Shrotriya S, et al. Navigating the Path to Ethical and Responsible AI Integration in Health and Life Sciences with Human and Machine Collaboration. *Journal of the Society for Clinical Data Management*. 2025; 5(1): 5, pp. 1–12. DOI: https://doi.org/10.47912/jscdm.412

REVIEW ARTICLE

Navigating the Path to Ethical and Responsible AI Integration in Health and Life Sciences with Human and Machine Collaboration

Shobhit Shrotriya*, P. K. Nizar Banu†, Avi Kulkarni‡ and Sujata Aiyangar§

Artificial Intelligence (AI) is bringing a transformative change to the way businesses strategize, operate, and work. Human and Machine working together in tandem to deliver critical business outcomes has become an expectation. The application of AI in the synergistic fields of Health and Life Sciences is changing the way clinical research is being done, with medical practitioners adopting AI-based solutions for the diagnosis of certain diseases, such as various types of cancer and clinicians adopting AI-based solutions in drug development and several downstream processes. As an example, the application of AI in radiology and imaging has yielded positive outcomes, however, it has been observed that these algorithms may acquire unwanted biases through the training datasets, which may lead to inaccurate diagnosis and potentially flawed care recommendations. These biases may prove to be fatal and hence it is imperative to institutionalize guidelines and governance for ethical and responsible use of AI. This paper is an attempt to understand the nuances of these biases and to build a sustainable framework that will help enable and scale the implementation of Responsible AI in the fields of Health and Life Sciences.

Keywords: Artificial Intelligence (AI); Generative AI (GenAI); Ethical AI; Responsible AI; datasets

1. Introduction

According to a Gartner report,[1] 80% of enterprises will have incorporated AI by 2026. MarketsandMarkets estimates that AI in health care is projected to grow from $14.6 billion in 2023 to $102.7 billion by 2028; it is expected to grow at a compound annual growth rate of 47.6% during the forecasted period.[2] AI, if used responsibly and wisely, is predicted to create a better future for society at large and would benefit industries, businesses, and governments. The people in the top positions in industry have realized the immense potential that AI offers. Humans and Machines working in synergy can transform the way work is being performed today across the spectrum of industry. The use of AI in addressing complex business problems has motivated organizations to incubate, scale, and industrialize its implementation.[15–20] Successful implementations and uses of AI in various forms and shapes, such as diagnosis of various types of cancer, have greatly benefitted those areas of the health and life sciences that have deployed them. However, at the same time, it is essential to ensure that the outcomes delivered by AI can be trusted, are unbiased, and do not pose risks that create adverse impacts. The journey to the ethical and responsible use of AI is a critical one and it is imperative to carefully strategize the design and development of AI algorithms that minimize the risks and challenges associated with it.

Understanding the concept of 'Ethics' is important. Ethics is defined as a set of moral principles to govern the behaviors or actions of an individual or a group of individuals.[3] When this definition of ethics is extended to AI, it relates to the moral obligations and duties of both the AI itself, and of its developers. Ethics of AI and Ethical AI are terms that are used interchangeably but there are certain nuances to these concepts that are depicted in **Table 1** below.

Table 1: The AI world of ethics.

	AI	Human	Society
Ethics of AI	Principles of developing AI to interact with other AIs ethically	Principles of developing AI to interact with Humans ethically	Principles of developing AI to function ethically in society
Ethical AI	How should AI interact with other AIs ethically?	How should AI interact with Humans ethically?	How should AI operate ethically in society?

* Department of Computer Science, CHRIST (Deemed to be University), Bangalore and Accenture, IN

† Department of Computer Science, CHRIST (Deemed to be University), Bangalore, IN

‡ ThoughtSphere, California, US

§ Accenture, IN

Corresponding author: Shobhit Shrotriya (shobhit.shrotriya@accenture.com)

2. Discussion

A. Understanding relation of humane-ness, human centeredness, human dignity in the application of AI

Before deep diving into concept of Responsible AI (RAI), it is crucial to understand the relationship between humane-ness, human-centeredness, human dignity, and the application of AI.[4–6] Below is a summary of each concept as it relates to AI that also shows their interrelated qualities.

Humane-ness

Humane-ness refers to the quality of being compassionate, kind, and considerate towards humans and their needs. In the context of AI, humane-ness implies developing and deploying AI technologies in a manner that prioritizes human well-being, safety, and dignity.

Humane AI seeks to minimize harm, promote fairness, and uphold ethical principles in the design, development, and deployment of AI systems.

Human-centeredness

Human-centeredness in AI emphasizes the importance of designing systems that prioritize human values, preferences, and experiences. It involves involving end-users in the design process, understanding their needs, and incorporating feedback to create AI technologies that align with human goals and aspirations. Human-centered AI aims to enhance human capabilities, improve user experience, and foster trust between humans and AI systems.

Human dignity

Human dignity is the inherent and inviolable worth and value of every human being. It encompasses principles such as autonomy, respect, fairness, and non-discrimination. In the context of AI, respecting human dignity means ensuring that AI systems do not infringe upon human rights, privacy, or autonomy. AI technologies should be developed and deployed in a manner that upholds human dignity and preserves human agency.

There is interconnectedness between application of AI and humane-ness, human-centeredness, human dignity. Humane-ness and human-centeredness guide the development and deployment of AI technologies to prioritize human interests, needs, and values. Human dignity serves as a foundational principle that underpins ethical AI practices, ensuring that AI systems respect and protect the rights, freedoms, and dignity of individuals. By integrating humane-ness, human-centeredness, and respect for human dignity into AI development processes, AI technologies can be built that are ethically sound, socially responsible, and beneficial to humanity.

B. Designing and operationalizing humane-ness values and ethical AI principles

The principles described above are essential to ensure that AI technologies align with societal values, respect human rights, and promote human well-being.[4] Steps to consider in this process are detailed below.

Identify Core Humane-ness Values

Begin by identifying the core humane-ness values that should guide the development and use of AI technologies. These may include principles such as fairness, transparency, accountability, privacy, inclusivity, and safety.

Develop Ethical AI Principles

Based on the identified humane-ness values, develop a set of ethical AI principles that reflect these values and serve as guiding frameworks for AI development, deployment, and use. Ethical AI principles should be comprehensive, clear, and actionable, providing guidance to developers, researchers, policymakers, and other stakeholders involved in AI ecosystems.

Embed Ethical Considerations in AI Design

Incorporate ethical considerations into the design process of AI systems from the outset. This involves integrating mechanisms for fairness, transparency, and accountability into algorithmic decision-making processes. Use techniques such as fairness-aware machine learning, transparent model documentation, and algorithmic impact assessments to address ethical concerns and biases in AI systems.

Promote Human-Centered Design

Adopt a human-centered design approach that places human needs, preferences, and experiences at the forefront of AI development. Engage with diverse stakeholders—including end-users, domain experts, ethicists, and impacted communities—to gather insights, feedback, and perspectives throughout the design process.

Establish Governance and Oversight Mechanisms

Establish governance structures and oversight mechanisms to ensure compliance with ethical AI principles and regulatory requirements. This may involve creating AI ethics committees, regulatory bodies, or industry standards organizations tasked with reviewing, monitoring, and enforcing ethical guidelines and best practices.

Promote Transparency and Accountability

Foster transparency and accountability in AI systems by ensuring that decision-making processes and outcomes are explainable, interpretable, and auditable. Implement mechanisms for documenting, tracing, and validating AI algorithms, data sources, and decision-making logic to enhance transparency and accountability.

Educate and Train Stakeholders

Provide education and training programs to stakeholders involved in AI development, deployment, and use. This includes developers, data scientists, policymakers, ethicists, and end-users, to raise awareness of ethical considerations, promote responsible AI practices, and foster a culture of ethical decision-making.

Iterate and Adapt

Continuously iterate and adapt ethical AI principles and practices in response to evolving technological advancements, societal needs, and ethical challenges. Regularly assess and reassess the impact of AI technologies on individuals, communities, and society at large, and adjust ethical frameworks accordingly.

By systematically integrating humane-ness values and ethical principles into the design and operation of AI systems, we can foster the development of AI technologies that are trustworthy, responsible, and beneficial to humanity.

C. Implementing responsible AI in health care

Implementing responsible AI in health and life sciences requires careful consideration of ethical, regulatory, and practical factors to ensure that AI technologies are used safely, ethically, and effectively.[7,8] Approaches to implement responsible AI are suggested below.

Ethical Guidelines and Frameworks

Develop and adhere to ethical guidelines and frameworks that govern the design, development, and deployment of AI technologies in health care. Ensure that AI systems prioritize patient safety, privacy, autonomy, and well-being while adhering to principles such as transparency, accountability, fairness, and non-discrimination.

Regulatory Compliance

Ensure compliance with health care regulations, data protection laws (such as HIPAA in the United States), and industry standards governing the use of AI in health care and drug development. Stay abreast of evolving regulatory requirements and ensure that AI systems adhere to legal and ethical standards to protect patient rights and confidentiality.

Clinical Validation and Evidence-Based Practice

Ensure humans are in the loop to validate AI algorithms and models through rigorous testing and validation to ensure their accuracy, reliability, and safety and to prevent any data leakage in both clinical trial and real-world health care settings. Emphasize evidence-based practice and ensure that AI-driven diagnostic, predictive, and treatment recommendations are based on sound scientific evidence and clinical guidelines.

Transparency and Explainability

Promote transparency and explainability in AI-driven solutions and systems by providing clear explanations of how AI algorithms work, how decisions are made, and how recommendations are generated. Ensure that there is an awareness amongst both practitioners and patients about the AI-driven solutions and systems being used.

Data Privacy and Security

Implement robust data privacy and security measures to protect patient health information and sensitive medical data from unauthorized access, breaches, and misuse.

Adhere to data protection regulations and industry best practices for data encryption, access controls, data anonymization, and secure data storage and transmission.

Bias Detection and Mitigation

Identify and mitigate biases in AI algorithms and datasets to ensure fair and equitable outcomes for diverse patient populations. Implement bias detection techniques, diverse training approaches, and algorithmic audits to address biases related to race, ethnicity, gender, socioeconomic status, and other factors.

Clinical Decision Support and Human Oversight

Use AI-driven clinical decision support systems to assist health care and clinical professionals in diagnosis, treatment planning, and patient management. Maintain human oversight and accountability by ensuring that AI recommendations are reviewed and validated by trained professionals before clinical or medical use.

Continuous Monitoring and Evaluation

Continuously monitor, evaluate and report the performance, safety, and impact of AI-driven solutions and systems in clinical practice. Collect feedback from health care providers, clinicians, patients, and other stakeholders to identify potential issues, improve system usability, and optimize clinical outcomes over time. Institutionalize a structured governance and escalation process to manage risk and ensure transparency at all levels.

By implementing these approaches, health care organizations can harness the potential of AI technologies to improve patient care, enhance clinical decision-making, and advance medical research while upholding ethical standards and protecting patient rights.

D. Deep dive into Responsible AI (RAI) – Understanding RAI

Responsible AI is the practice of using AI with good intentions to empower employees and businesses, and fairly impact customers and societies – allowing companies to engender trust and scale AI with confidence.[9] The consequences of AI bias can be severe, for example, an AI model diagnosing malignancy as non-malignancy or vice-versa. It is therefore imperative that AI is implemented responsibly to meet consumer expectations and remain relevant. Laws have not kept up with technology, so companies must recognize that they need to do more than avoiding illegal activities.

Managing AI and AI Bias requires soft skills (improving management and governance of AI), technical solutions (actions for engineers to reference during development), and thorough understanding of associated data together with business context. Neutral outcomes, data reliability, and unbiased AI are mandatory success criteria in the implementation of AI.

Businesses should use a human-centered approach to achieve RAI, based on a transparent, reliable, understandable, sustainable, and trainable (TRUST) framework that will enable them to shape their key priorities, implement governance strategy, build systems, and drive their businesses to success. **Figure 1** summarizes the TRUST concept.[9]

The TRUST equation is an amalgamation of attributes that drives openness and transparency across the entire journey of building processes and guidelines for the ethical and responsible use of AI. The primary intent is the 'transparent' use of technological tools to detect bias. There needs to be a commitment to eliminate biases through the implementation of robust and 'reliable' governance; there must also be the ability to open the 'black box' of AI models to make the algorithms, and their inputs and outputs, understandable. There is a requirement for consistent, scalable and manageable guardrails to make AI systems secure and 'sustainable'. Finally, a 'trainable' model for human and machine interaction needs to be developed.

E. Challenges during the RAI implementation journey in Health and Life Sciences

There are significant challenges in the implementation of RAI. A few of these are listed below.

Regulatory and legal issues[10]

Violation of anti-discrimination laws. An AI algorithm trained with biased datasets which may result in different treatment options for different racial or ethnic groups, could potentially violate anti-discrimination laws.

Uninformed consent. When patients are not informed that an AI model may be less accurate for their demographic group due to biases in the training data, they could argue that they could not provide fully informed consent for their treatment.

HIPAA violations. If bias in AI models leads to inappropriate treatment recommendations, this could potentially lead to a breach of the Health Insurance Portability and Accountability Act (HIPAA) in the US.

Product recalls. The US Food and Drug Administration (FDA) regulates medical devices, which include certain types of AI models. A biased AI model that may lead to harm, could result in the model being recalled or its FDA approval being revoked.

Medical harm. An incorrect treatment recommendation due to a biased AI may lead to harm to a patient.

Liability complications. Developers of a biased AI model, who were aware of the biases and did not take sufficient steps to address them, increase liability.

Data Infrastructure and Quality

The need to be able to access large sets of structured and highly standardized data from disparate systems across the clinical trial value chain and constraints attached to patient medical information/records.

Technical

Clinical platforms have legacy design constraints that pose challenges to integration on new AI models. Interoperability with multiple systems and platforms that are usually highly customized and validation is required for every upgrade.

People and Process

Data stewardship is needed, including clearly defined processes on how to treat, govern, and leverage data to make decisions. Poor data science strategies across the value chain that are designed only to derive maximum return on investment on AI products and services present a challenge to the implementation of RAI.

AI implementation requires specialized resources with skills and knowledge training needed to align the implementation with the regulatory pathways embedded across the IT landscapes. End- users also require training on the best use of the system. There also needs to be acceptance by the end-users of AI-based clinical and/or medical decisions.

F. The journey to RAI and Trusted AI (TAI) – Establishing a future-looking framework

To establish trust, RAI drivers should be woven into the AI architecture and operating capabilities. Three important attributes embed responsibility and establish trust.[9]

Govern

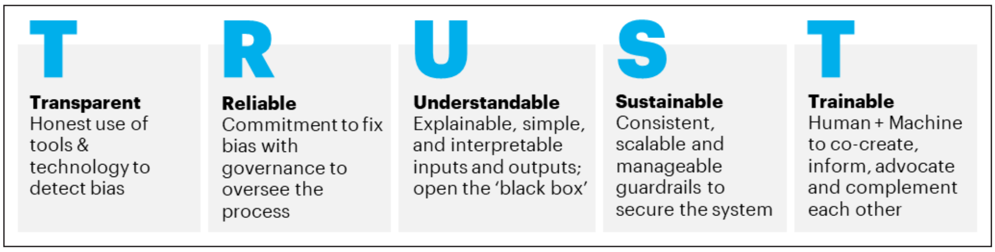
Create an internal governance framework and processes that are anchored to industry and societal shared values, regulations, ethical guardrails, and accountability. Promote clarity around decisions.

Design

Design and deploy AI with TRUST principles (e.g., privacy, transparency and security) by designing and building systems that lead to "explainable" AI. Empower project teams to understand and address issues of bias.

Monitor

Monitor and audit the performance of AI against key value-driven metrics, including algorithmic accountability, bias, and cybersecurity.



Transparent Honest use of tools & technology to detect bias

Reliable Commitment to fix bias with governance to oversee the process

Understandable Explainable, simple, and interpretable inputs and outputs; open the 'black box'

Sustainable Consistent, scalable and manageable guardrails to secure the system

Trainable Human + Machine to co-create, inform, advocate and complement each other

Figure 1: The TRUST Equation.

Combined with these attributes are the five RAI pillars depicted in **Figure 2**.[11,12]

Fairness

AI bears the risk of amplifying human bias, potentially resulting in unfair and unintended treatments that may put the entire solution in jeopardy.

Transparency

Given that adoption is directly linked to trust, it is imperative to be transparent about the use and decision-making agency of AI.

Data Stewardship

Given AI's unique reliance on large data sets, a comprehensive approach to data stewardship is required.

Accountability

Given its novelty, the potential risks associated with AI place increased pressure on organizations to properly govern and self-regulate responsible AI programs.

Community

Embracing human and machine interaction through talent sourcing, education, and empowerment will be critical in creating an AI community.

While assessing the five pillars of RAI, an important aspect to consider simultaneously is the application of **'FAREA'** methodology.[13]

Is the algorithm **fair**? Ensure the artificial intelligence tool cannot be used to propagate virtual discrimination.

Is the algorithm **auditable**? Do you fully understand what the AI tool is meant to do and how it operates? What data is it collecting?

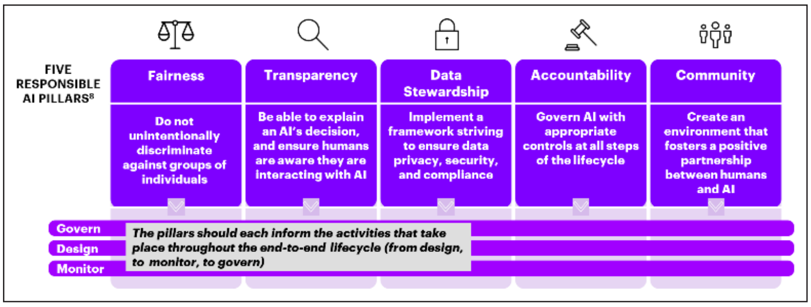
Who is **responsible** for the algorithm? Who is responsible for the digital decisions the algorithm makes? Who is responsible for auditing the algorithm?

Is the algorithm **explainable**? Are digital decisions and any data necessary available and explainable to end users and stakeholders?

Is the algorithm **accurate**? Have you tested the algorithm with test data? Do you know how the tool will react to novel situations?

Imagine implementing an industry-agnostic Trusted AI (TAI) framework from startup to advanced compliance monitoring, which allows companies to embrace a superpowered workforce without making costly missteps as a result of bias.[9] **Figure 3** shows such a framework.[14]

TAI would reside in a 'data lake' and would be powered by a human-machine operating engine to orchestrate the optimal synergy of data, applied intelligence and digital technologies to power the workforce for transforming business operations.



FIVE RESPONSIBLE AI PILLARS[8]

Fairness
Do not unintentionally discriminate against groups of individuals

Transparency
Be able to explain an AI's decision, and ensure humans are aware they are interacting with AI

Data Stewardship
Implement a framework striving to ensure data privacy, security, and compliance

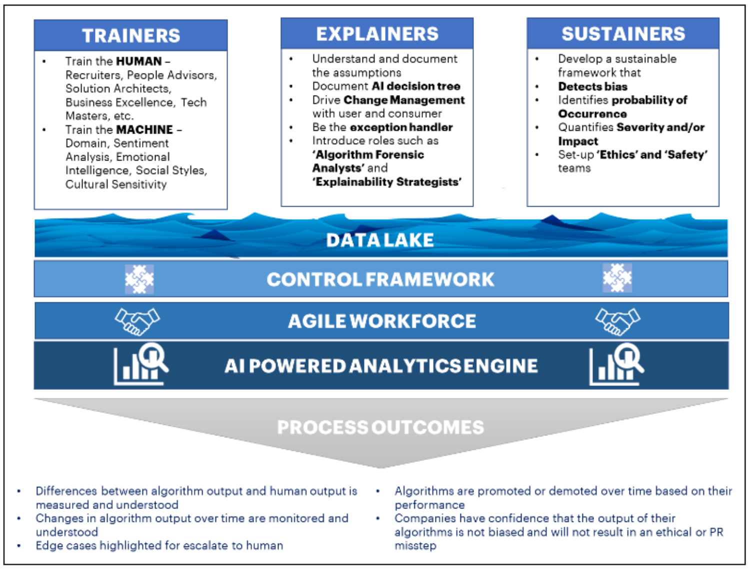
Accountability
Govern AI with appropriate controls at all steps of the lifecycle

Community
Create an environment that fosters a positive partnership between humans and AI

Govern
Design
Monitor

The pillars should each inform the activities that take place throughout the end-to-end lifecycle (from design, to monitor, to govern)

Figure 2: The RAI Pillars.



TRAINERS
- Train the **HUMAN** – Recruiters, People Advisors, Solution Architects, Business Excellence, Tech Masters, etc.
- Train the **MACHINE** – Domain, Sentiment Analysis, Emotional Intelligence, Social Styles, Cultural Sensitivity

EXPLAINERS
- Understand and document the assumptions
- Document **AI decision tree**
- Drive **Change Management** with user and consumer
- Be the **exception handler**
- Introduce roles such as **'Algorithm Forensic Analysts'** and **'Explainability Strategists'**

SUSTAINERS
- Develop a sustainable framework that
- **Detects bias**
- Identifies **probability of Occurrence**
- Quantifies **Severity and/or Impact**
- Set-up **'Ethics' and 'Safety'** teams

DATA LAKE

CONTROL FRAMEWORK

AGILE WORKFORCE

AI POWERED ANALYTICS ENGINE

PROCESS OUTCOMES

- Differences between algorithm output and human output is measured and understood
- Changes in algorithm output over time are monitored and understood
- Edge cases highlighted for escalate to human

- Algorithms are promoted or demoted over time based on their performance
- Companies have confidence that the output of their algorithms is not biased and will not result in an ethical or PR misstep

Figure 3: The TAI Framework.

The TAI framework would be driven by 'three key' components of Trainers, Explainers and Sustainers.[14]

The '**Trainers**' component would be a combination of human and machine. The human sub-component will involve a variety of individuals that perform various roles in the organization, such as recruiters, people advisors, solution architects, technologists, etc. who will form the link to the machine element, which will be trained by a set of AI algorithms on numerous dimensions, such as domain, sentiment, culture, emotional intelligence, social styles, etc.

The second component, the '**Explainers**' would include those people that understand and document the AI decision tree, drive change management, and be exception handlers.

The third and most critical component, the '**Sustainers**' would be responsible for testing and detecting biases in existing or new algorithms, and for the continuous monitoring of implemented algorithms for bias or unexpected results.

The TAI framework would be powered by a scorecard that would help to identify probability of occurrence of a particular bias (or a set of biases), to quantify the severity and/or impact of the bias, and finally to assign a score to a bias to help identify the best mitigation strategy to address it.

In summary, TAI would be expected to do the following-

- Test new algorithms for bias
- Monitor implemented algorithms for bias or unexpected results
- Continuous learning engine to learn from identified biases
- Delivered by a 'new skilled' agile workforce
- Leverages proven Applied Intelligence assets
- Powered by new Analytics Apps, AI Advisors, Responsible AI tools, and the TAI scorecard

TAI will allow for the monitoring of AI by its 'AI twin' to enable continuous bias retraining. Intelligent work distribution will escalate edge cases for a human to review and remediate potential biases. Companies will be able to explain to customers and to regulators with greater precision why a loan was approved or why a change in cells could be early-stage cancer. Companies using TAI can do this with confidence that the results have not been biased by, for example, race, gender, or class.

3. Encountering the AI Ethical Dilemma

A. Understanding the common dilemmas and challenges AI poses today

Over the years, it has been observed that while unlocking value, AI and analytics introduce new risks and challenges. There are numerous examples across industries that outline 'what went wrong'. A selection of these is given below.

Amazon's HR Hiring Tool

From 2014 to 2017, Amazon invested heavily in building an AI-enabled hiring tool. However, the existing gender imbalances for technical jobs were ingrained in the historical data that Amazon's AI used to make hiring decisions. The outcome was successful male candidates and unsuccessful female ones. After four years of significant investment, Amazon was forced to retire the application.[15]

Microsoft's Twitter Bot "Tay"

In 2016, shortly after its launch, Twitter (now X) users began tweeting politically incorrect phrases to Microsoft's Twitter chatbot, "Tay". Microsoft had not trained the chatbot to navigate these types of inappropriate behavior and so Tay responded by sending its own inflammatory tweets to the public. After 16 hours of operation, Microsoft was forced to shut the bot down.[16]

Winterlight Labs Auditory Testing tool

In 2016, a Canada-based start-up company focused its energies on developing AI-powered auditory tests for neurological diseases. The technology captured the person's dialect and analyzed the data to determine the onset of Alzheimer's disease. Though the test had a staggering accuracy of >90%, the datasets it was trained on were mainly of native English speakers only. When non-English speakers took the test, due to their non-fluency, pauses while speaking or mispronunciations were misconstrued as markers of the disease.[17,18]

Skin Cancer Detection

In several research efforts, AI algorithms have been developed for skin cancer detection and diagnosis. Because of the limited availability of diversified training datasets, it has been observed that AI has been more accurate in diagnosing malignancies of light-skinned patients compared with dark-skinned individuals. The current trend indicates that AI models do not usually consider clinical context or metadata. Several attributes that include but are not limited to genetics, medical history, lesion prognosis, etc. play an important role in accurate diagnosis. An MIT Media Lab study published in February 2018 found that AI-based skin cancer detection solutions were 11% to 19% more accurate on lighter-skinned individuals and 34% less accurate for darker-skinned individuals.[17–20]

Convolutional Neural Network for Common Thoracic Diseases

A 2020 PNAS study highlighted how gender imbalances inherited from the training data sets resulted in lower accuracy in the underrepresented group. The study illustrates that gender imbalance in medical imaging datasets produced biased classifiers for computer-aided diagnosis based on convolutional neural networks (CNNs). It is a known fact that CNNs and other similar models learn from techniques that are deployed for image segmentation, filtration, and structural changes that appear in the images. It therefore requires large and balanced datasets to ensure that the AI-model is bias-free and can diagnose disease conditions as expected.[21,22]

In general, bias in AI systems can have serious implications, including disparities in diagnosis, treatment, and patient outcomes.[23,24] Some examples for potential AI bias in the health care industry are outlined below:

Racial Bias in Dermatology Diagnoses

AI algorithms may be less accurate in diagnosing skin conditions in darker-skinned patients compared with lighter-skinned patients as a result of the training datasets fed in at the time of learning. This bias could lead to misdiagnosis and delayed treatment for patients from ethnic minority groups.[17–20]

Gender Bias in Cardiovascular Risk Assessment

Given that males are more prone to heart disease, AI algorithms may underestimate the risk of heart disease in women compared with men, leading to underdiagnosis and undertreatment of cardiovascular conditions in female patients.

Socioeconomic Bias in Predictive Models

Predictive models used to identify patients at high risk of hospital re-admission or poor health outcomes may exhibit socioeconomic biases. These models often rely on electronic health record (EHR) data, which may reflect biases related to socioeconomic status, access to health care, and social determinants of health. As a result, patients from marginalized or underserved communities may be disproportionately affected by biased predictions and recommendations.

Language Bias in Natural Language Processing (NLP) Systems

Natural Language Processing (NLP) systems that are used for analyzing clinical notes, patient records, and medical literature may exhibit language bias. These systems may perform poorly or inaccurately when interpreting text written in languages other than English or when reading medical jargon used by health care professionals. Language bias can lead to errors in clinical documentation, miscommunication between health care providers, and disparities in patient care.

Geographic Bias in Diagnostic Imaging Algorithms

AI algorithms used for analyzing medical imaging data, such as X-rays, CT scans, and MRI images, may exhibit geographic bias. These algorithms may be trained on datasets that primarily include patients from certain geographic regions or health care institutions, leading to performance disparities in different populations. Patients from underserved regions or with restricted access to high-quality health care facilities may receive less accurate diagnoses or interpretations from AI-driven imaging systems.

Addressing bias in AI systems used in health and life sciences requires careful consideration of data quality, model training methodologies, algorithmic transparency, and ongoing evaluation and validation processes. By proactively identifying and mitigating bias, organizations can ensure that AI technologies contribute to equitable and patient-centered research and care.

B. Application of Generative AI in health and life sciences

According to the latest report from Accenture,[25] advances in Large Language Models (LLMs) can revolutionize the health and life sciences industries. 98% of the providers and 89% of executives who participated in the study believe these advancements will play an important role in their organizations' strategies in the next three tofive years, as 40% of all working hours could be impacted by LLMs like Open AI's GPT-4.

Generative AI (Gen AI) is also demonstrating promising results in drug discovery and development.[26] Whether it is about efficient analysis of large datasets, identifying promising research candidates for clinical trials, or even predicting potential side effects and interactions, Gen AI is making a significant impact. Together with expediting the discovery process, it is also enhancing the precision and safety of drug development. In the clinical data management space, use of Gen AI is enabling faster, more efficient cleansing of clinical trial data. In the Pharmacovigilance space, adverse-event case processing, aggregate reporting, and narrative writing is becoming swifter and more efficient.[32] Product dossier completion and authoring of regulatory documents has become very efficient with the use of Gen AI.

Traditionally, patient treatment was administered based on broad population data, with limited consideration for individual variations.[26] GenAI is now making it possible to dive deep into patients' genetic profiles, medical histories, and real-time health data. Customized and personalized care is now possible that meets the unique needs and genetic makeup of each patient. This patient-centric approach is resulting in precise and effective medical care that improves outcomes and reduces the occurrence of adverse events significantly.[33]

Data-driven real-time insights are enabling public health strategies, optimizing hospital operations, and reshaping medical care at scale, leading to sustainable and industrialized health care systems. Gen AI holds immense promise and is already revolutionizing the health care industry.[33] While the future seems very promising, enabled as it is through this digital transformation, it also raises ethical and regulatory concerns and threats to patient privacy and security.

C. Understanding Ethical issues with AI and Gen AI

The examples in the preceding section are reflective of the unintended consequences of AI implementation. Launching AI without an understanding of its social impact can be risky to a company's reputation and brand. Furthermore, deploying AI without anchoring it to robust compliance and core values may expose a business to significant risks, including employment/HR issues, data privacy breaches, and health and safety problems.

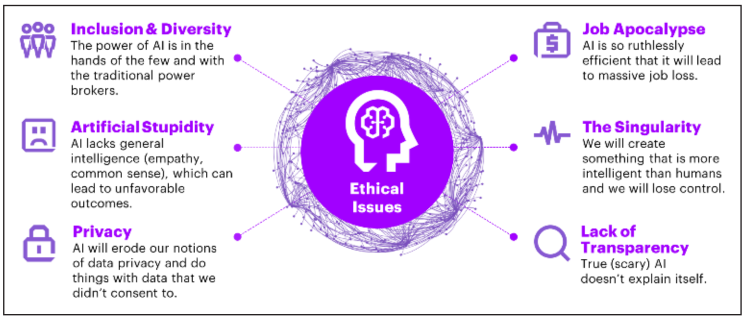
Figure 4 shows how the adoption and implementation of AI can potentially lead to ethical concerns taking center stage.[9] These concerns are outlined below.

Job Apocalypse

Once AI is implemented successfully for use and is delivering satisfactory outcomes, it will likely lead to massive job losses resulting in a 'Job Apocalypse'.

Singularity

There is a fear that humans will create something more intelligent than ourselves (AI) and we will lose control.



Inclusion & Diversity The power of AI is in the hands of the few and with the traditional power brokers.

Artificial Stupidity AI lacks general intelligence (empathy, common sense), which can lead to unfavorable outcomes.

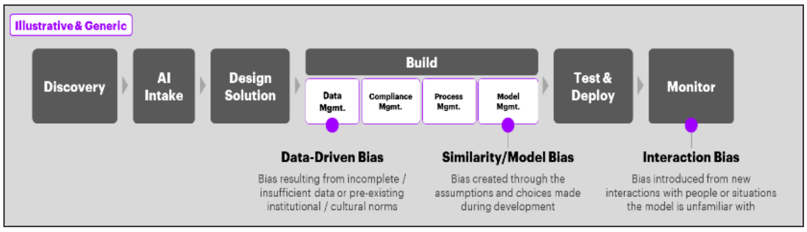
Privacy AI will erode our notions of data privacy and do things with data that we didn't consent to.

Job Apocalypse AI is so ruthlessly efficient that it will lead to massive job loss.

The Singularity We will create something that is more intelligent than humans and we will lose control.

Lack of Transparency True (scary) AI doesn't explain itself.

Figure 4: Ethical issues with AI/Gen AI.



Data-Driven Bias Bias resulting from incomplete / insufficient data or pre-existing institutional / cultural norms

Similarity/Model Bias Bias created through the assumptions and choices made during development

Interaction Bias Bias introduced from new interactions with people or situations the model is unfamiliar with

Figure 5: AI lifecycle and AI bias controls.

Lack of Transparency

AI, once developed and implemented, cannot (or will not) explain itself, creating a dangerous lack of transparency for the outcomes that it may impact and/or deliver.

Data Privacy

AI may do things with data that we may not have consented to do.

Artificial Stupidity

We understand that AI models, unlike humans, lack empathy, common sense and general intelligence. This absence may result in erroneous or unfavorable outcomes

Inclusion and Diversity

Finally, ignoring inclusion and diversity can have significant impacts on outcomes delivered by AI models.

A paper published in 2018 by Harvard Business Review on 'Auditing Algorithms for Bias' stated that "Data is not objective; it is reflective of pre-existing social and cultural biases".[27] It is therefore essential to understand how bias creeps in AI algorithms as they are trained using complex and large datasets.

Oxford Dictionary defines bias as an "inclination or prejudice for or against a person or group, especially in a way considered to be unfair". AI is susceptible to biases that emerge from its interactions with humans and the data it is given. As AI learns from biased people and data its ability to learn, predict, and surprise is tainted with the biases it learns and by the new biases it propagates.

Bias can be introduced into AI applications through data, models, or ongoing operations. All three components work in conjunction and can reinforce each other if biases are not addressed at the early design stage. The following high-level process flow shown in **Figure 5** illustrates where unintentional bias is most likely to arise in the context of a typical AI development and implementation process.[28]

While developing AI models, as one transitions from design stage to build stage, two types of biases can creep in. A data-driven bias results from incomplete or insufficient data or pre-existing cultural norms. For example, in a CNN model developed for skin cancer detection, if the datasets are not diverse enough with images of patients across color, age, and ethnicity, AI is most likely to make inaccurate and biased assessments. A Similarity/Model bias is created during the 'assumptions creation' process and is related to choices made during development. For example, in an AI model development for the detection of head and neck cancer, assuming that women are less likely to consume tobacco, or smoke may result in the model making biased decisions while processing datasets that are a combination of combined gender population.

Once the model is developed, tested and deployed for use, it is important to ensure continuous monitoring and assessment to prevent any interaction bias (which is the third type of bias) to creep in. This bias is introduced through new interactions with people or situations that the model is unfamiliar with.

Bias needs to be eliminated at every step of AI development. Whether it is an unconscious bias that creeps in from a human component or an unplanned bias that occurs because of training data (see **Figure 6**), it can proliferate through the implementation of AI.[13]

Training Data bias

All AI models learn decision-making through training datasets. It is therefore essential to assess the datasets for the presence of any bias. For example, training data for a facial recognition algorithm that over-represents white people may create errors when attempting facial recognition for people of color.[29] There are health apps that default to male symptoms for heart attacks as they are trained through predominantly male-specific datasets.[30]

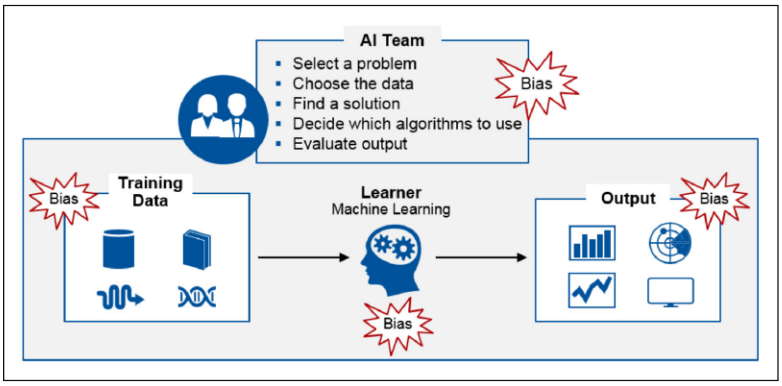


Figure 6: Unintentional AI bias introduction.

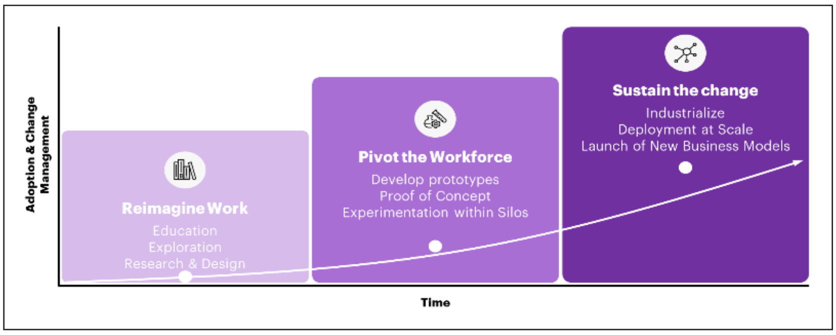


Figure 7: The AI Change Management.

Algorithm bias

AI models are developed by humans who are likely to make assumptions while programming and development that may be an outcome of a conscious or unconscious bias. These assumptions may result in algorithm bias which may further get amplified with the use of biased training data sets.[29] For example, an AI-enabled text-to-image system called Midjourney, when asked to create images of smart or influential people, often displays pictures of old white men with glasses, showing a lack of representation for other races.[30]

Cognitive bias

All humans are influenced by their experiences, situations, preferences, and choices. These biases inadvertently make their way into AI models and are termed 'cognitive biases'. For example, this type of bias may lead to favoring datasets gathered from one section of society rather than sampling from a range of populations around the globe.[29]

Addressing all the challenges associated with AI requires taking all of these biases into account.

3. Conclusion

In this paper, we have reviewed the applications of ethical and responsible AI in the health and life sciences industry. The advent of AI introduces new intersections between humans and machines, reshaping the activities that each has traditionally been known to do in a way that unlocks greater value. As an example, in the drug development process, the application of AI will help to accelerate clinical development, improve productivity, and expedite regulatory submissions. Developing and deploying AI models that do not inherit biases from training datasets and which are governed by RAI and TAI frameworks will help clinicians to draw insights from large complex datasets, and make better-informed decisions; it will help health care professionals diagnose and predict the onset of diseases, and will provide options for treatment and care. It is, however, important to note that managing this AI change is critical. Actions that would help fulfill the RAI mandate follow the glidepath shown in **Figure 7**, with each element driving a different type of value for businesses and industries.[31]

It all would start with 'reimagining work', which is a combination of evolution of work and the elevation of workers. As an example, a pharmacovigilance professional processing huge volumes of safety cases to assess safety issues related to a drug, determining causality and composing a safety narrative for submission to the regulatory authority would be assisted in the future by an AI using Natural Language Processing and Machine Learning to determine safety issues related to the drug. It would perform causality analysis and write the regulatory submission-ready narrative, thereby freeing up the time of the pharmacovigilance professional to focus on high-risk cases and cater to the growth in adverse events cases. This development, as an example of a re-configuration of a job profile, will enable employees to take on work that is of higher value, providing them with the opportunity to contribute to strategic priorities of the organization. The second tenet of the glidepath is 'pivoting the workforce' to areas that create new forms of value with a significant impact on how organizations conduct their businesses. This tenet enables collaboration between humans and machines that would help to create new business models, thereby driving positive customer experience. As is evident, the most valuable skill required to collaborate with AI is that of judgement, which is needed in situations when a machine is unable to make the right decision. For example, when an AI-enabled algorithm processes an adverse event case to determine whether the event was caused as a result of the consumption of the drug or not and makes an incorrect assessment, the pharmacovigilance specialist makes an intervention to correct that inaccurate assessment and overrides the outcome through their fair judgment, thereby helping the AI algorithm to learn that the next time it assesses a case with similar conditions, its ability to make an accurate assessment is improved. This is a classic example of how human and machine can collaborate to instill the culture responsible use of AI. The last (but the most important) tenet in this glidepath is 'Sustenance of Change,' which is about humans helping AI to help humans over time. More and more datasets being fed to the AI algorithm for it to learn from a growing variety of cases will help it to improve its judgement, thereby improving overall accuracy of assessment. This, in turn, would ensure higher quality of any safety reports submitted to regulatory authorities. Adoption and scaling the approach will help achieve sustained growth for organizations and help deliver better outcomes for customers and society at large.

Sustained success largely depends on adopting, embracing and then practicing Responsible AI to ensurethat data and associated systems are fair, transparent and accountable. Organizations globally can establish themselves as RAI and TAI industry leaders by taking steps to improve and implement best practices related to managing bias in AI. It is imperative that business leaders get immersed and involved in the RAI and TAI development and deployment process, thereby building trust in the entire ecosystem.

Competing Interests

The authors have no competing interests to declare.

References

1. **Ortiz S.** *80% of enterprises will have incorporated AI by 2026, according to a Gartner report.* www.zdnet.com. October 12 2023. Accessed December 30, 2023. https://www.zdnet.com/article/80-of-enterprises-will-have-incorporated-ai-by-2026-according-to-a-gartner-report/. DOI: https://doi.org/10.1007/s40278-024-66943-8
2. *Artificial intelligence (AI) in healthcare Market Growth, drivers, and opportunities.* (n.d.). MarketsandMarkets. Retrieved July, 2024, from https://www.marketsandmarkets.com/Market-Reports/artificial-intelligence-healthcare-market-54679303.html
3. **Siau K, Wang W.** Artificial Intelligence (AI) Ethics. *Journal of Database Management.* 31(2): 74–87. DOI: https://doi.org/10.4018/JDM.2020040105
4. **Ozmen Garibay O, Winslow B, Andolina S, Antona M, Bodenschatz A, Coursaris C, … Xu W.** Six Human-Centered Artificial Intelligence Grand Challenges. *International Journal of Human–Computer Interaction.* 39(3): 391–437. DOI: https://doi.org/10.1080/10447318.2022.2153320
5. **Alakwe K.** Human Dignity in the Era of Artificial Intelligence and Robotics: Issues and Prospects. *Journal of Humanities and Social Sciences Studies.* 5(6): 87–97. DOI: https://doi.org/10.32996/jhsss.2023.5.6.10
6. **Crowley J.** *Toward AI Systems that Augment and Empower Humans by Understanding Us, our Society and the World Around Us.* https://www.humane-ai.eu/wp-content/uploads/2019/11/D21-HumaneAI-Concept.pdf.
7. **Naik N, Hameed BMZ, Shetty DK, Swain D, Shah M, Paul R, Aggarwal K, Ibrahim S, Patil V, Smriti K, Shetty S, Rai BP, Chlosta P, Somani BK.** Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility? *Front Surg.* 2022 Mar 14; 9: 862322. DOI: https://doi.org/10.3389/fsurg.2022.862322
8. **Farhud DD, Zokaei S.** Ethical Issues of Artificial Intelligence in Medicine and Healthcare. *Iran J Public Health.* 2021 Nov; 50(11): i–v. DOI: https://doi.org/10.18502/ijph.v50i11.7600
9. *Responsible AI Governance consulting & Solutions.* (n.d.). Accenture. Retrieved July 6, 2024, from https://www.accenture.com/bg-en/services/applied-intelligence/ai-ethics-governance
10. **Nipko, J.** (2023, October 5). Overcoming AI bias within life sciences. Retrieved July 7, 2024, from *PharmExec.* https://www.pharmexec.com/view/overcoming-ai-bias-within-life-sciences
11. *Facebook's five pillars of Responsible AI.* https://ai.meta.com/blog/facebooks-five-pillars-of-responsible-ai/. June 21 2021. Retrieved June 3, 2024, from https://ai.meta.com/blog/facebooks-five-pillars-of-responsible-ai/
12. **Eitel-Porter R, Corcoran M, Connolly P.** *Responsible AI: From Principle to Practice.* Accenture.com. March 30, 2021. https://www.accenture.com/us-en/insights/artificial-intelligence/responsible-ai-principles-practice
13. **Smith G, Rustagi I, Berkeley Haas Center for Equity, Gender and Leadership.** (2020). Mitigating Bias in Artificial Intelligence: An Equity Fluent Leadership Playbook (A. Morse, Humans for AI, C. Carson, AI Ethics Lab, C. Jeanmaire, Center for Human-Compatible AI, UC Berkeley, G. Neff, Oxford Internet Institute, University of Oxford, J. Zou, Stanford University, L. Schiebinger, Gendered Innovations in Science, Health & Medicine, Engineering, and Environment, Stanford University, P. Gertler, Berkeley Haas, S. Russell, Computer Science, UC Berkeley, S. West, & AI Now Institute, Interviewers). In A. Melgoza, A. Mackey, Google, D. Wimmer, Google, F. LeBaron, Berkeley Haas EGAL, J. Ellenbogen, Google, J. Wells, Berkeley Haas EGAL, J. Deutsch, BCG, J. Kaiser, BCG, K. Walsh, Levi Strauss & Co., K. McElhaney, Berkeley Haas EGAL, UC Berkeley (Eds.), *Berkeley Haas Center for Equity, Gender and*

Leadership. Retrieved July 21, 2024, from https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf.

14. **Daugherty PR, Wilson HJ.** *Human + Machine*. Harvard Business Press. March, 2018. http://books.google.ie/books?id=wpY4DwAAQBAJ&printsec=frontcover&dq=Human+%2B+Machine,+Reimagining+Work+In+The+Age+of+AI&hl=&cd=1&source=gbs_api
15. **Dastin J.** *Insight – Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. October, 2018. Retrieved December 30, 2023, from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/. DOI: https://doi.org/10.1201/9781003278290-44
16. **Vincent J.** (2016, March 24). Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge*. Retrieved July 7, 2024, from https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist
17. **Gershgorn D.** (2022, July 20). If AI is going to be the world's doctor, it needs better textbooks. *Quartz*. Retrieved July 13, 2024, from https://qz.com/1367177/if-ai-is-going-to-be-the-worlds-doctor-it-needs-better-textbooks
18. **M S.** (2021, December 14). The pitfalls of AI bias in Healthcare – Siddharth M – Medium. *Medium*. Retrieved July 13, 2024, from https://medium.com/@sidart.m92/the-pitfalls-of-ai-bias-in-healthcare-6a68e35b67a7
19. **Beltrami EJ, Brown AC, Salmon PR, Leffell DJ, Ko J, Grant-Kels JM.** Artificial intelligence in the detection of skin cancer. *Journal of the American Academy of Dermatology*. 1 December, 2022; 87(6): 1336–1342. DOI: https://doi.org/10.1016/j.jaad.2022.08.028
20. **Esmo.** (2023, May 9). Man Against Machine: Artificial Intelligence is Better than Dermatologists at Diagnosing Skin Cancer. *ESMO*. Retrieved July 14, 2024, from https://www.esmo.org/newsroom/press-and-media-hub/esmo-media-releases/artificial-intelligence-skin-cancer-diagnosis
21. **Larrazabal AJ, Nieto, N, Peterson V, Milone DH, Ferrante E.** Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*. 26 May, 2020; 117(23): 12592–12594. DOI: https://doi.org/10.1073/pnas.1919012117
22. **Fulmer, J.** (2024, August 22). *Addressing AI and implicit bias in healthcare*. TechnologyAdvice. Retrieved July 14, 2024, from https://technologyadvice.com/blog/healthcare/ai-bias-in-healthcare/
23. **Gichoya JW, Thomas, K, Celi, LA., Safdar NM, Banerjee I, Banja JD, Seyyed-Kalantari L, Trivedi H, Purkayastha S.** AI pitfalls and what not to do: Mitigating bias in AI. *British Journal of Radiology*. 1 October, 2023; 96: 1150. DOI: https://doi.org/10.1259/bjr.20230023
24. **Abràmoff MD, Tarver ME, Loyo-Berríos N, Trujillo S, Char D, Obermeyer Z, Eydelman MB, Maisel WH.** Considerations for addressing bias in artificial intelligence for health equity. *Npj Digital Medicine*. 12 September, 2023. DOI: https://doi.org/10.1038/s41746-023-00913-9
25. **Ghosh B, Narain K, Guan L, Wilson J.** (2023, March 23). *Generative AI technology in business*. Accenture. Retrieved July 20, 2024, from https://www.accenture.com/us-en/insights/technology/generative-ai?c=acn_glb_largelanguagemomediarelations_13427684&n=mrl_0323
26. **Hafke T.** (2024, June 19). Generative AI in healthcare: use cases, benefits, and drawbacks. *AlphaSense*. Retrieved July 21, 2024, from https://www.alpha-sense.com/blog/trends/generative-ai-healthcare/
27. **Chowdhury R, Mulani N.** Auditing Algorithms for Bias. *Harvard Business Review*. 25 October, 2018. https://hbr.org/2018/10/auditing-algorithms-for-bias#:~:text=Data%20is%20not%20objective%2C%20is,negative%20consequences%20and%20inequitable%20outcomes.
28. **Gichoya JW, Thomas K, Celi LA, Safdar N, Banerjee I, Banja JD, Seyyed-Kalantari L, Trivedi H, Purkayastha S.** AI pitfalls and what not to do: mitigating bias in AI. *Br J Radiol*. 2023 Oct; 96(1150): 20230023. Epub 12 September 2023. DOI: https://doi.org/10.1259/bjr.20230023
29. **IBM Data and AI Team.** *Shedding light on AI bias with real world examples*. IBM Blog. 16 October, 2023. Retrieved July 21, 2024, from https://www.ibm.com/blog/shedding-light-on-ai-bias-with-real-world-examples/
30. **Haritonova A.** *AI Bias Examples: From Ageism to Racism and Mitigation Strategies*. PixelPlex. 14 November, 2023. https://pixelplex.io/blog/ai-bias-examples/
31. **Romanhuk A.** *Reworking the Revolution 2019*. Academia.edu. 27 December, 2018. Retrieved July 21, 2024, from https://www.academia.edu/38048342/Reworking_the_Revolution_2019
32. **Desai MK.** Artificial intelligence in pharmacovigilance – Opportunities and challenges. *Perspect Clin Res*. 2024 Jul–Sep; 15(3): 116–121. Epub 2024 Mar 27. PMID: 39140015; PMCID: PMC11318788. DOI: https://doi.org/10.4103/picr.picr_290_23
33. **Praveen J, Krishna Cm, Channappa, A.** Transforming Pharmacovigilance Using Gen AI: Innovations in Aggregate Reporting, Signal Detection, and Safety Surveillance. *The Journal of Multidisciplinary Research*. 2023; 3: 9–16. DOI: https://doi.org/10.37022/tjmdr.v3i3.484

How to cite this article: Shrotriya S, Banu PKN, Kulkarni A, Aiyangar S. Navigating the Path to Ethical and Responsible AI Integration in Health and Life Sciences with Human and Machine Collaboration. *Journal of the Society for Clinical Data Management*. 2025; 5(1): 5, pp. 1–12. DOI: https://doi.org/10.47912/jscdm.412

Submitted: 17 November 2024 **Accepted:** 03 February 2025 **Published:** 13 March 2025

JSCDM

Journal of the Society for Clinical Data Management is a peer-reviewed open access journal published by Society for Clinical Data Management.

OPEN ACCESS