare as we begin to embrace AI as a tool. If a healthcare professional is unaware of AI’s limitations (i.e. AI hallucinations), they may inadvertently cause harm to patients due to inaccurate claims.

In an editorial written by Dr. Hussam Alkaissi and Dr. Samy McFarlane to Cureus, they highlight several instances of AI hallucinations and how this can have implications in healthcare [2]. For example, when tasked to provide information on homocystinuria-associated osteoporosis and, on a separate occasion, late-onset Pompe disease, the AI provided a thorough paper with several citations with PubMed IDs. However, after fact-checking, it was found that the provided paper titles were fabricated and the PubMed IDs were associated with other papers [2]. Moreover, in an interview on CBS News’ “60 Minutes”, Google’s developers give their take on the future of AI and AI hallucinations. When Google’s Bard was tasked with writing an essay and recommending five books, it produced five seemingly realistic titles. However, the book titles were nothing but figments of an imaginative realm. They referred to these “hallucinations” as “errors with confidence” [3].

Alternative to the terminology of AI hallucinations

Initially, we proposed the term "AI confabulations" in a letter responding to Dr. Teva Brender’s Hey Chatbot, Write Me an H&P. In his reply, he agreed with the sentiment of abandoning the term "AI hallucinations", however, he thoughtfully noted that the word "confabulations" would arguably animate or attribute lifelike properties, further raising concerns about sentience [4-5]. We appreciate Dr. Brender for his comments and commitment to not perpetuating harmful stereotypes. We offer the alternative term "AI misinformation" as we believe it to be more accurate and in line with a shared vocabulary that individuals without AI knowledge can understand. Moreover, this term refrains from stigmatizing a community and avoids attributing lifelike characteristics to AI.

ChatGPT’s Response

The following is an excerpt from a ChatGPT3.5 conversation.

> "Q: What would be an accurate word to describe when a large language model (LLM) generates false information?
>
> A: An accurate word to describe when a large language model (LLM) generates false information is "misinformation." Misinformation refers to the dissemination of false or incorrect information, whether intentionally or unintentionally. If the false information is spread with the intention to deceive, the term "disinformation" could also be used."

How can healthcare professionals mitigate AI misinformation?

When healthcare professionals incorporate AI into their services, they must exercise caution and recognize that AI alone cannot replace the individualized care that professionals offer. Even ChatGPT stresses the importance of “the human element of healthcare” and recognizes AI as a valuable tool in healthcare, not a replacement.

> "AI can complement healthcare professionals by handling routine tasks, assisting in diagnosis, and improving treatment plans. However, the human element of healthcare-empathy, judgment, ethical decision-making, and adaptability-remains irreplaceable. The optimal approach is likely to involve a synergy between AI and human healthcare providers, with AI supporting and enhancing the work of professionals while leaving the core aspects of patient care in the hands of trained experts." - ChatGPT 3.5, August 16, 2023

To minimize the AI misinformation phenomenon, healthcare professionals should be aware of AI’s limitations and latest advancements. Some universities and hospitals provide free online continuing medical education (CME) courses, which now expand to include the scope of AI in healthcare. In clinical settings, AI-generated responses should be verified with reliable peer-reviewed medical sources. If any instances of AI misinformation are noted, they should be reported to the appropriate entity to ensure that these inaccuracies can be corrected.

Large language models (LLM) such as ChatGPT 3.5 are only as accurate as the information provided to them; therefore, clinicians and AI scientists need to work together to continuously improve the data available to these systems. AI misinformation can be further mitigated by advocating for more diverse and representative datasets that would provide more generalizable data for these LLMs. While embracing the potential impact of AI in healthcare, it is important that we reiterate the importance of collaborating with patients and other colleagues to continue to bolster shared decision-making.

Conclusion

We continue to believe the term "AI hallucination" is inaccurate and stigmatizing to both AI systems and individuals who experience hallucinations. Because of this, we suggest the alternative term "AI misinformation" as we feel this is an appropriate term to describe the phenomenon at hand without attributing lifelike characteristics to AI. As healthcare professionals begin to explore AI for clinical use, it is important we use it responsibly to ensure we do no harm.

Additional Information

Disclosures

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

Some of the text within the article has been generated with the help of ChatGPT 3.5.

References

1. Rajpurkar P, Chen E, Banerjee O, Topol EJ: AI in health and medicine. Nat Med. 2022, 28:31-8. 10.1038/s41591-021-01614-0
2. Alkaissi H, McFarlane SI: Artificial hallucinations in ChatGPT: implications in scientific writing . Cureus. 2023, 15:e35179. 10.7759/cureus.35179
3. Is artificial intelligence advancing too quickly? what AI leaders at Google say. (2023). Accessed: July 28, 2023: https://www.cbsnews.com/news/google-artificial-intelligence-future-60-minutes-transcript-2023-04-16/.
4. Hatem R, Simmons B, Thornton JE: Chatbot confabulations are not hallucinations [IN PRESS]. JAMA Intern Med. 2023, 10.1001/jamainternmed.2023.4231
5. Brender TD: Chatbot confabulations are not hallucinations-Reply [IN PRESS] . JAMA Intern Med. 2023, 10.1001/jamainternmed.2023.3875