AI Governance at Stanford

David Magnus, PhD
Thomas A. Raffin Professor of Medicine and Biomedical Ethics, Professor of Pediatrics
Associate Dean of Research
Director Emeritus, Laure J. Girand Center for Biomedical Ethics
Stanford University

Governance of Clinical Applications

NEJM Catalyst

JOURNAL E

ARTICLE

Standing on FURM Ground: A Framework for Evaluating Fair, Useful, and Reliable AI Models in Health Care Systems

Authors: Alison Callahan, PhD, MIS , Duncan McElfresh, PhD, Juan M. Banda, PhD, Gabrielle Bunney, MD, MBA, Danton Char, MD, Jonathan Chen, MD, PhD, Conor K. Corbin, PhD, +16 , and Nigam H. Shah, MBBS, PhD Author Info & Affiliations

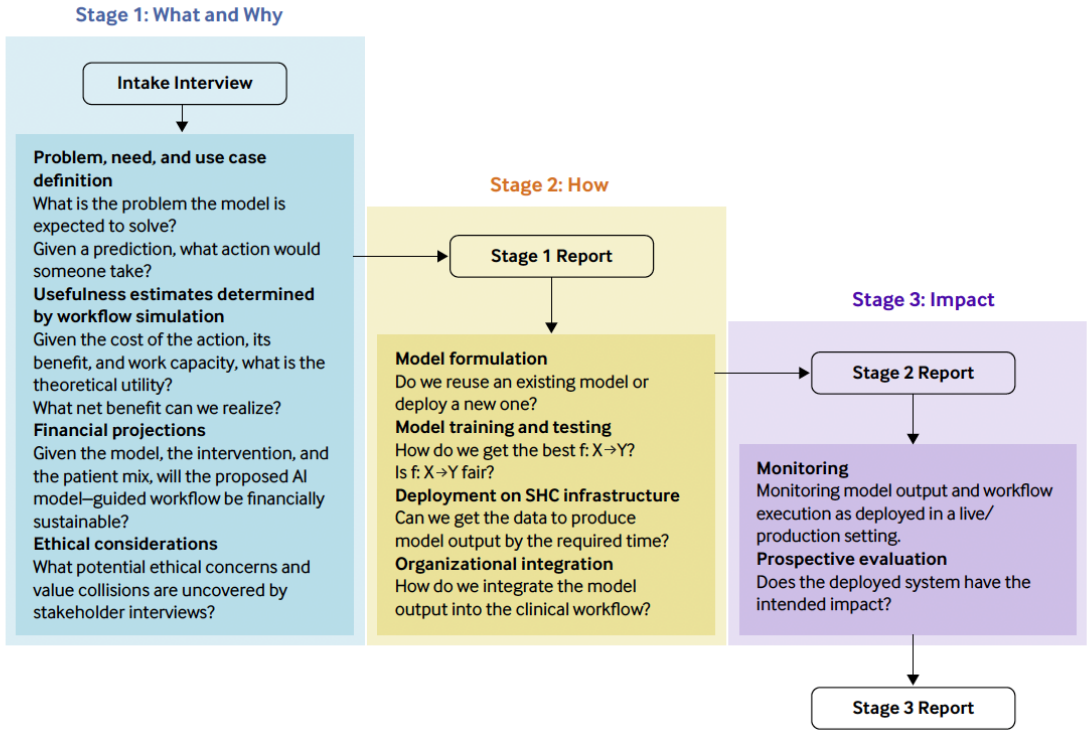
Published September 18, 2024 | NEJM Catal Innov Care Deliv 2024;5(10) | DOI: 10.1056/CAT.24.0131

VOL. 5 NO. 10 | Copyright © 2024

FIGURE 1

The Fair, Useful, and Reliable AI Model Assessment Process

Each Fair, Useful, and Reliable AI Model (FURM) assessment consists of three stages: evaluating the what and why behind a particular AI use case; how a given AI model will be formulated, evaluated, and integrated into a given health care system workflow; and the impact of the proposed implementation. Each of these three stages has multiple components (as shown in the boxes), each addressing a different aspect of the use case. After each of the three stages, a report is prepared, which recommends whether or not to proceed.



AI = artificial intelligence, SHC = Stanford Health Care.

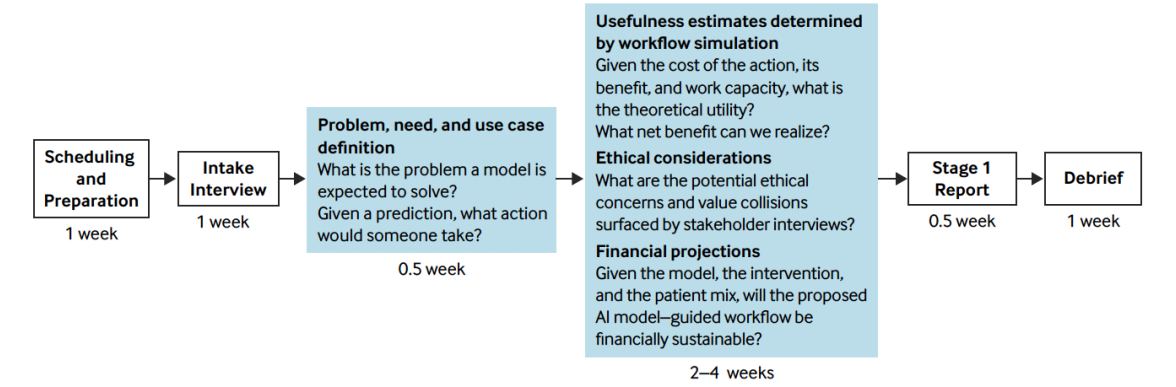
Note: f: X → Y denotes a function that maps some values of X to some values of Y.

Source: The authors

NEJM Catalyst (catalyst.nejm.org) © Massachusetts Medical Society

A Fair, Useful, and Reliable AI Model Assessment: An Estimated Timeline for Stage 1

This figure depicts the estimated time required to complete each component of a Stage 1 Fair, Useful, and Reliable AI Model (FURM) assessment and the application of Little's Law to arrive at a desired capacity of two simultaneous FURM assessments, enabling the completion of one FURM assessment per month, with the ultimate goal of seeing three to five potential AI model–guided workflows through to deployment in the health care system per year. The time spent by the Data Science Team on each request is approximately 6–8 weeks.



Little's Law: $L = \lambda \cdot W$

L – average # of items in queuing system – # of simultaneous FURM assessments
λ – average # of items arriving in the system per unit time – # of assessment requests
W – average waiting time an item spends in queuing system – time for one assessment from start to finish

Fix $\lambda = 1$ per month, and $W = 2$ months

↓

$L = 1 \cdot 2 =$ **2 simultaneous FURM assessments**

FURM = Fair, Useful, and Reliable AI Model.

Source: The authors

NEJM Catalyst (catalyst.nejm.org) © Massachusetts Medical Society

Stanford Health Care

Every AI system proposed for deployment has a required FURM assessment

Ethics team carries out interviews with stakeholders and conduct ethical analysis for each application

Committee that reviews all ethical assessments and interview summaries

AI Research

Limitations of Human Subjects Protection

Focus on Participants

- Risks are reasonable in relation to anticipated benefits
- Risks are minimized as much as possible (consistent with sound research)
- Informed Consent
- Equitable subject selection (at least in principle)

What IRB's don't do

- Common Rule 45 CFR 46.111 (a)(2)
- "The IRB should not consider possible long-range effects of applying knowledge gained in the research (e.g., the possible effects of the research on public policy) as among those research risks that fall within the purview of its responsibility."
- Downstream impact of AI is outside of the purview of the IRB and REC

Ethics and society review: Ethics reflection as a precondition to research funding

Michael S. Bernstein[a], Margaret Levi[b,c,1], David Magnus[d], Betsy A. Rajala[b], Debra Satz[e], and Charla Waeiss[b]

[a]

ESR Process


Grant application and ESR statement submitted to funding program
Merit review by funding program
ESR panel review
Recommendatic to funding program
Feedback and iteration
Name the risks, articulate principles for mitigation, instantiate those principles in the research design
Interdisciplinary panel, including Anthropology, Communication, CS, History, MS&E, Medicine, Philosophy, Political Science, and Sociology
Feedback raises other risks, suggests mitigations, identifies stakeholders, & etc.

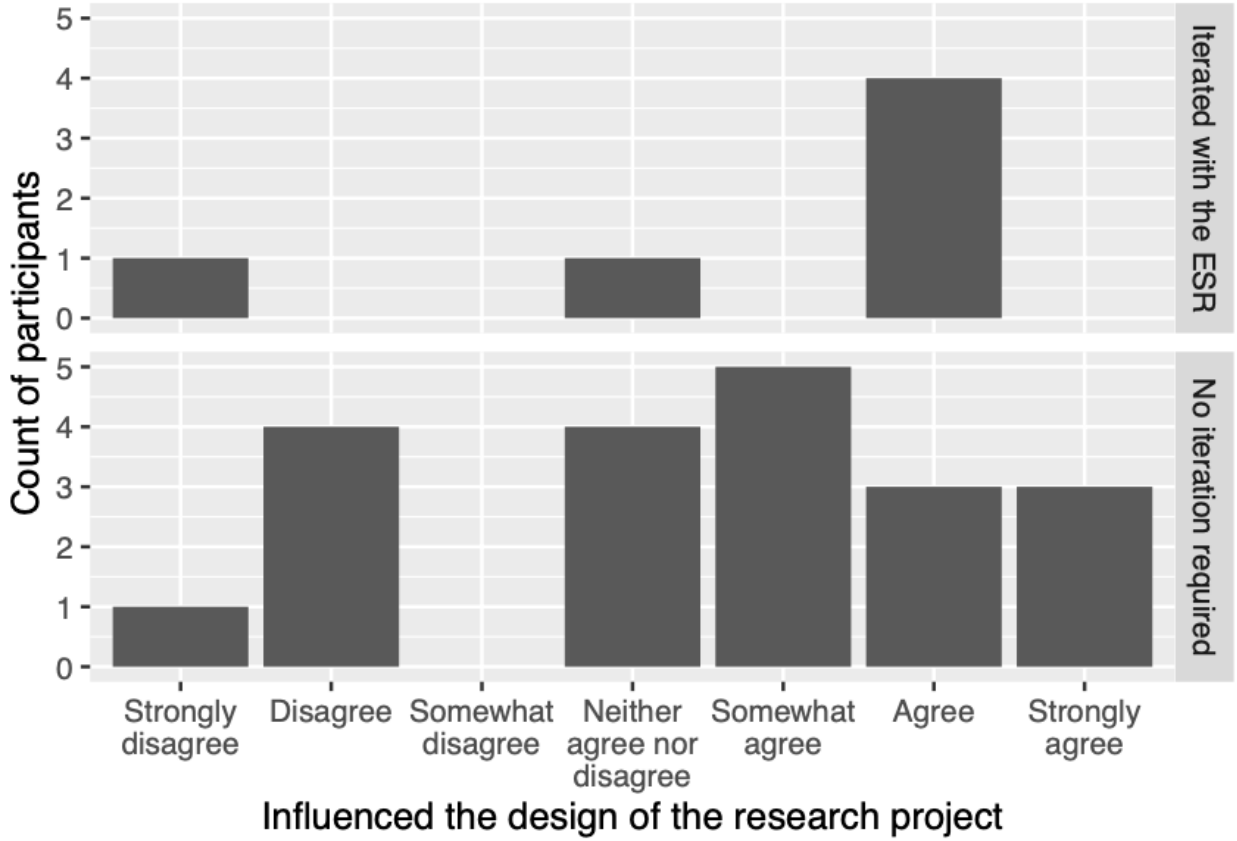


Figure 4: 67% of researchers who iterated with the ESR, and 58% of all researchers, felt that the ESR process had influenced the design of their project.

"The ESR requirement ... led me to engage with my co-PI ... because, as a psychologist, I ... wasn't aware of some of the potential ethical implications that this ... AI work may have, and it helped me to engage with my co-PI as part of this requirement."
— PI, Social Science

"In fact, **we might flip our whole research approach** to being about privacy. [The] pretty strong reaction from the [ESR made] us rethink, to lead with ... privacy. ... **We don't have answers yet, but ...** it's definitely helped us think about a better way to approach the research, how we're doing it and how we're talking about it.
— PI, Engineering

ESR common issues

Dual Use

Representativeness

Stakeholder involvement in design

Harms to subgroups

The New York Times

Account

A.I. Bots Told Scientists How to Make Biological Weapons

Scientists shared transcripts with The Times in which chatbots described how to assemble deadly pathogens and unleash them in public spaces.

nature machine intelligence

Perspective

https://doi.org/10.1038/s42256-024-00926-3

Toward a framework for risk mitigation of potential misuse of artificial intelligence in biomedical research

Received: 2 April 2024

Accepted: 16 October 2024

Published online: 26 November 2024

Check for updates

Artem A. Trotsyuk[1,17], Quinn Waeiss[1,2,17] ✉, Raina Talwar Bhatia[1], Brandon J. Aponte[1], Isabella M. L. Heffernan[1], Devika Madgavkar[1], Ryan Marshall Felder[3], Lisa Soleymani Lehmann[4,5], Megan J. Palmer[6], Hank Greely[7], Russell Wald[8], Lea Goetz[9], Markus Trengove[9], Robert Vandersluis[9], Herbert Lin[10,11], Mildred K. Cho[1], Russ B. Altman[6,12], Drew Endy[6], David A. Relman[10,13,14], Margaret Levi[10,15], Debra Satz[16] & David Magnus[1]

The rapid advancement of artificial intelligence (AI) in biomedical research presents considerable potential for misuse, including authoritarian surveillance, data misuse, bioweapon development, increase in inequity and abuse of privacy. We propose a multi-pronged framework for researchers

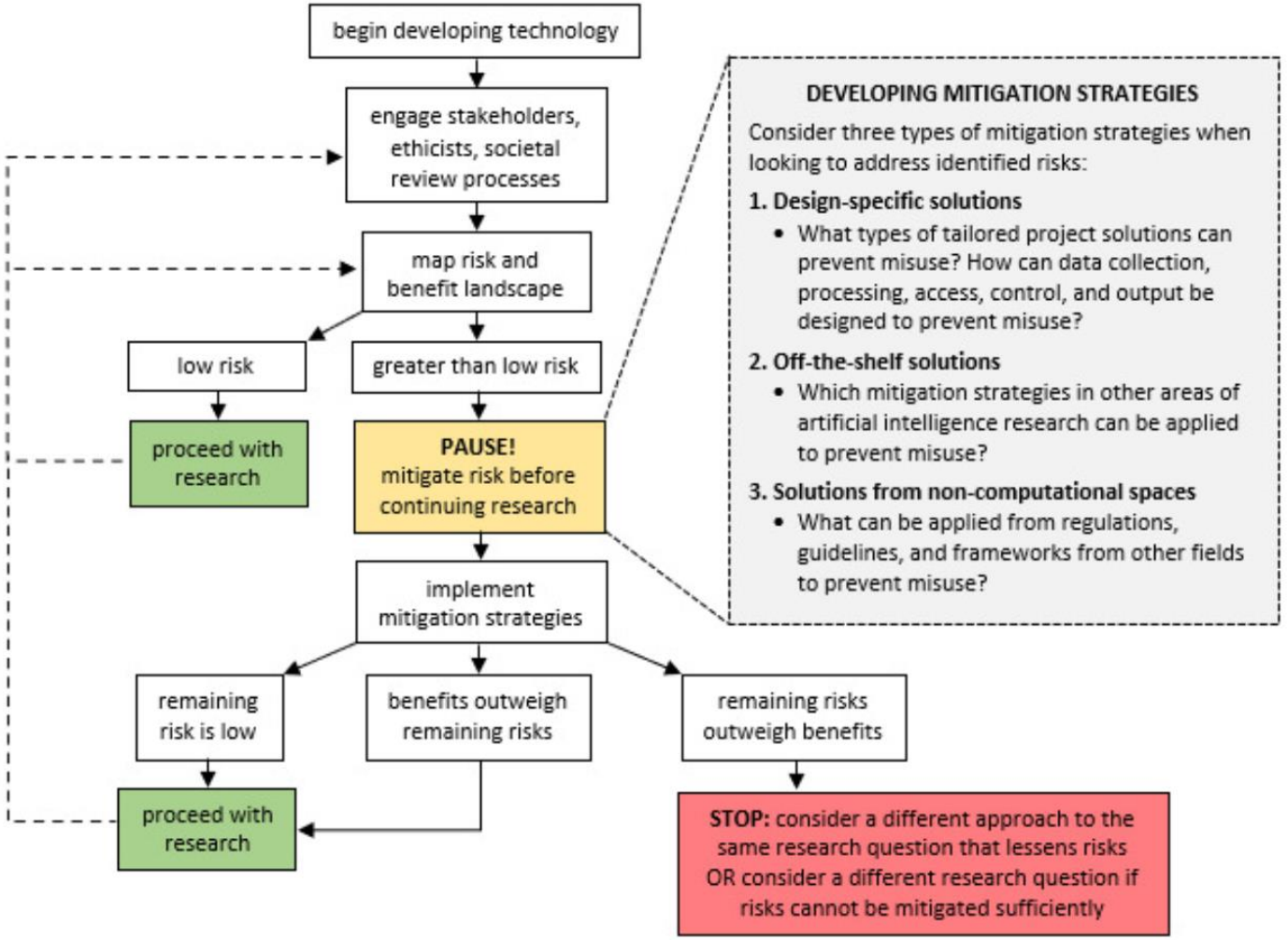


Figure 1. A framework for determining mitigation strategies for stratified risk of AI in biomedicine.

Table 1: Design-specific strategies for mitigating misuse risks

Mitigation strategy	Purpose
data fiduciary	prioritize user autonomy by allowing users to customize which data are collected, how it is used, who can access it, etc[31].
digital watermarking	enable researchers to track unauthorized data use or sharing[32]
face-blurring in images and video data	ensure individuals cannot be easily identified in collected data[33]
differential privacy	allow data analysis while ensuring results do not reveal specific information about individual observations in data[34]
homomorphic encryption	allow for computations on encrypted data without need for decryption[35]
federated learning	enable model training across multiple devices/servers while keeping data localized[36]
execute model on safeguards	enable researchers to limit certain types of model use and flag inappropriate use patterns
tailored evaluations	examine how the raw or synthetic data corresponds to target populations and other metrics[37-39]

Mitigation strategy	Purpose
fairness-aware machine-learning techniques	allow researchers to assess whether AI model prioritizes one group over another[49]
adversarial testing	reveal biases within data or model by examining model outputs to range of unanticipated inputs[6,49,50]; compare algorithmic objective function to its intended use[3,51]
red-teaming	experts interact with an AI in attempts to reveal undesirable outcomes[52]
blue-teaming	experts interact with an AI in attempts to resolve undesirable outcomes
AI transparency/explainability methods	enable human intellectual oversight of AI decisions and predictions to provide users with appropriate information on data[53-55]
system of qualified data storage and tracking	provide infrastructure for oversight of access and use of data (e.g., structured access to data from the All of Us Research Program)

Table 3: Strategies from ethical frameworks and regulatory measures for mitigating misuse risks

Mitigation strategy	Purpose
Establish institutional review process for specified research	Process for oversight of research activities with high risk of misuse harms emphasizing communication and safety protocols between researchers, institutions, funding, and regulatory agencies[12,13,59,60]
Provide training and education to researchers and other stakeholders on misuse risks	Improve identification of misuse risks and development of mitigation strategies[12,13]
Establish principles for ethical considerations of AI research in biomedicine	Guide the ethical assessment of misuse risks; principles from synthetic biology include: public beneficence, responsible stewardship, intellectual freedom and responsibility, democratic deliberation, and justice and fairness[59]; trustworthiness[61]
Establish recommendations for weighing misuse risks against benefits of AI in biomedicine	Guide researchers' navigation of misuse risks, how to weigh risks against benefits[60,62], researchers' responsibilities, and reporting requirements[61]