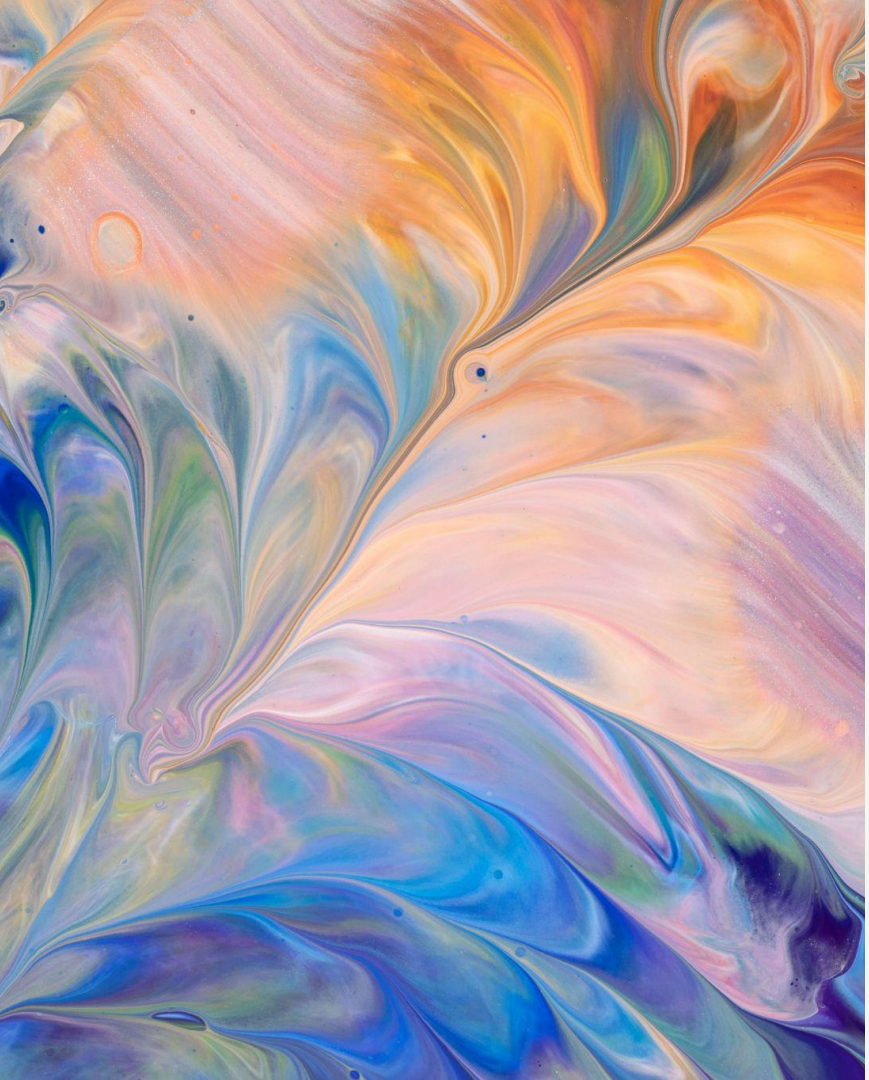


🔍AI / ML - Enabled Medical Devices: Further Challenges that Generative AI Poses

AIRIS 2024

Johan Ordish, Global Head of Digital Health and Innovation Policy, Roche

February 2024 | External Release

Thanks for including me

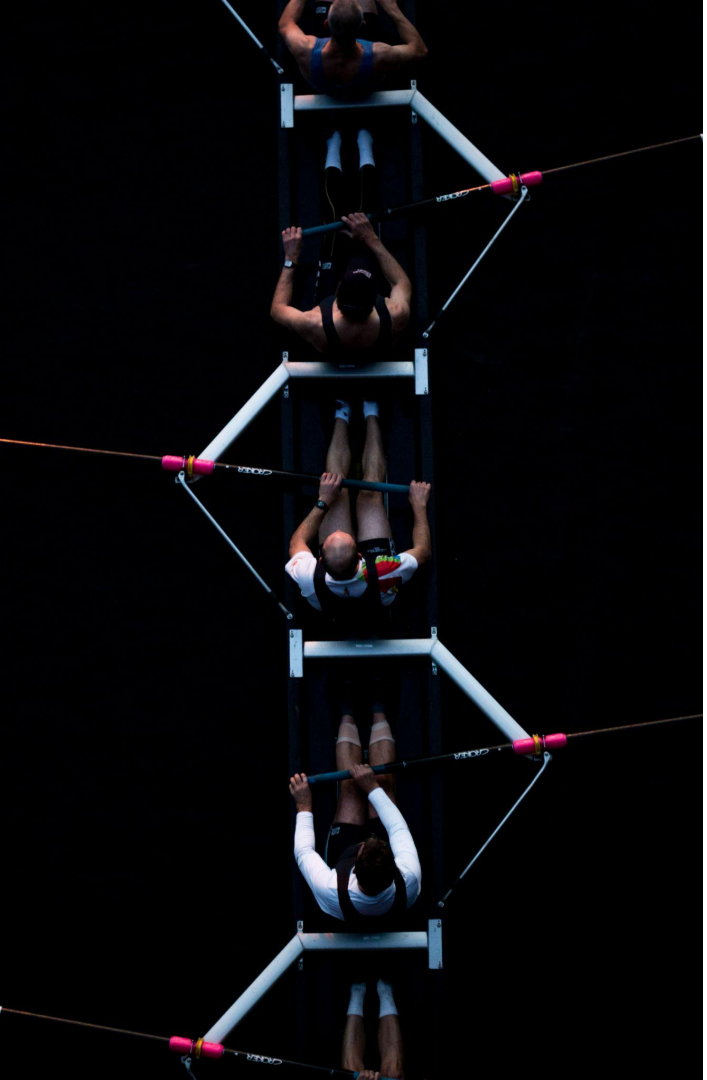
Global Head of Digital Health and Innovation Policy, Roche Diagnostics

Previously, Head of Software and AI at the UK's Medicines and Healthcare products Agency (MHRA)

Honorary Associate Professor, College of Medical and Dental Science, University of Birmingham

By-Fellow, Hughes Hall, University of Cambridge

Table of contents



Digital Health is more important than ever

Advances such as generative AI pose challenges that only international cooperation can overcome

It is extremely encouraging to see such cooperation between MFDS and the US FDA

Only together can we effectively protect patients and ensure access to innovative medical devices

What happened to the AI revolution?

AI promises to intervene across patient pathways, yet the promise of a revolution is yet to come to pass

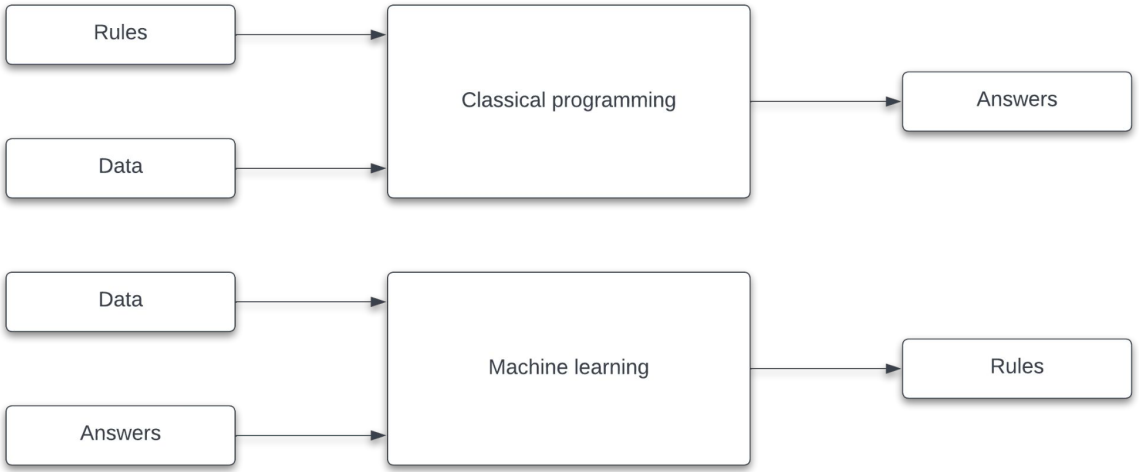
AI has made it to clinical practice but transformative change has been stymied by: difficulties with assuring models, a lack of acceptance in clinical practice, and even basic blockers such as failure to digitise services

📌Three key challenges posed by *discriminative* AI

What's special about AI with respect to medical device regulation?

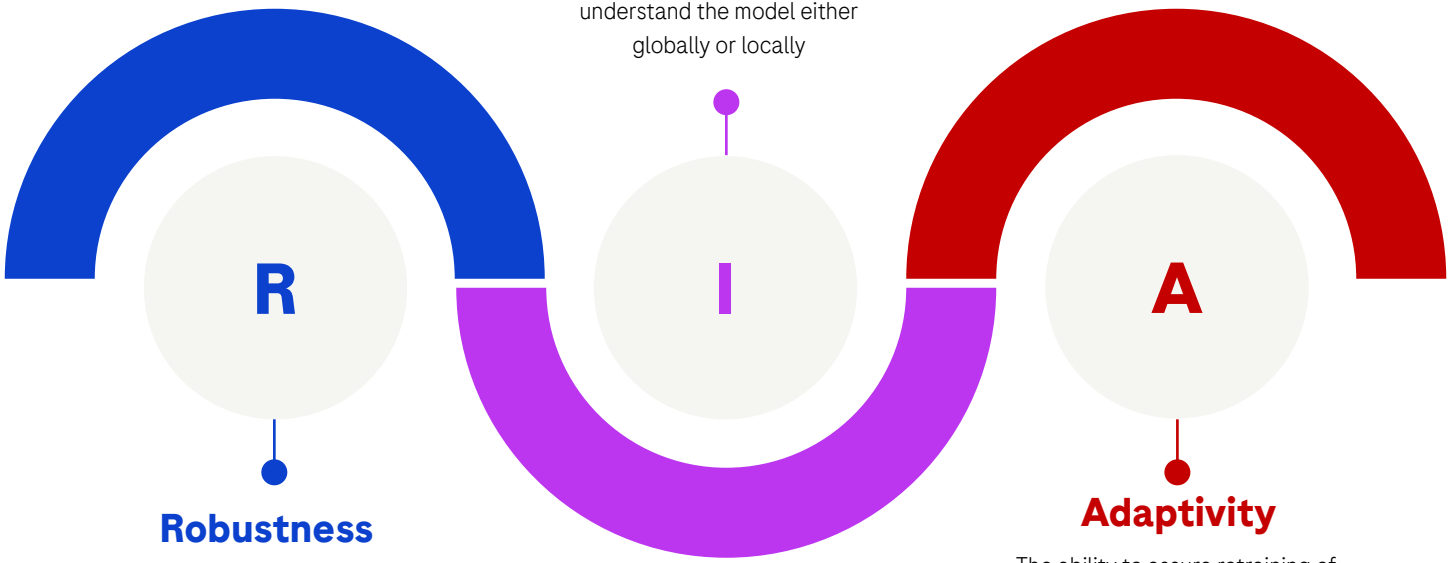
My favoured definition of machine learning in particular comes from Chollet (2018),[1] this definition emphasises that machine learning systems are trained rather than being explicitly programmed

This aspect of machine learning best frames the challenges it can pose for medical device regulation



[1] Chollet, Deep Learning With Python (2018)

Three key challenges for discriminative AI



Interpretability

The ability for intended users to understand the model either globally or locally

Robustness

The ability to maintain performance across the intended populations and deployment locations

Adaptivity

The ability to assure retraining of a model across time

State of the art is progressing to address many of these challenges, for example

Robustness

- Better understanding of what drives data variation
- Better understanding of what representative data means for given intended uses
- Better testing methodologies to signal robustness of models

Interpretability

- Better understanding of what the clinical community wants (usually, transparency)
- Better understanding of what is technically feasible and XAI methods
- Progressing methodologies to analyse the usability of AI with a medical purpose

Adaptivity

- Better understanding of the drivers for data drift
- Crystallisation of predetermined change control plans to assure modifications

As a community: regulators, industry, academia, and clinicians we're addressing the challenges that discriminative AI poses

🚨 Further challenges posed by *generative* AI et al

Roughly, here's what I mean by 'generative AI' in this context

Generative AI (GenAI) is the "use of AI to create new content like text, images, music, audio, and videos" [3]

Gen AI is powered by Foundation Models

Foundation Models are a new paradigm of deep learning that that utilise transfer learning at scale [4]

"Transfer Learning is what makes foundation models possible, but scale is what makes them powerful"[4]

4

On the Opportunities and Risks of Foundation Models

Rishi Bommasani* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kuditipudi Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou Percy Liang*‡

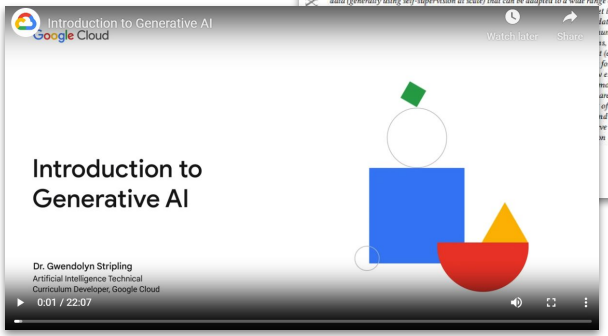
Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

AI is undergoing a paradigm shift with the rise of models (e.g., BERT, DALL-E, GPT-3) trained on broad data (generally using self-supervision at scale) that can be adapted to a wide range of downstream tasks.

*Equal contribution.

arXiv:2108.07258v3 [cs.LG] 12 Jul 2022

3



Three key challenges that GenAI et al can pose for medical device regulation

Robustness[2]

State of the art to measure the performance of *discriminative* AI is likely insufficient for *generative* AI

SaMD Qualification

It is likely that GenAI and Foundation Models pose further challenges in determining the regulatory status of products

SOUP

Fine turning base models for a medical purpose often results in the fine-tuner not having sufficient documentation of the base model

GenAI and SaMD Qualification

It is characteristic of Foundation Models that they utilise Transfer Learning, which is “improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned”[5]

Given the scale of Foundation Models as well as the usage of Transfer Learning, this new paradigm likely further test how we assess what software functions qualify as software as a medical device

There are at least two sub-challenges for qualification:

1. Foundation Models in particular are intended to perform a wide set of functions; they are general purpose models
   - If a Foundation Model performs a broad range of functions and one of those functions is a medical purpose, is that Foundation Model function now a medical device?
2. It is common to finetune a base Foundation Model for more specific tasks, creating a finetuned model
   - Do we assume that the base Foundation Model does not have a medical purpose (per the above problem) and only the fine-tuned model has a medical purpose function? Is this always the case?
   - If so, we still have the SOUP and robustness challenges, which I discuss next

[5] Olivas et al, Handbook of Research on Machine Learning Applications and Trends (2009)

GenAI and Software of Unknown Provenance (SOUP)

As noted, it is common in healthcare to finetune a base Foundation Model for more specific tasks

This creates an opportunity to harness these powerful models for more specific tasks

However, it also creates challenges, for example:

- Often the developer of the finetuned model will not have documentation for the base Foundation Model
- That is, it is likely that there's Software of Unknown Provenance (SOUP) at the heart of the SaMD
- SOUP is common in SaMD but the amount of SOUP and the centrality of that SOUP is poses a challenge, for instance:
  - It may be difficult to deploy sufficient controls for risk management purposes
  - Even if the fine tuning developer had access to documentation, we still likely have state of the art challenges with respect to methods to test robustness in GenAI, which I discuss next

[6] MHRA, Large Language Models and software as a medical device (2023); [7] Gilbert et al, Large language model AI chatbots require approval as medical devices (2023)

GenAI et al and Robustness[2]

GenAI likely also poses challenges with respect to how we measure performance, methods to test robustness, and characterising errors

For instance, consider the problem of sample sizes with respect to functional testing:

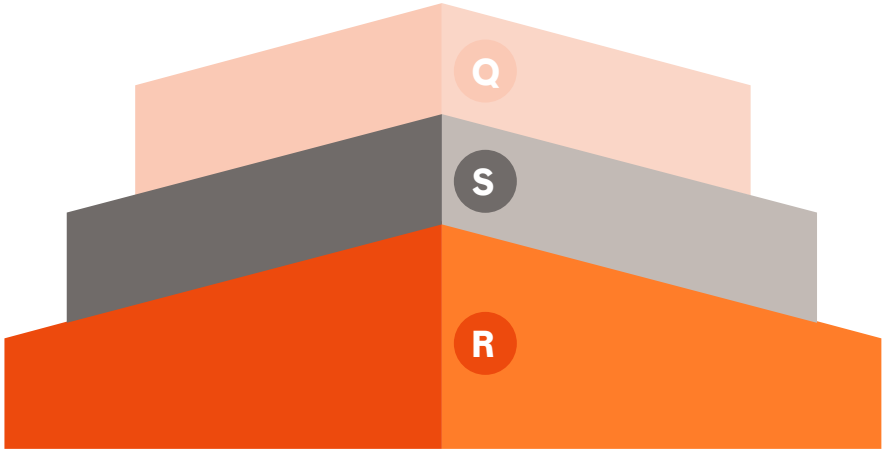
- Functional Testing, that is, pushing inputs in and observing outputs
- Given that it common to have a lack of constraints in terms of: functions, inputs, and outputs, the sample size to provide adequate assurance quickly balloons

In addition, it seems the state of the art to detect hallucinations is yet to crystalise, meaning there's little consensus on how best to detect and characterise errors in these models[8]

We'll likely have to come together as a community to address the issue of GenAI and robustness if they're to be safe and effective for a medical purpose

[8] Tonmoy et al, A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models (2024)

A hierarchy of challenges posed by GenAI



SaMD Qualification

- A known challenge further amplified in GenAI
- Likely solvable via clear guidance

SOUP

- A known challenge but a step change
- Solutions either require disclosure of Foundation Model documentation or new methods to constrain models and control risk

Robustness[2]

- A fundamental challenge to the methods and state of the art for how we measure the performance of SaMD
- Solutions yet to crystalise or are currently beyond our reach

The rise of GenAI requires cooperation across the international community of: regulators, academia, clinicians, and industry

Some of this is already underway with the IMDRF AI/ML-Enabled Gen AI Sub Group

Doing now what patients need next