NYU Langone Health

APIBQ Congrès 2022

ETHICS AND AI IN RADIOLOGY AND RADIATION ONCOLOGY

Sunshine Osterman, PhD

Noah Bice, PhD

Conflicts of Interest / Disclosures

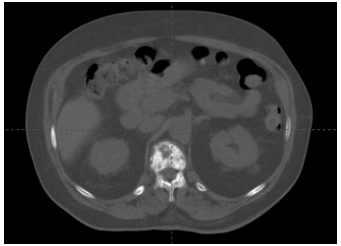
- Co-author of the current AAPM Code of Ethics
- Vice Chair of the AAPM Ethics Committee
- Clinical Therapy Physicist
- Member of AAPM Diversity & Inclusion Subcommittee
- Not a trained ethicist!

We're all biased. We apply different weight to different aspects of an event, because of our life, professional experiences, and training.

vs.

vs.

vs.



Outline

- ‘Elevator Pitch’ : How to apply biomedical ethics to AI
  - Our professional Codes of Ethics and applicability to AI
  - Bioethics literature and Society Position Statements
- Overview of Artificial Intelligence – what is it, and how can we as scientists and clinicians can guide its development and use in Radiology and Radiation Therapy
- Describe how a bioethical framework can be used in the AI lifecycle – Inception → Deployment

Ethical Dilemma / Case exploration

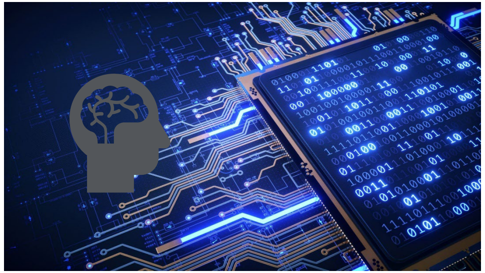
What is ethics?

- Distinct from civil, moral or religious law
- Behavioral – making “right” choices
- Transactional – right relationship with others
- Personal – discipline, awareness and responsibility

Codes of ethics?

Our behavior influences how our professions are perceived and how effective we can be at improving health as a professional society and as individuals.

We're involved with AI



Professional Codes of Ethics

Our behavior influences how AI is developed, perceived and utilized in improving and protecting healthcare.

Trust

NYU Langone Health

Value as Physicists, Clinicians, Engineers

1. Know our fields in depth

   EXPERTISE

2. Accept Responsibility for our Decisions

   ACCOUNTABILITY

Professional Codes of Ethics

The Canadian Medical and Biological Engineering Society (CMBES/SCGB)

CMBES/SCGB

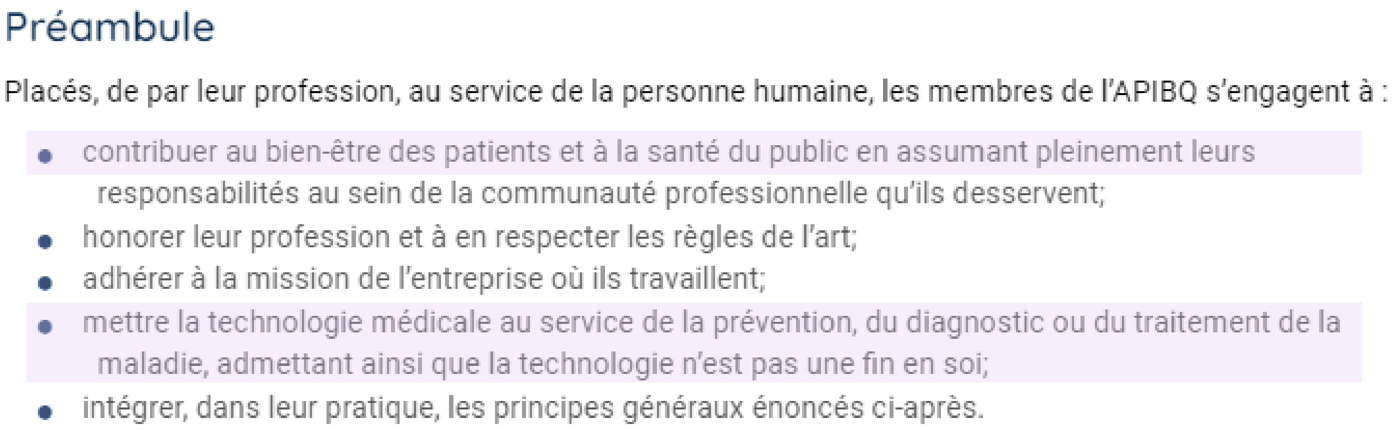
COMP/OCPM

COMP
Canadian Organization
of Medical Physicists

OCPM
Organisation canadienne
des physiciens médicaux

All give profession-specific context and have a longer list (13) of principles

APIBQ – Code of Ethics



Préambule

Placés, de par leur profession, au service de la personne humaine, les membres de l'APIBQ s'engagent à :

- contribuer au bien-être des patients et à la santé du public en assumant pleinement leurs responsabilités au sein de la communauté professionnelle qu'ils desservent;
- honorer leur profession et à en respecter les règles de l'art;
- adhérer à la mission de l'entreprise où ils travaillent;
- mettre la technologie médicale au service de la prévention, du diagnostic ou du traitement de la maladie, admettant ainsi que la technologie n'est pas une fin en soi;
- intégrer, dans leur pratique, les principes généraux énoncés ci-après.

APIBQ Members have a responsibility to make sure technology is being developed for Patient and Society well-being, not just as an end in itself. **TRUST**

Artificial Intelligence

- Adjuvant Tool used by Professionals
  - Efficiency
  - Cost
  - Consistency
  - Accuracy
  - Insight
- Replacement of Professionals
  - Automated task completion
  - Decision Making

Artificial Intelligence

- If something goes wrong, who will assume the responsibility?
- Who will apologize & fix it?

The Programmer who wrote the code?

The Researcher who chose the data and tested the system?

The Vendor who sells the code?

The Physicist/Physician who uses the tool?

The patient who consented to the use of AI in their care?

Even Better …

What decisions can we do to prevent or minimize the chances that something will go wrong in the first place?

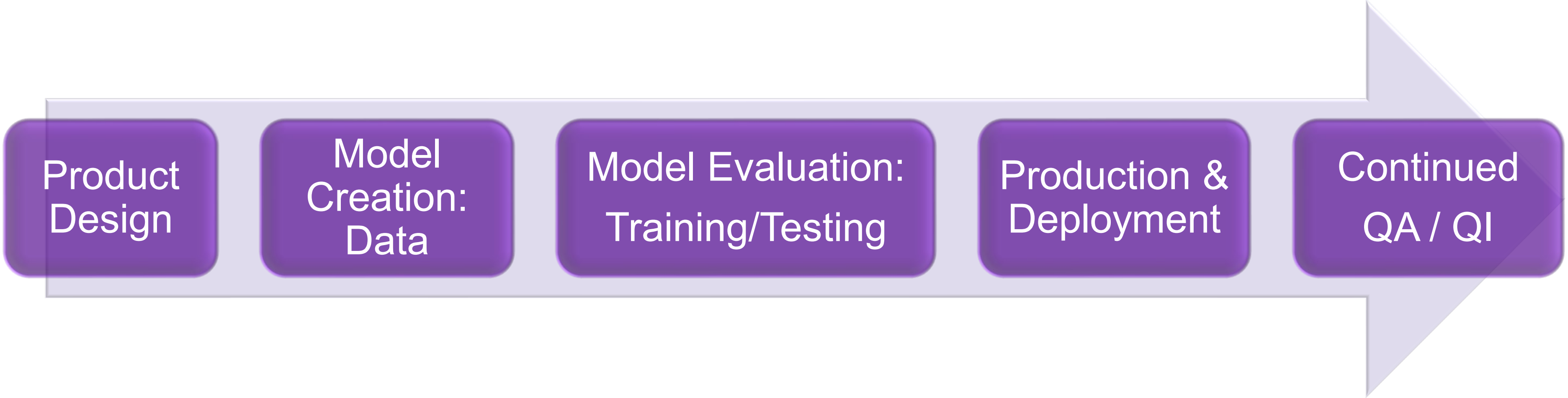
Using Failure Mode and Effects Analysis to Evaluate Risk in the Clinical Adoption of Automated Contouring and Treatment Planning Tools, Nealson KA et al. Pract Radiat Oncol. Jul-Aug 2022.

I don’t know, but physicists and engineers will be involved, and the decisions are:

- Distinct from civil, moral or religious law
- Behavioral – making “right” choices
- Transactional – right relationship with others
- Personal – discipline, awareness and responsibility

Ethics

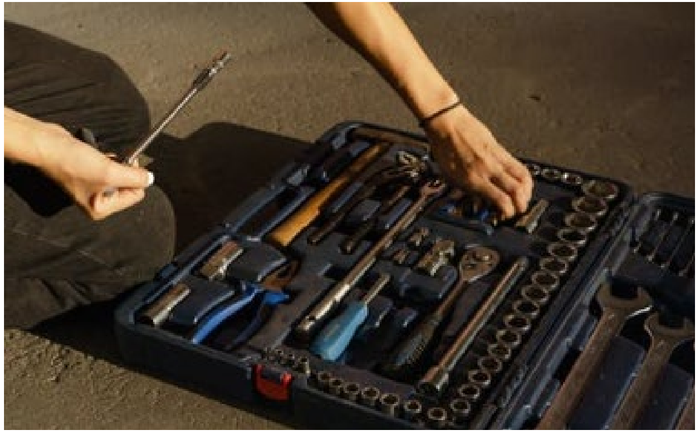
AI “Life-Cycle”



Ethics is an essential component at all stages of AI

Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics and effectiveness, Vollmer S et al. 2020 https://www.bmj.com/content/368/bmj.l6927

What do we have to work with in our Biomedical Ethics toolbox?

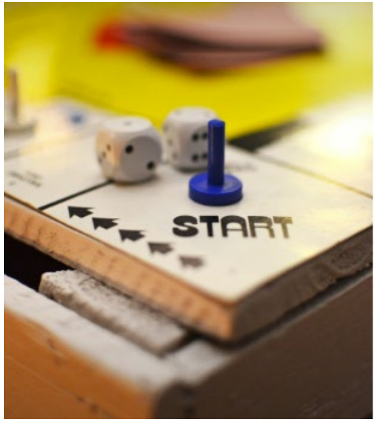


Ethics Building Blocks: Common Morality

“Set of universal norms shared by all persons committed to morality”

Principles of obligation which are:

- Broad
- Abstract
- Content-thin



Principles of Biomedical Ethics, Beauchamp TL, Childress JF. 8th Ed. NY: Oxford University Press; 2019. 1st edition published in 1979.

Principles of Biomedical Ethics

1. Respect for Autonomy

Respecting and supporting autonomous decisions

2. Nonmaleficence

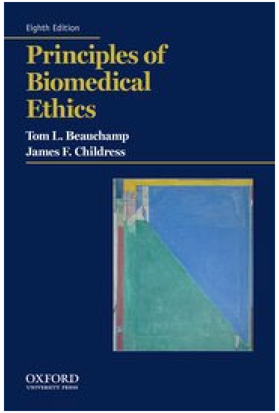
Avoiding the causation of harm

3. Beneficence

Relieving, lessening, or preventing harm and providing benefit; weighing benefits vs. risks and cost

4. Justice

Fairly distributing benefits, risks and costs



Eighth Edition

Principles of Biomedical Ethics

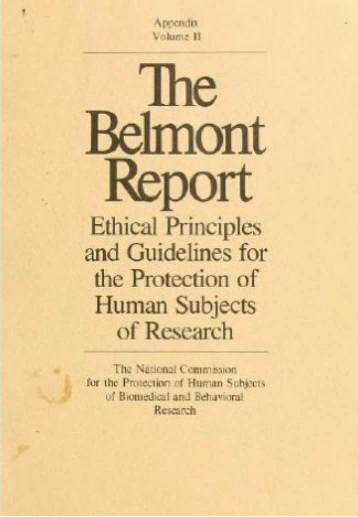
Tom L. Beauchamp
James F. Childress

OXFORD
UNIVERSITY PRESS

Ethical Framework

A similar set of principles has been laid out in the Belmont Report.

1. Respect for Persons
2. Beneficence
3. Justice



Appendix
Volume II

The Belmont Report

Ethical Principles and Guidelines for the Protection of Human Subjects of Research

The National Commission
for the Protection of Human Subjects
of Biomedical and Behavioral
Research

NYU Langone Health

Prima facie duty

A duty that is binding or obligatory…. ALL OTHER THINGS BEING EQUAL

Step 1:

- Respect for Autonomy
- Nonmaleficence
- Beneficence
- Justice

Ethics in Theory vs. Practical Tool

Step 2:

- Specification
- Balancing of Principles in the Given Context

An iterative problem which relies on the perspective of all the stakeholders to determine which principle(s) take priority

Step 1:

- Autonomy
- Nonmaleficence
- Beneficence
- Justice

Step 2: (specification)

<u>What? Why? How?</u>

Diagnosis. Time consuming & variable. AI.

<u>Who?</u>

Patients, Radiologists, Oncologists, Administrators, Coders, Vendors, Regulators

Patient: Cure, Quality of Life (Beneficence, Justice)

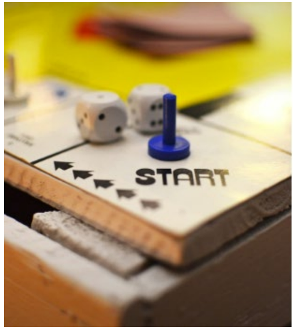
Physicist/Physician: Explainability (Nonmaleficence, Autonomy)

Developer: IP protection, security (Nonmaleficence, Beneficence)

Applying Frameworks for ethical decision-making

Step 1:

- Choose a set of building blocks (principles)
  - Broad
  - Abstract
  - Relatively independent of context



Step 2:

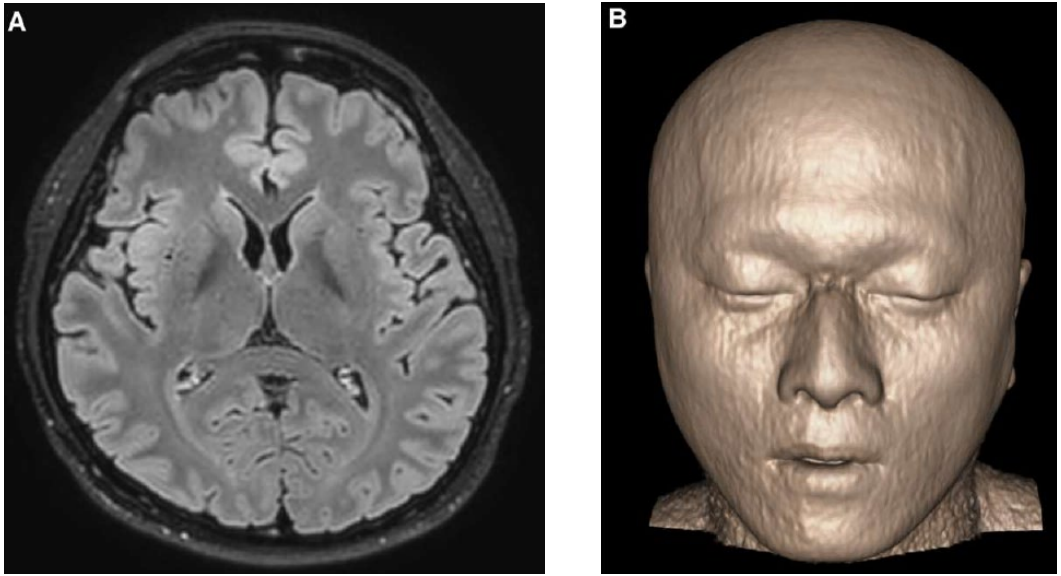
- Specification
- Balancing of Principles in the Given Context
- Bring in the Stakeholders: Minimize Bias and Blindspots
- Adjust and Modify, Test for Inconsistencies

Canadian Association of Radiologists White Paper on Ethical and Legal Issues Related to Artificial Intelligence in Radiology

Respect for Privacy

Anonymization

Hidden PHI in anonymized data



Jaremko JL, Azar M, Bromwich R, et al. Canadian Association of Radiologists White Paper on Ethical and Legal Issues Related to Artificial Intelligence in Radiology. *Can Assoc Radiol J*. 2019;70(2):107-118. doi:10.1016/j.carj.2019.03.001

Building or Curating a dataset

- Are the data truly anonymous?
- Is **autonomy** compromised in broad- or implied-consent data collection?
- Is it **beneficent** to gather as much data as possible to build a good model?
- Data regulations (varies by country)
  - Who owns the data? How much is the data worth? What happens if data is sold privately?
  - If we don't own the data, but act as a data custodian, what are our rights & obligations?
  - Who advocates for / protects the patient and society?

Compassionate Human-Centric AI

Keep the patient (complex and unique) at the center of the discussion/care.

Compliance goes up, outcomes improve, and patient satisfaction with their treatments increases when patients feel respected, heard and are given information about their healthcare options that they can understand.

Concordant Care (language, race, gender) => Better Outcomes (↑ Survival, ↑ QOL)

Is AI compatible with the concept of a patient as a unique individual?

Lukowicz, P. The challenge of human centric AI. *Digitale Welt*, *(2019)*

Next:

Artificial Intelligence in more detail

- Basic terminology
- Ways to test and analyze an AI tool

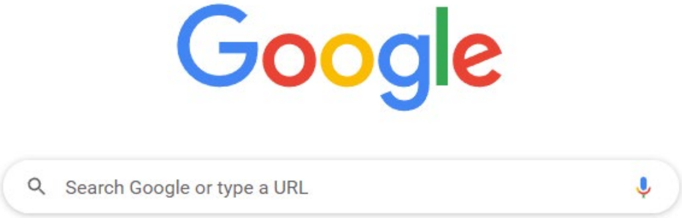
Recommendation for Ethical AI Design

Disclosures

- Medical Physics Resident at NYU
- Also not an ethicist!

AI in the history of computation

- Church-Turing Thesis (1936)
  - Any “effectively computable” function can be computed with a Turing machine.
  - Universal programmable computers can exist.
    - We don’t need separate Turing machines to calculate π to 1,000 digits, to send emails, to play Minecraft, etc.
- If these statements are true, to what extent can we reproduce human intelligence with machines?



The first attempts at AI were rule-based.

- IBM’s Deep Blue (1997) – brute force search
  - Purpose-built hardware to evaluate millions of positions per second
  - An 8000-term “evaluation function” based on hundreds of thousands of master and grandmaster games

NYU Langone Health

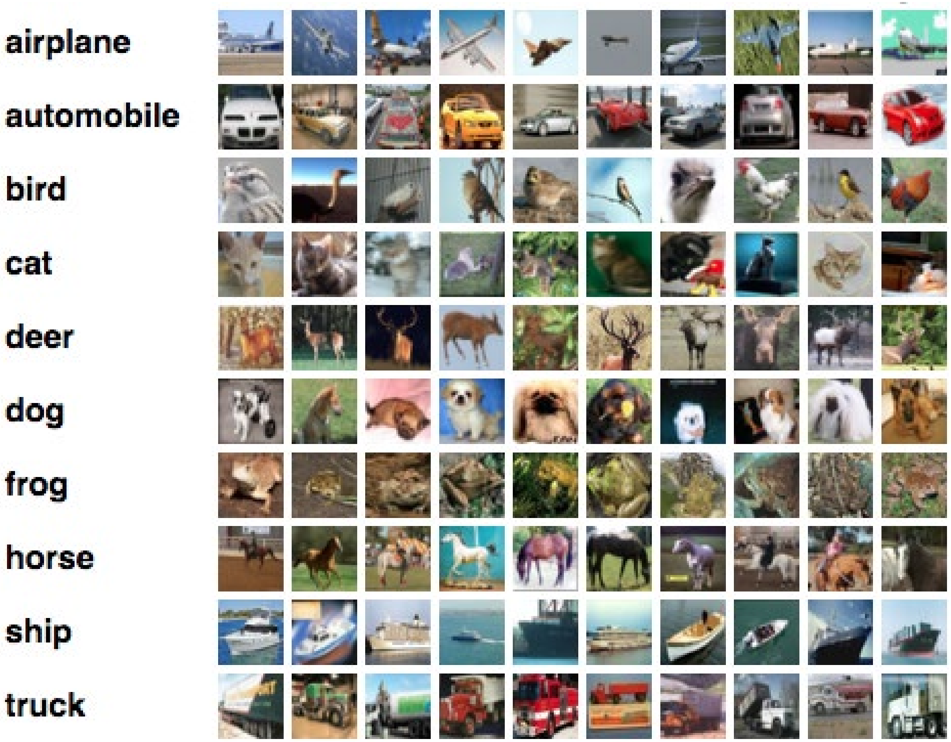
Major challenges in the development of AI are subjective, intuitive problems, that are easy for a 5-year-old child, but difficult to describe with formal rules.

Major challenges in the development of AI are subjective, intuitive problems, that are easy for a 5-year-old child, but difficult to describe with formal rules.

- Identifying that this is a dog, despite different lighting conditions and orientations.
- Being able to describe the dog or tell a story about it.
- Being able to draw new dogs based on your knowledge of what a dog looks like.

NYU Langone Health

ImageNet Large Scale Visual Recognition Challenge

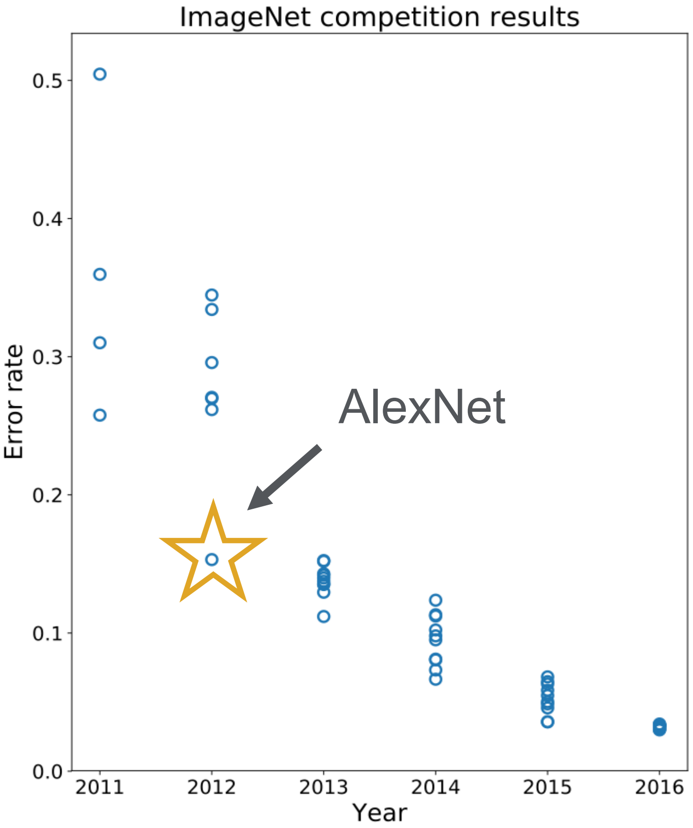


An annual image classification competition beginning in 2010.

14 million images belonging to more than 20,000 classes.

Deep Learning

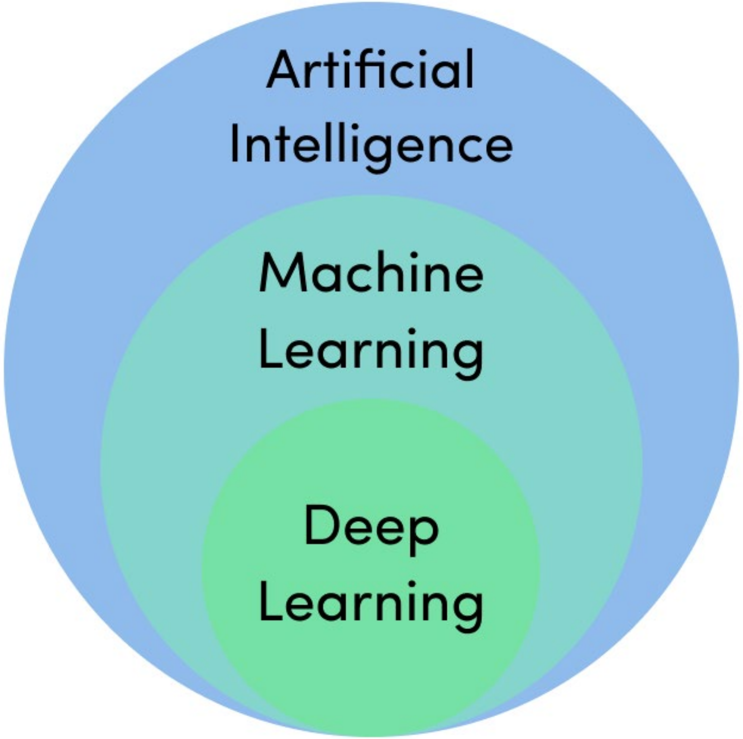
- 2012 ImageNet dataset and competition and AlexNet with Convolutional Neural Network (CNN).
- **AlexNet = CNN = Deep Learning**.
- Deep learning is an old technology with recent major progress in implementation.


ImageNet competition results

Deep learning ⊂ machine learning ⊂ AI

What is deep learning?

Our best attempt (so far) at mimicking the tricky computations that are hardwired by evolution.



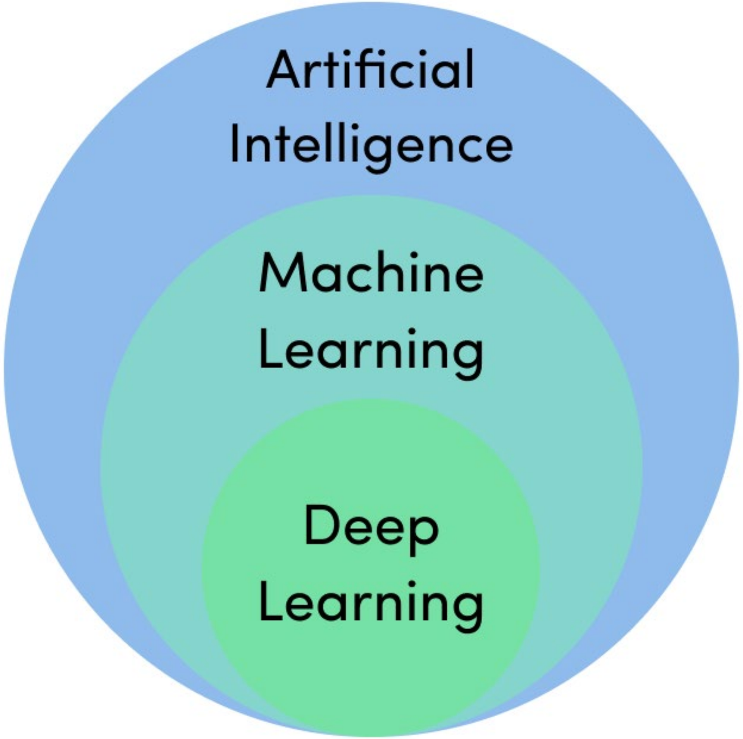
NYU Langone Health

Deep learning ⊂ machine learning ⊂ AI

Artificial Intelligence – broad term to describe the intelligence of machines, as opposed to natural intelligence of animals.

Machine learning – AI strategy that leverages examples from data to draw inferences.

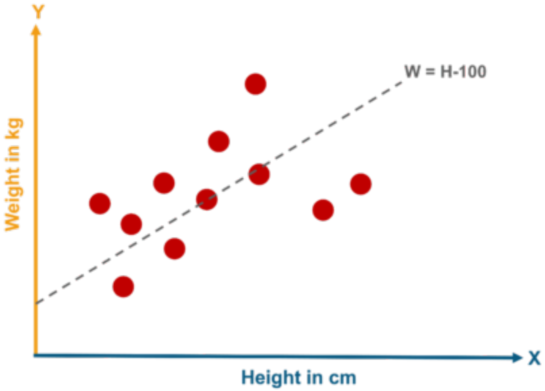
Deep learning – Machine learning with deep neural networks.

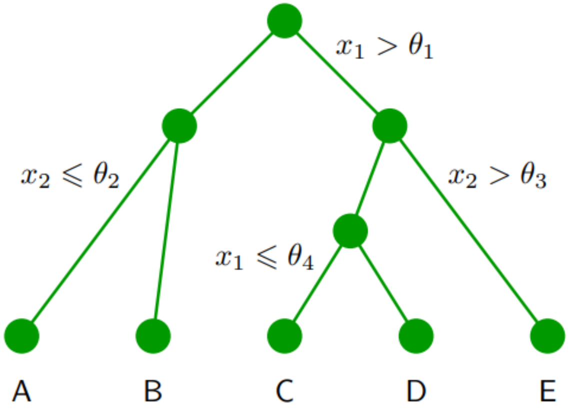


NYU Langone Health

What is machine learning?

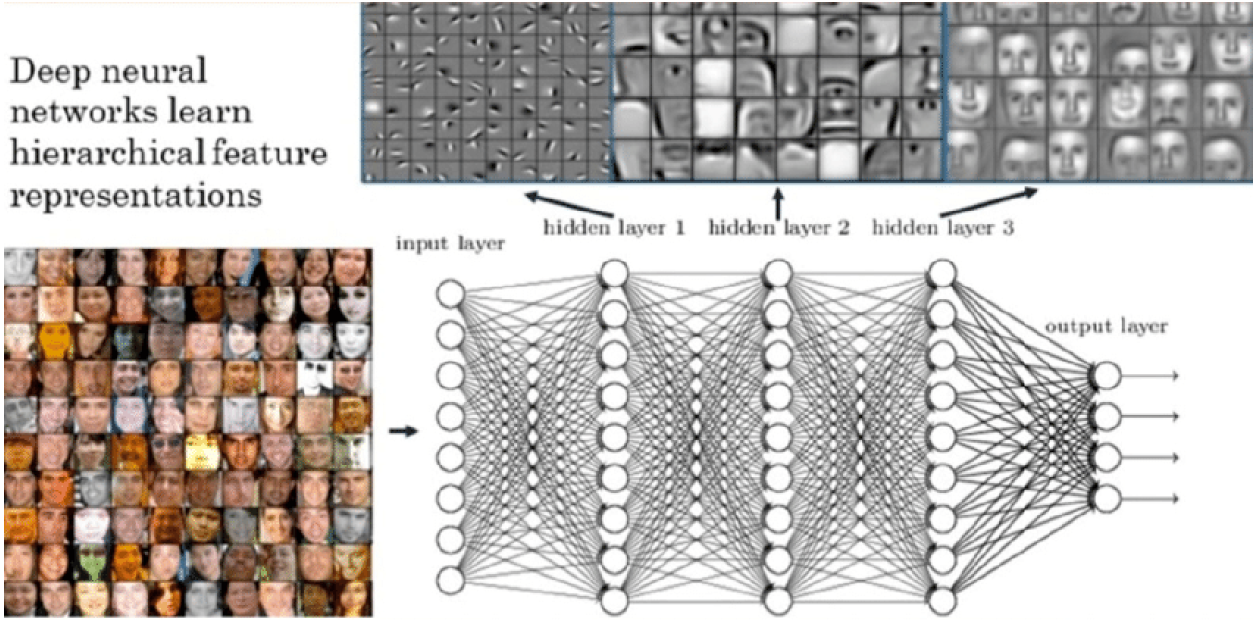
"Machine learning is essentially a form of applied statistics with increased emphasis on the **use of computers to statistically estimate complicated functions** and a decreased emphasis on proving confidence intervals around these functions." – Ian Goodfellow, *Deep Learning*, 2015.





Christopher Bishop. *Pattern Recognition and Machine Learning*, 2006.

What is deep learning?



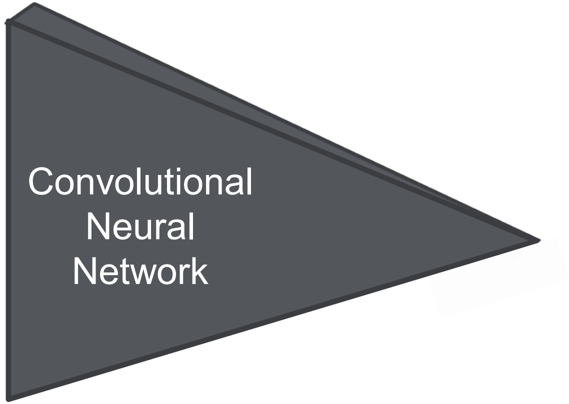
Deep neural networks learn hierarchical feature representations

Raphael et al. Diversity 2020 12(1):29

Deep learning models are machine learning models which use **neural networks** to exploit a **hierarchy of information**.

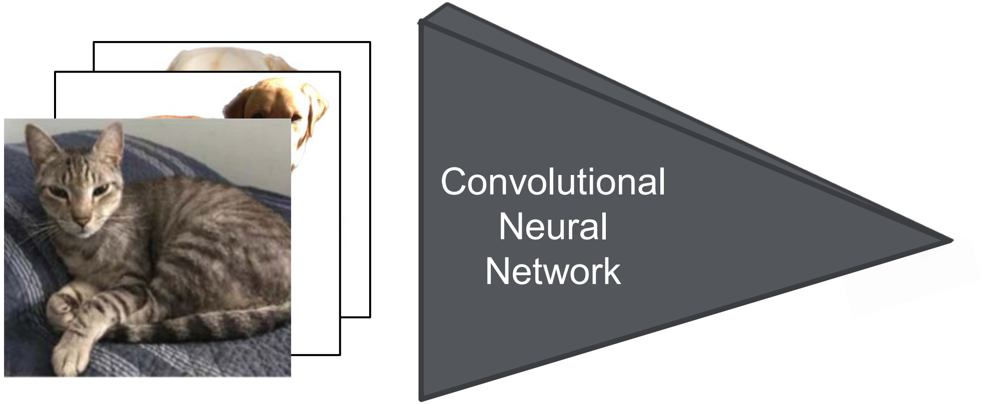
Training a neural network: example of a cat/dog binary classifier

1. Create a architecture with the capacity to classify images, takes image data as an input and outputs {0, 1}.



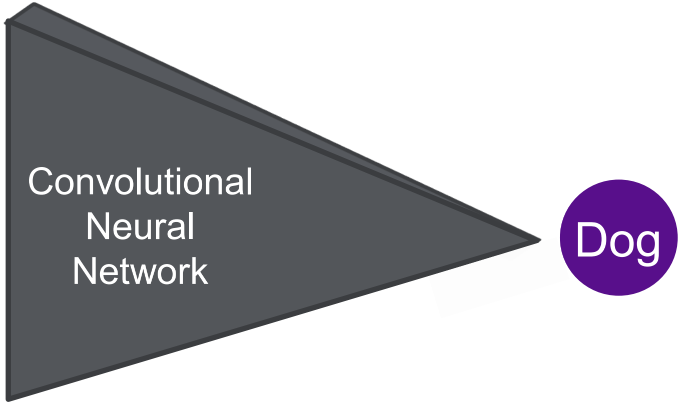
Training a neural network: example of a cat/dog binary classifier

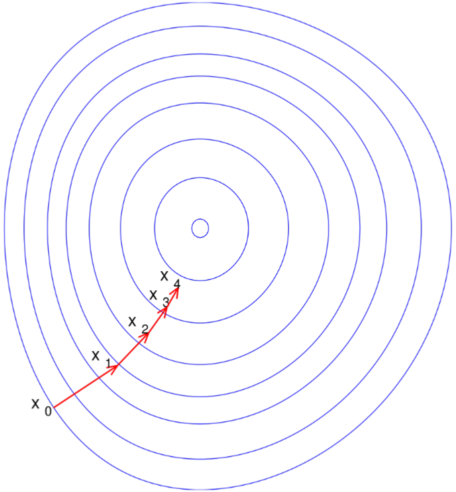
2. Show the model many examples of cats and dogs. **Large** datasets are required to train **large** models.



Training a neural network: example of a cat/dog binary classifier

3. Correct the model’s parameters based on its classification accuracy with respect to labels.





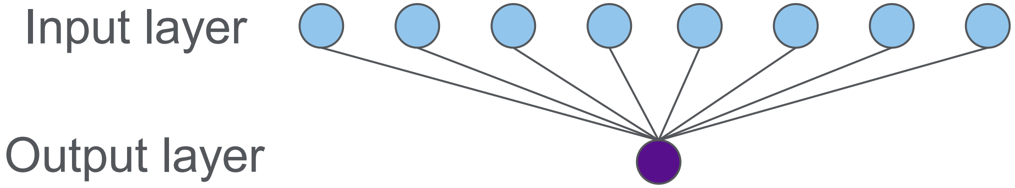
Very high-dimensional space

$$L(\hat{\mathbf{y}}, \mathbf{y}) = \frac{-1}{N} \sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)$$

High-order feature interactions and non-linearity

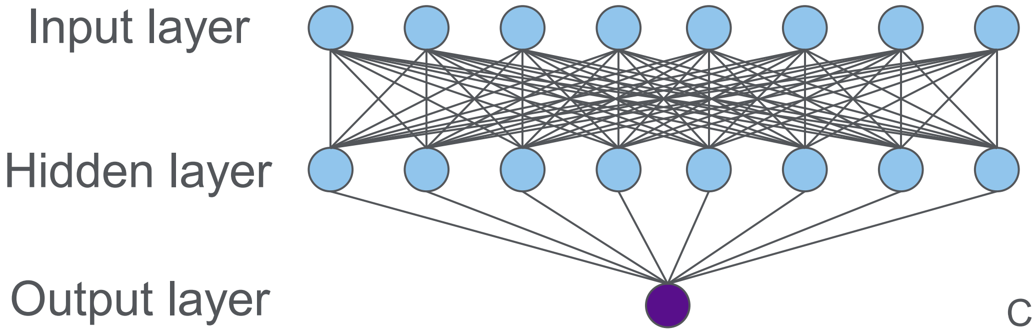
- One of the great strengths and weaknesses of deep learning.
- Allows us to learn complicated functions at the cost of **interpretability**.

Multi-variate linear regression/classification



High-order feature interactions and non-linearity

- One of the great strengths and weaknesses of deep learning.
- Allows us to learn complicated functions at the cost of **interpretability**.



Neural network regression/classification

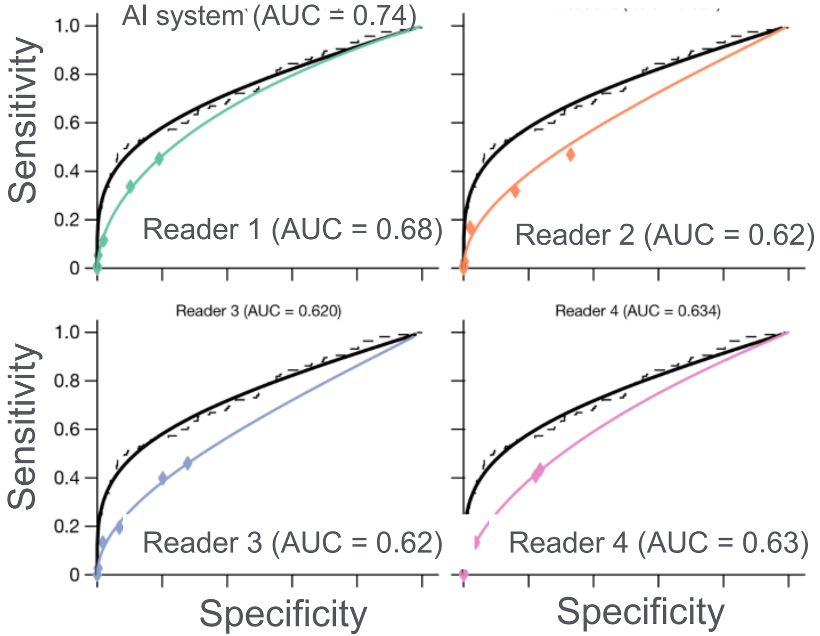
Comes with concerns about overfitting and model generalization.

AI Applications in Healthcare

- Cancer screening
- Treatment recommendation
- Prognostication
- Drug discovery
- Image segmentation
- Image registration
- Treatment planning
- Treatment delivery
- Plan checking
- Quality assurance

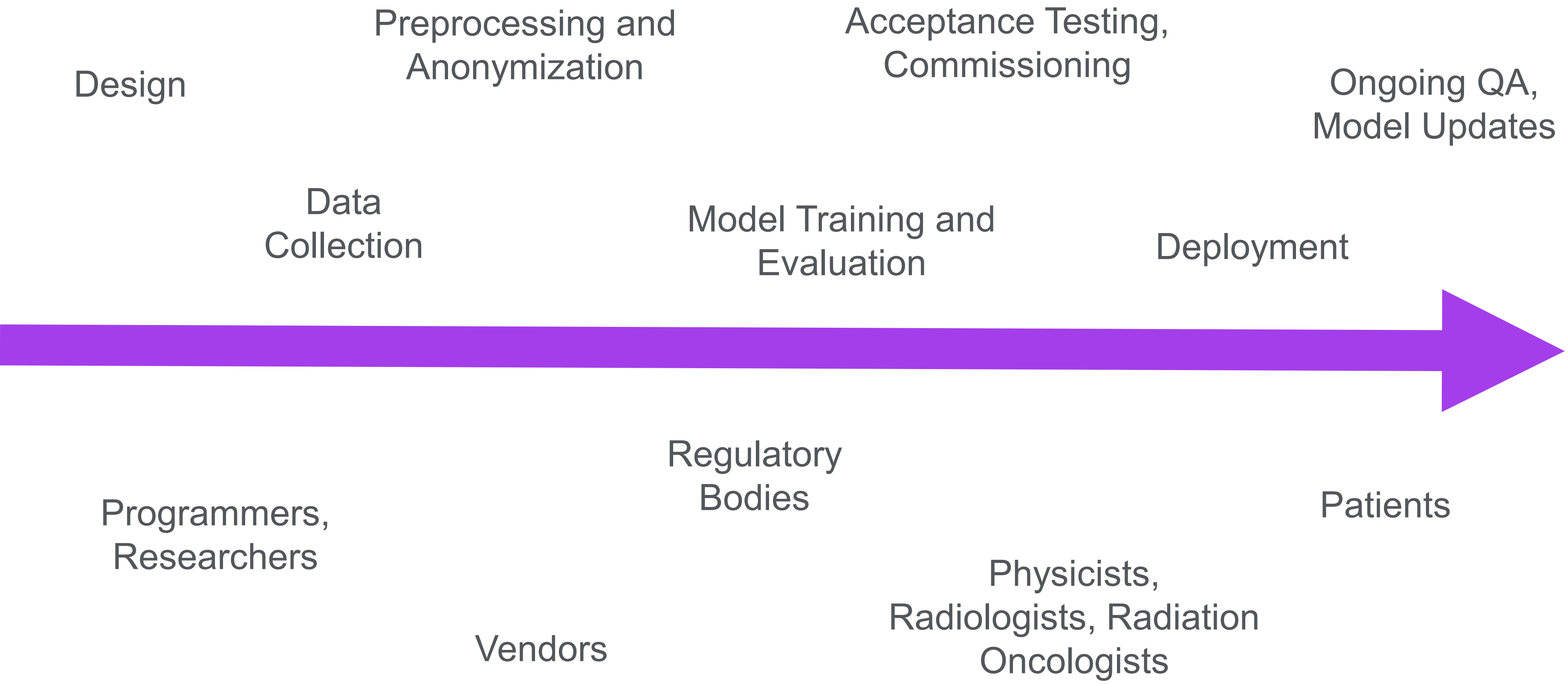
black = AI,
other colors = radiologists

Breast Cancer Prediction (2 yrs, USA)



McKinney et al. “International evaluation of an AI system for breast cancer screening.” Nature (2020).

What?



Design

Data Collection

Preprocessing and Anonymization

Model Training and Evaluation

Acceptance Testing, Commissioning

Deployment

Ongoing QA, Model Updates

Programmers, Researchers

Vendors

Regulatory Bodies

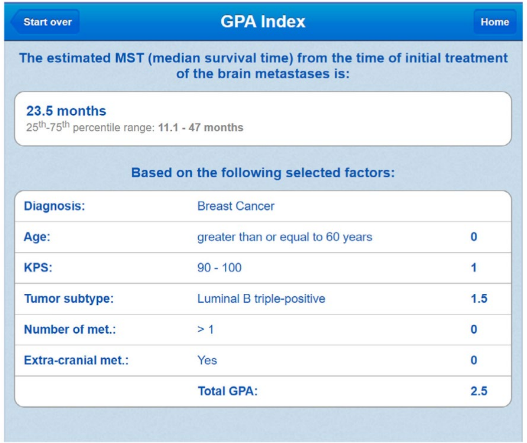
Physicists, Radiologists, Radiation Oncologists

Patients

Who?

Is deep learning necessary?

There are many cases in which the potential benefit, maybe a few percent increase in accuracy, might be outweighed by the need for interpretability.



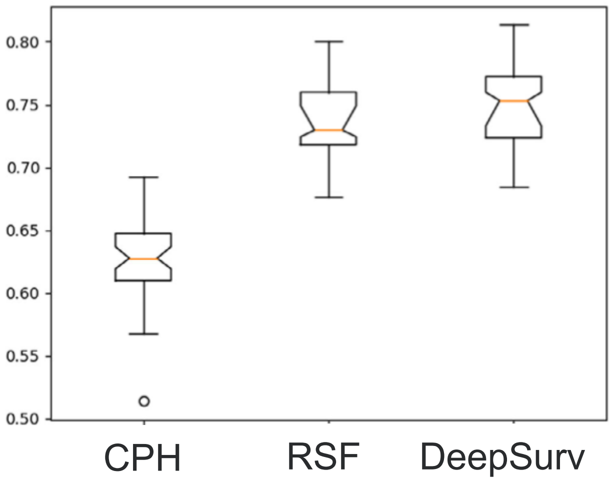
The estimated MST (median survival time) from the time of initial treatment of the brain metastases is:

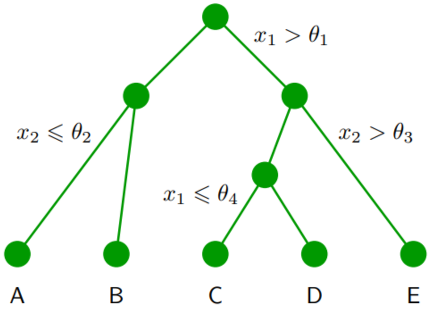
23.5 months
25th-75th percentile range: **11.1 - 47 months**

Based on the following selected factors:

Diagnosis:	Breast Cancer	
Age:	greater than or equal to 60 years	0
KPS:	90 - 100	1
Tumor subtype:	Luminal B triple-positive	1.5
Number of met.:	> 1	0
Extra-cranial met.:	Yes	0
	Total GPA:	2.5

Test set concordance indices





Bice et al. “Deep-learning based survival analysis for brain metastasis patients with the national cancer database.” JACMP (2020).

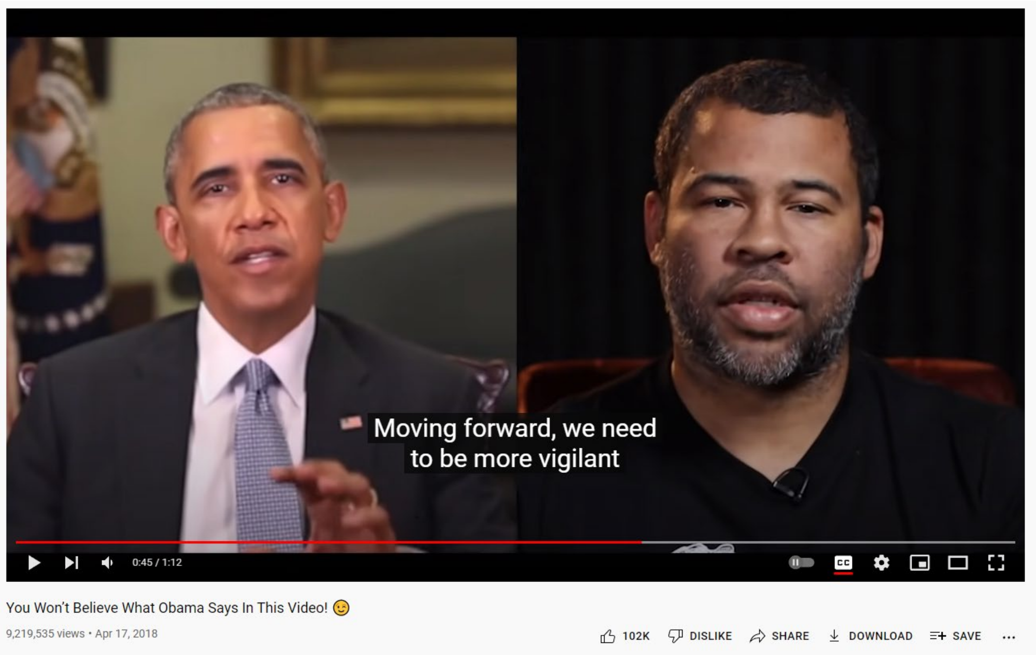
What are the needs/objectives for the tool?

- Same as any other research and development project.
- Patient care, societal benefit, monetary or **professional** gain

The international journal of science / 24 October 2019
index
Young universities
nature
QUANTUM SUPREMACY
Classical supercomputer outperformed by quantum chip for the first time
Secret history
Fossils rescue mammals from the shadow of dinosaurs
History of science
Industrial research – from Nylon to Nobels to now
Bacterial activity
How gut microbiota influences fear-related learning

Design

Are ethical principles, beneficence, justice, and autonomy, supported?


Moving forward, we need to be more vigilant

You Won't Believe What Obama Says In This Video! 😉

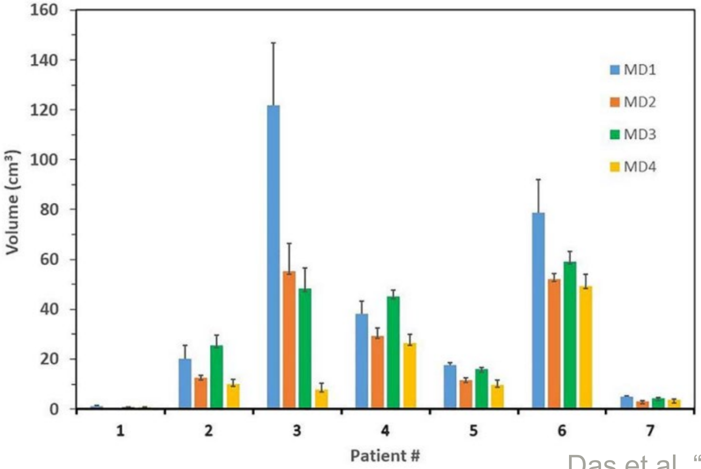
9,219,535 views • Apr 17, 2018

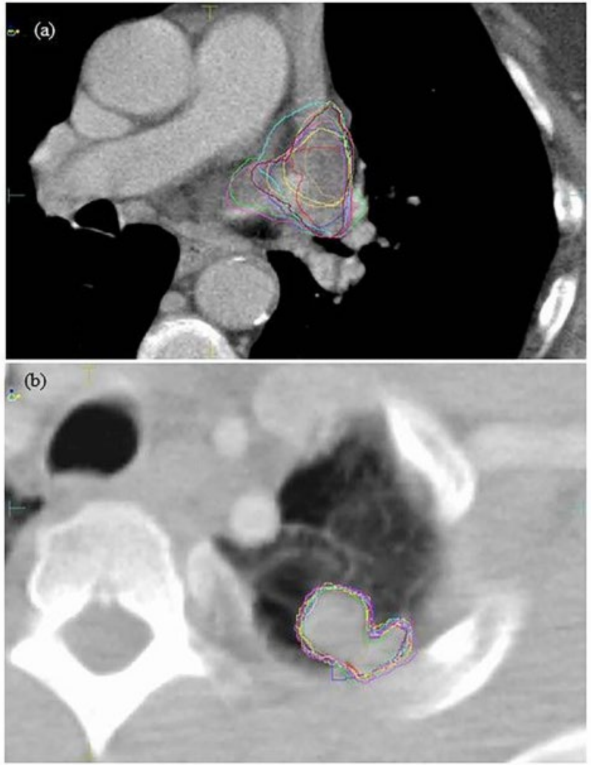
- Deepfakes
  - ☺ AI voice and video actor in-painting, superhero children's hospital visits
  - ☹ Fake news, celebrity pornographic videos, scams, financial fraud

https://www.youtube.com/watch?v=cQ54GDm1eL0

Labeling errors

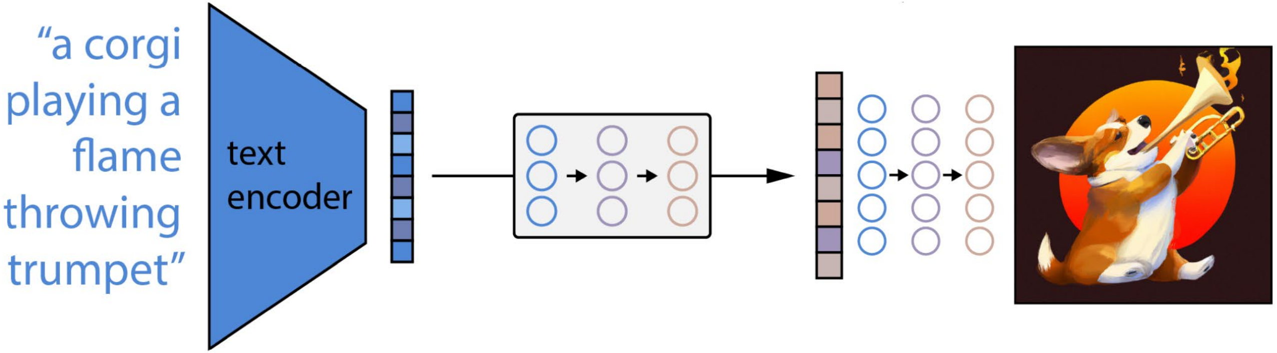
- Your model is only as good as your dataset.
- When using a labeled dataset, it is important to be conscientious of the labeling process.
- When a deep learning model is trained, it is taught to **reproduce the training labels**.





Das et al. “Intra- and inter-physician variability in target volume delineation in radiotherapy.” Journal of Radiation Research (2021).

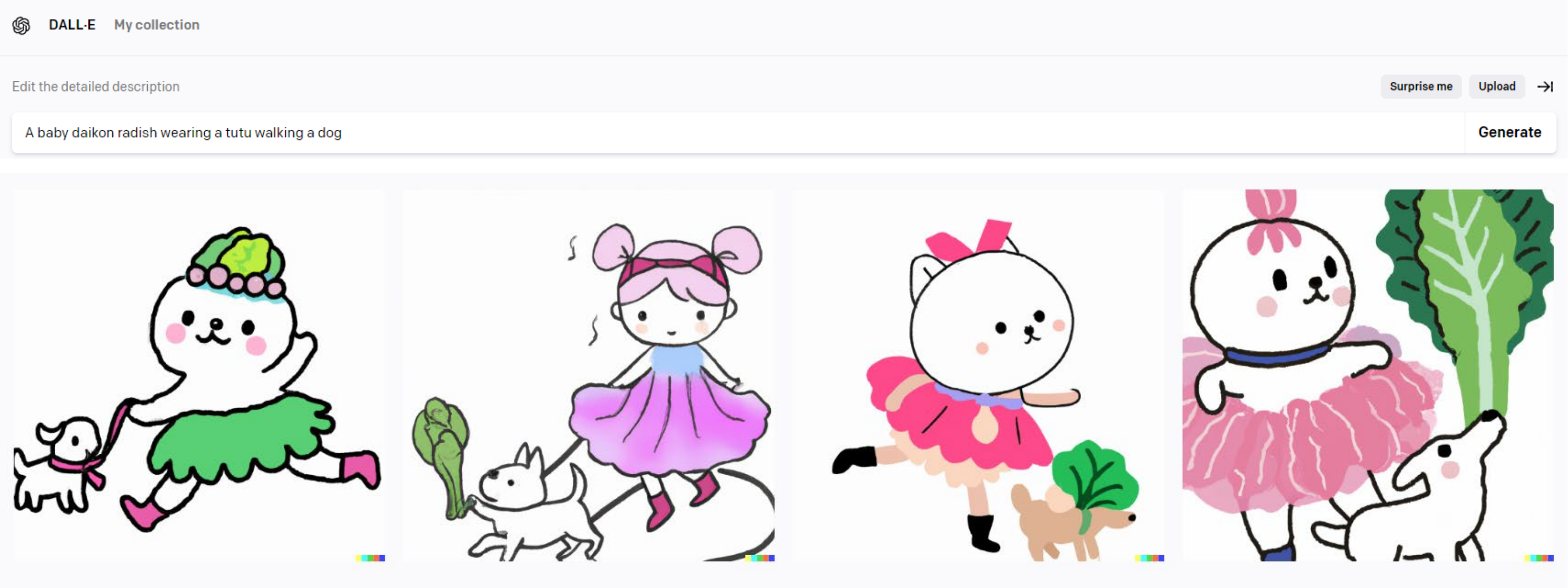
Representation Bias



12 billion parameters. Trained on 250 million image-text pairs curated from the internet.

Ramesh et al. "Zero-shot text-to-image generation." arXiv:2102.12092v2 (2021).

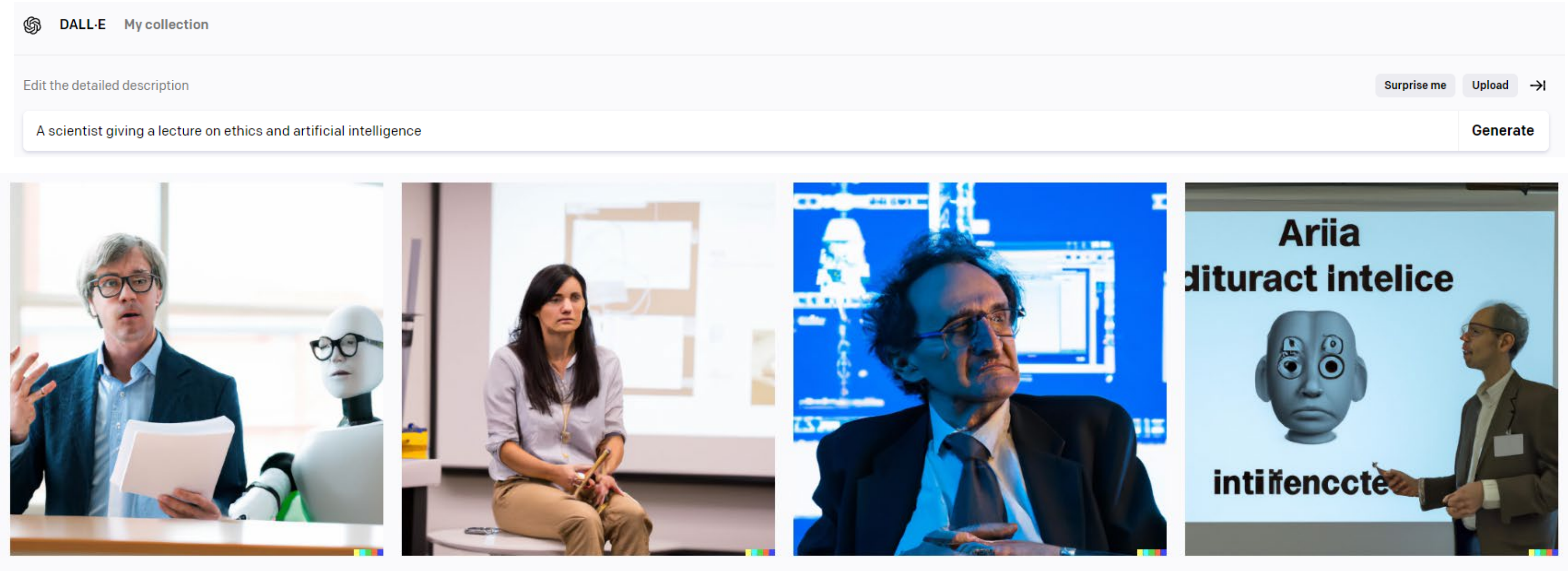
Representation Bias



A baby daikon radish wearing a tutu walking a dog

Generated with DALL-E 2, October 2022

Representation Bias



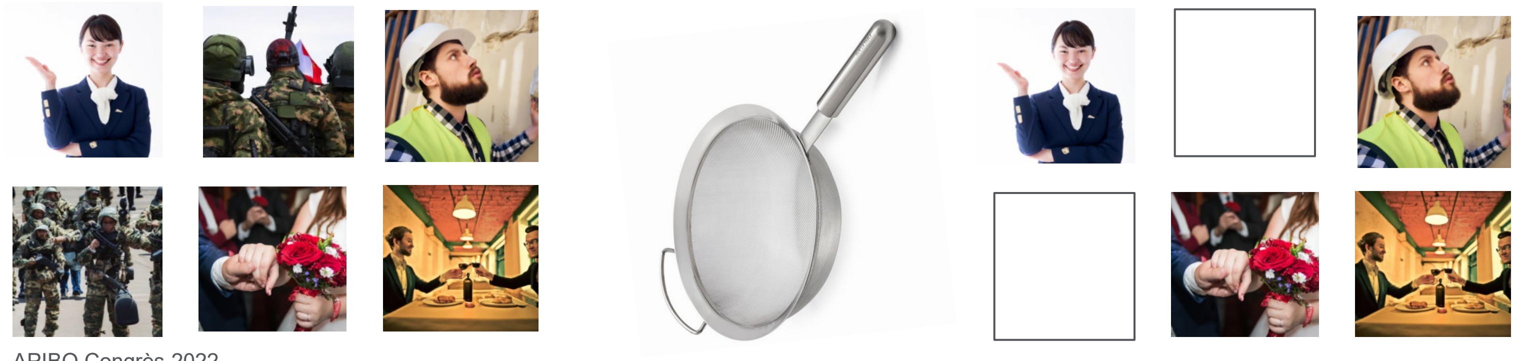
Edit the detailed description

A scientist giving a lecture on ethics and artificial intelligence

Generated with DALL-E 2, October 2022

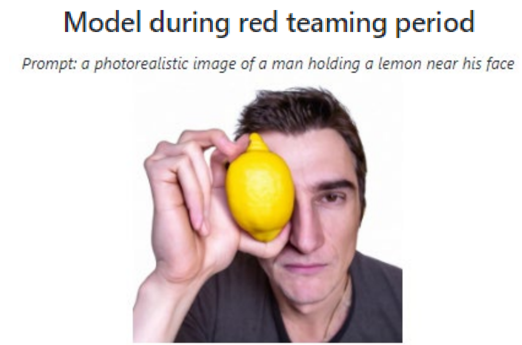
Representation Bias

- This is a **statistical** model that draws conclusions from the training dataset.
- How to we create a model which is representative of the training data distribution without harmful biases?
- Maybe it’s fair that DALL-E associates “scientist” with an older person, but not fair that it associates “scientist” with “white” or “man”.
- What if we filter the dataset, or process the outputs, to make the system more ethical?



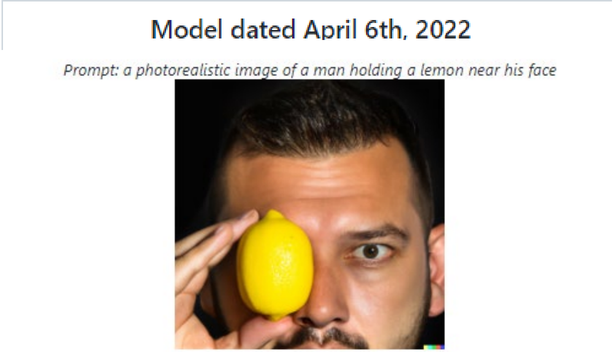
Representation Bias

- OpenAI tried this! “Red team” with diverse make-up to filter the dataset and introduce post-processing, improves but does not fix the model.
- **“You can’t be what you can’t see.”**
  - If we want an unbiased model, we need an unbiased dataset.
- What are the implications of representation bias in healthcare AI? Might traditionally marginalized groups continue to be marginalized because they are underrepresented in training data?



Model during red teaming period

Prompt: a photorealistic image of a man holding a lemon near his face



Model dated April 6th, 2022

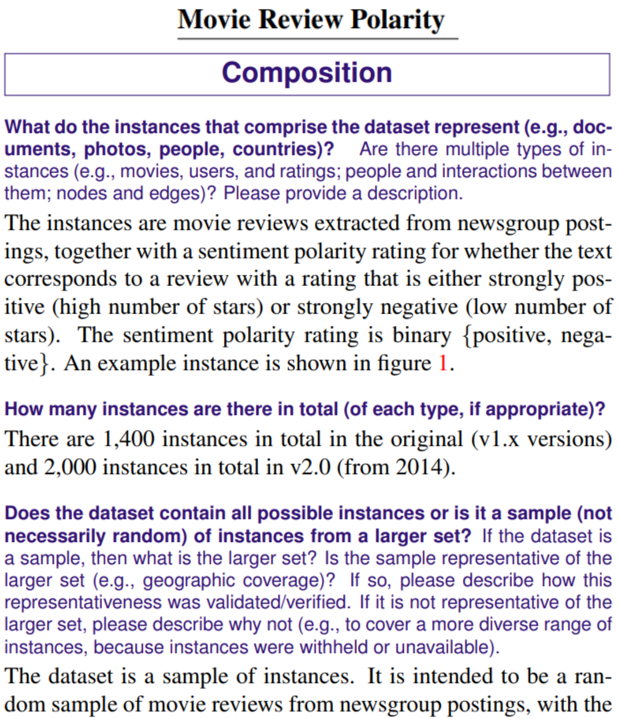
Prompt: a photorealistic image of a man holding a lemon near his face

https://github.com/openai/dalle-2-preview/blob/main/system-card.md

NYU Langone Health

Dataset Datasheets

- As with DALL-E, it might not be possible feasible to curtail all of the dataset’s potential biases.
- Data curators can at least make an effort at transparency with **Dataset Datasheets**.
  - Addresses **motivation, composition, collection process, preprocessing, uses, distribution, and maintenance** of the dataset.



Movie Review Polarity

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are movie reviews extracted from newsgroup postings, together with a sentiment polarity rating for whether the text corresponds to a review with a rating that is either strongly positive (high number of stars) or strongly negative (low number of stars). The sentiment polarity rating is binary {positive, negative}. An example instance is shown in figure 1.

How many instances are there in total (of each type, if appropriate)?

There are 1,400 instances in total in the original (v1.x versions) and 2,000 instances in total in v2.0 (from 2014).

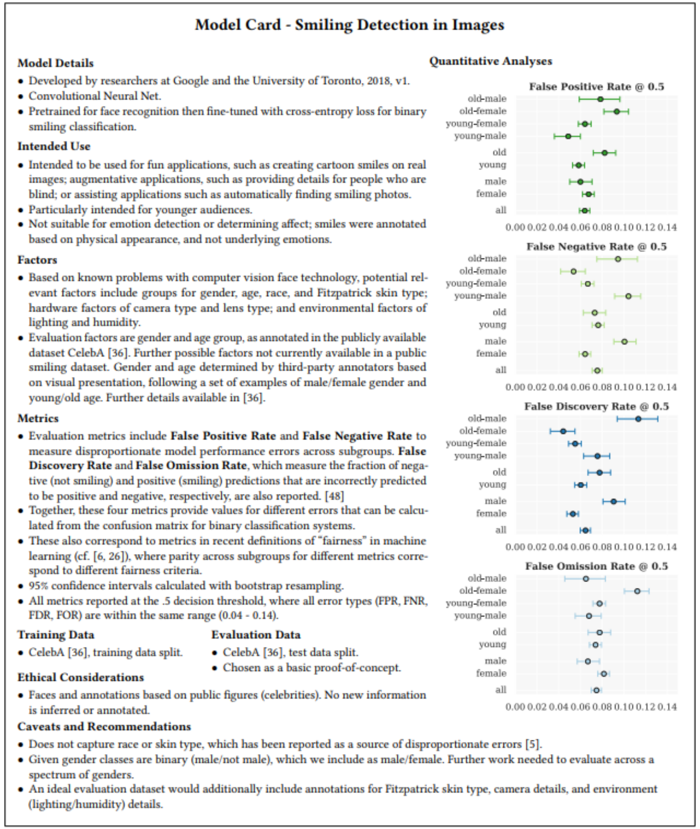
Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of instances. It is intended to be a random sample of movie reviews from newsgroup postings, with the

Gebru et al. “Datasheets for datasets.” arXiv:1803.09010v8 (2021)

Model Cards

Transparent account of model training, intended use, metrics, ethical considerations, and potential shortcomings.



Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data

- CelebA [36], training data split.

Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Quantitative Analyses

False Positive Rate @ 0.5
False Negative Rate @ 0.5
False Discovery Rate @ 0.5
False Omission Rate @ 0.5
old-male
old-female
young-female
young-male
old
young
male
female
all
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14

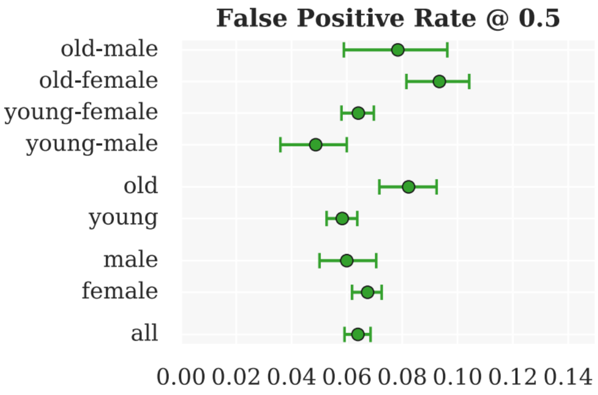
Caveats and Recommendations

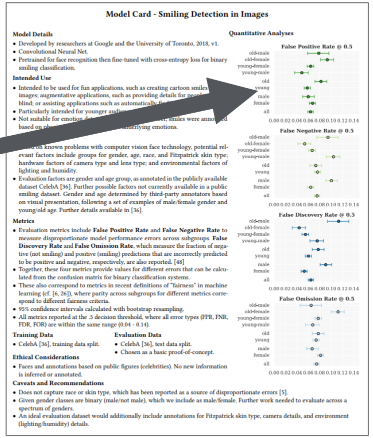
- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Mitchell et al. "Model cards for model reporting." arXiv:1810.03993v2 (2019).

Deep Learning Interpretability

- Is it enough for the developers to be transparent about the dataset/model?
- “Global interpretability” is difficult for neural networks.
- *Modular* global interpretability: how does the model respond, on average, to different inputs? Partial dependence plotting, ablation analysis, etc.





Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles images; augmentative applications, such as providing details for peopl blind; or assisting applications such as automatically fi
- Particularly intended for younger audien
- Not suitable for emotion de smiles were anno ed based on ph nderlying emotions.

- on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of “fairness” in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data

- CelebA [36], training data split.

Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

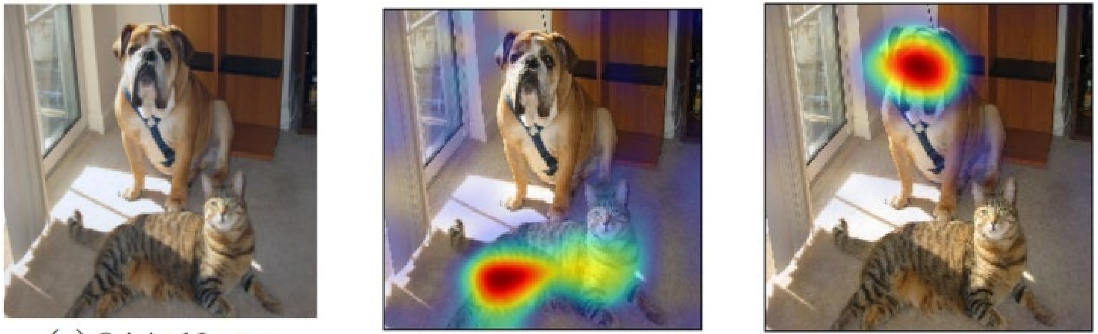
Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Christoph Molnar. *Interpretable Machine Learning*, 2022.

Local Interpretability

- Local interpretability: why did the model respond in a particular way to a specific input?
- Saliency mapping for image interpretably.



(a) Original Image

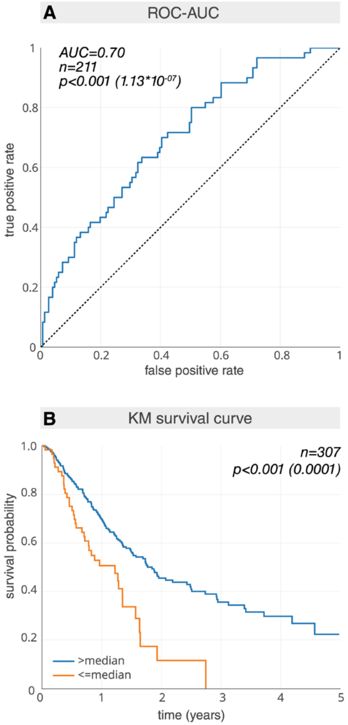
(c) Grad-CAM ‘Cat’

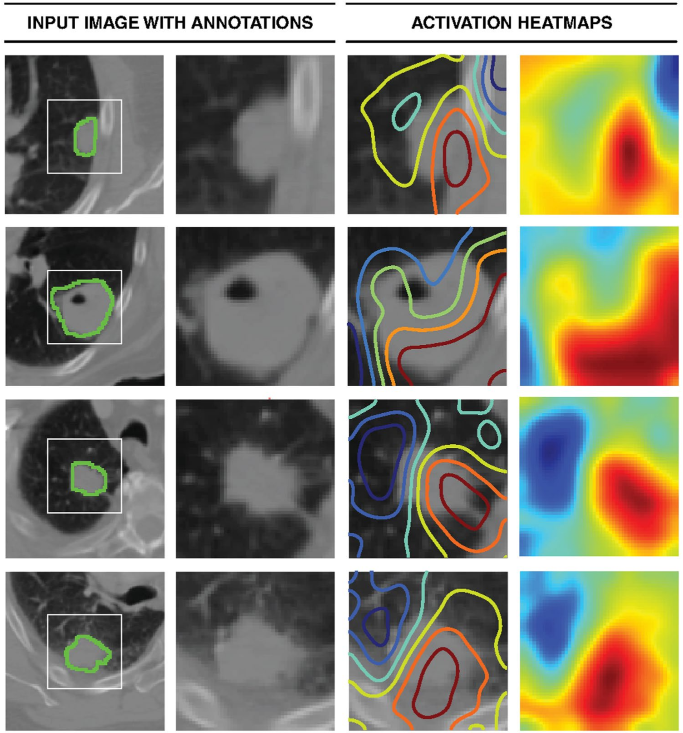
(i) Grad-CAM ‘Dog’

Ramprasaath et al. “Grad-CAM: visual explanations from deep networks via gradient-based localization.” arXiv:1610.02391v4 (2019)

Importance of Interpretability

- Consider a risk stratification model that estimates prognosis based on image data, such as Hosny et al. (2018).
- Is the model’s decision informed by relevant clinical information or some confounding variable such as COVID?

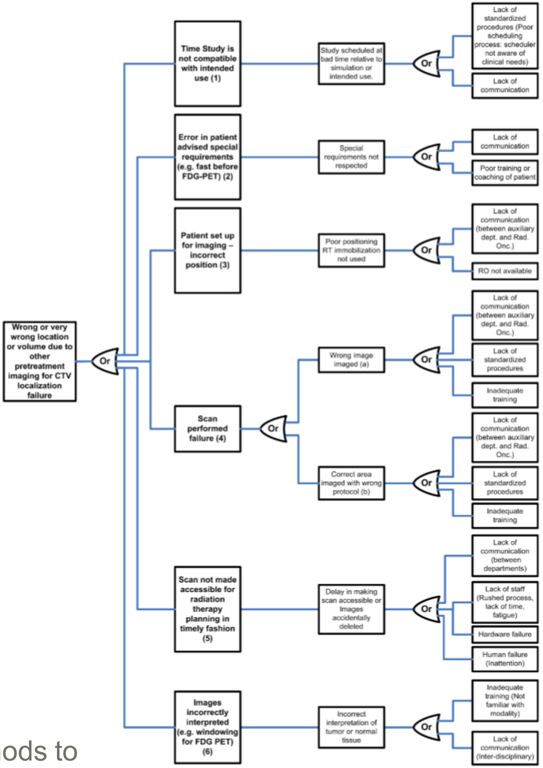




Hosny et al. “Deep learning for lung cancer prognostication: a retrospective radiomics study.” PLOS Medicine, (2018)

Acceptance Testing, Commissioning, and Ongoing QA

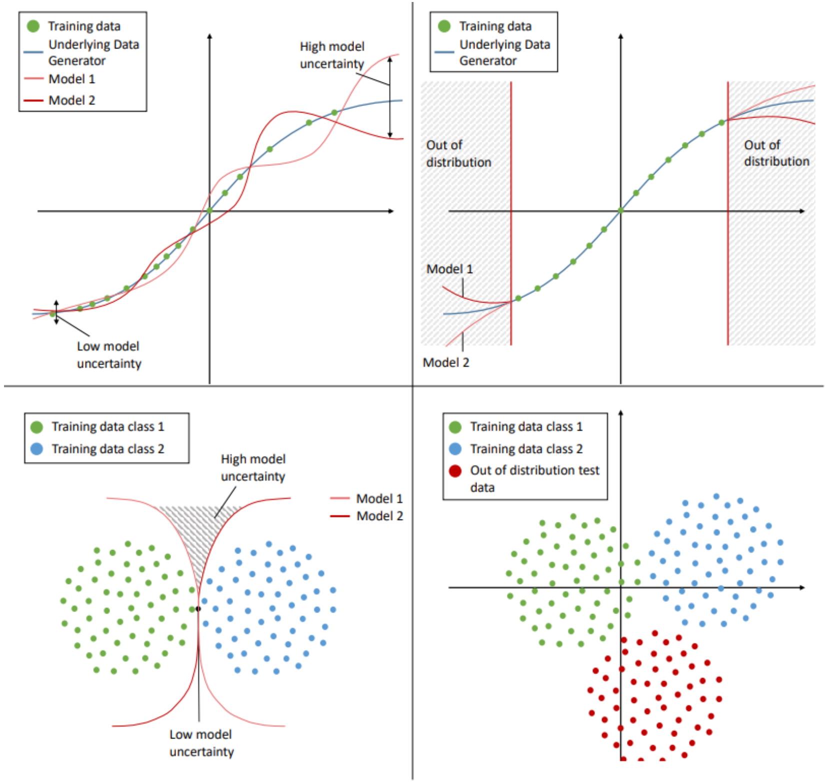
- Our role as experts to
  - Validate that technology is performing as specified by the vendor
  - Identify and prepare for possible failure modes (FMEA).
  - Create checklists including ethical interrogation.
  - Establish baseline performance, verify that AI performance is stable over time and after software (model) updates


Wrong or very wrong location or volume due to other pretreatment imaging for CTV localization failure
Or
Time Study is not compatible with intended use (1)
Study scheduled at bad time relative to simulation or intended use.
Or
Lack of standardized procedures (Poor scheduling process: scheduler not aware of clinical needs)
Lack of communication
Error in patient advised special requirements (e.g. fast before FDG-PET) (2)
Special requirements not respected
Or
Lack of communication
Poor training or coaching of patient
Patient set up for imaging – incorrect position (3)
Poor positioning RT immobilization not used
Or
Lack of communication (between auxiliary dept. and Rad. Onc.)
RO not available
Scan performed failure (4)
Or
Wrong image imaged (a)
Or
Lack of communication (between auxiliary dept. and Rad. Onc.)
Lack of standardized procedures
Inadequate training
Correct area imaged with wrong protocol (b)
Or
Lack of communication (between auxiliary dept. and Rad. Onc.)
Lack of standardized procedures
Inadequate training
Scan not made accessible for radiation therapy planning in timely fashion (5)
Delay in making scan accessible or Images accidentally deleted
Or
Lack of communication (between departments)
Lack of staff (Rushed process, lack of time, fatigue)
Hardware failure
Human failure (Inattention)
Images incorrectly interpreted (e.g. windowing for FDG PET) (6)
Incorrect interpretation of tumor or normal tissue
Or
Inadequate training (Not familiar with modality)
Lack of communication (Inter-disciplinary)

Huq et al. “TG-100: Application of risk analysis methods to radiation therapy quality management.” MedPhys, 2016.

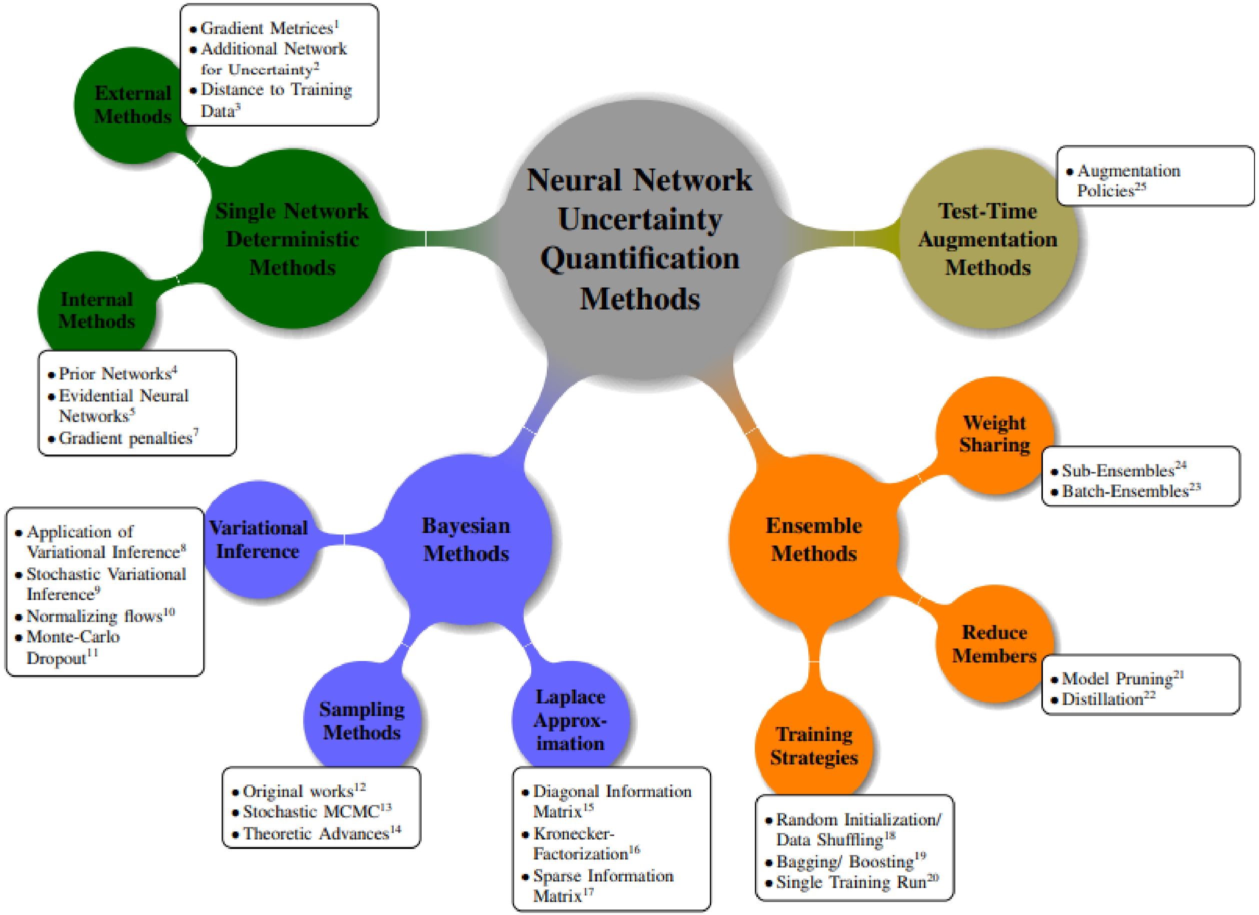
Uncertainty Quantification

- What is our cue to step in and override a decision?
- Uncertainty quantification –
  - Model performance varying over dataset?
  - Is training dataset population representative of deployment population?



Gawlikowski et al. "A survey of uncertainty in deep neural networks." arXiv:2107.03342v3 (2022).

Uncertainty Quantification



Neural Network Uncertainty Quantification Methods

Single Network Deterministic Methods

External Methods
- Gradient Metrices[1]
- Additional Network for Uncertainty[2]
- Distance to Training Data[3]

Internal Methods
- Prior Networks[4]
- Evidential Neural Networks[5]
- Gradient penalties[7]

Test-Time Augmentation Methods
- Augmentation Policies[25]

Bayesian Methods

Variational Inference
- Application of Variational Inference[8]
- Stochastic Variational Inference[9]
- Normalizing flows[10]
- Monte-Carlo Dropout[11]

Sampling Methods
- Original works[12]
- Stochastic MCMC[13]
- Theoretic Advances[14]

Laplace Approximation
- Diagonal Information Matrix[15]
- Kronecker-Factorization[16]
- Sparse Information Matrix[17]

Ensemble Methods

Weight Sharing
- Sub-Ensembles[24]
- Batch-Ensembles[23]

Reduce Members
- Model Pruning[21]
- Distillation[22]

Training Strategies
- Random Initialization/ Data Shuffling[18]
- Bagging/ Boosting[19]
- Single Training Run[20]

Gawlikowski et al. "A survey of uncertainty in deep neural networks." arXiv:2107.03342v3 (2022).

Summary of Ethical AI Toolkit

- Autonomy, beneficence, justice
- Ethics Codes
- Transparent development
  - Dataset Datasheets
  - Model Cards
- Quality assurance, quality improvement
  - Uncertainty quantification
  - Average model performance / bias testing
  - Saliency mapping

CASE STUDY / DILEMMAS

NYU Langone Health

Scenario 1:

Our department purchased a commercial AI package for contouring. In-house testing shows that it does a poor job of contouring the optic chiasm, an adequate job in some other sites, and does well in others + it saves time + more consistent. What do we do now?

1. Use it **Potential for patient injury?**
2. Ask for a refund **Failure to provide a service to our patients?**
3. Partial use: use for certain anatomical sites and require human oversight **Automation Bias?**
4. Delayed use: wait until the vendor comes out with a new version that corrects the issue
5. Collaboration: work with the vendor (share data?) so the AI can be improved for all patients

Justice, privacy, compliance, IP protection?

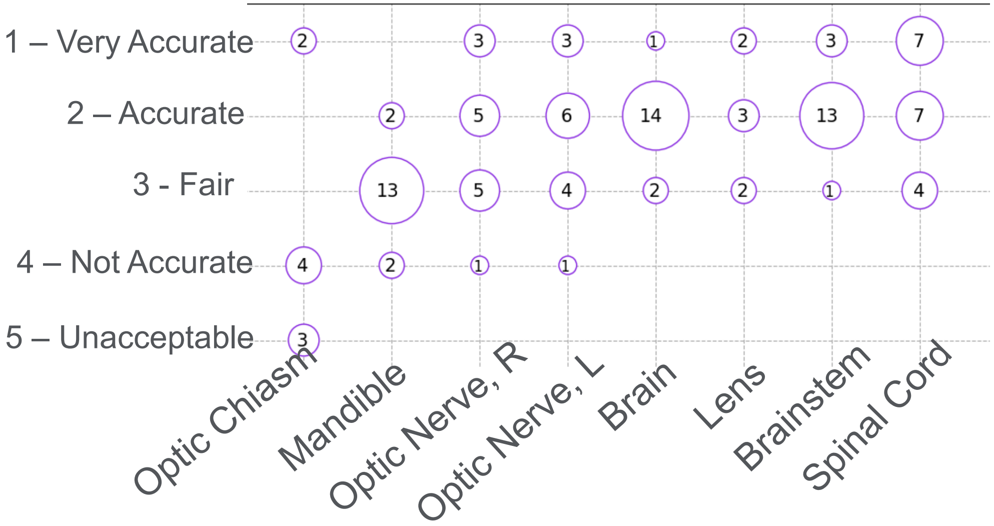
Manage Automation Bias - Part of the QA / QI stage of AI

1. Know your data and share with those doing / overseeing the work

Transparency + Uncertainty

2. Incentivize oversight
   Reward 'catches'
   Consider time pressures

AI Contouring Assessment by Dosimetrists



Scenario:

Our facility has identified LGBT patient experience and care as an area for improvement. We use AI in Radiology (diagnosis, prognosis). We test the diagnosis tool on a small group of transgendered patients and find it performs badly for one patient. What do we do?

1. Continue to use it with oversight **Automation bias?**
2. Delay use until you test on a larger cohort of patients **Delays and disruption of existing workflow?**
3. Stop using it in patients who have undergone gender affirming surgery in the ROI **Resources? Inequity?**
4. Stop using it for all patients **Delays and disruption of existing workflow? Reduction in Quality?**
5. Work with vendor to update their training data and testing (or update our model) **Resources? HIPPA?**
6. Include transgendered patients/experts in the conversation **Education?**
7. Discuss limitations of the software with all patients **Time? Inequity?**

Hybrid Approach: AI + Expert Review

1. Educate on the topic.. Include patients in the discussion
   - Multidisciplinary teams needed to provide comprehensive care for these patients
   - Need to know about neo-anatomical alterations and complications from surtery
   - Understand different risk factors in this patient group

2. Since the institution has identified **health equity** as a need, ask for resource support, collaborate with vendor, share findings.

Stowell et al. "Gender-affirming surgical techniques, complications, and imaging considerations for the abdominal radiologist" *Abdom Radiol* (2020) https://doi.org/10.1007/s00261-019-02398-1

Related: Racial Bias in Healthcare AI?

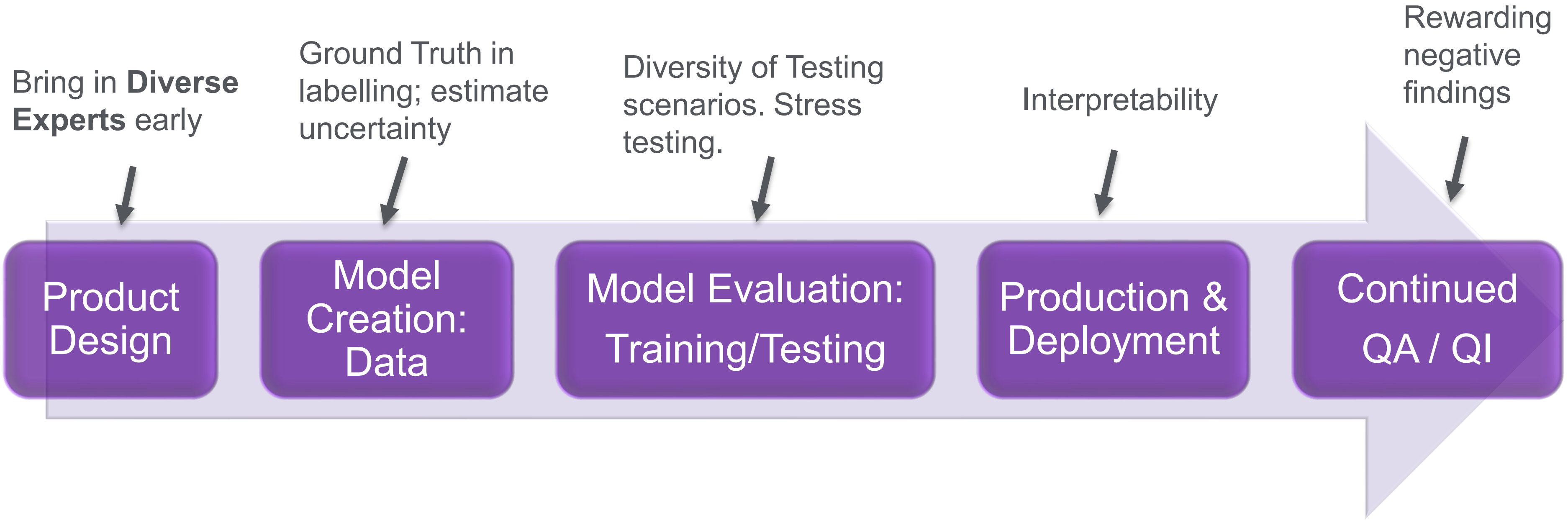
AI *risk-prediction* algorithms are currently being used for ~200 million Americans to improve the care of patients with complex needs. One impacting 70 millions was found to be inherently biased against African Americans.

17.7% patients identified as needing extra support were African Americans; eliminating this bias would cause the number to jump to 46.5%.

Healthcare expenditure was used a proxy for patient health. The average African American spent the same as the average white, but they were significantly sicker: hypertension, diabetes, renal failure, bad cholesterol, anemia.

Obermeyer et al. “Dissecting racial bias in an algorithm used to manage the health of populations” Science (2019) https://doi.org/10.1126/science.aax2342.

AI “Life-Cycle” … could this been found earlier?



Bring in **Diverse Experts** early

Ground Truth in labelling; estimate uncertainty

Diversity of Testing scenarios. Stress testing.

Interpretability

Rewarding negative findings

Product Design

Model Creation: Data

Model Evaluation: Training/Testing

Production & Deployment

Continued QA / QI

Autonomy √
Beneficence √
Nonmaleficence √
Justice √

Stop and ask.
Reward inquiry from diverse perspectives.

APIBQ CODE D’ÉTHIQUE

Responsabilité sociale

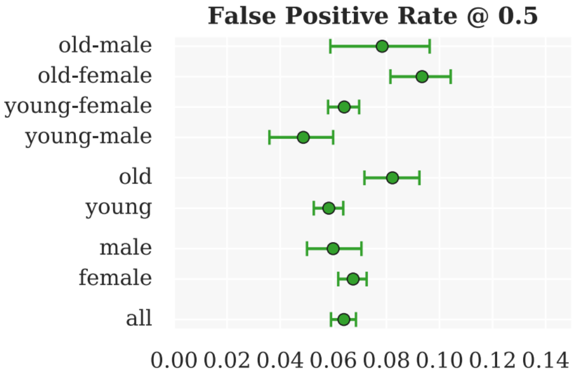
Tout member de l’APIBQ est encouragé *à* offrir ses services et son expertise professionnelle gracieusement, pour l’advancement de la science ou la défense de causes humananitaires.

Wow !!

84% reduction in bias after fixing labelling
No monetary reimbursement or support from the vendor. NIH grant

Model Cards – Smile Detection – Prior Example

Transparency: model training, intended use, metrics, ethical considerations, potential shortcomings.



Ethical Considerations

Faces and annotations based on public figures (celebrities). No new information is inferred or annotated

Perkowitz (2021) “The Bias in the Machine: Facial Recognition Technology and Racial Disparities.” https://doi.org/10.21428/2c646de5.62272586

Mitchell et al. “Model cards for model reporting.” arXiv:1810.03993v2 (2019).

Thank You, Team!

AAPM Colleagues:

Meghan, Hyun, PhD
Assoc. Prof.
Radiation Oncology
University of Nebraska Med. Center

Dandan Zheng, PhD
Prof., Dir. Of Physics
Radiation Oncology
University of Rochester
Rochester, NY

Steve G. Langer, PhD
Professor
Radiology Dept.
Mayo Clinic
Rochester, MN

NYU Colleagues:
Mario Serrano Sosa, PhD
Jason Domogauer, MD, PhD
David Barbee, PhD
Allison McCarthy
Peter Milien
Dina Patel
Pine Cheng
Elena Cantonjos

Thank You, APIBQ!

NYU Langone Health

THANK YOU

NYU Langone Health