Ethics in deployed AI systems

Taylor Kessler Faulkner

University of Washington · February 18, 2026

Some slide content inspired by Timnit Gebru and Emily Denton

Tell Your Friends!

Our applications are open for 2026–2027!

Spread the word if you have colleagues/friends who are interested in developing their AI skills

If you have a quote to share about your experience in the certificate, let us know!

Thank you for your hard work and dedication to the program!

AI as a tool

> *"AI is neither good nor evil. It's a tool, a technology for us to use."*
>
> *Oren Etzioni — Professor Emeritus, University of Washington; Founding CEO, Allen Institute for AI*

> *"Any new technology, if it's used by evil people, bad things can happen. But that's more a question of the politics of the technology."*
>
> *Geoffrey Hinton — Professor Emeritus, University of Toronto; Former VP & Engineering Fellow, Google; "Godfather of AI"*

Application: Smart Offices

Image source: integrated.com/solutions/smart-office-iot

An Optimistic Vision

> *"Imagine for a moment that you're in an office, hard at work. But it's no ordinary office. By observing cues like your posture, tone of voice, and breathing patterns, it can sense your mood and tailor the lighting and sound accordingly. Through gradual ambient shifts, the space around you can take the edge off when you're stressed, or boost your creativity when you hit a lull. Imagine further that you're a designer, using tools with equally perceptive abilities: at each step in the process, they riff on your ideas based on their knowledge of your own creative persona, contrasted with features from the best work of others."*

Landay (2019). "Smart Interfaces for Human-Centered AI"

A concern for the future

> *"Someday you may have to work in an office where the lights are carefully programmed and tested by your employer to hack your body's natural production of melatonin through the use of blue light, eking out every drop of energy you have while you're on the clock, leaving you physically and emotionally drained when you leave work. Your eye movements may someday come under the scrutiny of algorithms unknown to you that classifies you on dimensions such as 'narcissism' and 'psychopathy', determining your career and indeed your life prospects."*

Alkhatib (2019). "Anthropological/Artificial Intelligence & the HAI"

Application: What does your face say about you?

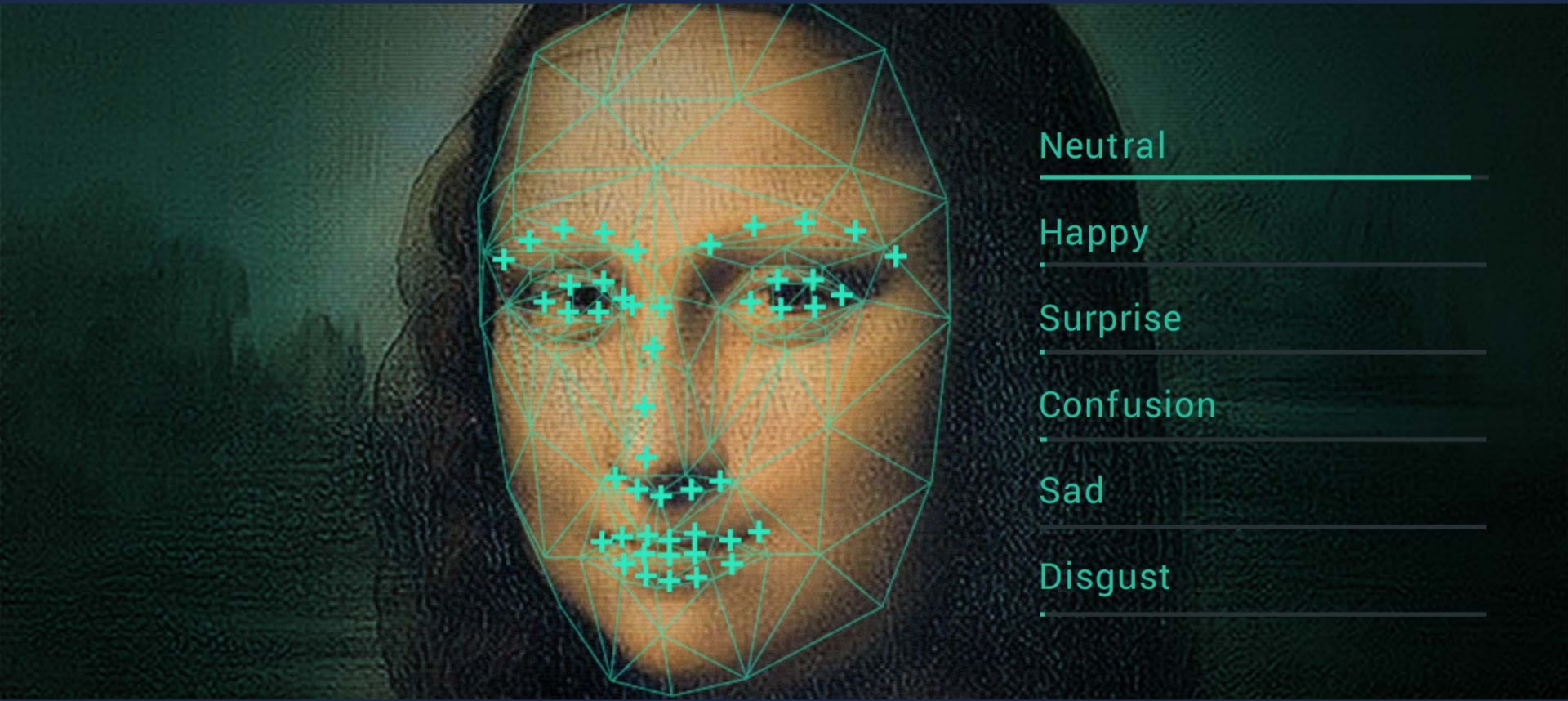


Image source: https://adverteyes.ai/vision-ai-decodes-the-perception-of-mona-lisa-enigmatic-smile/

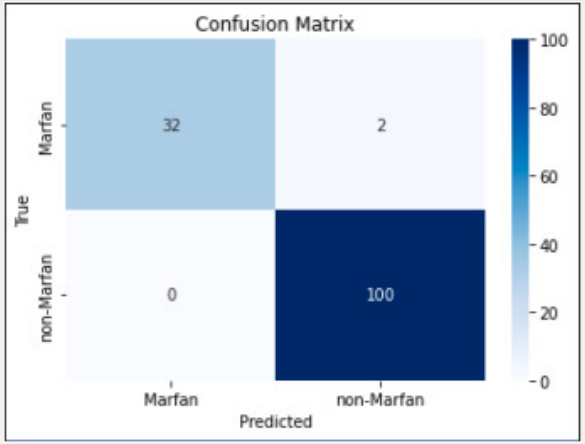
Marfan Syndrome diagnosis via AI face analysis

> *"Patients living with Marfan Syndrome are usually very tall and thin… They have long faces and are prone to spine and joint issues. However, many are not diagnosed… Being able to identify individuals from a photograph with AI will enhance diagnosis and enable protective therapies."*

John Elefteriades, MD — Professor of Surgery, Yale School of Medicine

Study Results

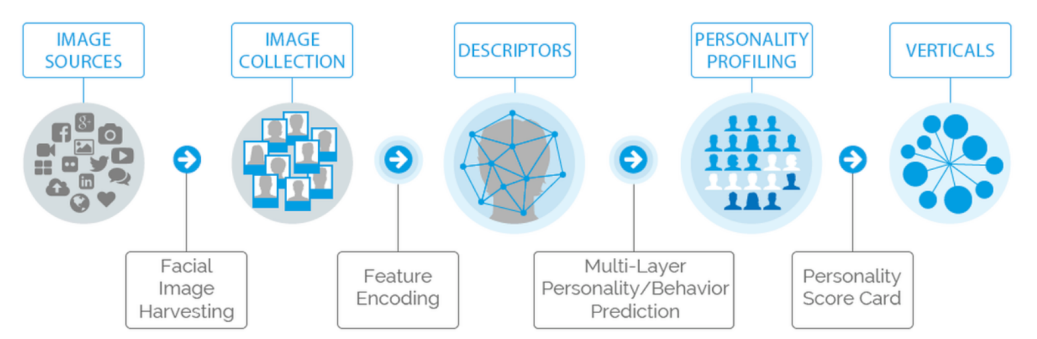
98.5%	Overall accuracy
0%	False positives
2%	False negatives



Saksenberg et al. (2024). "Pilot study exploring artificial intelligence for facial-image-based diagnosis of Marfan syndrome." Heliyon 10.13.

Faception: Personality Profiling from Faces

> *"Faception is first-to-technology and first-to-market with proprietary computer vision and machine learning technology for* ***profiling people*** *and revealing their personality* ***based only on their facial image****."*
>
> *"We live in a dangerous world... What if it was possible to know whether an anonymous individual is a potential terrorist, an aggressive person, or a potential criminal? Better yet, what if that information could be obtained and used in real-time, when it matters the most?"*
>
> — *Faception startup*



Marfan Syndrome diagnosis via AI face analysis

> *"Patients living with Marfan Syndrome are usually very tall and thin... They have long faces and are prone to spine and joint issues. However, many are not diagnosed... Being able to identify individuals from a photograph with AI will enhance diagnosis and enable protective therapies."*

John Elefteriades, MD — Professor of Surgery, Yale School of Medicine

Study Results

98.5%	Overall accuracy
0%	False positives
2%	False negatives

⚠ **Dataset:**

- **83% Caucasian/Hispanic**
- **10% Black**
- **7% Asian**

Saksenberg et al. (2024). "Pilot study exploring artificial intelligence for facial-image-based diagnosis of Marfan syndrome." Heliyon 10.13.

Application: Deepfakes

Image source: https://www.youtube.com/watch?v=cQ54GDm1eL0&t=10s

Bringing Dalí back to life

“Dalí was prophetic in many ways and understood his historical importance… He wrote, If someday I may die, though it is unlikely, I hope the people in the cafes will say, ‘Dalí has died, but not entirely.’ This technology lets visitors experience his bigger-than-life personality in addition to our unparalleled collection of his works.”

Dr. Hank Hine— Executive Director, The Dalí

https://thedali.org/exhibit/dali-lives/

AI deepfakes and workers' rights

> *"We're not going to consent to a contract that allows companies to abuse A.I. to the detriment of our members. Enough is enough."*

Fran Drescher — Actor; President, SAG-AFTRA

> *"The technology is complex, but regardless of the generative A.I. system in use, if your performance is being replicated, you deserve access to essential information, you deserve to give or not give your consent, and you need to be paid properly."*

Sarah Elmaleh — Actor; Committee Chair, SAG-AFTRA

Image of Frances Fisher by DAVID LIVINGSTON/GETTY IMAGES

Frameworks for Assessing Potential Harm

Model Cards for Model Reporting

Model cards: a standardized framework for transparent model reporting

For Model Creators

- Encourages thorough and critical evaluations
- Outlines potential risks or harms and implications of use

For Model Consumers

- Provides information to facilitate informed decision-making

Mitchell et al. (2019). Model Cards for Model Reporting.

Slide content credit: Timnit Gebru and Emily Denton

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.

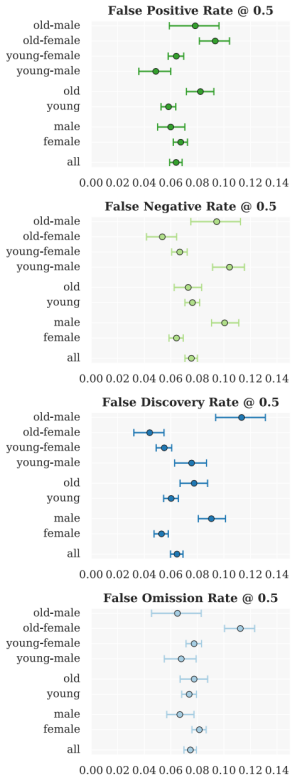
Evaluation Data

- 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



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.

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Mitchell et al. (2019). Model Cards for Model Reporting.

NIST AI Risk Management Framework: Potential Harms

Artificial Intelligence Risk Management Framework: Generative AI Profile (July 2024)

1. CBRN Information or Capabilities
2. Confabulation
3. Dangerous, Violent, or Hateful Content
4. Data Privacy
5. Environmental Impacts
6. Harmful Bias or Homogenization
7. Human-AI Configuration
8. Information Integrity
9. Information Security
10. Intellectual Property
11. Obscene, Degrading, or Abusive Content
12. Value Chain & Component Integration

EXERCISE 1 · WellScan

"Your health, understood at a glance."

WellScan is a smartphone app that uses your phone's camera to monitor your health in real time — no wearable required. Point your camera at your face for 30 seconds and WellScan uses **computer vision and deep learning** to estimate your heart rate, respiratory rate, stress level, sleep quality score, and early signs of conditions like anemia, jaundice, and hypertension.

Results are stored in your personal health timeline and can be shared with your doctor. **WellScan Premium integrates with your employer's wellness program** to earn points toward reduced health insurance premiums. Available on iOS and Android. **FDA-registered. HIPAA-compliant.**

Not a real company: pitched by Claude

WellScan · Some Identified Problems

Coercive consent

Premium discounts make the app financially mandatory. Opt-out means a real cost. That's not consent — it's coercion.

Skin tone accuracy gap

Photoplethysmography performs significantly worse on darker skin. WellScan encodes existing healthcare disparities at scale.

The regulatory gap

"FDA-registered" means self-certification. No body audits algorithmic accuracy across demographic groups for consumer health apps.

WellScan Framing

"Empowering" framing

Patients get more data about themselves — framed as autonomy, not surveillance.

"FDA-registered"

Sounds like regulatory approval. It isn't. Class I registration is self-certification — no algorithmic fairness review required.

"HIPAA-compliant"

Covers data transmission to providers, not fairness across skin tones or downstream insurer access.

Employer wellness framing

The premium discount sounds like a benefit. It makes the product functionally mandatory for those who can't afford to opt out.

Practical Steps

As a developer	Require disaggregated performance metrics by skin tone, age, and disability status before shipping. Document known failure modes.
As a product manager	Push back on employer-integration features that compromise voluntary consent. Flag "HIPAA-compliant" as a floor, not a ceiling.
As a citizen	Support algorithmic accountability bills that extend anti-discrimination law to automated health tools.
As a researcher	Publish subgroup accuracy data (and other metrics) even when it's unflattering.

EXERCISE 2 · FlowCity

"Smarter streets. Safer neighborhoods. A city that works for everyone."

FlowCity is an intelligent urban management platform using computer vision sensors mounted on existing city infrastructure — streetlights, traffic signals, transit stops — to make cities more livable in real time.

FlowCity can: detect and reroute traffic · identify overcrowded spaces · alert emergency services to accidents before a 911 call · help city planners understand how public spaces are actually used.

Does not use facial recognition. All data is processed on-device and immediately anonymized. Deployed in partnership with the City. Fully compliant with local privacy ordinances. **Trusted by 47 municipalities across North America.**

Not a real company: pitched by Claude

FlowCity Framing

"No facial recognition"

A meaningful-sounding guarantee. But it doesn't prevent inference about group behavior, crowd formation, or protest mapping.

"Anonymized"

Research shows mobility data is often re-identifiable with very few data points. The guarantee may not be technically enforceable.

Democratic framing

"The city decided" implies accountability. A 6-3 council vote is the only consent mechanism for thousands of residents.

"Trusted by 47 municipalities"

Volume of deployment implies vetting. In practice it mostly means the sales cycle worked 47 times.

FlowCity · Some Identified Problems

Who gets piloted on?

Sensor rollouts can begin in lower-income areas, framed as investment — but functioning as testing grounds before wealthier neighborhoods.

"Pedestrian flow" = encampment removal

Crowd detection tools have been used to identify and disperse unhoused communities. Neutral language, targeted impact.

Protest surveillance without FRT

No facial recognition needed to reconstruct protest movements. Crowd formation and density data alone is sufficient.

Re-identification risk

Anonymized mobility data is often re-identifiable with 3–4 data points. The guarantee may not be technically meaningful.

Practical Steps

As a practitioner	Insist on independent auditability of anonymization claims before deployment. "Trust us" is not a technical guarantee.
Working with a city	Advocate for algorithmic impact assessments before deployment — similar to environmental impact assessments. Seattle's Surveillance Ordinance (2017) is a model.
As a citizen	Seattle's Surveillance Ordinance requires city council approval for surveillance tech. You can read more about what is covered under this ordinance.