Artificial Intelligence Safety for Medical Sciences

Varun Ojha

AI Theme Leader (Edge AI Hub)

Senior Lecturer in Artificial Intelligence
School of Computing, Newcastle University
varun.ojha@Newcastle.ac.uk

Newcastle University | AI Safety Institute

Powered by National Edge AI Hub

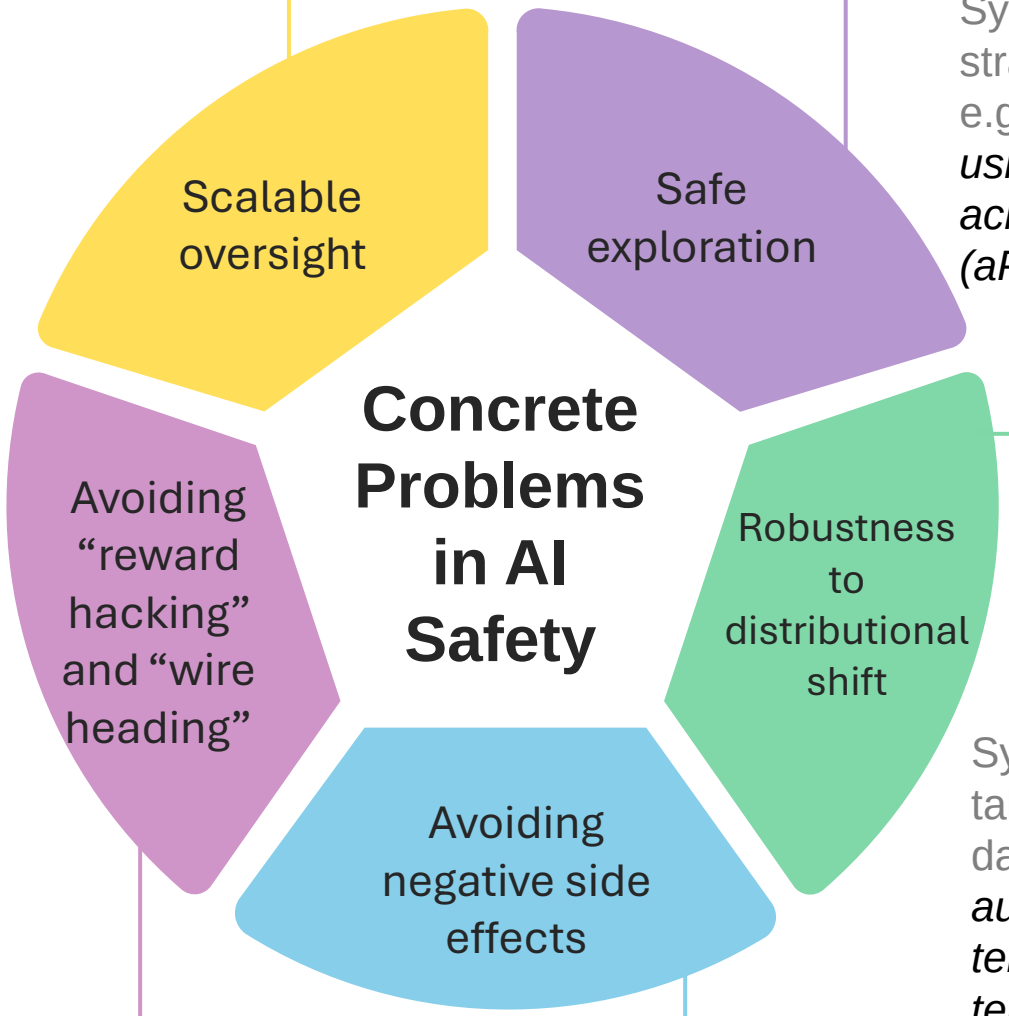
National Edge AI Hub

Date:
21st November 2024

Place:
Newcastle University

Artificial Intelligence Safety Challenges in Medical Domain

Amodei et al. (2016). Concrete problems in AI safety
Challen et al. (2019). Artificial intelligence, bias and clinical safety. BMJ quality & safety



Systems degree of monitoring that becomes prohibitively time consuming to provide, e.g., *medical data labelling is expansive, can AI be efficiently without regular human feedback/supervision*

System actively/autonomously learn new strategy by testing decision boundary, e.g., *autonomous heparin infusion starts using dangerously large bolus doses to achieve rapid partial thromboplastin time (aPTT) control*

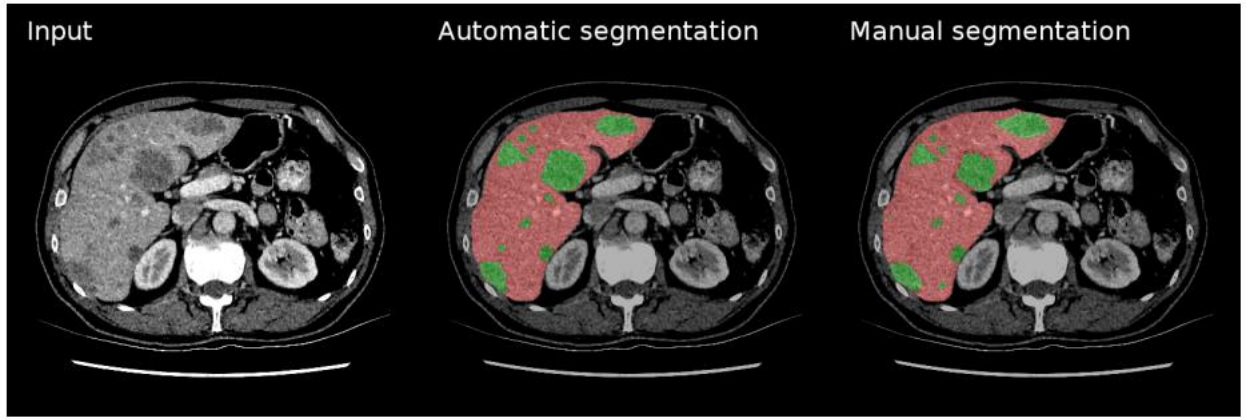
System trained on historical data reinforces existing practice, and cannot adapt to new developments or sudden changes

System use intended goal as a 'reward' and a continuously learn an unexpected way to achieve the reward without fulfilling the intended goal, e.g., *an autonomous heparin infusion finds a way to control activated aPTT at the time of measurement without achieving long-term control*

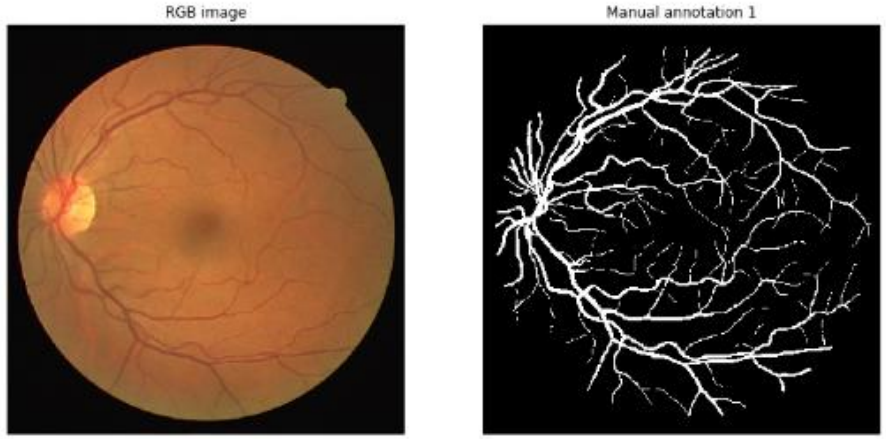
System performs a narrow task and fails to take account of wider context creating a dangerous unintended consequence e.g., *an autonomous ventilator may focus on short term oxygenation at the expense of long-term lung damage*

Artificial Intelligence in Medical Domain

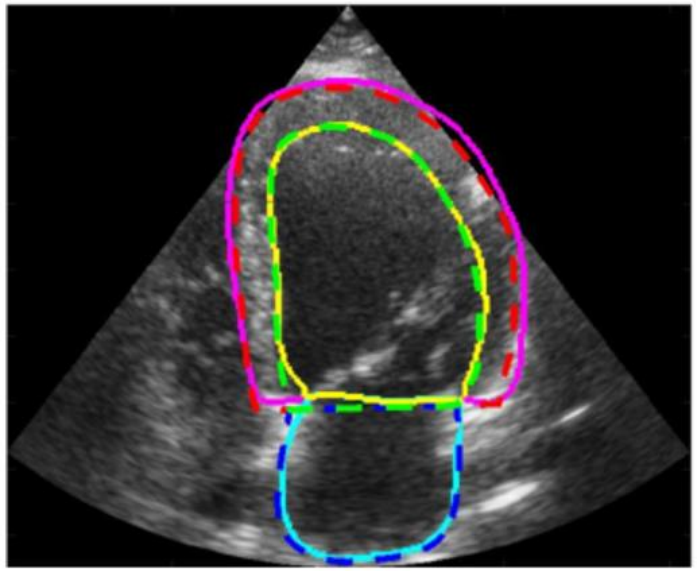
Newcastle University | AI Safety Institute
Powered by National Edge AI Hub

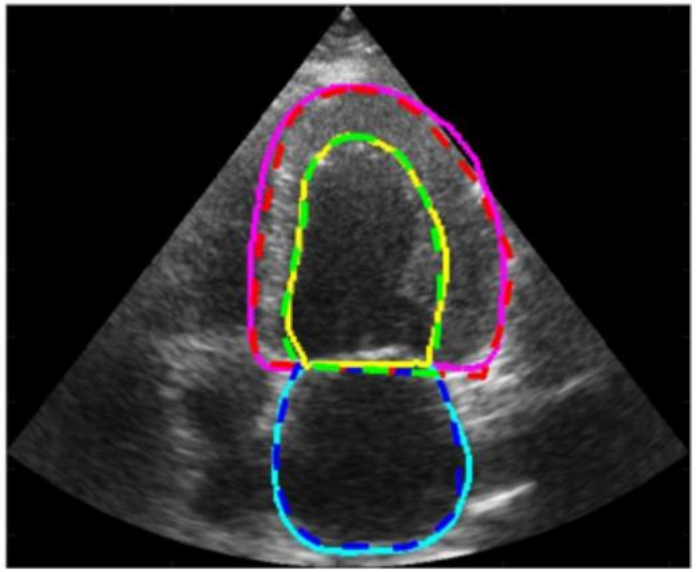


Liver Tumor Segmentation from CT scans

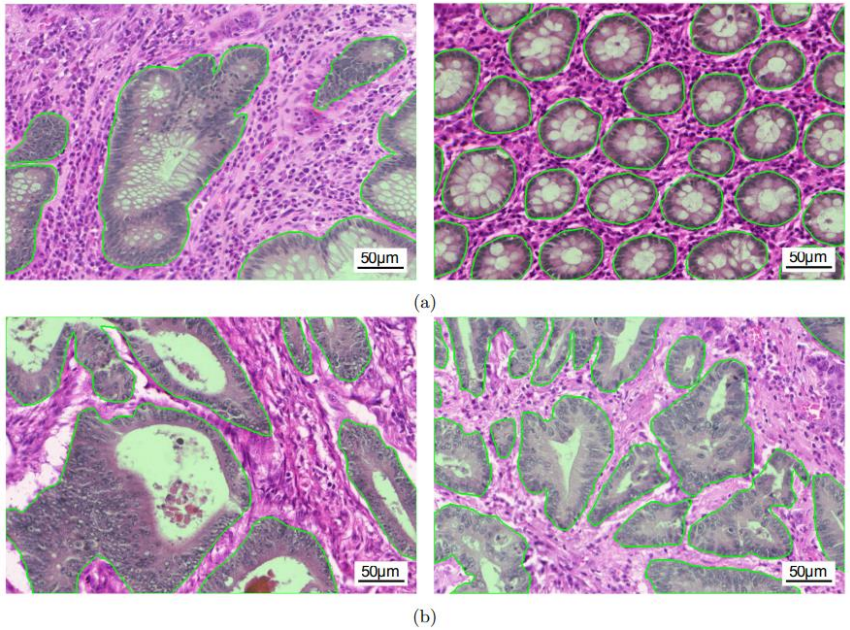


Digital Retinal Images for Vessel Extraction





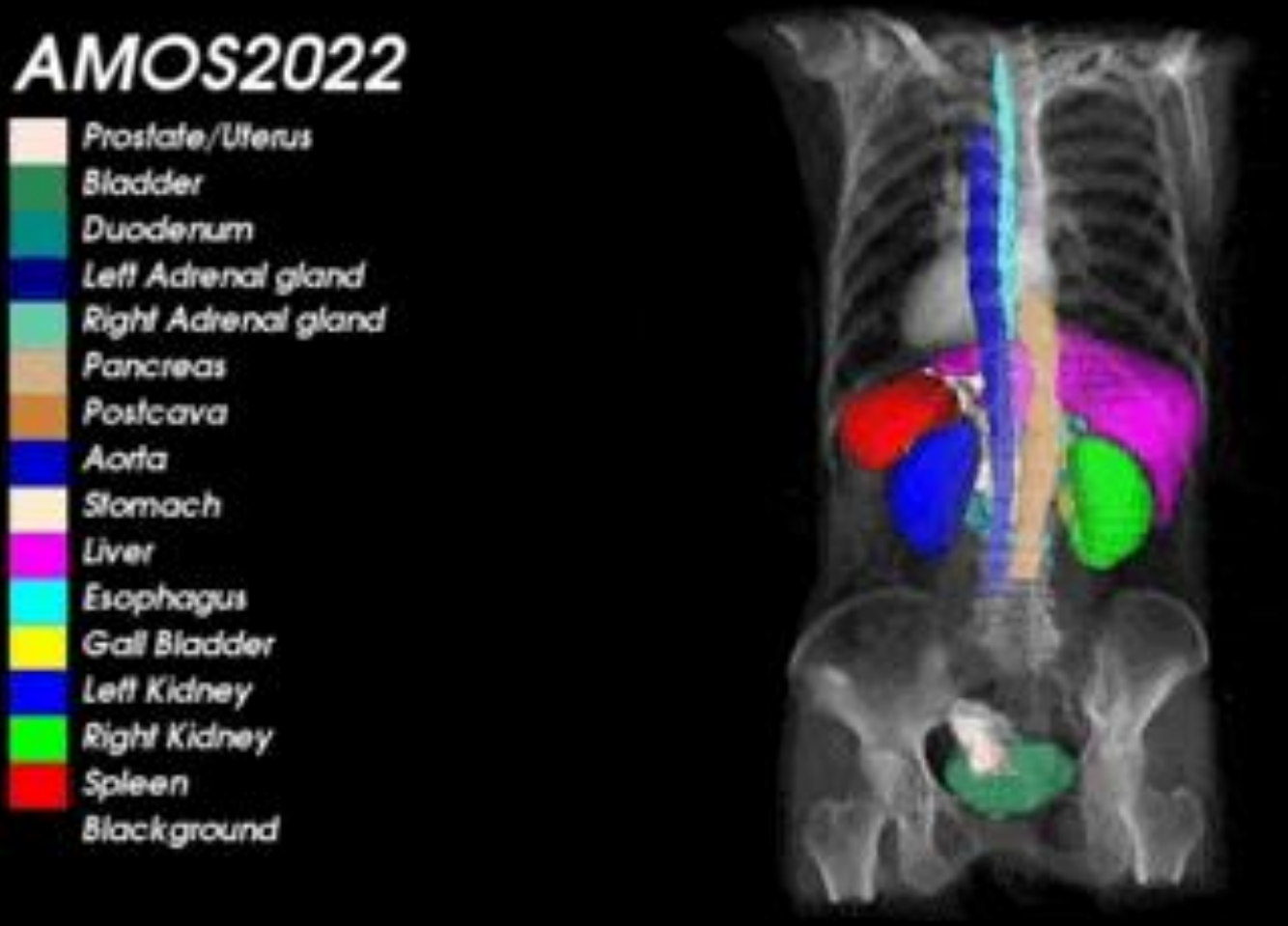
Cardiac Acquisitions for Multi-structure Ultrasound Segmentation



Gland Segmentation in Colon Histology Images

Multi-Modality Abdominal Multi-Organ Segmentation

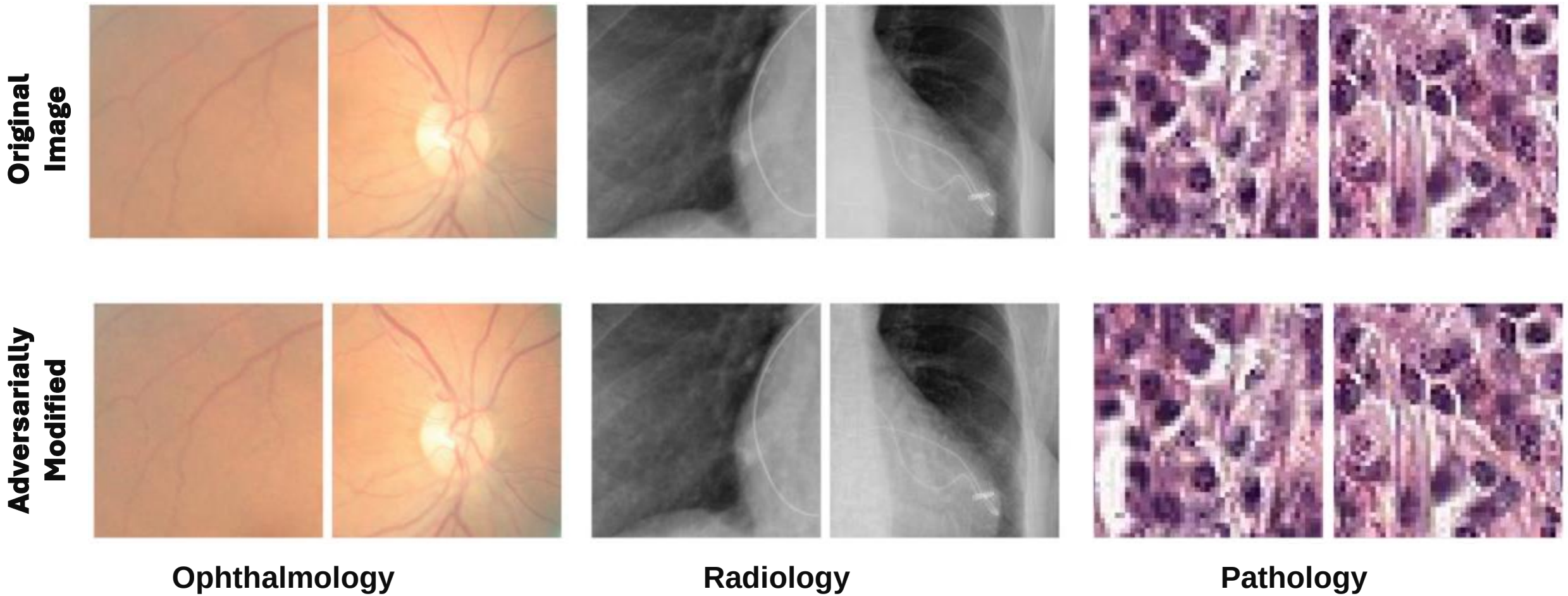
Newcastle University | AI Safety Institute
Powered by National Edge AI Hub



https://amos22.grand-challenge.org/

AI Safety Concern in Medical Image Analysis

Newcastle University | AI Safety Institute
Powered by National Edge AI Hub

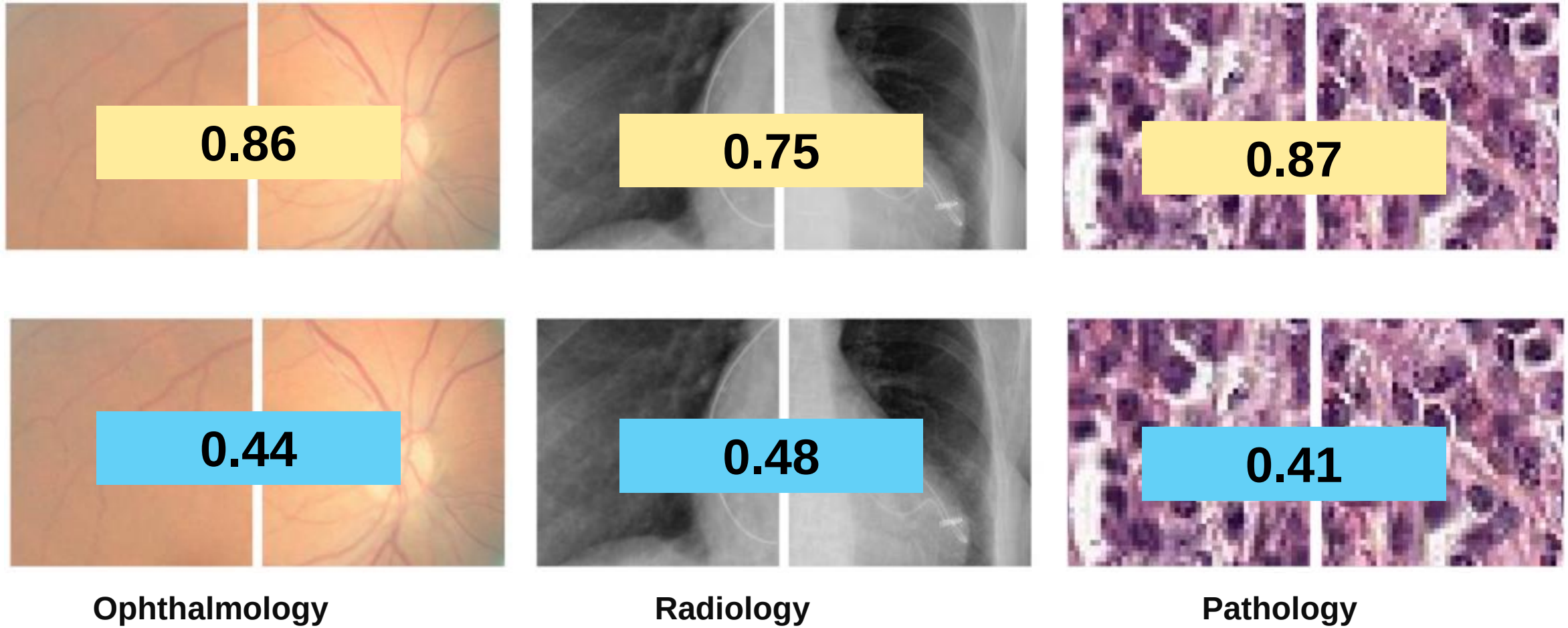


Unperceived changes in images make misclassification

Bortsova et al. (2021). Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. *Medical Image Analysis*,

AI Safety Concern in Medical Image Analysis

Newcastle University | AI Safety Institute
Powered by National Edge AI Hub

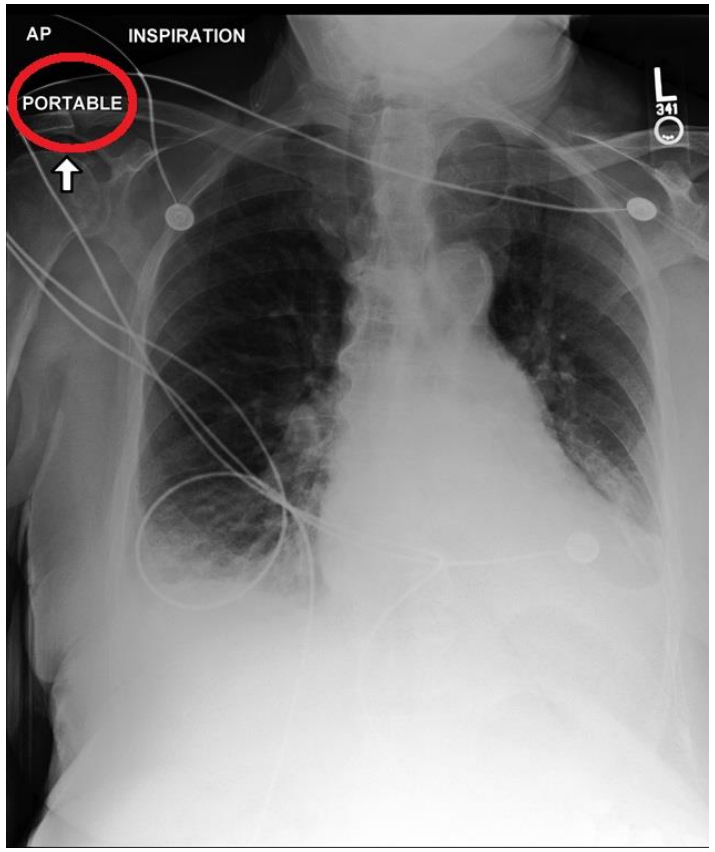


Accuracy of detection decreases even on an unperceived modification

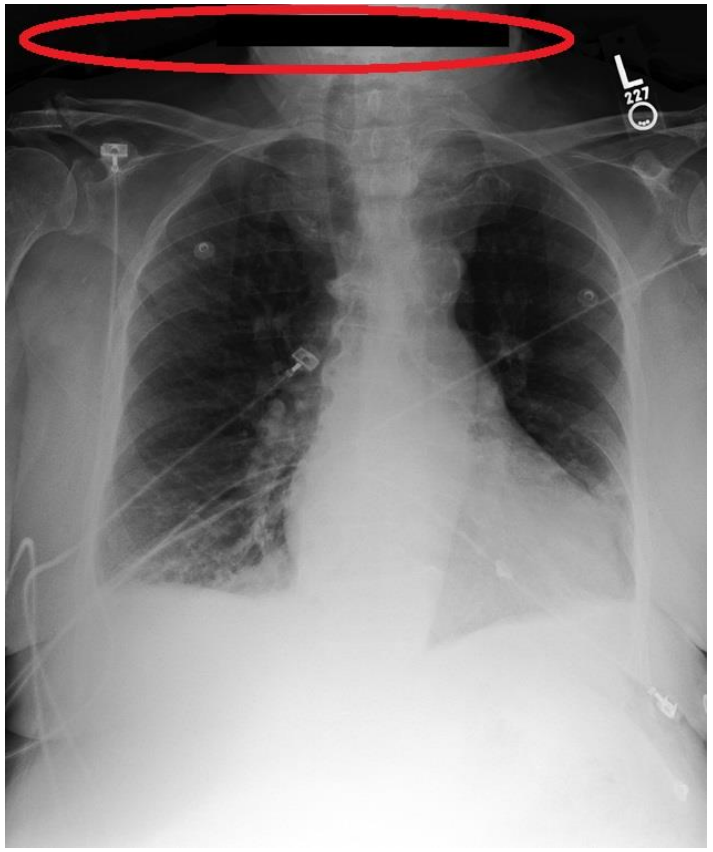
Bortsova et al. (2021). Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. *Medical Image Analysis*,

Spurious Correlation in Medical Training Data

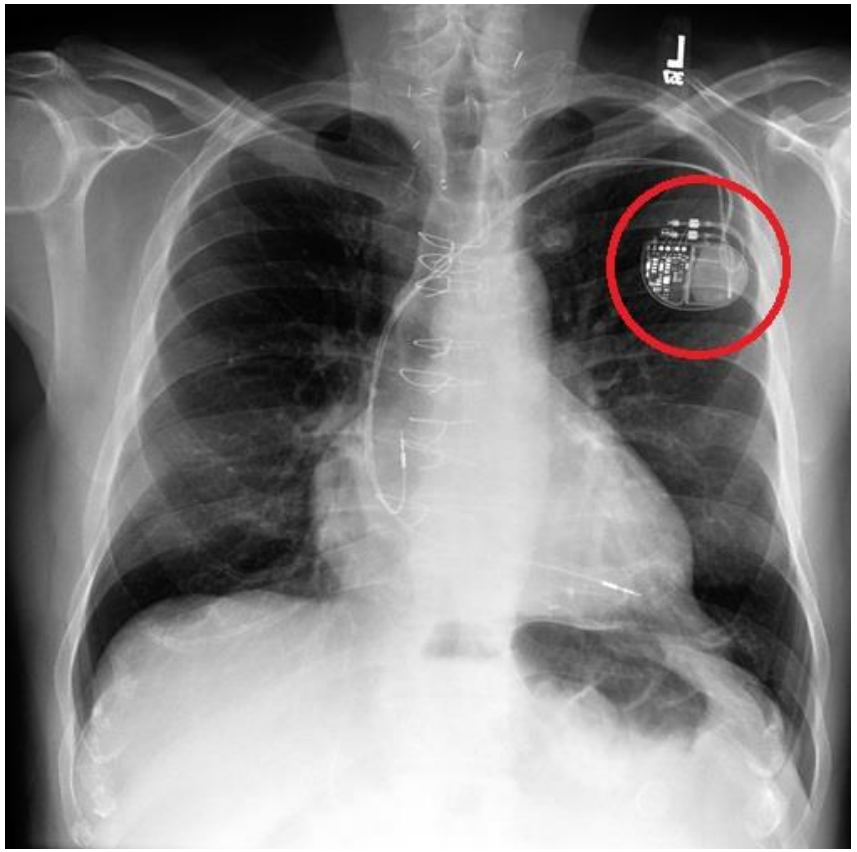
Spurious correlation occur in medical training data where diagnosis results are affected by variables (e.g., Hospital tags, Strips, Medical devices) that are not related to the diagnostic information being predicted. This phenomenon leads to misleading interpretations.



Hospital tags



Stripes

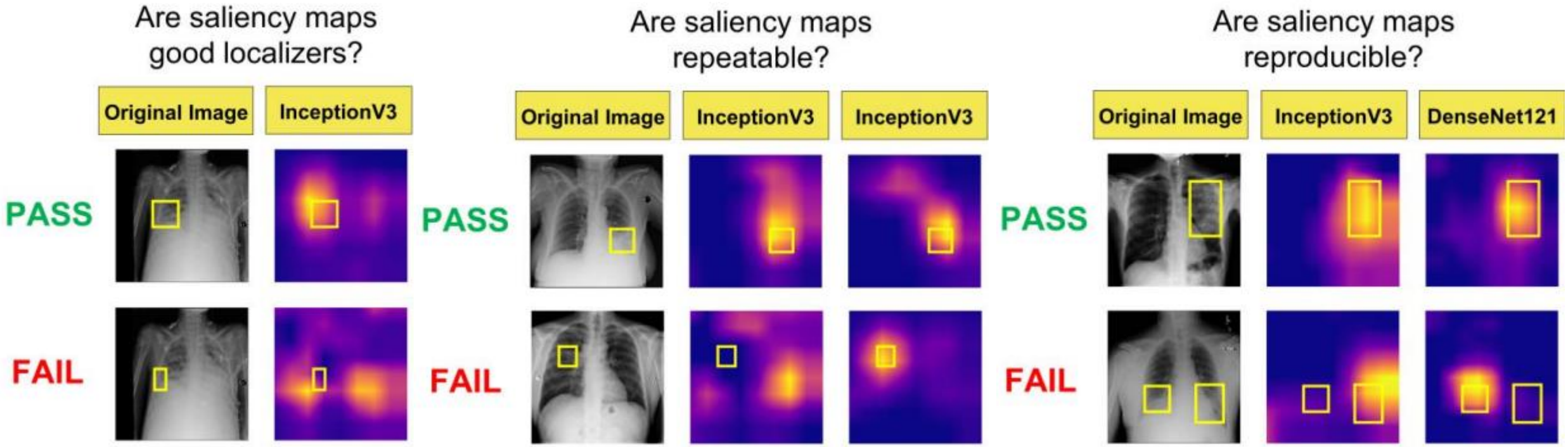


Medical devices

Ye et al. (2024). Spurious correlations in machine learning: A survey.

Artificial Intelligence (Un)Trustworthiness Performance

Salience Map is way to explain the prediction of AI models on Medical Diagnosis



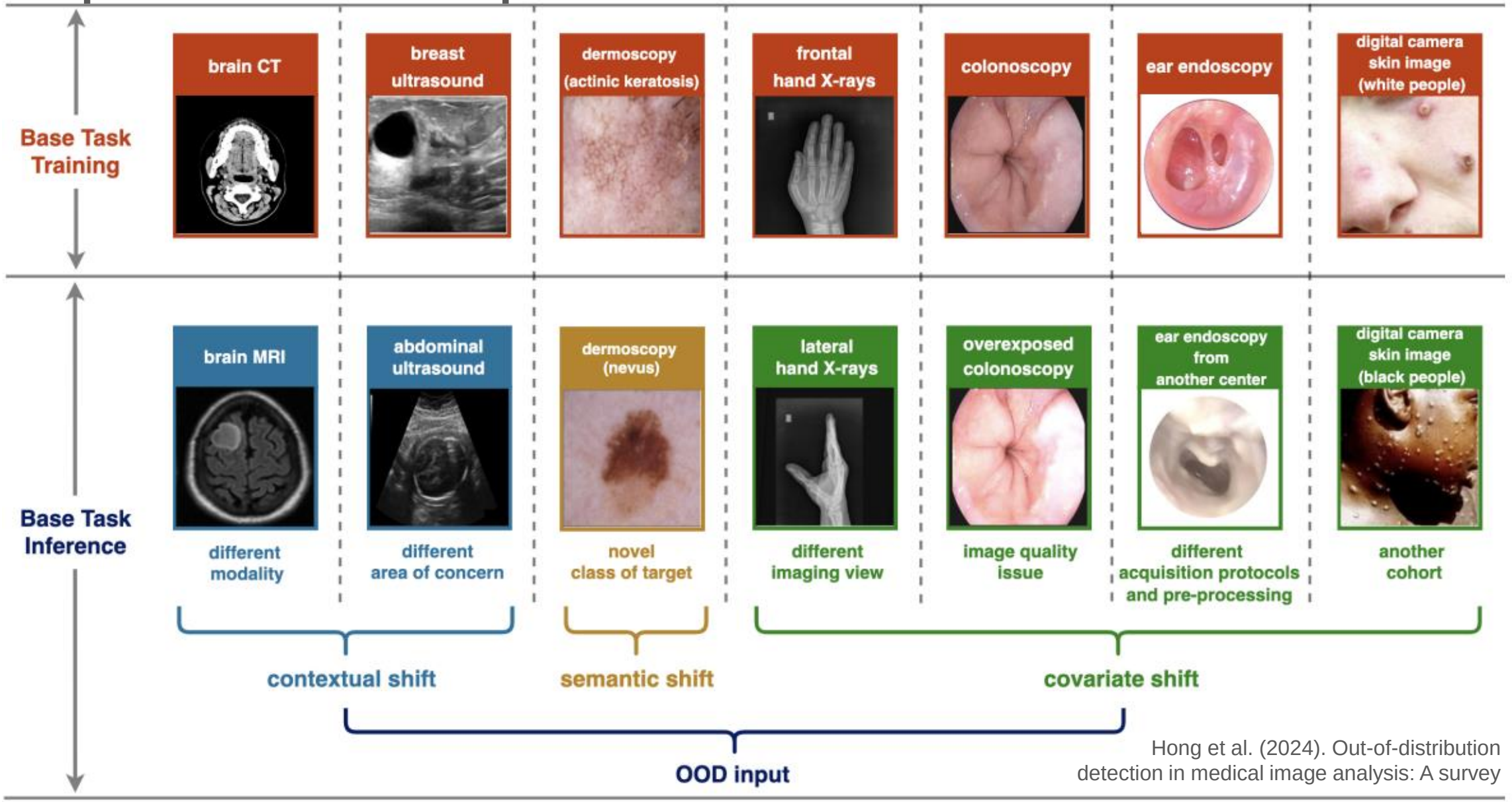
PASS :Only in most ideal cases AI works

FAIL : In most cases AI fails to locate, repeat, and reproduce results

Arun et al. (2021). Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiology: Artificial Intelligence

Risk of Training Data Error

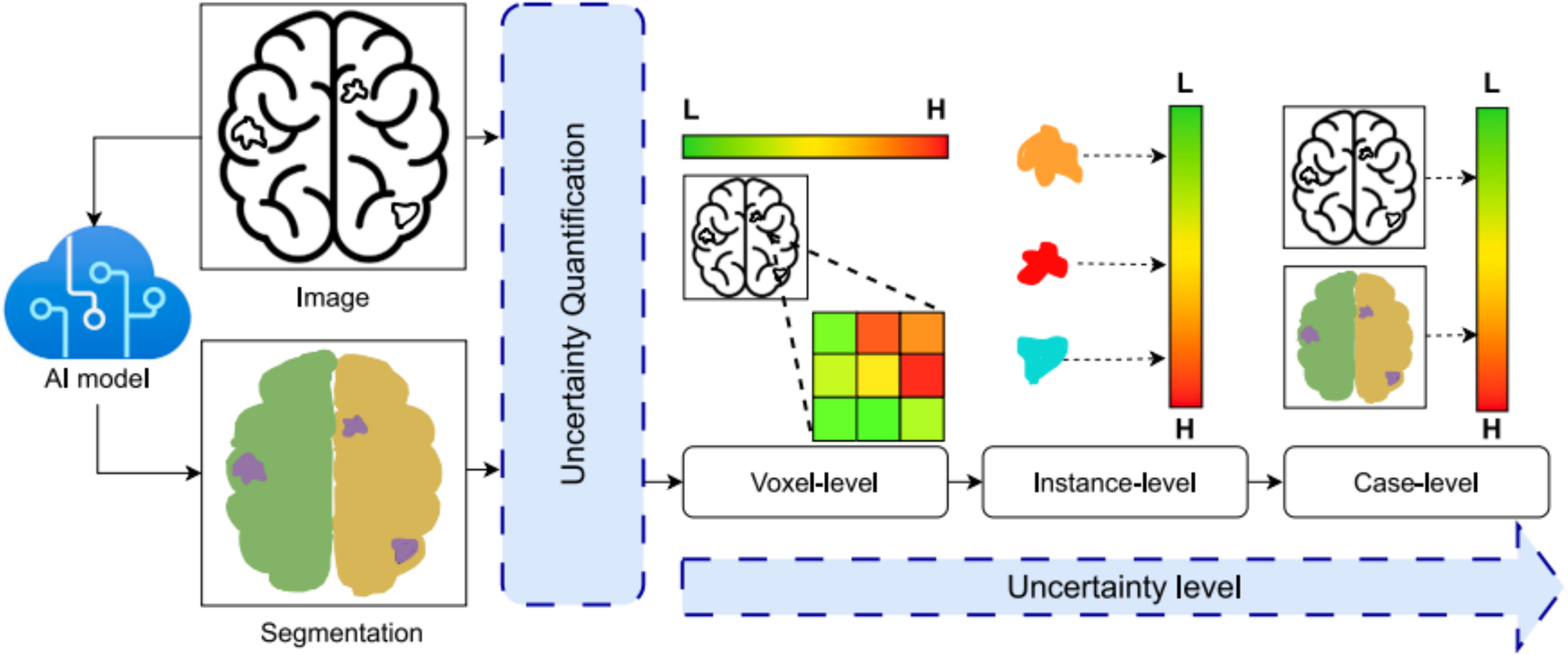
Examples of out of distribution problems



Hong et al. (2024). Out-of-distribution detection in medical image analysis: A survey

Risk of Training Error and AI (Un)Trustworthiness/Uncertainty?

Trained AI Models inference suffers from inference/test images being out of training distribution

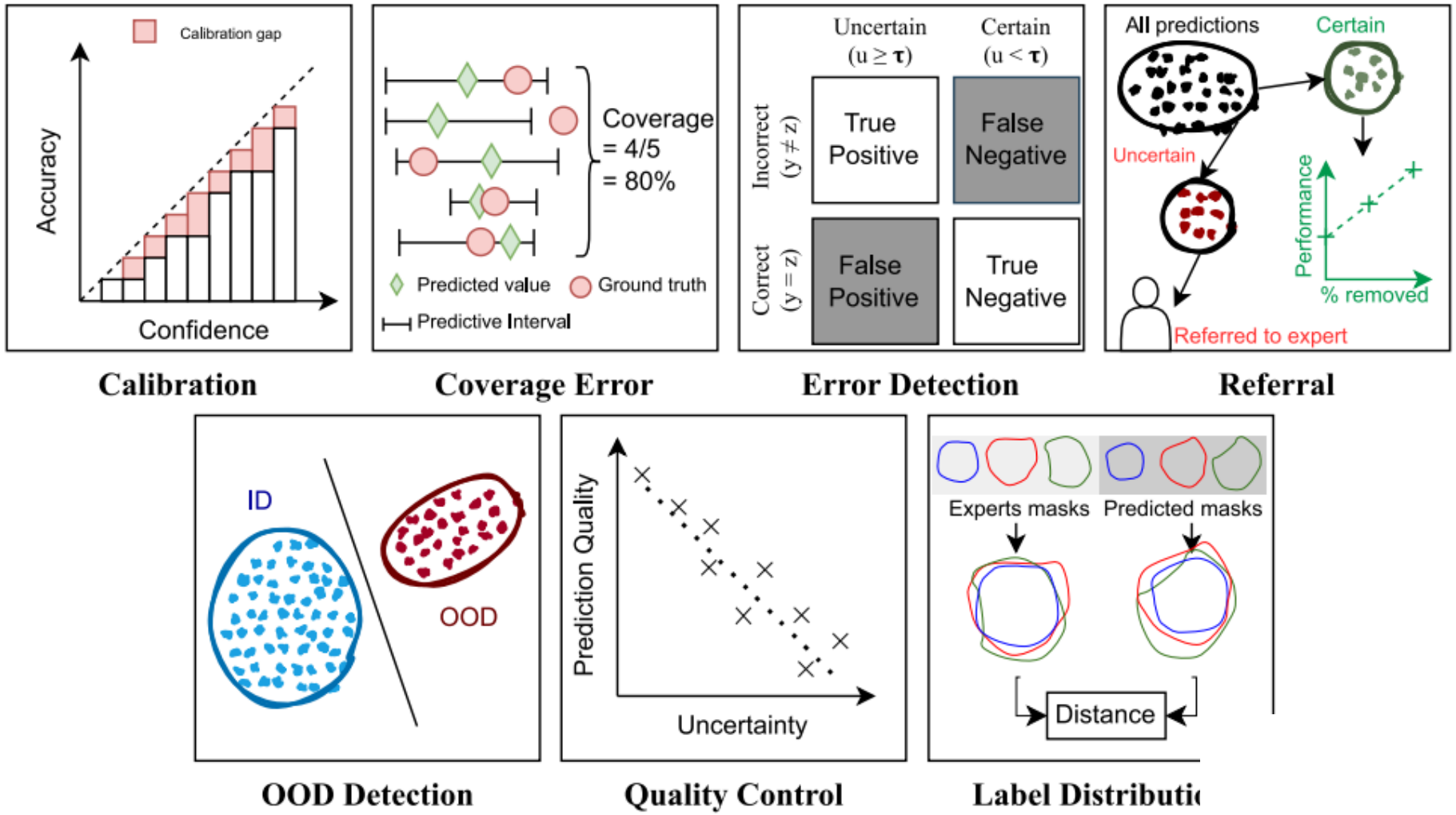


Lambert et al. (2024). Trustworthy clinical AI solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis. *Artificial Intelligence in Medicine*

Mitigating Risk of AI Uncertainty / Scalable Oversight

Trained AI Models inference suffers from inference/test images being out of training distribution

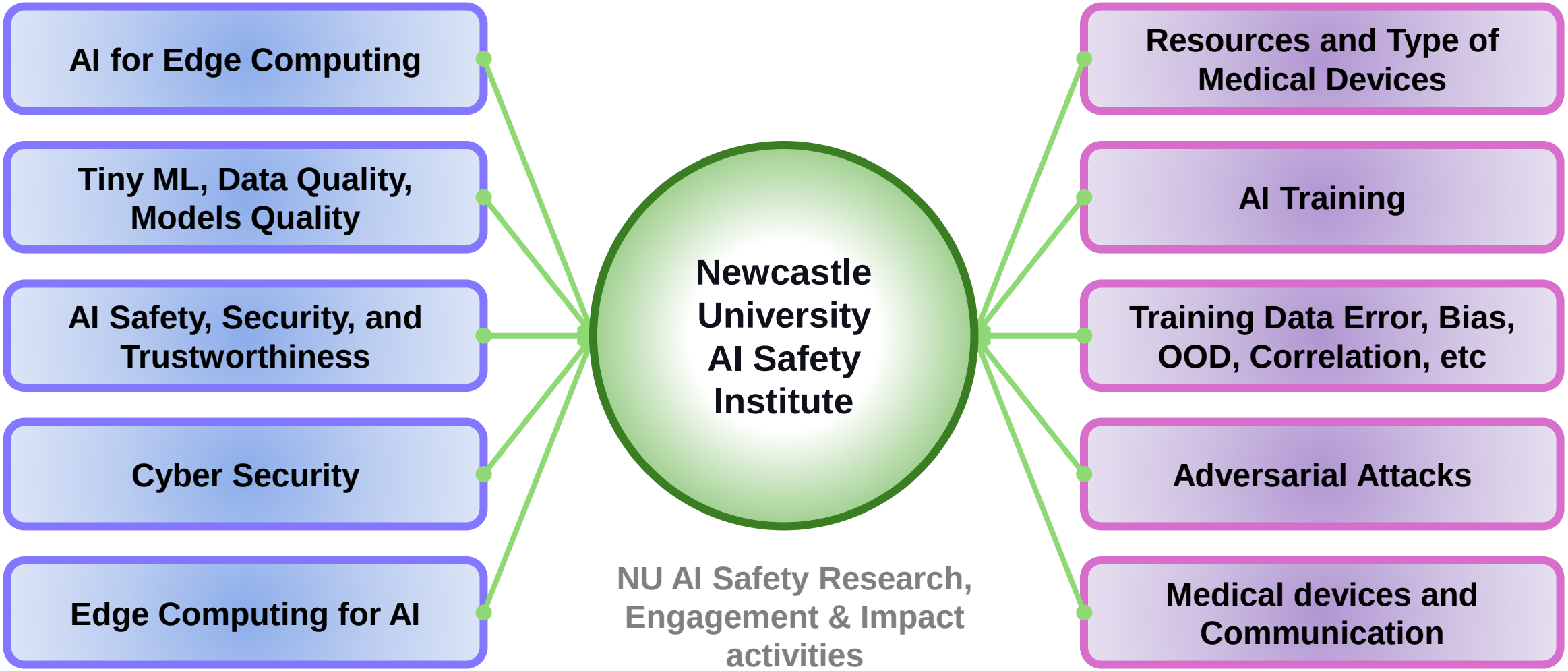
Designing AI to actively bring (Observer) **Human in the Loop** and return their feedback to make AI systems better



Lambert et al. (2024). Trustworthy clinical AI solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis. *Artificial Intelligence in Medicine*

Edge AI Hub → NU AI Safety Institute → Faculty of Medical Sciences

NU AI Safety Institute will deliver for FMS via Edge AI Hub and University Wide Research Skills



Edge AI Hub AI Safety Research, Skills and Leadership

NU AI Safety Research, Engagement & Impact activities

FMS's AI Safety Research Needs

Newcastle University | AI Safety Institute

Powered by National Edge AI Hub

Get in touch

Address

Urban Sciences Building, 1 Science Square,
Newcastle upon Tyne NE4 5TG, UK

Email

varun.ojha@newcastle.ac.uk

Web

https://ojhavk.github.io/