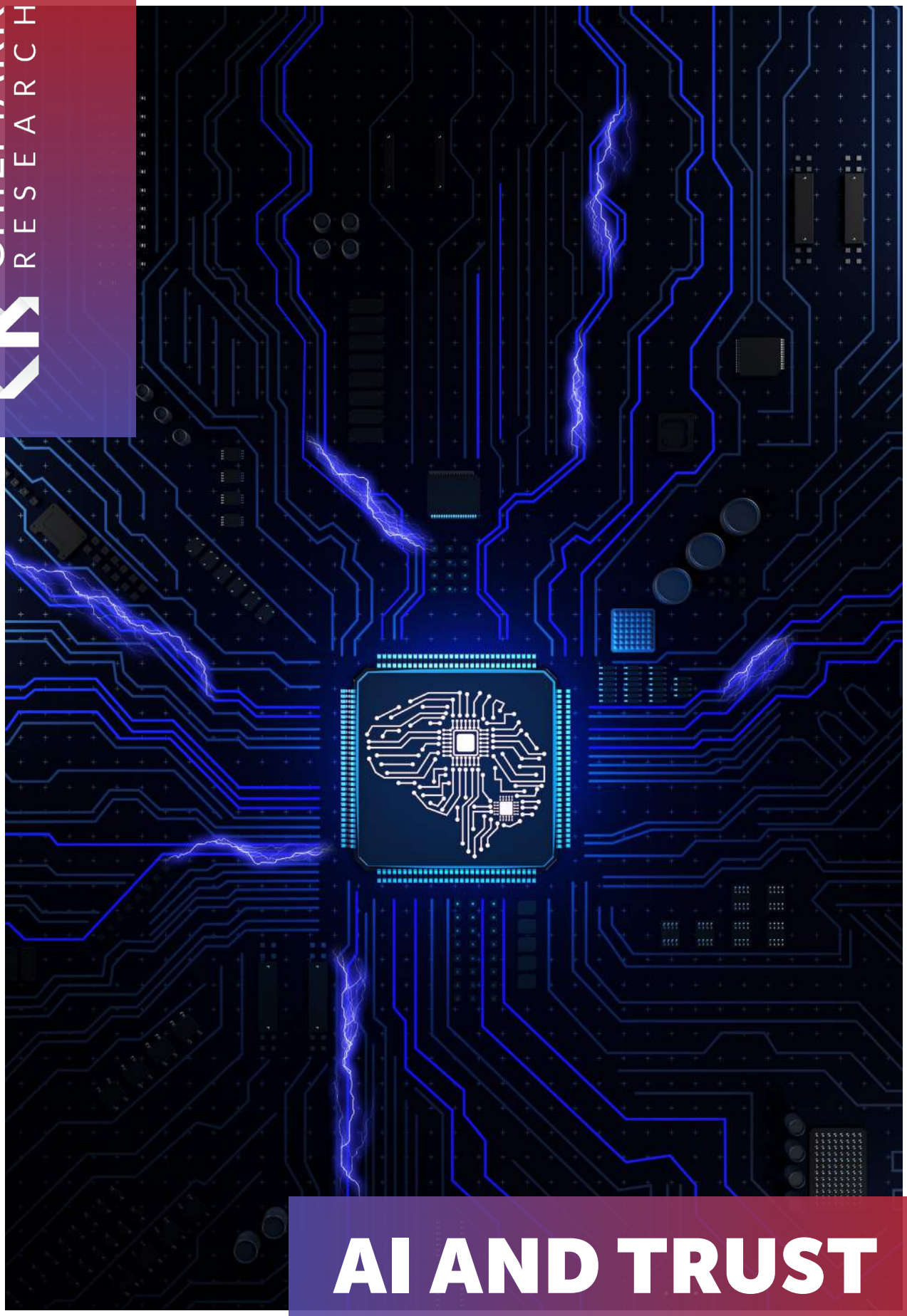
CHILMARK RESEARCH



AI AND TRUST IN HEALTHCARE

ABOUT CHILMARK RESEARCH

Founded in 2007, Chilmark Research is a preeminent global research and advisory firm focused exclusively on tracking the market evolution of healthcare information technologies (health IT) and use cases.

Our team is united by the belief that new health IT tools are critical for improving the quality and efficiency of care in a modern world. It is therefore our mission to foster the effective adoption, deployment, and use of these new solutions (and enabled services) through objective, high-quality research into those technologies with the greatest potential to impact care delivery.

This laser-sharp focus allows us to provide our community with the most in-depth, future-forward research on the critical technology and adoption trends occurring throughout the healthcare sector.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Credit.
Metamorworks

This report addresses one of the most important issues of our time—how to evaluate and build trust for AI applications in health and avoid harm to patients and society. From finance to education to health, we are seeing dramatic growth in the use of AI models and algorithms to augment decision-making and automate some processes.

However, along with the growth in AI-based products and services we frequently find concerning ethical issues and a landscape of many failures. Trust is becoming a foundational component of successful AI applications. But trust is not purely a question of singular companies. It takes an ecosystem with established protocols, standards, and best practices to create responsible AI applications that put patient safety first.

The report outlines processes that are critical to responsible AI and building trust. From assessments of validation and verification efforts to explainability and other mechanisms for transparency to social impact, fairness and safety, we provide a set of building blocks that will require additional work beyond the development of ethical guidelines. The FDA is currently lagging behind in addressing these issues and in many cases may not require enough from companies to garner physician and patient trust. Furthermore, AI applications will require monitoring over their lifecycles due to the unique nature of AI that differs from traditional software—the continuous learning and modifications to models as they learn from data and the need to be curated over time.

We look at approaches to AI design that come from the field of "human centered AI design" that is growing in some academic research centers. Key insights from the design world can help address some of the trust issues we outline in this report. We also provide some case studies in an appendix of human centered design for AI in practice.

One of the most important insights is the need for more intra-industry and public-private cooperation to build best practices and standards for the most important components of trust-building in AI. This involves developing a consensus for the most important use cases for each component and creating a type of "nutrition-label" that defines how these best practices are implemented in each case. This will help facilitate what some thought leaders have recently proposed—a market for liability insurance for AI.[1] The insurance industry has a role to play in incentivizing use of the best in class models.

Our report concludes with the recommendations for actionable policies that organizations can implement to drive safe, reliable, and relevant AI or what one can refer to as "trustworthy AI". Without a concerted effort to realize these policies in the coming years AI could suffer from even slower adoption than the current languid pace and growing risk for doing harm to patients.

[1] See https://catalyst.nejm.org/doi/full/10.1056/CAT.21.0242

INTRODUCTION

Trust is becoming one of the most critical components for the adoption of emerging technologies but is often poorly defined. When we think about AI and trust it may be wise to view AI as only one conspiracy theory away from experiencing a similar backlash as vaccines if we continue to encounter flawed algorithms in the marketplace or they become viewed as a source of unfairness and inequality in access to healthcare. The stakes are high.

Trust has many layers and facets as the partial glimpse in Figure 1 below illustrates. Building trust means instilling confidence that tools do what they were designed to do, are reliable over time and users have confidence that there will be redress if things go wrong.

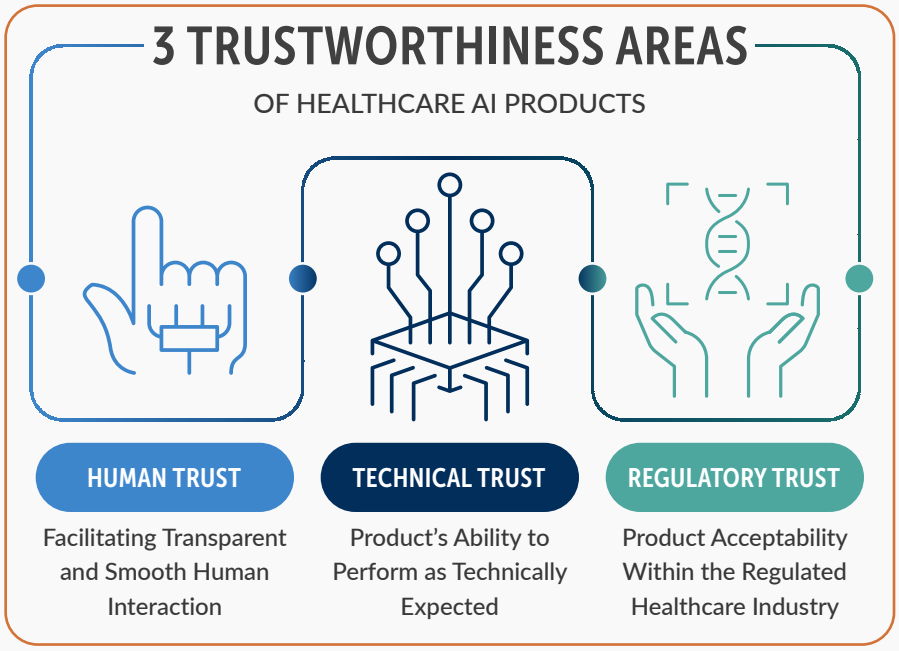


Figure 1.
Consumer Technology Association's breakdown of layers of trust in AI in Healthcare

Further complicating the trust equation is the experience of patients, providers, citizens with the so-called big tech companies (Facebook, Apple, Amazon, Google, etc.) and the rise of surveillance capitalism and AI's role in platform economies.[2] Surveillance capitalism involves the massive collection of data from our devices that is used to build predictive models to drive commerce, build recommendation engines, and what Zuboff calls behavioral futures markets.

In healthcare, we collect an increasing amount of data from patients in order to manage their care, predict outcomes and risks and better understand the efficacy of clinical pathways. Patients expect the industry to manage data securely, protect their privacy, and utilize effective and safe models that minimize the risk of harm to patients. This unspoken social pact -that data governance is paramount- can be damaged by growing distrust from consumers and citizens in how their data are handled by companies.

AI demands very large datasets to train models and this will increase the need for data governance to ensure privacy and security. We are seeing more interest in federated learning and differential privacy efforts to achieve these goals. But still, we hear of cybersecurity breaches constantly, mishandling of privacy of users in apps[3], and racial bias in algorithms[4].

Ian Corbin and Joe Waters observed an important shift in consumer perceptions of healthcare in the US. In 1966, more than three-quarters of Americans reported having high confidence in our medical leaders. By 2018, that number had fallen to 34 percent. One of the important shifts over the decades has been the perception of with whom the consumer interacts: in the 1960s people viewed their interactions

2 https://en.wikipedia.org/wiki/The_Age_of_Surveillance_Capitalism
3 See the example of some femtech apps https://www.thelawyersdaily.ca/articles/36505/data-privacy-invasions-in-femtech-a-threat-to-long-term-women-s-health-care-
4 https://www.science.org/doi/10.1126/science.aax2342

in the doctor-patient relationship whereas in recent decades they now describe their primary interaction as with a SYSTEM. This may impact how emerging technologies are adopted as well if trust in systems ends up being undermined by unfair or unsafe algorithms.

Trust: A Social Currency

Trust is becoming an important social currency in the context of new technologies such as AI. The predictive nature of AI and the scale of platforms has created a number of concerns across society about how we will need to regulate AI and mitigate risks. Mistrust can also be contagious when one company introduces faulty AI products into the market and other companies are damaged from the perception that the entire AI field is too risky or under-developed. It is to industry's benefit that risk mitigation and trust are carefully practiced and maintained. Collective efforts from industry stakeholders are required to confront the risks we face.

This report will provide insights and approaches on the following aspects of trust:

- AI Ethics and Responsible AI: why principles are only a beginning and more granular practices that are ethically informed are necessary
- Validation of AI: a brief analysis of validation in AI and digital health and the need to audit for biases
- Explainable AI (XAI) how does XAI differ from validation and what should it accomplish and what are some of the criticisms?
- AI and new challenges for regulators: identification of AI specific challenges and what is the current state of legal and regulatory reasoning to address?
- Intra-industry forms of cooperation to help de-risk AI in healthcare algorithms and to create an innovative ecosystem with patient safety and health equity goals.

In conclusion, a proposal is proffered for an eco-system of institutions and new instruments to provide better standards across the validation, bias, explainability and patient safety set of concerns. This will in-turn help to improve adoption of the best AI tools. Today's sole reliance on the FDA will not go far enough. Industry stakeholders must step in to build trust and create mechanisms to improve the quality of AI products that are entering the market at an ever increasing pace.

CHAPTER 1: WHY TRUST MATTERS NOW AND ETHICAL GUIDELINES

From the early days of AI/ML as a concept in computer science there have been concerns over the autonomous nature of the technology that can evolve without human control. This has been the portrayal of AI in science fiction and film where both utopian and dystopian imaginaries inform public opinion of AI. Beyond these popular portrayals there are concerns over the impact on employment, and in healthcare specifically, patient safety, data integrity and bias in access to services are critically important real-world challenges that designers and users of AI/ML systems must contend with today.

Recently the FDA released their "Guiding Principles for Good Machine Learning Practice" co-developed with Health Canada, and the United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA) to help the AI/ML industry navigate both patient safety and continuing innovation in new devices and AI/ML algorithms. These principles are derived from those used in the medical device and other sectors to help guide best practices.

The principles include the following:

1. The total product life cycle uses multidisciplinary expertise.
2. The model design is implemented with good software engineering and security practices.
3. Participants and data sets represent the intended patient population.
4. Training data sets are independent of test data sets.
5. Selected reference data sets are based upon best available methods.
6. Model design is tailored to the available data and reflects intended device use.
7. Focus is placed on the performance of the human-AI team.
8. Testing demonstrates device performance during clinically relevant conditions.
9. Users are provided clear, essential information.
10. Deployed models are monitored for performance, and retraining risks are managed.

To clarify, these are guidelines for software that is intended to treat, diagnose, cure, mitigate, or prevent diseases or other conditions that fall under the FDA's regulatory umbrella.[5] Some principles (numbers 1, 3, 4, 5) are at least partly intended to address the bias issue that can impact the performance of AI/ML tools across diverse populations.

From these guidelines and a plethora of universities, think tanks and industry organizations' pronouncements on the ethics of AI and healthcare we see a broad consensus on basic principles that constitute a Responsible AI in Healthcare approach. In Figure 2 we have consolidated these norms across institutions.

5 US FDA, "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-based Software as a Medical Device (SaMD)."

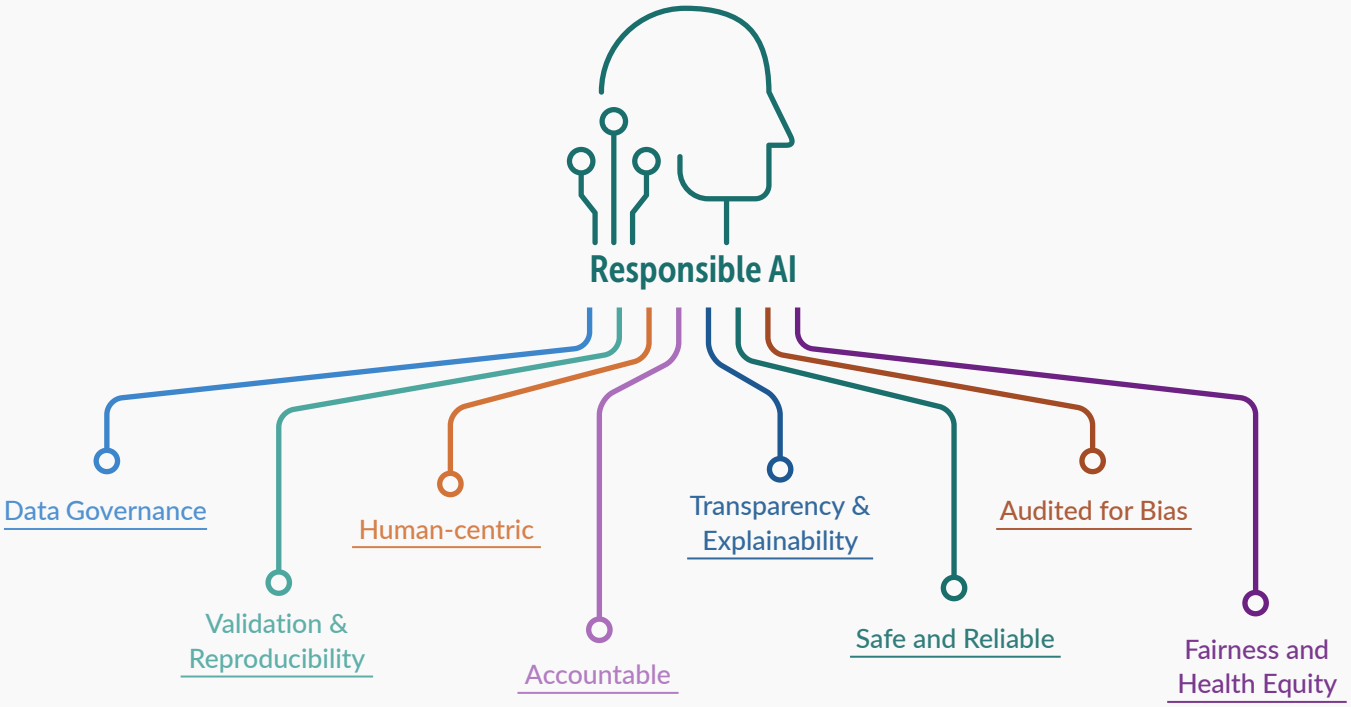


Figure 2.
Components of Responsible AI

Principles provide broad guardrails to guide development of ethical or responsible AI. However, they often lack the granularity to provide practical steps for software developers and practitioners. They also often fall short when institutional politics create gray areas or debates over privileging one specific principle over another. In the following chapters we look at the various steps to mitigate the most important risks and the practices that developers and organizations can utilize to better ensure responsible AI.

CHAPTER 2: VALIDATION OF MODELS

Does the model perform at scale beyond the original training set?

It has become almost commonplace to learn that a model developed by a vendor later turned out to suffer from significant bias. Validation of algorithms is necessary to address this issue. The healthcare industry will need to develop standards for validation of algorithms for specific use cases.

The FDA is developing their guidance for validation of Software-as-a-Medical Device (SAMD) and has the three steps as outlined below:

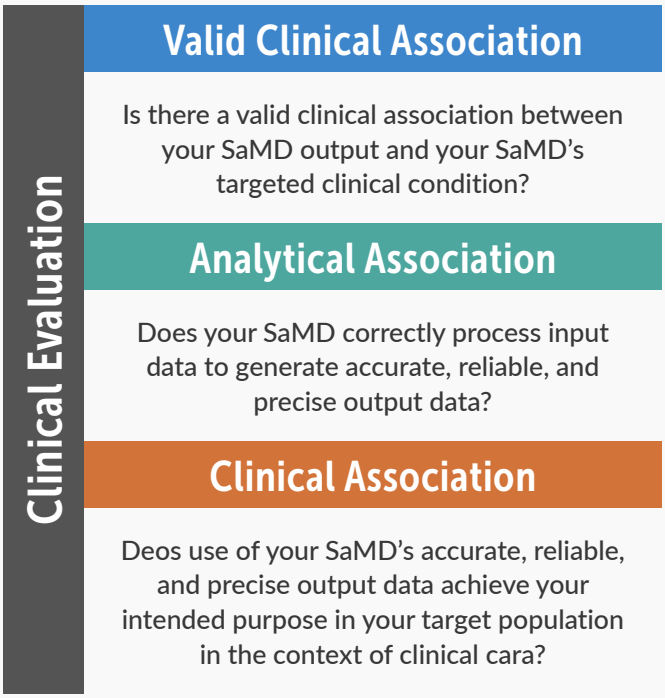


Figure 3.
FDA Guidance on Validation

Verification is typically conducted by software engineers in AI and the validation steps above require both clinical and engineering expertise combined. We rarely see double-blind RCTs in SAMD. Furthermore, a large number of AI applications are classified as "wellness" apps and do not require as rigorous a validation process. This is one of the gray areas where things can go awry and further erode trust in AI/ML.

The current state of validation of AI in healthcare is quite problematic, contributing to slow adoption and a lack of trust. *A recent study of algorithms approved by the FDA between 2008 to 2021 found a total of 118 approvals.[6] The study found that 17/118 posted no validation claims or data. Only 9/118 had validation dataset sizes over 1000 patients. This means that it is not really possible to justify clinical use of the algorithms or infer generalizability or presence of bias, the authors assert.*

The validation problem is worse than it may appear from this study alone given that many algorithms do not require FDA approval and there are some significant data science challenges that we will discuss below.

Dimensionality in digital health data [7]

One of the major challenges AI faces in clinical applications currently is the issue of drift or the declining performance of models as they are deployed on across newer and broader populations beyond the original training datasets. Today data can come from numerous sources including EHRs, imaging data, speech samples, wearable data on activities, clinical variables, genomic (-omic) data.

Innovations in hardware and software are dramatically improving the resolution of many of these data sources. This adds to the **dimensionality** of data and the richness that has the potential to improve clinical insights. However, some of the issues with calibration drift in AI models may be stemming from the dimensionality of data.

The curse of dimensionality occurs when models are trained on small population sizes and these models have many features or high dimensionality. High dimensionality can lead to hidden information in the data or blind spots. As clinical variables increase there is an exponential increase in combinatorial values within these datasets that requires a rather large increase in population size of the training models. The curse of dimensionality happens when inadequate sample sizes are used for complex health phenomena and lead to researchers having an inaccurate reading of how well the model is actually performing during training.

6 https://www.sciencedirect.com/science/article/abs/pii/S1076633221004153
7 For a full discussion of this challenge see "Digital Medicine and the curse of dimensionality." Berisha et al (2021) https://www.nature.com/articles/s41746-021-00521-5#:~:text=Digital%20health%20data%20are%20multimodal%20and%20high%2Ddimensional.&text=This%20paper%20considers%20how%20the,dimensionality%E2%80%9D%20in%20statistical%20learning%20theory.

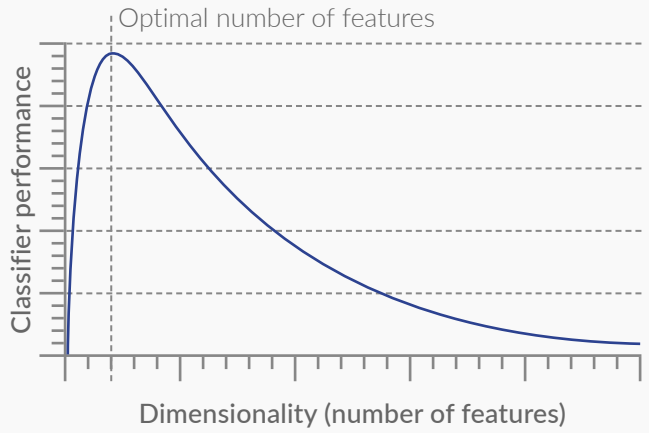


Figure 4.
Curse of Dimensionality
(Source: Vision Dummy)

Berisha et al. point to the case of IBM Watson for Oncology for treatment recommendations for multiple forms of cancer. The training sample size was 106 ovarian cancer patients and 635 lung cancer cases. One study found that 71% of all healthcare AI models were trained on data collected from California, Massachusetts and New York. They warn this could lead to massive blind spots for AI models as they are deployed in other contexts. Most often the developer of the model only becomes aware of the blind spots after deployment when the exponential growth in the error rate has already occurred.

Bias mitigation

As AI has become more prevalent a number of implicit bias concerning race have arisen with adverse consequences for people of color. One important example is the algorithm used for kidney transplant triage that measures kidney health, the glomerular filtration rate (eGFR). The problem rests with the fact that the levels that determine thresholds for access to medicines or transplants are race adjusted due to a biased 1999 study that treats race as a biological category. The result is that thousands of African-Americans are arbitrarily denied treatment due to a flawed, racially biased measure.[8]

A 2020 article by Vyas et al., in the New England Journal of Medicine identified a wide number of clinical decision support tools that include racial bias largely by conflating the notion of race as a social construct with race as rooted in biology. This neglects the fact that genetic variability within racial groups is greater than the variability across racial categories. Neglecting this fact can lead to problematic algorithms with the ability to do harm.

	Systemic Biases	Statistical and Computational Biases	Human Biases
Datasets Who is counted, and who is not counted?	› Issues with latent variables › Underrepresentation of marginalized groups	› Sampling and selection bias › Using proxy variables because they are easier to measure › Automation bias	› Observational bias (streetlight effect) › Availability bias (Anchoring) › McNamara fallacy
Processes and Human Factors What is important?	› Automation of inequalities › Underrepresentation in determining utility function › Processes that favor the majority/ minority › Cultural bias in the objective function (best for individuals vs best for the group)	› Automation bias › Likert scale (categorical to ordinal to cardinal) › Nonlinear vs linear › Ecological fallacy › Minimizing the L1 vs L2 norm › General difficulty in quantifying contextual phenomena	› McNamara fallacy › Groupthink leads to narrow choices › Rashomon effect leads to subjective advocacy › Difficulty in quantifying objectives may lead to McNamara fallacy
TEVV How do we know what is right?	› Reinforcement of inequalities (groups are impacted more with higher use of AI) › Predictive policing more negatively › Widespread adoption of ridesharing/ self-driving cars/etc. may change policies that impact population based on use	› Lack of adequate cross-validation › Survivorship bias › Difficulty with fairness	› Confirmation bias › Automation bias

Figure 5.
Bias and Patient Harm (Source: NIST)

8 See presentation by Dr. Tania Martin-Mercado at HIMSS 2022 session on "Implicit Bias and AI" and the Chilmark Podcast with her at: https://www.chilmarkresearch.com/chilcast-healthcare-tech-talks-dr-tania/

A recent National Institute for Standards and Technology report, "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence" provides a useful framework for mitigating risk of bias and the need for humans in the loop to think critically about model design, testing, development and deployment. The NIST report also raises the problem of what is termed the McNamara Fallacy where quantitative measurements and metrics are considered more objective or better than other observations (such as ethnographic data). As healthcare continually engages more with social determinants of health (SDoH) data there will be an equally increasing need for qualitative data on social contexts to better guide analytics.

NIST has broken down bias into three primary categories: systemic bias, human bias, and statistical/computation bias as seen in Figure 6 below.

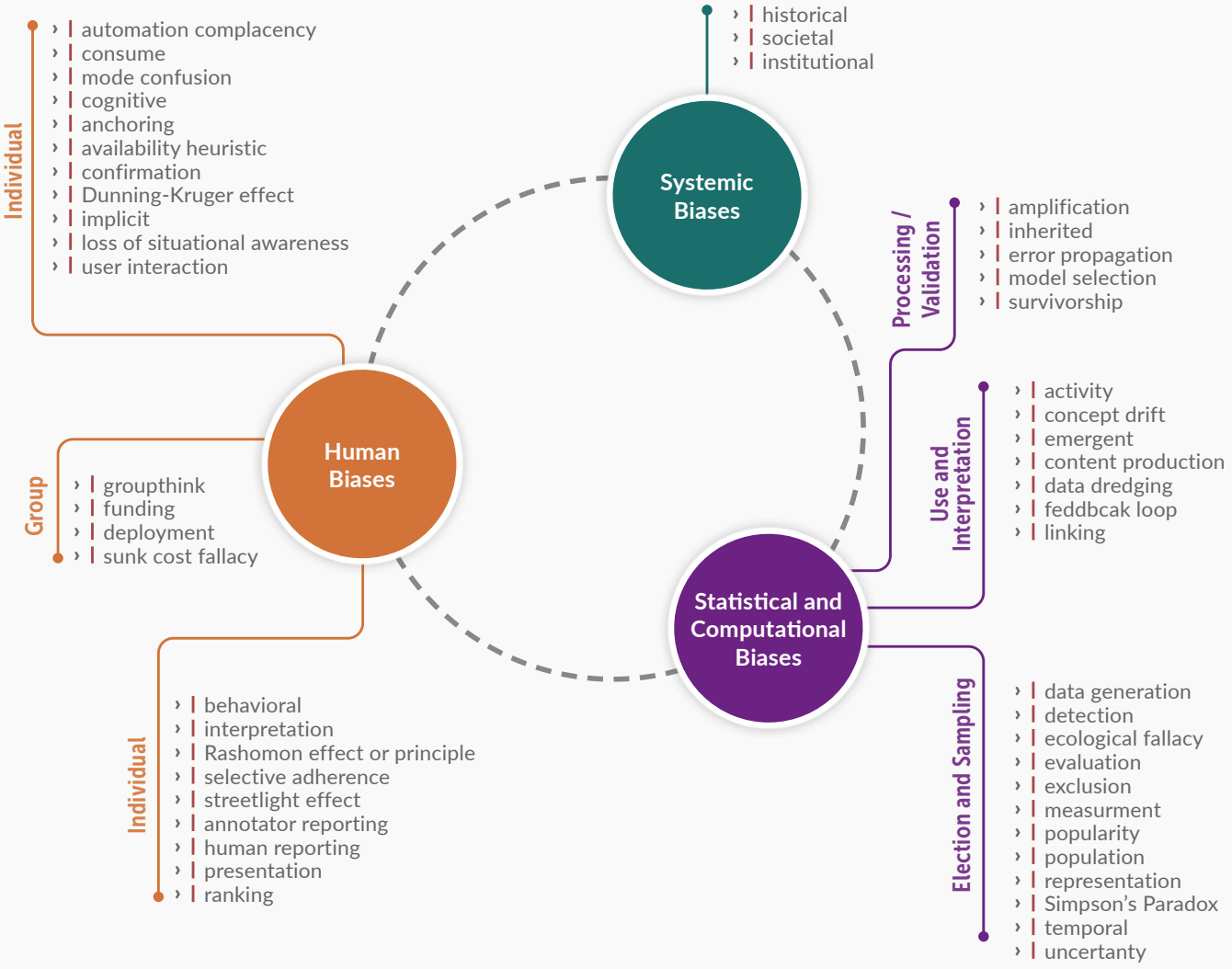


Figure 6.
Types of Bias (Source: NIST)

The Center for Applied Artificial Intelligence at the University of Chicago has also developed an Algorithmic Bias Playbook.[9] This publication is one of the best general manuals for bias detection in the field and informs the Brookings Institute's template. The key steps they recommend for rooting out bias include:

- Conducting simple statistical tests to see if the algorithm is performing well on the variables it is predicting
- Hold the algorithm accountable for predicting the ideal target for the underserved populations
- Create accountability structures and stewards within organizations responsible for oversight of bias and with the power to challenge and change processes when necessary. This is not a human resources position but a technology executive role
- Create sound and thorough documentation processes of the entire lifecycle of the algorithm development and deployment. This is also just sound scientific practice.

The documents cited in this section of the report are all important guides to mitigating bias and there is a continually growing literature of case studies and methods published almost weekly that data scientist teams will need to track.

AI can also be used with human-in-the-loop approaches for auditing models for bias and correcting the models when necessary. There is a growing number of companies that are offering bias auditing tools as well as the use of synthetic data to address bias that may be inherent to some data sets. This is an emerging science and will require close observation by data scientists to evaluate which tools are performing at the highest level. This is an area in need of improvement in standards for evaluation and performance metrics of the tools.[10]

Some companies are offering technological approaches to bias auditing and mitigation in models and apply various statistical validation tests to double check models. Below is a list of current companies in the market offering auditing bias tools.

- VirtuousAI spans the continuum of explainability to diagnostics of models that provide insights on validation and reliability
- Biasfix provides a validation service for security, quality and discrimination in risk assessment tools
- Eticas Research and Consulting provides bias audits
- Aequitas is a tool developed by Carnegie Mellon University's Data Science and Public Policy program to audit bias in algorithms

9 See https://www.chicagobooth.edu/research/center-for-applied-artificial-intelligence/research/algorithmic-bias
10 See https://www.vanderschaar-lab.com/synthetic-data-breaking-the-data-logjam-in-machine-learning-for-healthcare/ and https://www.nature.com/articles/s41551-021-00751-8

- ForHumanity is a non-profit that provides audits rules for all algorithms in the public sphere involved with trust, security, ethics and bias
- Parity was developed by Rumman Chowdhury and acquired by Twitter for auditing purposes and the IP is now used for a service they provide in the area of governance and compliance and mitigating risk of AI models
- Arthur.ai has a platform that spans performance monitoring, explainability and bias mitigation
- IBM AI Fairness 360
- Google's What If Tool
- Oracle/Skater Python Library for model interpretation

One final note on technological approaches is that they do not negate the need for domain experts to critically examine models for bias as well. Simply throwing technology at technology rarely solves the problem entirely. Organizations should be wary of vendors offering a purely technological approach to bias mitigation and the need for diverse data science teams and an organizational commitment to Responsible AI and health equity should be viewed as mandatory.

Key Takeaways on Validation and Bias:

- Lack of documentation on model development and design as well as small sample sizes are contributing to flawed AI models and much better design protocols are needed
- Pay attention to the dimensionality of digital health data and take the necessary steps to attend to this aspect during the validation process
- Build diverse data science teams and a steward for accountability in AI with the power to challenge and change processes
- Checklists have been developed to help facilitate bias mitigation and these provide fairly comprehensive approaches, but organizations will still need to think beyond checklists to avoid falling behind the emerging science and practices that the broader Responsible AI community is continuously developing

CHAPTER 3: EXPLAINABILITY AND TRANSPARENCY

AI/ML poses a significant challenge in providing transparency given the continuous learning of AI/ML models that can be at least somewhat autonomous of human agents. The so-called black box phenomenon - whereby developers of models may not fully understand the details of how a model has changed over time - is a challenge to standard notions of transparency. Another layer of complexity is the fact that legal, regulatory, clinical and social science perspectives on transparency and explainability often differ.

In the past several years a number of companies have created tools for XAI to address the problem of black box algorithms, to justify decisions being made, to improve interpretability or readability of models. A number of companies have been investing in XAI beyond healthcare including:

- Google Cloud's XAI: utilizes a "What If" tool to score how different factors contributed to results
- Tableau's Explain Data helps users understand "the why" behind a data point
- IBM's AI Explainability 360 is an open source toolkit for algorithms that support interpretability and explainability by surfacing the majoring factors that contribute to outcomes and can help uncover biases
- Microsoft Azure offers a set of tools for providing interpretability frameworks for models built with Azure
- Kyndi has developed an explainable AI platform for NLP (documentation provenance), machine learning and knowledge graphs across finance, government and pharmaceutical sectors
- Darwin AI is an explainable AI platform for automation and human in the loop decision-making use cases
- Fiddler.ai offers an explainable AI engine as well as model validation tools

This represents only a partial list of companies developing tools and platforms to address the challenge of black box algorithms. Other examples include Factmata, Logical Glue, Flowcast, Imandra and even DARPA has added been building an XAI program. The DARPA model has been viewed as almost a standard approach to explainable AI for many companies.

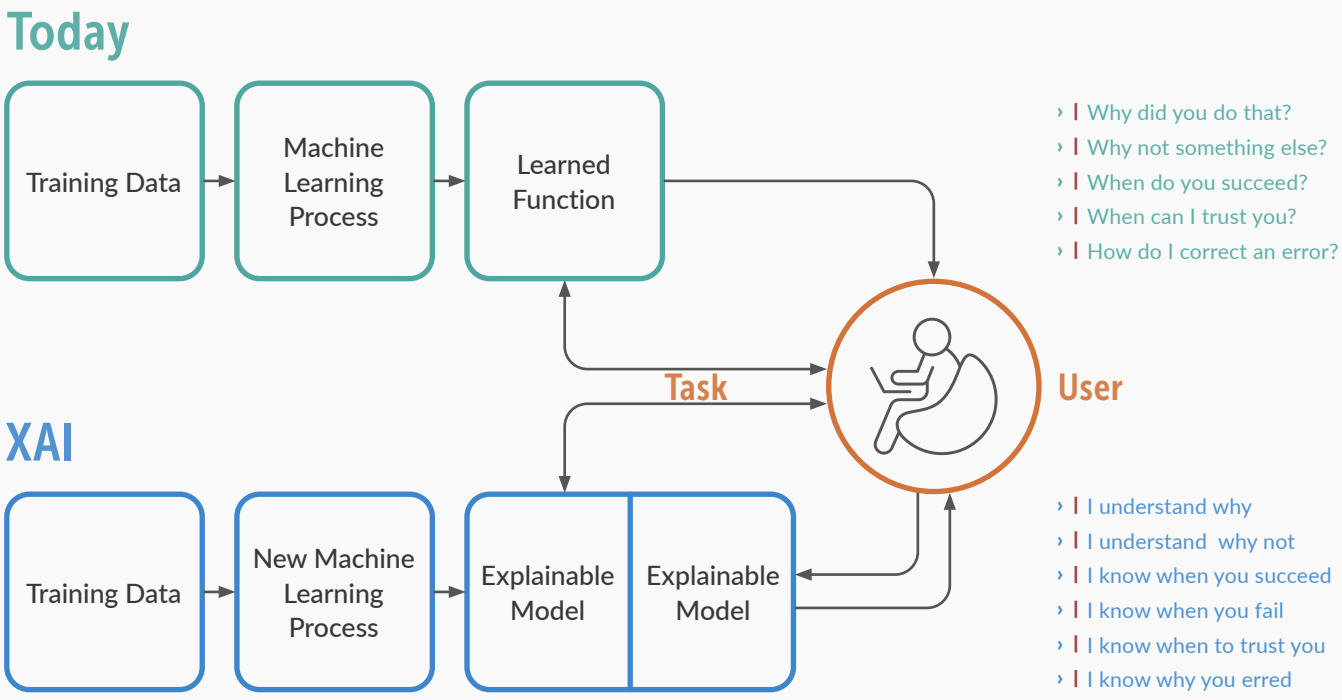


Figure 7.
What questions XAI attempts to answer (Source: DARPA)

In the EU the General Data Protection Regulation (GDPR) has an explanation- and information-based approach to transparency for AI embodied as a core principle in Article 5(1)a that states:

> *"...personal data must be processed lawfully, fairly and in a transparent manner in relation to the data subject."*

This requirement means that data controllers (companies utilizing the data via algorithmic manipulation) must provide explanations in clear and plain language in an accessible format. However, there are currently **no clear standards as to what "plain" and "accessible" mean in practice** and development of standards is needed to help guide practice.

Explainability is coming under criticism from some quarters.[11] Some critics point out that it is important to rule out biases and phenomena such as the "Clever Hans" problem in explainable AI.[12] The Clever Hans problem concerns the issue of prediction performance based on meta-data rather than data itself. A famous example is when a model was used to discern between huskies and wolves and the prediction was found to be driven by identification of a snowy background rather than actual differences between wolves and huskies.

The authors cited above use the medical example of an algorithm used at Mount Sinai hospital that identified high risk patients vs low risk patients (asthmatics) based on x-ray imaging. When the algorithm was used outside of Mount Sinai it performed poorly. Later, it was found that the algorithm analyzed meta-data based on the x-ray machine used in the ICU and was actually identifying different x-ray machines rather than features of images in x-rays.

This example highlights the need to differentiate between validation and explainability. Adequately validated algorithms would not have made it to the explainability step until the bias was corrected for during a validation process. The validation process should be focused on identifying bias and confounders upstream whereas XAI is a downstream step for transparency after validation.

Emerging Critiques of Explainable AI

In recent months a number of additional critiques of existing approaches to explainable AI have emerged in the pages of Lancet and Science. Ghassemi et al. raise questions about the possibilities of providing a local explanation – i.e. an explanation for an individual case . They argue that XAI techniques can offer very general understandings of how an AI system works but can be very misleading when applied to an individual case. Instead, they argue, the emphasis should be on validation rather than XAI.

The reason for the limited use of XAI rests in the possibilities of unrecognized confounders in the model. They also cite research that demonstrates that data scientists routinely "over-trust and misuse interpretability tools" using

11 https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00208-9/fulltext
12 https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-01332-6

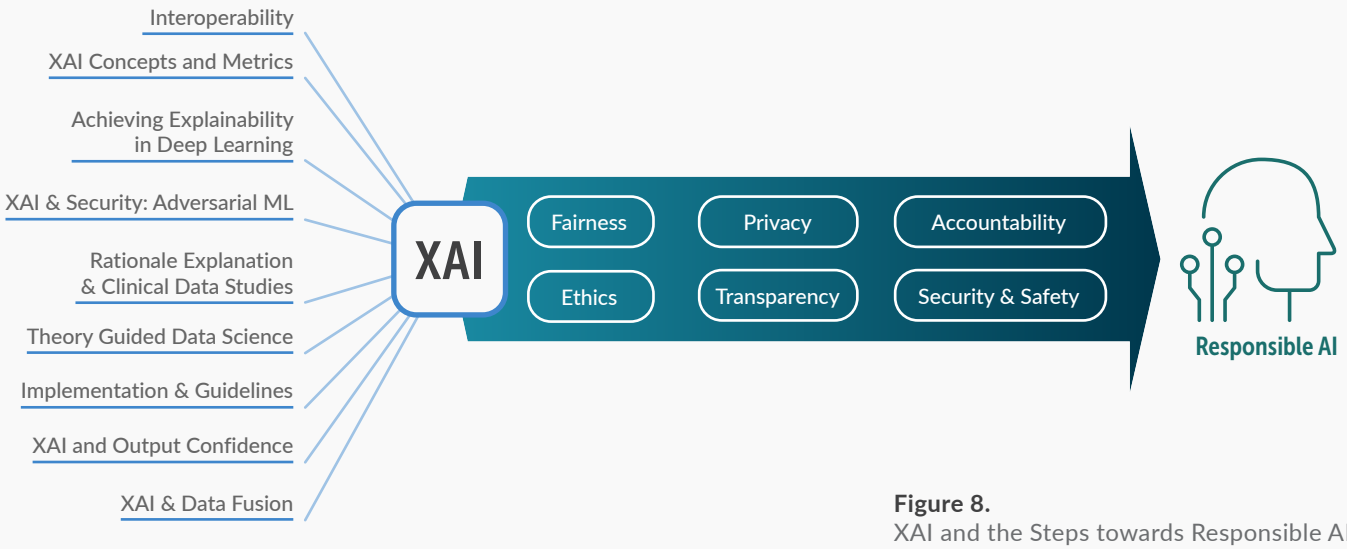


Figure 8.
XAI and the Steps towards Responsible AI

visualization aids with interpretability tools incorrectly. In radiology, heat maps are common as part of the XAI toolkit where the heat map can illustrate how much each region of a medical image contributed to the decision made by the model. The XAI method in this case is called "post-hoc explainability". However, as Ghassemi et al. observed, it is difficult to make such clear-cut distinctions with heat maps as there can be both useful and non-useful information in any given hot spot.

Rebuttal to the XAI Critique: ClosedLoop.AI Case Study

In May 2021 CMS, announced the winner of the $1M AI Health Outcomes Challenge as ClosedLoop.ai. The Challenge attracted over 300 companies to predict health outcomes (hospitalizations, adverse events and skilled nursing facilities admissions) for Medicare patients over a 12 month period. Two of the key criteria used included how well the algorithms rooted out bias as well as how the vendor addressed transparency defined as how well the solution explained to doctors and nurses how the algorithms arrived at a specific predictive risk profile for a patient.

After interviewing the ClosedLoop.ai team we realized that the critiques of XAI above provided a somewhat limited take on both the role of XAI as well as the range of uses of XAI currently being used.[13]

First, regarding the issue of asking too much of XAI, the critics fall into the trap of a techno-utopian vision for XAI that places it at the center of the trust equation. XAI alone cannot be asked to root out bias in data sources or model development. That belongs to the validation exercise. For example, XAI would not have identified bias in the well-known Optum example of racial bias in identification of patients for additional clinical or disease management services, however, validation would have.[14]

Secondly, focusing critiques on medical imaging and CDS alone overlooks the complexity of a wider range of use cases of AI in development for issues such as predictive risk modeling (the focus of ClosedLoop.ai) where social determinants and more multi-factorial types of modeling are more prevalent. A rigid biomedical model assuming a more deterministic type of model with binary decisions, for example, is a completely different type of AI from predictive modeling for these use cases.

Finally, when it comes to CDS and XAI, McCall and DeCaprio of ClosedLoop.ai make a more general point around the nature of medical decision making that overlaps with general issues of evidence-based medicine and clinical practice. The critiques that focus on bias and

13 The discussion that follows is based on an interview with Carol McCall and Dave DeCaprio, executives with ClosedLoop.ai with deep experience in actuarial sciences, population health, and machine learning.
14 In this case the critical error in model development was the use of claims as proxy for health status.

confounders in black box algorithms may also overlook the black box of physician decision making which is often a juggling act of using data and statistics with a physician's experience and judgement, often referred to as the art of medicine.

There are gray zones and degrees of freedom where clinicians draw upon their experience of other patients plus what the algorithm offers as a clinical pathway or solution. This epistemic muck of clinical decision making is often referred to as clinical judgment.[15] For this reason we are beginning to see clinical decision support tools be viewed more as "wayfinding" tools to facilitate stepwise decision-making rather than definitive decision-making tools.[16]

Key Takeaways on Explainability:

- Avoid conflating explainability and validation. Rigorous validation standards upstream that root out bias, problematic dataset limitations, and model development are distinct from tools to facilitate understanding of model performance. However, XAI also should point out areas of potential bias and the limitations of models.
- The industry needs to develop some form of standards for explainability methods that clarify what plain and transparent language means for whom. Explainability for physicians will mean something different for patients and both have a spectrum of knowledge bases and medical literacies that need to be attended to for understanding AI tools.
- A huge gap in the market is public engagement with AI and healthcare and the need to rethink general patient engagement in the era of AI.

There is not clear standard as to what "plain" and "accessible" mean across different contexts where XAI is used."

15 More on the n=1 vs population averages in clinical studies can be found in Herbert Weisberg, "Willful Ignorance: The Mismeasure of Uncertainty", 2009.

16 https://jamanetwork.com/journals/jama/article-abstract/2787207

CHAPTER 4:
DATA ETHICS AND AI

In addition to the bias, validation and explainability there are a number of issues concerning data poverty and data collection that AI developers need to consider. One issue that feeds into the bias problem is data poverty for some populations that are not included in many datasets used to train algorithms. Researchers need to address these inequities around datasets while balancing the need to not harm underserved areas with data collection methods that could stigmatize communities.

A recent article notes that we have systemic differences in the quantity or quality of data representing different groups or populations that may introduce bias.[17] For example, as of 2018 genomic wide association studies (GWAS) were 78% European, 10% Asian, 2% African, 1% Hispanic. AI models used in medical imaging are disproportionately trained on data from California, Massachusetts, and New York with little data from the other 47 states.

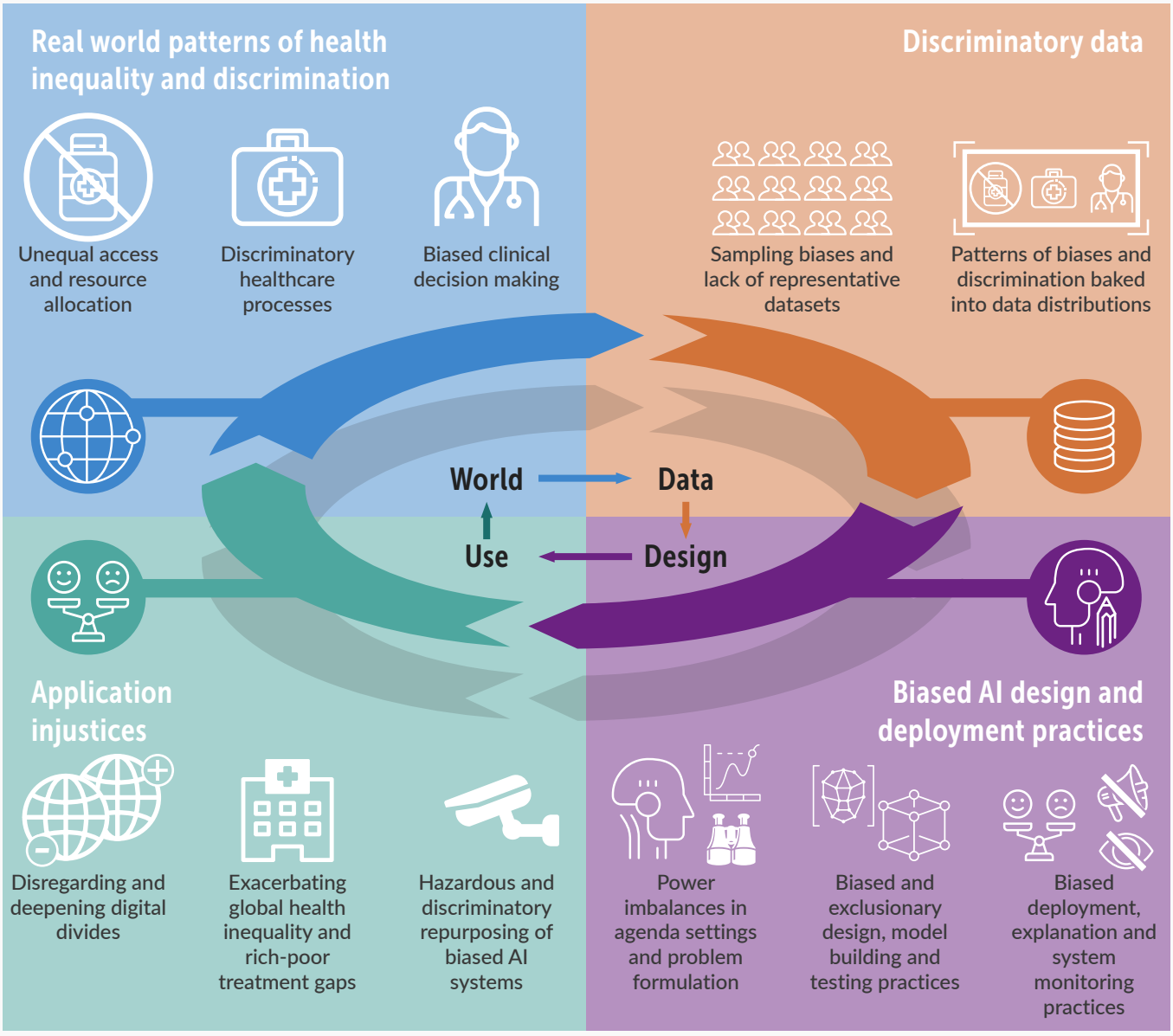


Figure 9.
Inequalities and Health Data Nexus (Source: Leslie, et al, BMJ)

17 https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30317-4/fulltext

The authors of this study point to the need for more citizen engagement about data and how data about these individuals and communities can be utilized for building better tools and healthcare delivery systems. But engagement rests on a bedrock of privacy protections and confidence that the systems will utilize data appropriately. Very strong data governance policies need to be in place.[18]

Due to the privacy and security challenges of working with large volumes of data required for AI model development there is a growing market for companies that can improve data governance as well as facilitate sharing of data across institutional collaborations. Tools such as differential privacy and federated learning are growing in both sophistication and prevalence in the market. Below is a partial list of companies offering services in these critical areas.

- OneTrust provides privacy, data governance and security services including clients such as Aetna
- Privacera offers data access management and encryption for data security and privacy
- Acuratio helps break silos for collaborating organizations through their federated learning platform

Some organizations are beginning to engage with synthetic data vendors as a privacy protection measure. Digital twins of data sets or synthetic data techniques can be deployed to create data set clones of original data sets but without the identities. Some of the vendors are either focused on healthcare or have partnered with healthcare organizations that have offerings in the market:

- Statice.ai – a data anonymization service for researchers. One can download Statice from a cloud provider and the client runs the program on their data to train a model based on the data features and then generates data. The privacy and utility of the dataset is then assessed.
- Ai.Reverie (acquired by Facebook)– a multisectoral vendor with specialized focus on unstructured data and computer vision. The platform is relevant to medical imaging, pathology types of applications which they have applied to one pharmaceutical application for pill or medication identification in ERs
- MDClone – Israeli data platform company with a Synthetic Data Engine exclusively focused on healthcare and working with hospitals and researchers
- Unlearn.AI – the focus is on computational clinical trials utilizing synthetic data
- Tonic – works across industries including healthcare for creating environments for product testing and development

Synthetic data has its own potential bias issues and a number of academic labs are working on techniques to mitigate bias. Today, there is also a lack of standards for assessing synthetic data.[19]

Conclusion: Health Equity and the Value of Third-Party Standards Organizations for Building an Innovative, Trustworthy Ecosystem

In healthcare we have a regulatory body such as the FDA that oversees the safety component, however we are still lacking an oversight organization that can certify AI applications as worthy of a seal of approval according to established guidelines. This would help build trustworthiness in the eyes of end users. A consortia of stakeholders and other third-party organizations is needed to develop a framework of standards for validation, explainability, patient safety and health equity.

Drawing upon insights from human centered design for computing, Ben Schneiderman has developed a framework for reliable, safe and trustworthy AI.[20] Reliability is defined as:

- Having audit trails and analysis tools
- Software engineering workflows documented
- Verification and validation testing
- Bias testing to enhance fairness
- Explainable user interfaces

18 This subject is worthy of a white paper itself but a basic framework is provided here https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7647243/
19 See the following article for a discussion of the issues: https://www.nature.com/articles/s41551-021-00751-8
20 https://hcil.umd.edu/human-centered-ai/ and See "Human-Centered AI" by Ben Schneiderman, chapter 7.

Organizations should also create cultures of safety that leadership espouses to instill confidence and trust by providing extensive reporting and discussion of failures, internal review boards and align with industry best practices. Academic medical centers already have institutional review boards (IRBs) for clinical research. IRBs could be a model for data science as well.

Organizations should also do a pre-implementation validation check and ask whether the problem they are trying to solve is amenable to AI models or even if this is the best approach. End users need to assist in validating whether the proposed future model is appropriate for the problem and be responsible for supervising and utilizing the model over time.

Plans for how it will be sustained over the lifecycle of its application are important as well. A model may be mathematically robust but the human side of the equation is lacking in implementation. That is, the validation of the technology fit with the problem is lacking.

These components offer a good start. However, in healthcare we need to go deeper and broader. Health equity is extremely important and AI can have both positive and negative effects due to the bias issues as well as the ability to scale data-driven efforts.

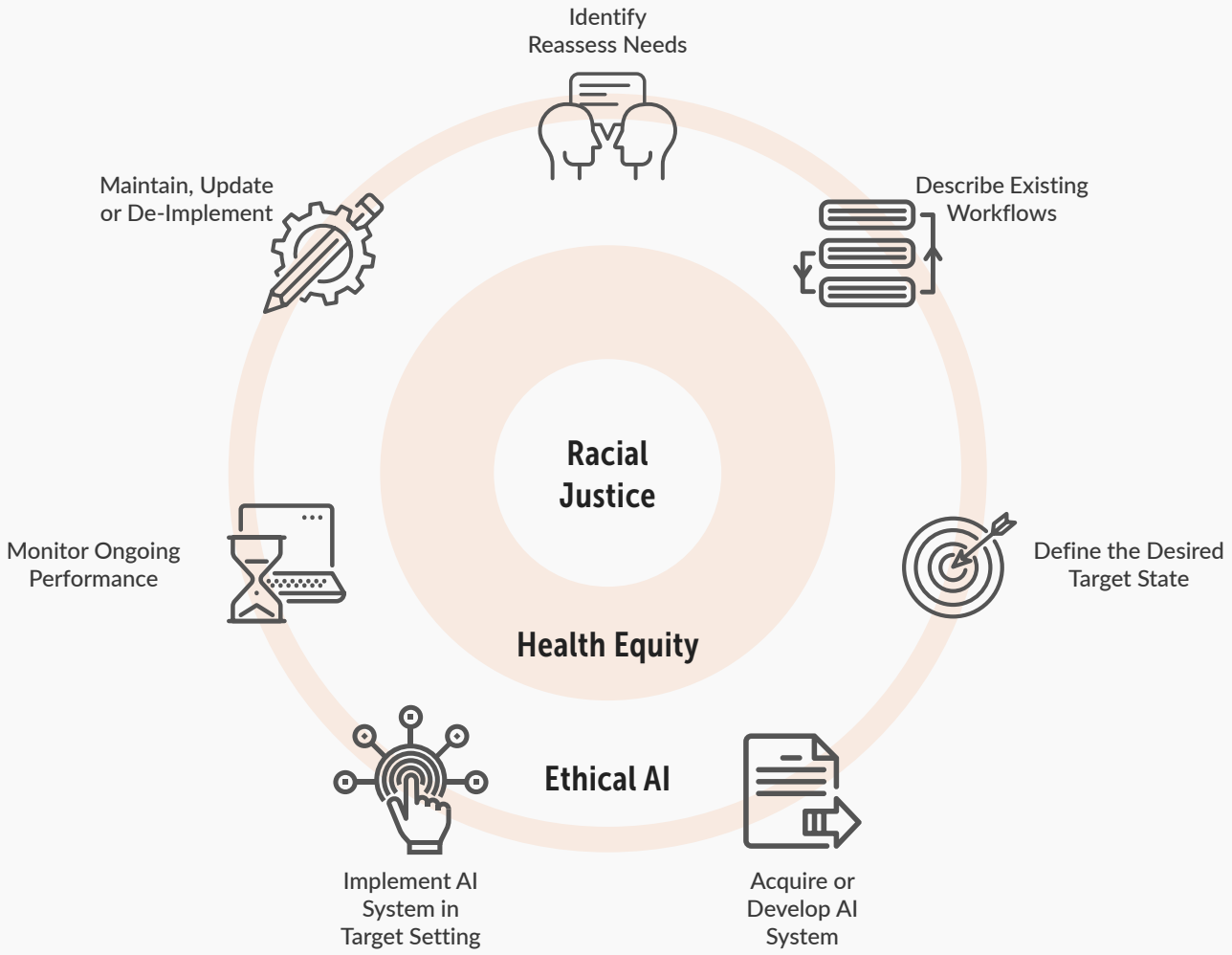


Figure 10.
Ethical AI, Health Equity and Racial Justice Integrated Across the Lifecycle of AI Development (Source: National Academy of Medicine and Dankwa-Mullen et al. 2021)

The framing of AI-based interventions matters. Dankwa-Mullen et al. have extended a National Academy of Medicine framework to include racial justice and health equity along with the need to monitor model performance over the lifetime of use. One component we would add is greater reflexivity on how data are collected and processed into action and implementation. A growing number of technology and AI researchers note that targeted interventions that are top-down can have the detrimental effect of stigmatizing communities.

In the broader ethics and AI community we hear discussions about using a kind of nutrition label for AI applications. The scorecard for applications would address how they perform across multiple assessment criteria. Below is an assessment framework that would be governed by a standards body or consortium.

Trust as an Intangible Asset: Building an Innovative Ecosystem

Economists Jonathan Haskel and Stian Westlake highlight the role of investment in intangibles that have become neglected over the past decade, but are an important part of the innovation equation. If we think about vaccine development for COVID, tangible assets such as vaccine production facilities, R&D labs, syringes, etc. are critical. But if trust is absent, vaccines can still fail.

Developing and distributing vaccines required large public-private partnerships and advance purchase commitments that ensured predictable markets downstream. These partnerships and advance commitments were instrumental in the overall success of the COVID-19 vaccine development process.[21] **Trust and risk-sharing were critical intangible assets in the creation of the market.**

We have seen a plethora of AI and ethics organizations created to develop and disseminate guidelines for ethical or responsible AI. **Yet there is a dearth of investment in the intangibles, such as guidelines, with the granularity to be effective in use.** Elinor Ostrom's work on the commons (management of the commons) and the insights on intangibles help provide some insights into the institutions that need to be created between companies and the market or the government

Figure 11.
Scorecards for Trustworthy AI

21 See Tim Harford, "From Software to soft power, intangibles still matter." Financial Times, 16/17 April 2022. https://www.ft.com/content/a14263cd-35e0-4f34-b10e-ae8523fad8d5

and the market to ensure best in class AI products enter the market. In an era when the FDA has insufficient resources to meet this need across the spectrum of AI and healthcare applications a commons-based approach to building risk mitigation frameworks is a likely path forward.

IEEE/ISTO have a successful history of creating standards and best practices for many different industries. If the healthcare industry could create industry consortia across multiple verticals within healthcare and partner with an organization such as IEEE/ISTO to create standards for validation, XAI, patient safety, data governance, etc. this would facilitate an ecosystem where "certified" products could also be eligible for liability insurance.

Innovation in healthcare in the era of AI will increasingly need innovation in governance and intangibles to ensure patient safety, health equity goals and development of responsible technologies. The Figure 12 below illustrates the approaches to numerous governance gaps that the World Economic Forum has observed are impacting many emerging technologies such as AI.

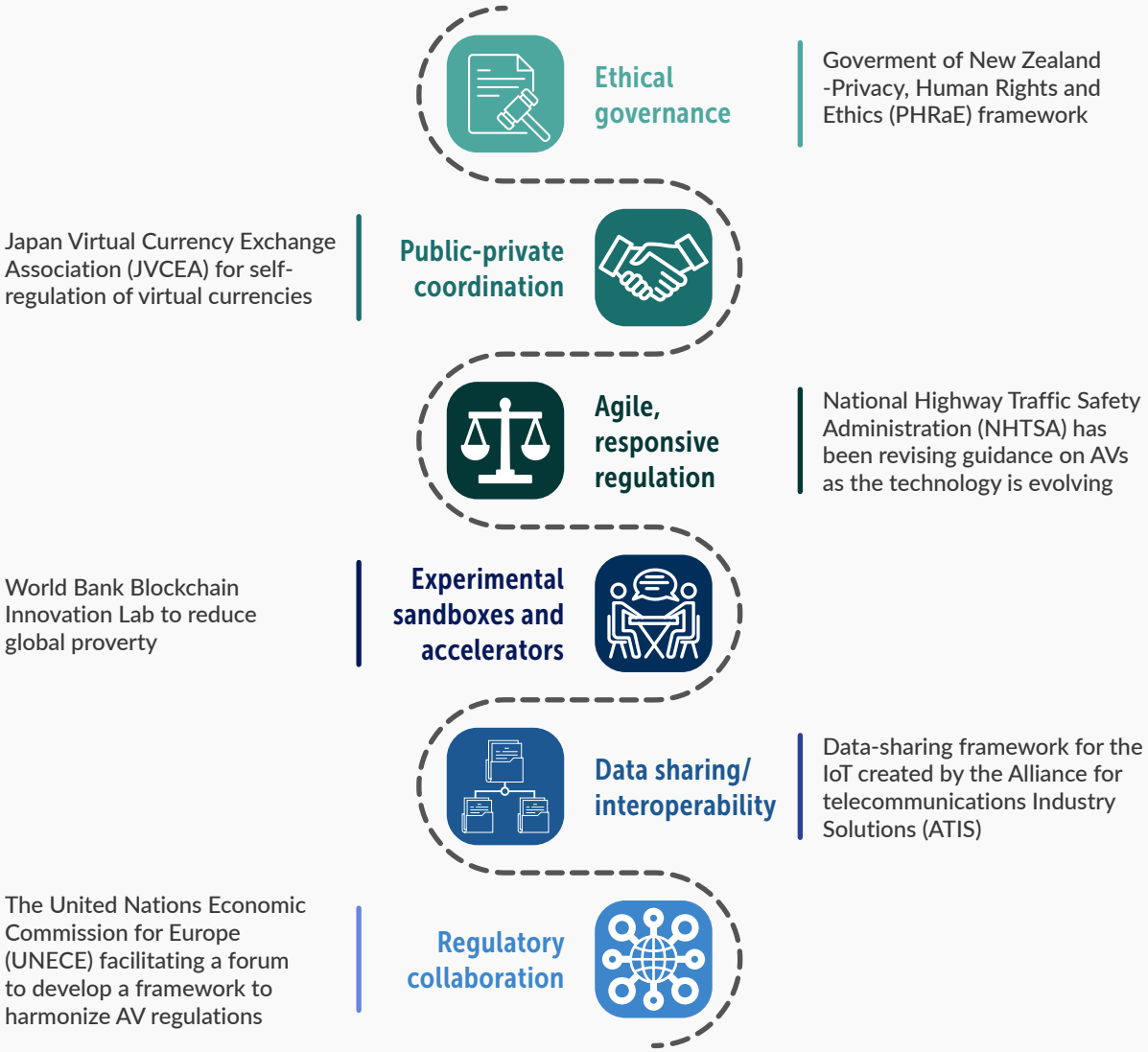


Figure 12.
WEF overview of innovative governance frameworks for emerging technology

AI Liability Insurance and Consortia

Recently Ariel Dora Stern and colleagues have proposed the creation of a form of liability insurance for AI products that could accelerate the adoption of effective AI tools.[22] In an earlier article they made the call for treating algorithms like drugs with similar warning systems for adverse events, side effects, as well as standards similar to the pharmaceutical industry with Good Manufacturing Processes (GMP) , Good Clinical Processes (GCP) and Good Laboratory Practices (GLP).[23] Why not Good Algorithmic Practices (GAP)?

These authors note that the slow adoption of AI is partly due to legal uncertainty and unpredictability in healthcare. They argue that AI liability insurance can facilitate adoption of the highest quality AI tools that reduce the margin of error of humans and be in alignment with patient, providers, and healthcare organization leadership interests. Insurance underwriters will only underwrite AI products that are demonstrably safe and follow evidence-based medicine. Virtuous cycles can be created that impact the cost of insurance for the best products leading to increased adoption.

Stern et. al. do not discuss the use of consortia and investment in intangibles to create an ecosystem that may also de-risk new AI innovations to some extent. As argued throughout this report, industry stakeholders must take the lead. Mayo Clinic, Mitre and several academic organizations have created the Coalition for Health AI that is focused on developing guidelines and guardrails for responsible AI in healthcare. This type of consortia partnered with industry and non-profit entities as discussed above with more robust experience in standards and certification development could have the potential to build a more robust ecosystem of trustworthy and safe AI products. There is a great deal of thought leadership and critical thinking about AI/ML that draws upon fields outside of healthcare that should also inform these coalitions. Together, liability insurance and intra-industry consortia will drive more effective AI tools that stakeholders can trust.

Recommendations

The following provides an overview of broader policy recommendations for healthcare organizations.

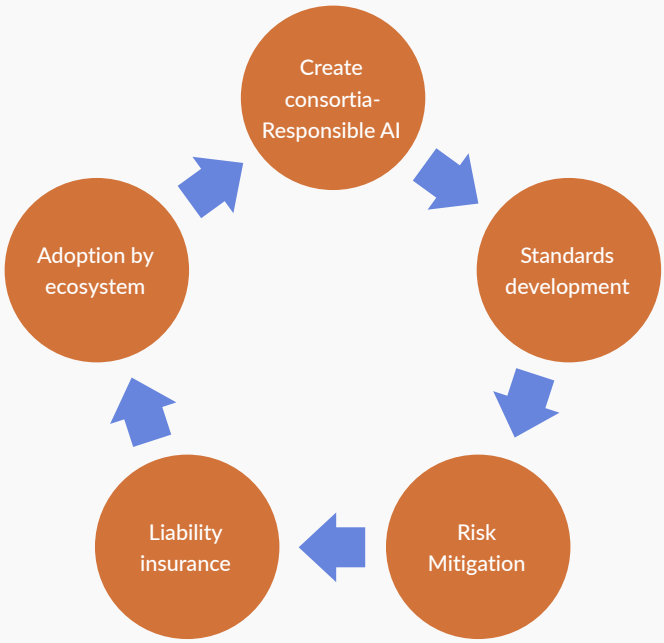


Figure 13.
Virtuous Cycle of standards, Risk Mitigation and Liability Insurance for Adoption of Innovations

- Organizations should create awareness of the implicit biases such as race, gender, sexuality and class and impact on model development and outcomes. Diversity in both socio-cultural backgrounds as well as disciplinary backgrounds is of increasing importance. There are a number of checklists available for walking through critical questions teams need to ask to ensure bias issues are surfaced and addressed, but social science expertise should not be overlooked. Pre-mortem templates can also be used in project plans to guide teams through a process that can facilitate the elimination of bias. Brookings Institute has developed one template that is particularly useful for data science teams.

- To mitigate this risk it is important to thoroughly examine dimensionality of data from the early stages and design the verification and validation processes with sufficient sample sizes. Below are some actions for addressing dimensionality.

- Model training and tuning can utilize models from supervised learning such as logistic regression, decision trees, and k-nearest neighbor classifiers that are more sensitive to dimensionality.

22 https://catalyst.nejm.org/doi/full/10.1056/CAT.21.0242
23 https://qz.com/1540594/treating-algorithms-like-prescription-drugs-could-reduce-ai-bias/

- Passively collected data from sensors will need to assess the context of data collection and issues such as noise or adopt methods that can reduce the dimensionality of data (ie. noise) and may need to consider using "active, maximum performance tasks" rather than passive data to reduce dimensionality issues (Berisha et al).
- Utilize much larger sample sizes for training with high dimensionality data and design representative data sets (but this can be very difficult and may require synthetic data that also has challenges).
- Data scientists should consult this paper and the references to dive deeper into this issue given the number of examples of models not performing well on non-training set populations.
- A comprehensive responsible AI program that can build trust will need to follow best practices across data management, validation/verification, bias audits, checking for drift and model performance over lifecycle, impact assessments for fairness and differential impact on populations and health equity.
- Follow emerging standards for explainability of algorithms without confusing explainability with validation. This is an area in flux so it will be important to stay current with new tools and debates on levels of transparency required by patients and providers to adequately understand how models work and arrive at a particular decision.
- Develop policies for redress and accountability for when models underperform or deliver erroneous results with the potential to harm patients.
- Take opportunities to improve patient literacy of AI. The nature of patient/consumer/citizen engagement will change as AI matures and the public will need a robust knowledge of algorithms to be informed at the level required for informed consent.
- Organize as an industry to improve transparency, quality and the overall market for AI through consortia for best practices that are in alignment with responsible AI practices.

ABOUT THE AUTHOR

Jody Ranck, DrPH.
Senior Analyst, Chilmark Research
jody@chilmarkresearch.com

Dr. Ranck has nearly 30 years of experience working in the global health arena and has helped lead a number of major health technology initiatives throughout his career. Author of two books on digital health, he is a globally recognized thought leader on digital health and has been listed in the "Always On" top 100 minds in Global mHealth (2013). His past clients have included Humana, TM Forum, CLSA, T-Systems, Stanford University's School of Medicine, UC Berkeley, the UN, and ARM to name a few. He has been a frequent advisor to large healthcare companies and startups focused on providing more patient-centric care and transitioning to value-based care. In the past he has been appointed as a member of an Institute of Medicine Committee on ICTs in global health/violence prevention and helped launch a major global eHealth initiative with the Rockefeller Foundation. He has been a frequent keynote speaker at health IT conferences and recently organized and chaired the Healthcare Blockchain Summit (2017-18).

Jody has written and worked extensively on mobile innovations, the Internet of Things (IoT), wearables, blockchain and the analytics market in healthcare. He is also working with cutting edge startups on next generation biosensor platforms, patient generated data for clinical research, and emerging blockchain applications in healthcare. His education includes a Doctorate in Public Health (University of California, Berkeley), MA in International Relations and Economics (Johns Hopkins University) and a BA in Biology (Ithaca College).

APPENDIX: CASE STUDIES FOR BUILDING RESPONSIBLE AI

Case Study 1
Design and Evaluation of AI – Including the Users in Project Design

Sepsis Predictive Model Development at UCHealth

Designing AI based services also requires substantial work in designing the overall service to fit with clinical workflows. Even technical aspects of a model can change when feedback from clinician users may require modifications to the model, but also the service that wraps around the model. We provide an example of CT Lin's work with a sepsis model at UCHealth.

When CT Lin's team began looking for problems that could be addressed with machine learning predictive models they began with a number of questions that could be asked to target a specific problem amenable to ML insights. First, is it a problem that prediction is possible to deploy because not all medical issues have the same level of predictability. If the problem is predictable; are the data available? If insights are created are there effective actions that can be taken? Is there a clear view of the outcomes and dependent variable(s)? Are there passionate, engaged leaders among users who will become operational leaders?

The data science team met with doctors, computational experts, EHR architects and sorted through 3 years of EHR data after deciding to focus on sepsis based on the filters above. From their discussions they initially decided that sensitivity of the model was the most important feature so that they would not miss a single case of sepsis and they began building a model based on ensemble methods for decision trees. Emphasizing sensitivity meant sacrificing some accuracy on the signal/noise front. This project was also launched before UCHealth implemented Epic's Cognitive Computing Platform.

The first trial of the model demonstrated 93% sensitivity, 85% specificity and utilized a dashboard with a color-coded score for non-ICU patients. After deploying the dashboard with nurses and assuming that they had received buy-in from the users during the exploratory phase, they soon realized the system was not being utilized by nurses. The reason was that the nurses had too many devices and dashboards. The system sent alerts 8-12 hours in advance of onset of symptoms with a high alert volume that amounted to 61 alerts for only 1-2 true cases that advanced to sepsis.

After 6 weeks the results showed no change in the number of sepsis cases and no change in the speed of pharmacy orders that should have been triggered by the alerts. Clinicians were simply too busy to even look and alerts often came after they already knew which patients were at risk of sepsis. The low signal/noise ratio (30:1) resulted in bedside fatigue and the alerts came too late. What they really needed was a 12 hour window to alert rather than 8. The team went back to the drawing board to re-design the system.

The second trial was far more successful and included the use of a Virtual Health Center (VHC) that would act as the monitor across 12 hospitals and had a team with a critical care physician and nurse. The VHC also had access to video of patient rooms, Epic EHR chart, communications to bedside nurses and authority to place orders to pharmacies and escalate care. A critical component was to recognize that the sepsis bundle (algorithm) had over 100 tasks so they needed a focused algorithm deployed for a small, highly trained virtual team responsible for a centralized response rather than the decentralized approach of the first trial.

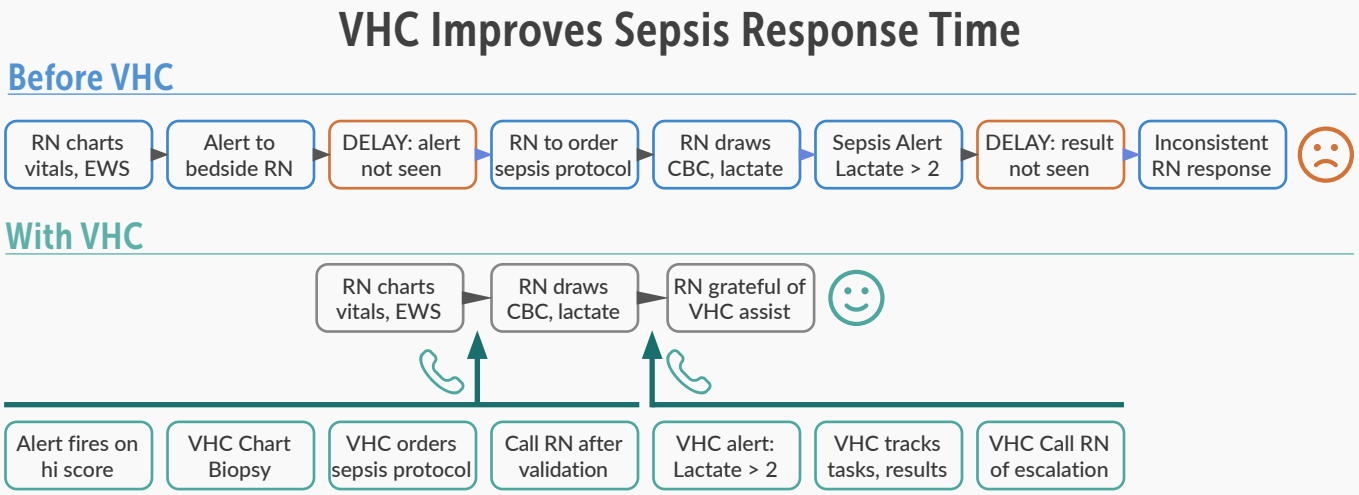


Figure 14.
Comparison of first and second trials (pre/post VHC) Source: CT Lin

The results of the second trial were much better. In the first trial the average time to fluids and 49 minute faster time to antibiotics. Having the VHC team filter out false positives and a better signal to noise ratio dramatically improved response times. In terms of impact on outcomes the 49 minute response time for antibiotics saves 211 lives per year!

Lessons Learned: Hard Questions for Proposed ML projects

	ML Project Hard Questions	Details
Should we do this?	WHY this project?	Describe the Strategic Priority. Is there operational an clinical leadership support and resources?
	WHO is the champion?	Who is the passionate clinical champion? Does this person have clinical and operational support?
	WHAT are you wanting to predict?	Have you defined your outcome / dependent variable?
	WHERE is the data?	Is the data in the EHR? If not, can we access?
	> Is data timely or delayed?	Is the data timely (e.g. RN scribbles vitals on paper and charts too late, at end of shift?)
	> How much data is missing?	Is there consistent data for all patients? (eg not mandatory charting)
	WHO are the patients?	What is the target patient population?
	WHAT is the action?	Is there a clear action you can take on the prediction?
How do we design?	WHEN: how far in the future?	How far in advance do you need prediction?
	HOW MANY patients?	How broadly will you implement your project? One unit? 50? Do operational leaders agree?
	WHO will act on prediction?	Be careful: bedside clinicians are already busy and may not respond to future predictions.
	HOW sensitive/specific?	Do you prioritize sensitivity (don't miss any cases) or signal/noise (more cases found per alerts)? Risk of overtreatment, team burnout, missing cases, alerting too early, too late?
How do we improve?	WHEN will you know you've succeeded?	What time frame is your pilot test intervention? What are your success metrics?

Figure 15.
Lessons Learned from Trials (Source: CT Lin)

The above chart outlines a number of lessons they learned about the challenges of designing machine learning programs for clinical use. Getting busy bedside clinicians to engage with the tool is not easy and may require some form of facilitation such as the VHC. Lin points to the 80-20 Pareto Rule that states that human factors and social skills constitute 80% of the implementation of new technologies. This case study incorporates lessons learned from design and evaluation to build more trustworthy models.

Case Study 2
RPA and Data Integrity Case Study

A growing number of remote process automation (RPA) vendors are emerging to automate processing and analysis of various types of health data, from revenue cycle management to even clinical data. These vendors promise efficiencies that can reduce administrative waste and costs and also have the potential to improve patient experiences by cutting delays caused by errors in paperwork, for example. In general, these vendors are providing an important service and innovation that addresses serious pain points in the system.

However, there are growing concerns about what happens when RPA goes awry? Who is held accountable and what are the risks to the institution, to the patient, or for the provider? The biggest risk posed by RPA applied in healthcare so far appears to be focused on the issue of data integrity. Bots can be run on different types of data to process and analyze this data. However, on occasion the bots can corrupt data in data lakes or databases. The risks to patients are less, perhaps, in revenue cycle management, but on the clinical record side this poses a serious patient safety issue.

In our research for this report we have learned that EHR vendors have had experiences with bots that in a matter of minutes have corrupted tens of thousands of pharmacy records of patients. Damage of this magnitude is expensive and time consuming to repair. Imagine having to flag 50,000 patient records so that entire medication histories of those patients are double checked on the next clinical encounter. Just one episode like this can destroy trust in RPA from EHR vendors, hospital systems and other vendors using the RPA service.

Mitigating the risk involved with bots and data integrity really requires cooperation across industry and competitors to resolve. In the current situation an EHR vendor could refuse to allow an RPA vendor to access medical records. In theory this is possible but EHR vendors risk charges of “data blocking” from the ONC. Due to interoperability requirements across EHRs the problem can spread from one EHR vendor to another and cause additional damage. One vendor’s corrupted medical data records can become another’s problem. Cooperation across the health IT vendor landscape is necessary.

This is an important data integrity issue that can cause reputation harm to both EHR and RPA vendors in addition to posing a patient safety issue. In order to build and maintain trust the ONC can theoretically drive standards and norms creation that ensure the highest level of safety and quality for RPA bots. However, at present the ONC has not issued any guidance so the responsibility of addressing the issue will fall on vendors.

A solution can be framed by consideration of data integrity, patient safety issues for AI applications and health data as a kind of public good across the health IT ecosystem. Both AI vendors and EHR vendors would be wise to help facilitate a third space or non-profit entity that would work with these vendors and organizations such as IEEE/ISTO to develop norms and standards for RPA (bots) and rules of the road that could also govern the issue of responsibility when harm is caused to medical records. Who has the fiduciary responsibility to cover the financial costs of repairing the damage and contacting patients and physicians whose records were damaged? The conclusion of this report details our recommended approach for consortia that could address this issue.

1 Beacon St, 15th Floor
Boston, MA 02108
www.ChilmarkResearch.com
info@chilmarkresearch.com
PH: 617.615.9344