Unintended consequences of AI

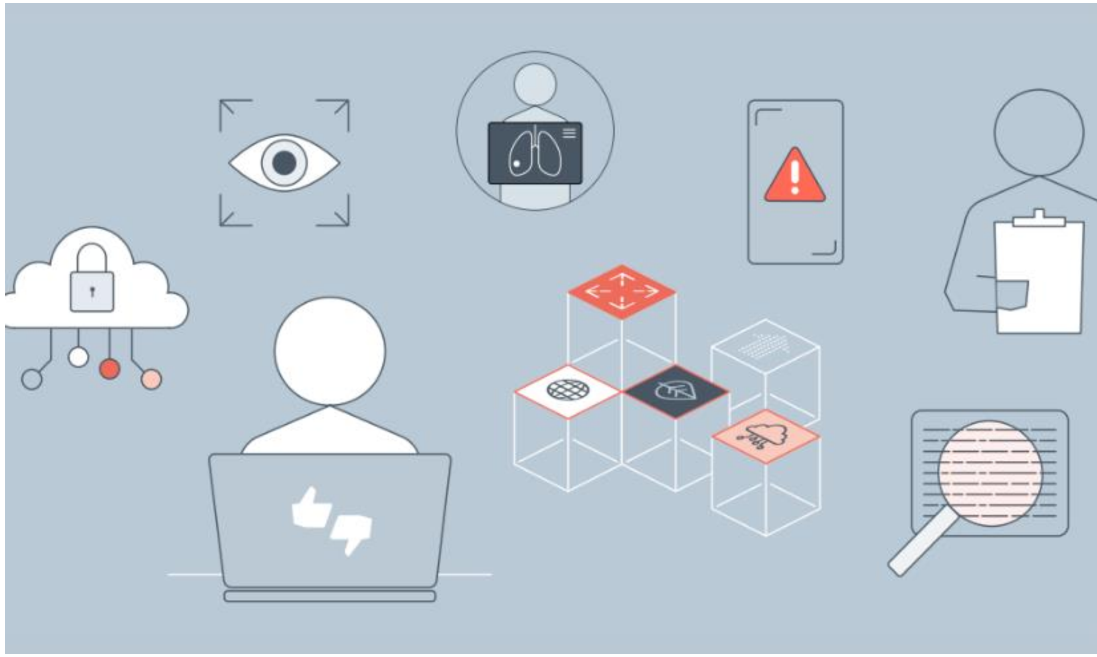
May 2024

Fran Bennett

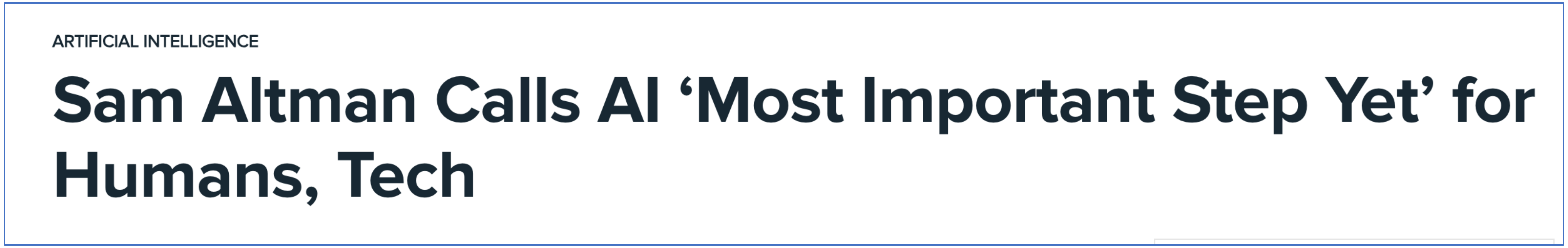
Ada Lovelace Institute

About the Ada Lovelace Institute

The Ada Lovelace Institute is an independent research institute with a mission to make data and AI work for people and society.

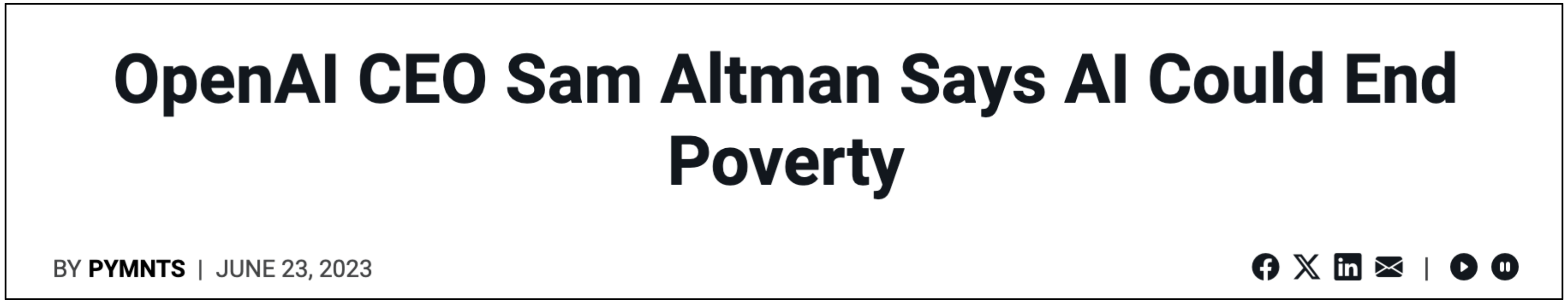


Tina Ricky



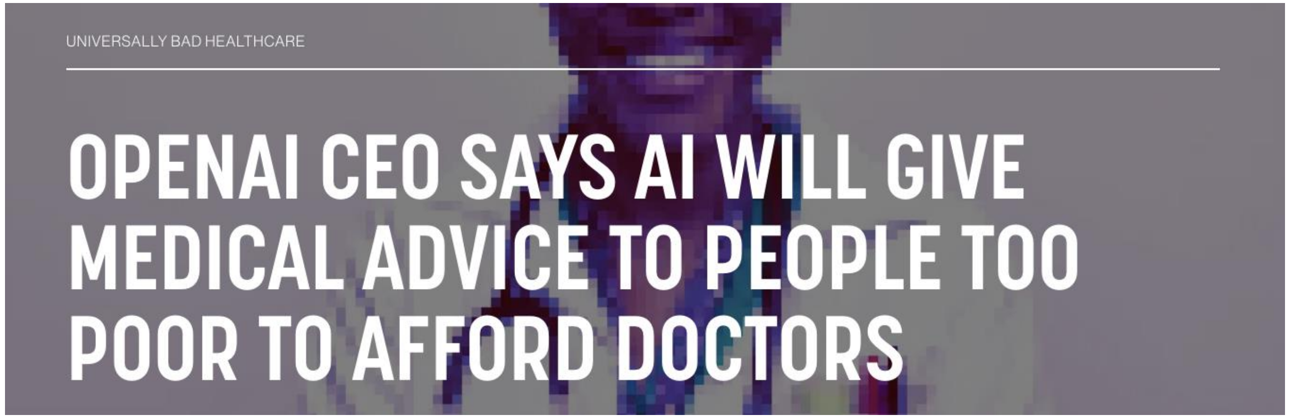
ARTIFICIAL INTELLIGENCE

Sam Altman Calls AI 'Most Important Step Yet' for Humans, Tech



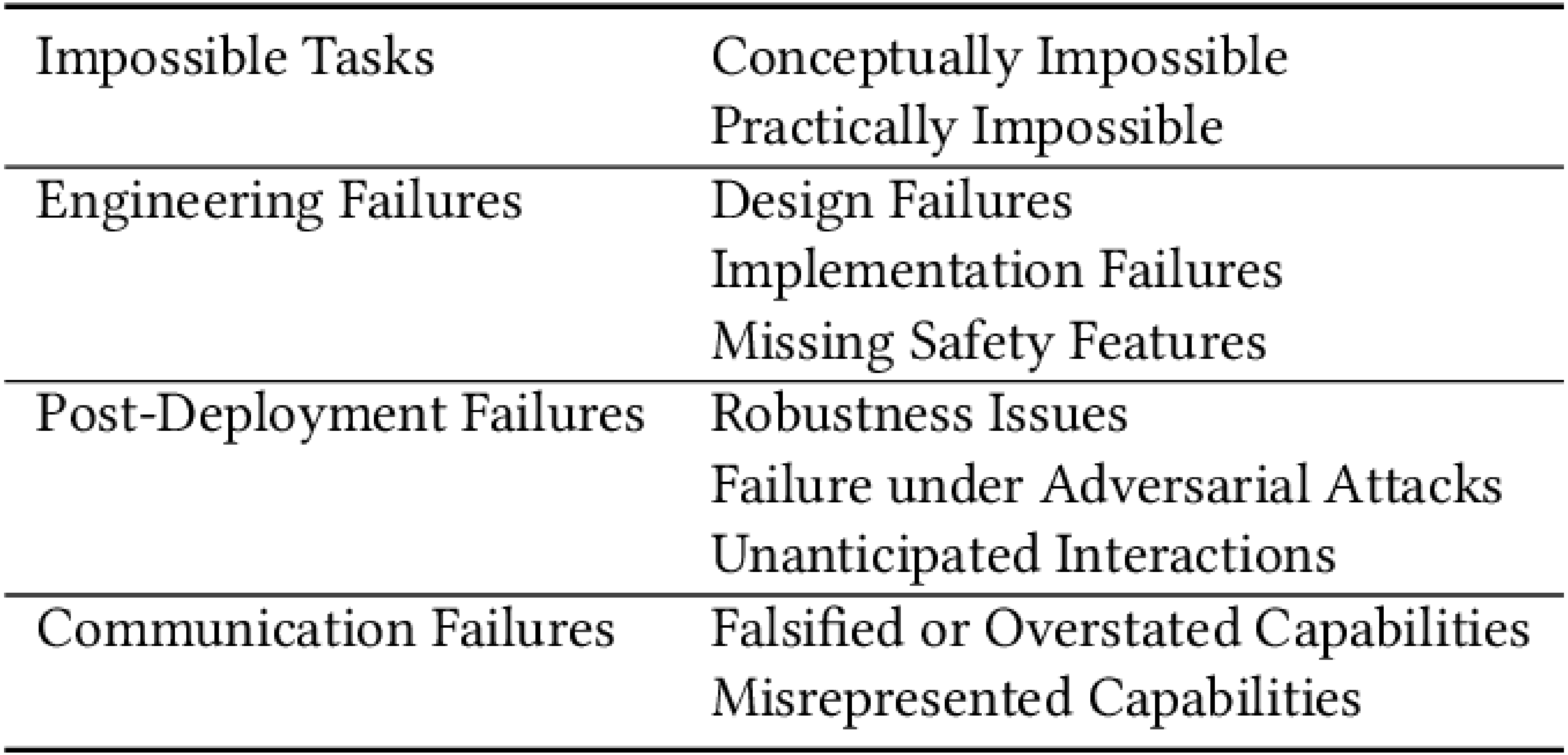
OpenAI CEO Sam Altman Says AI Could End Poverty

BY **PYMNTS** | JUNE 23, 2023



UNIVERSALLY BAD HEALTHCARE

OPENAI CEO SAYS AI WILL GIVE MEDICAL ADVICE TO PEOPLE TOO POOR TO AFFORD DOCTORS

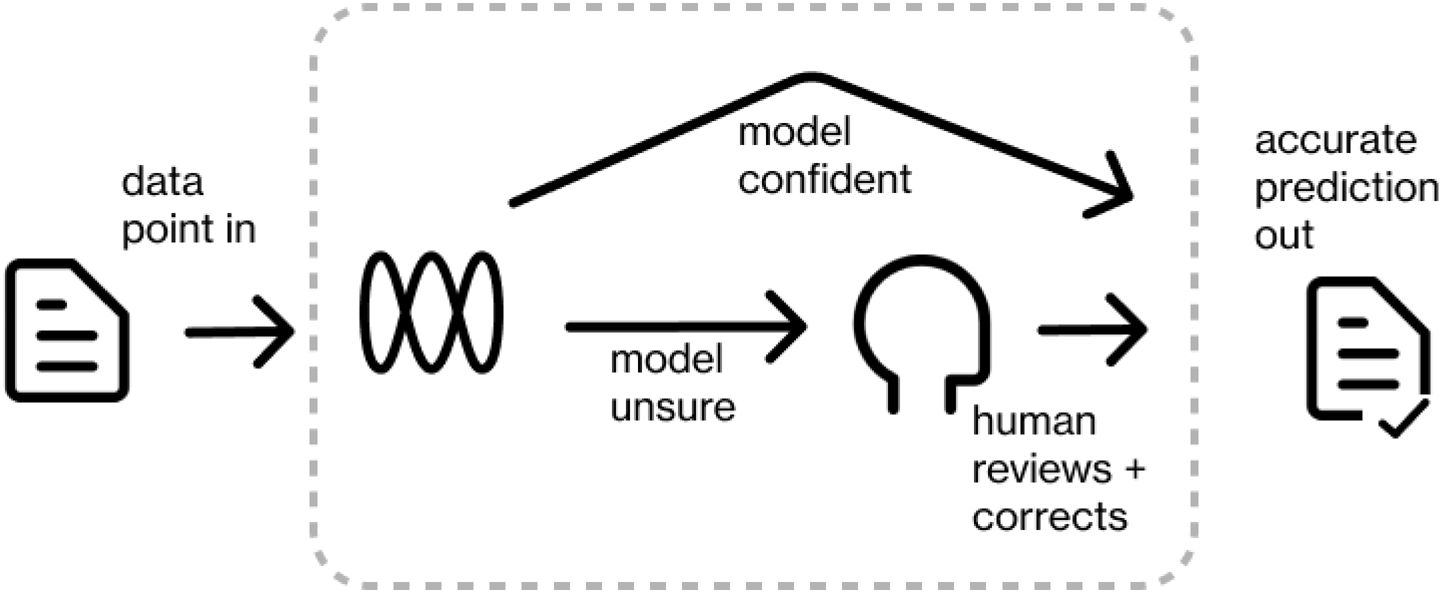


Impossible Tasks	Conceptually Impossible Practically Impossible
Engineering Failures	Design Failures Implementation Failures Missing Safety Features
Post-Deployment Failures	Robustness Issues Failure under Adversarial Attacks Unanticipated Interactions
Communication Failures	Falsified or Overstated Capabilities Misrepresented Capabilities

Source: The Fallacy of AI Functionality, Raji et al, June 2022

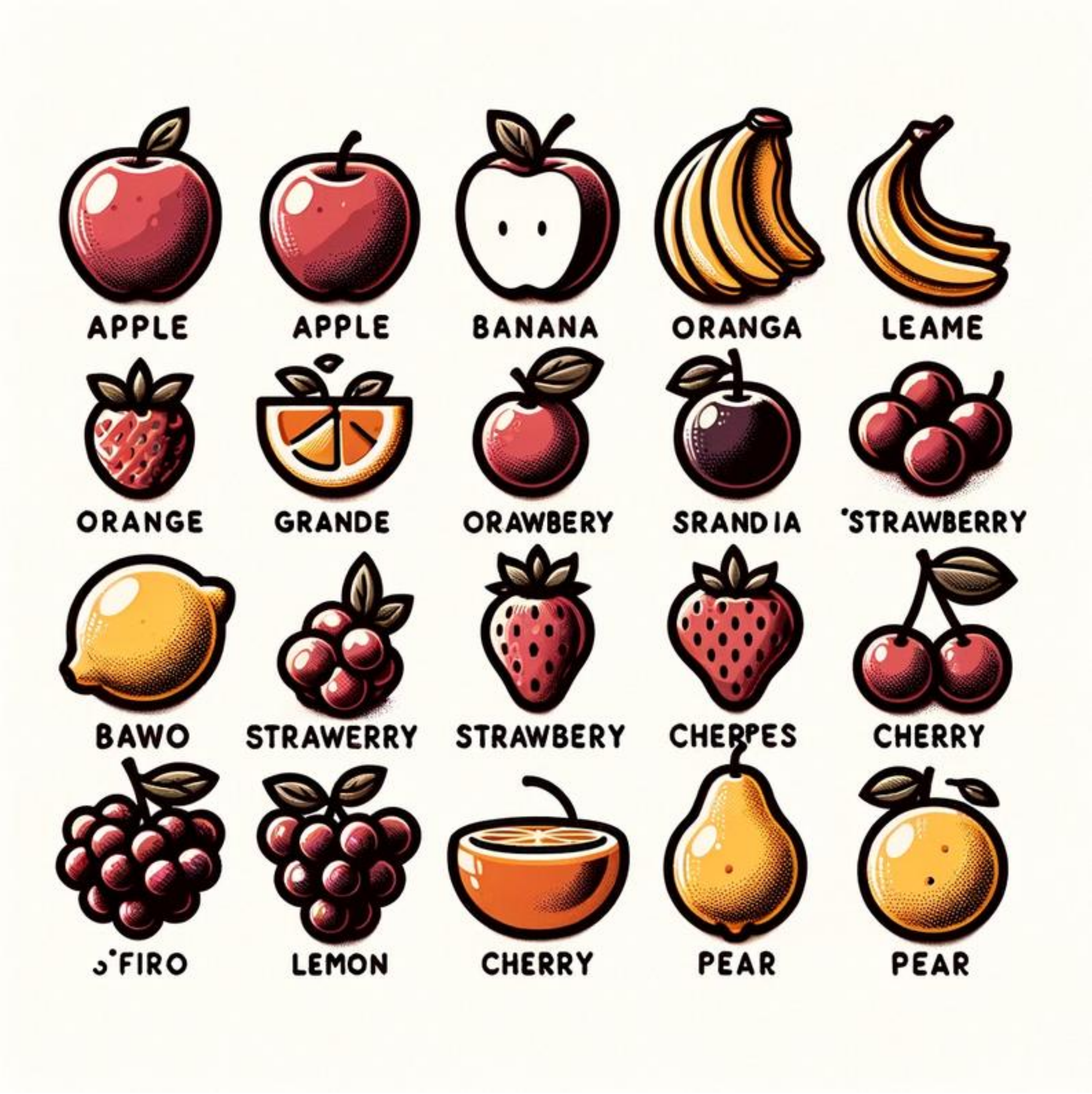
Justice For Subpostm Allia The Fight F Co
SOS: SUPPORT OUR SUB - POSTMASTERS
SOS: SUB
Justice F Subpostm Allia The Fight Co
JUSTICE FOR POST OFFICE VICTIMS
e For masters ce
SOS: SUPPORT OUR SUB - POSTMASTERS

Source: ITV



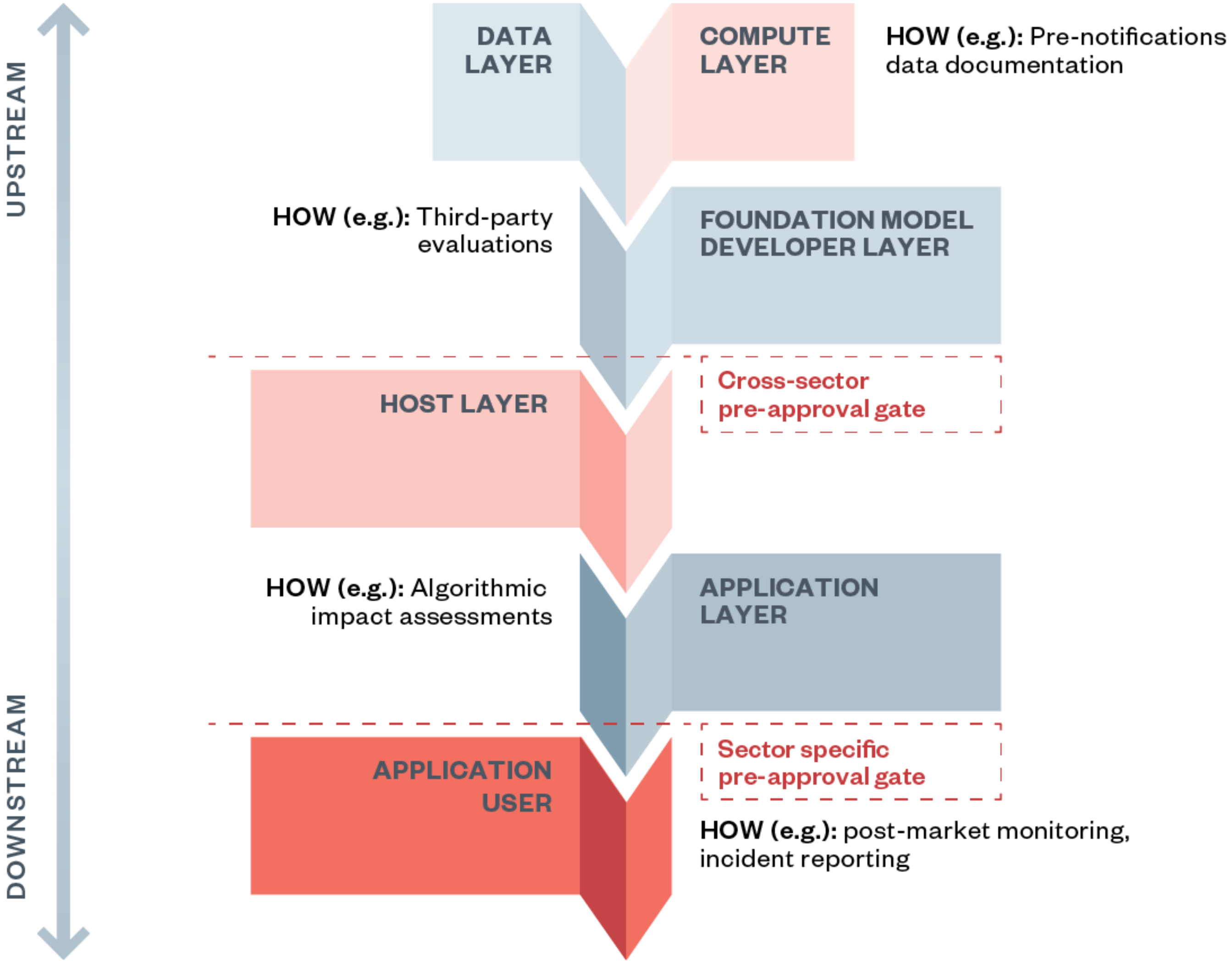
**Workers-in-the-Loop
AI deployment**

Humanloop



Source: AIWeirdness.com / Janelle Shane

Figure 2: FDA-style oversight for foundation models


UPSTREAM
DOWNSTREAM
DATA LAYER
COMPUTE LAYER
HOW (e.g.): Pre-notifications data documentation
HOW (e.g.): Third-party evaluations
FOUNDATION MODEL DEVELOPER LAYER
Cross-sector pre-approval gate
HOST LAYER
HOW (e.g.): Algorithmic impact assessments
APPLICATION LAYER
Sector specific pre-approval gate
APPLICATION USER
HOW (e.g.): post-market monitoring, incident reporting

Source: Ada Lovelace Institute

- **Biases** from the training data or feedback process

- **Black box** and **stochastic** models – can't investigate why they do what they do, and outputs are a little different each time

- **Non-transparent supply chain –** what was the input data? What was the feedback process? Are there extra control layers?

- **Test safety in a crash** – as we do with cars
- **Assign liability for harm** – as we do with food poisoning
- **Hear voices of those affected** – as we do with planning permission
- **Right to contest decisions and request explanations** – as we do with human bureaucracies

Thank you

Fran Bennett, Ada Lovelace Institute
@AdaLovelaceInstitute

Ada
Lovelace
Institute