AI Ethics

By Ashish Pal

Human-Centered Design for AI

Approach

HCD involves people in every step of the design process. Your team should adopt an HCD approach to AI as early as possible - ideally, from when you begin to entertain the possibility of building an AI system.

The following six steps are intended to help you get started with applying HCD to the design of AI systems. That said, what HCD means for you will depend on your industry, your resources, your organization and the people you seek to serve.

1. Understand people's needs to define the problem

Working with people to understand the pain points in their current journeys can help find unaddressed needs. This can be done by observing people as they navigate existing tools, conducting interviews, assembling focus groups, reading user feedback and other methods. Your entire team – including data scientists and engineers – should be involved in this step, so that every team member gains an understanding of the people they hope to serve. Your team should include and involve people with diverse perspectives and backgrounds, along race, gender, and other characteristics. Sharpen your problem definition and brainstorm creative and inclusive solutions together.

A company wants to address the problem of dosage errors for immunosuppressant drugs given to patients after liver transplants. The company starts by observing physicians, nurses and other hospital staff throughout the liver transplant process. It also interviews them about the current dosage determination process - which relies on published guidelines and human judgment - and

shares video clips from the interviews with the entire development team. The company also reviews research studies and assembles focus groups of former patients and their families. All team members participate in a freewheeling brainstorming session for potential solutions.

2. Ask if AI adds value to any potential solution

Once you are clear about which need you are addressing and how, consider whether AI adds value.

- Would people generally agree that what you are trying to achieve is a good outcome?
- Would non-AI systems - such as rule-based solutions, which are easier to create, audit and maintain - be significantly less effective than an AI system?
- Is the task that you are using AI for one that people would find boring, repetitive or otherwise difficult to concentrate on?
- Have AI solutions proven to be better than other solutions for similar use cases in the past?

If you answered no to any of these questions, an AI solution may not be necessary or appropriate.

A disaster response agency is working with first responders to reduce the time it takes to rescue people from disasters, like floods. The time- and labor-intensive human review of drone and satellite photos to find stranded people increases rescue time. Everybody agrees that speeding up photo review would be a good outcome, since faster rescues could save more lives. The agency determines that an AI image recognition system would likely be more effective than a non-AI automated system for this task. It is also aware that AI-based image recognition tools have been applied successfully to review aerial footage in other industries, like agriculture. The agency therefore decides to further explore the possibility of an AI-based solution.

3. Consider the potential harms that the AI system could cause

Weigh the benefits of using AI against the potential harms, throughout the design pipeline: from collecting and labeling data, to training a model, to deploying the AI system. Consider the impact on users and on society. Your privacy team can help uncover hidden privacy issues and determine whether privacy-preserving techniques like differential privacy or federated learning may be appropriate. Take steps to reduce harms, including by embedding people - and therefore human judgment - more effectively in data selection, in model training and in the operation of the system. If you estimate that the harms are likely to outweigh the benefits, do not build the system.

An online education company wants to use an AI system to ‘read’ and automatically assign scores to student essays, while redirecting company staff to double-check random essays and to review essays that the AI system has trouble with. The system would enable the company to quickly get scores back to students. The company creates a harms review committee, which recommends that the system not be built. Some of the major harms flagged by the committee include: the potential for the AI system to pick up bias against certain patterns of language from training data and amplify it (harming people in the groups that use those patterns of language), to encourage students to ‘game’ the algorithm rather than improve their essays and to reduce the classroom role of education experts while increasing the role of technology experts.

4. Prototype, starting with non-AI solutions

Develop a non-AI prototype of your AI system quickly to see how people interact with it. This makes prototyping easier, faster and less expensive. It also gives you early information about what users expect from your system and how to make their interactions more rewarding and meaningful.

Design your prototype’s user interface to make it easy for people to learn how your system works, to toggle settings and to provide feedback.

The people giving feedback should have diverse backgrounds – including along race, gender, expertise and other characteristics. They should also understand and consent to what they are helping with and how.

A movie streaming startup wants to use AI to recommend movies to users, based on their stated preferences and viewing history. The team first invites a diverse group of users to share their stated preferences and viewing history with a movie enthusiast, who then recommends movies that the users might like. Based on these conversations and on feedback about which recommended movies users enjoyed, the team changes its approach to how movies are categorized. Getting feedback from a diverse group of users early and iterating often allows the team to improve its product early, rather than making expensive corrections later.

5. Provide ways for people to challenge the system

People who use your AI system once it is live should be able to challenge its recommendations or easily opt out of using it. Put systems and tools in place to accept, monitor and address challenges.

Talk to users and think from the perspective of a user: if you are curious or dissatisfied with the system's recommendations, would you want to challenge it by:

- Requesting an explanation of how it arrived at its recommendation?
- Requesting a change in the information you input?
- Turning off certain features?
- Reaching out to the product team on social media?
- Taking some other action?

An online video conferencing company uses AI to automatically blur the background during video calls. The company has successfully tested its product with a diverse group of people from different ethnicities. Still, it knows that there could be instances in which the video may not properly focus on a person's face. So, it makes the background blurring feature optional and adds a button for customers to report issues. The company also creates a customer service team to monitor social media and other online forums for user complaints.

6. Build in safety measures

Safety measures protect users against harm. They seek to limit unintended behavior and accidents, by ensuring that a system reliably delivers high-quality outcomes. This can only be achieved through extensive and continuous evaluation and testing. Design processes around your AI system to continuously monitor performance, delivery of intended benefits, reduction of harms, fairness metrics and any changes in how people are *actually* using it.

The kind of safety measures your system needs depends on its purpose and on the types of harms it could cause. Start by reviewing the list of safety measures built into similar non-AI products or services. Then, review your earlier analysis of the potential harms of using AI in your system (see Step 3).

Human oversight of your AI system is crucial:

- Create a human ‘red team’ to play the role of a person trying to manipulate your system into unintended behavior. Then, strengthen your system against any such manipulation.
- Determine how people in your organization can best monitor the system’s safety once it is live.
- Explore ways for your AI system to quickly alert a human when it is faced with a challenging case.
- Create ways for users and others to flag potential safety issues.

To bolster the safety of its product, a company that develops a widely-used AI-enabled voice assistant creates a permanent internal ‘red team’ to play the role of bad actors that want to manipulate the voice assistant. The red team develops adversarial inputs to fool the voice assistant. The company then uses ‘adversarial training’ to guard the product against similar adversarial inputs, improving its safety.

References

To dive deeper into the application of HCD to AI, check out these resources:

- Lex Fridman’s introductory lecture on Human-Centered Artificial Intelligence
- Google’s People + AI Research (PAIR) Guidebook
- Stanford Human-Centered Artificial Intelligence (HAI) research

Q&A (Exercise)

1) Reducing plastic waste

A Cambodian organization wants to help reduce the significant amounts of plastic waste that pollute the Mekong River System. Which of the following would be an appropriate way to start? (Your answer might use more than one option.)

- Watch the people currently addressing the problem as they navigate existing tools and processes.
- Conduct individual interviews with the people currently addressing the problem.
- Assemble focus groups that consist of people currently addressing the problem.

2) Detecting breast cancer

Pathologists try to detect breast cancer by examining cells on tissue slides under microscopes. This tiring and repetitive work requires an expert eye. Your team wants to create a technology solution that helps pathologists with this task in real-time, using a camera. However, due to the complexity of the work, your team has not found rule-based systems to be capable of adding value to the review of images.

Would AI add value to a potential solution? Why or why not?

3) Flagging suspicious activity

A bank is using AI to flag suspicious international money transfers for potential money laundering, anti-terrorist financing or sanctions concerns. Though the system has proven more effective than the bank's current processes, it still frequently flags legitimate transactions for review.

What are some potential harms that the system could cause, and how can the bank reduce the impacts of these potential harms?

4) Prototyping a chatbot¶

During an ongoing pandemic outbreak, a country's public health agency is facing a large volume of phone calls and e-mails from people looking for health information. The agency has determined that an AI-powered interactive chatbot that answers pandemic-related questions would help people get the specific information they want quickly, while reducing the burden on the agency's employees. How should the agency start prototyping the chatbot?

- Build out the AI solution to the best of its ability before testing it with a diverse group of potential users.
- Build a non-AI prototype quickly and start testing it with a diverse group of potential users.

5) Detecting misinformation

A social media platform is planning to deploy a new AI system to flag and remove social media messages containing misinformation. Though the system has proven effective in tests, it sometimes flags non-objectionable content as misinformation.

What are some ways in which the social media platform could allow someone whose message has been flagged to contest the misinformation designation?

6) Improving autonomous vehicles

What are some of the ways to improve the safety of autonomous vehicles? (You might pick more than one option.)

- Incorporate the safety features of regular vehicles.
- Test the system in a variety of environments.
- Hire an internal ‘red team’ to play the role of bad actors seeking to manipulate the autonomous driving system. Strengthen the system against the team’s attacks on an ongoing basis.

Identifying Bias in AI

Six types of bias

Once we’re aware of the different types of bias, we are more likely to detect them in ML projects. Furthermore, with a common vocabulary, we can have fruitful conversations about how to mitigate (or reduce) the bias.

We will closely follow a research paper from early 2020 that characterizes six different types of bias.

Historical bias

Historical bias occurs when the state of the world in which the data was generated is flawed.

As of 2020, only 7.4% of Fortune 500 CEOs are women. Research has shown that companies with female CEOs or CFOs are generally more profitable than companies with men in the same position, suggesting that women are held to higher hiring standards than men. In order to fix this, we might consider removing human input and using AI to make the hiring process more equitable. But this can prove unproductive if data from past hiring decisions is used to train a model, because the model will likely learn to demonstrate the same biases that are present in the data.

Representation bias

Representation bias occurs when building datasets for training a model, if those datasets poorly represent the people that the model will serve.

Data collected through smartphone apps will under-represent groups that are less likely to own smartphones. For instance, if collecting data in the USA, individuals over the age of 65 will be under-represented. If the data is used to inform design of a city transportation system, this will be disastrous, since older people have important needs to ensure that the system is accessible.

Measurement bias

Measurement bias occurs when the accuracy of the data varies across groups. This can happen when working with proxy variables (variables that take the place of a variable that cannot be directly measured), if the quality of the proxy varies in different groups.

Your local hospital uses a model to identify high-risk patients before they develop serious conditions, based on information like past diagnoses, medications, and demographic data. The model uses this information to predict health care costs, the idea being that patients **with higher**

costs likely correspond to high-risk patients. Despite the fact that the model specifically excludes race, it seems to demonstrate racial discrimination: the algorithm is less likely to select eligible Black patients. How can this be the case? It is because cost was used as a proxy for risk, and the relationship between these variables varies with race: Black patients experience increased barriers to care, have less trust in the health care system, and therefore have lower medical costs, on average, when compared to non-Black patients with the same health conditions.

Aggregation bias

Aggregation bias occurs when groups are inappropriately combined, resulting in a model that does not perform well for any group or only performs well for the majority group. (This is often not an issue, but most commonly arises in medical applications.)

Hispanics have higher rates of diabetes and diabetes-related complications than non-Hispanic whites. If building AI to diagnose or monitor diabetes, it is important to make the system sensitive to these ethnic differences, by either including ethnicity as a feature in the data, or building separate models for different ethnic groups.

Evaluation bias

Evaluation bias occurs when evaluating a model, if the benchmark data (used to compare the model to other models that perform similar tasks) does not represent the population that the model will serve.

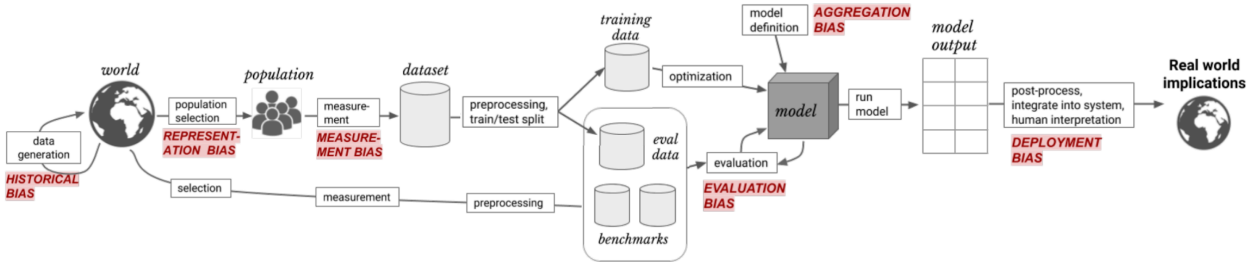
The Gender Shades paper discovered that two widely used facial analysis benchmark datasets (IJB-A and Adience) were primarily composed of lighter-skinned subjects (79.6% and 86.2%, respectively). Commercial gender classification AI showed state-of-the-art performance on these benchmarks, but experienced disproportionately high error rates with people of color.

Deployment bias

Deployment bias occurs when the problem the model is intended to solve is different from the way it is actually used. If the end users don't use the model in the way it is intended, there is no guarantee that the model will perform well.

The criminal justice system uses tools to predict the likelihood that a convicted criminal will relapse into criminal behavior. The predictions are not designed for judges when deciding appropriate punishments at the time of sentencing.

We can visually represent these different types of bias, which occur at different stages in the ML workflow:



Note that these are *not mutually exclusive*: that is, an ML application can easily suffer from more than one type of bias. For example, as Rachel Thomas describes in a recent research talk, ML applications in wearable fitness devices can suffer from:

- **Representation bias (if the dataset used to train the models exclude darker skin tones),**
- **Measurement bias (if the measurement apparatus shows reduced performance with dark skin tones), and**
- **Evaluation bias (if the dataset used to benchmark the model excludes darker skin tones).**

AI Fairness

Introduction

There are many different ways of defining what we might look for in a fair machine learning (ML) model. For instance, say we're working with a model that approves (or denies) credit card applications. Is it:

- fair if the approval rate is equal across genders, or is it
- better if gender is removed from the dataset and hidden from the model?

Four fairness criteria

These four fairness criteria are a useful starting point, but it's important to note that there are more ways of formalizing fairness, which you are encouraged to explore.

Assume we're working with a model that selects individuals to receive some outcome. For instance, the model could select people who should be approved for a loan, accepted to a university, or offered a job opportunity. (*So, we don't consider models that perform tasks like facial recognition or text generation, among other things.*)

1. Demographic parity / statistical parity

Demographic parity says the model is fair if the composition of people who are selected by the model matches the group membership percentages of the applicants.

A nonprofit is organizing an international conference, and 20,000 people have signed up to attend. The organizers write a ML model to select 100 attendees who could potentially give interesting talks at the conference. Since 50% of the attendees will be women (10,000 out of 20,000), they design the model so that 50% of the selected speaker candidates are women.

2. Equal opportunity

Equal opportunity fairness ensures that the proportion of people who should be selected by the model ("positives") that are correctly selected by the model is the same for each group. We refer to this proportion as the true positive rate (TPR) or sensitivity of the model.

A doctor uses a tool to identify patients in need of extra care, who could be at risk for developing serious medical conditions. (This tool is used only to supplement the doctor's practice, as a second opinion.) It is designed to have a high TPR that is equal for each demographic group.

3. Equal accuracy

Alternatively, we could check that the model has equal accuracy for each group. That is, the percentage of correct classifications (people who should be denied and are denied, and people who should be approved who are approved) should be the same for each group. If the model is 98% accurate for individuals in one group, it should be 98% accurate for other groups.

A bank uses a model to approve people for a loan. The model is designed to be equally accurate for each demographic group: this way, the bank avoids approving people who should be rejected (which would be financially damaging for both the applicant and the bank) and avoid rejecting people who should be approved (which would be a failed opportunity for the applicant and reduce the bank's revenue).

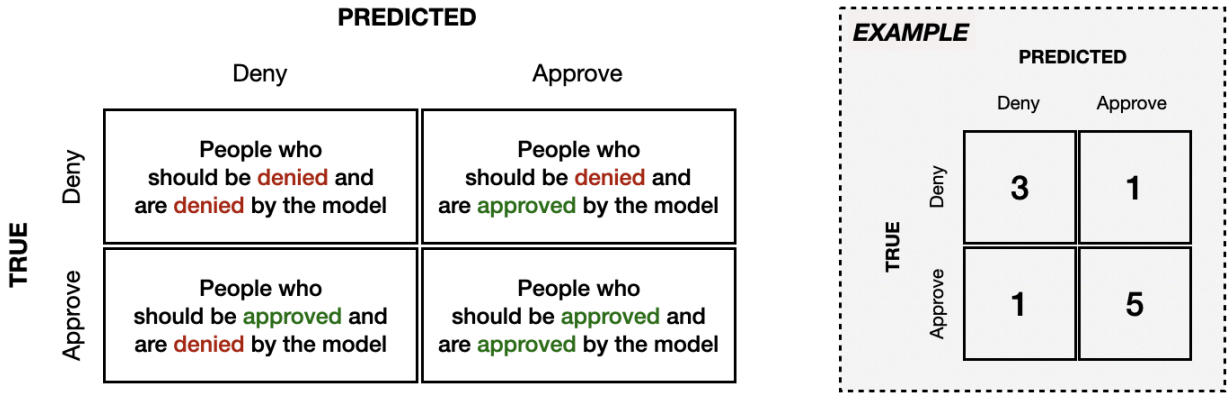
4. Group unaware / "Fairness through unawareness"

Group unaware fairness removes all group membership information from the dataset. For instance, we can remove gender data to try to make the model fair to different gender groups. Similarly, we can remove information about race or age.

One difficulty of applying this approach in practice is that one has to be careful to identify and remove proxies for the group membership data. For instance, in cities that are racially segregated, zip code is a strong proxy for race. That is, when the race data is removed, the zip code data should also be removed, or else the ML application may still be able to infer an individual's race from the data. Additionally, group unaware fairness is unlikely to be a good solution for historical bias.

Example

We'll work with a small example to illustrate the differences between the four different types of fairness. We'll use a confusion matrix, which is a common tool used to understand the performance of a ML model. This tool is depicted in the example below, which depicts a model with 80% accuracy (since 8/10 people were correctly classified) and has an 83% true positive rate (since 5/6 "positives" were correctly classified).

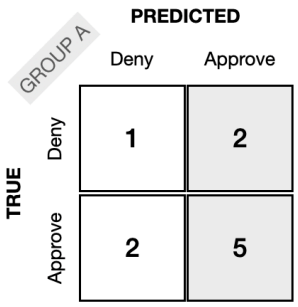


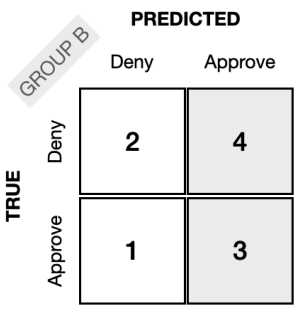
To understand how a model's performance varies across groups, we can construct a different confusion matrix for each group. In this small example, we'll assume that we have data from only 20 people, equally split between two groups (10 from Group A, and 10 from Group B).

The next image shows what the confusion matrices could look like, if the model satisfies demographic parity fairness. 10 people from each group (50% from Group A, and 50% from Group B) were considered by the model. 14 people, also equally split across groups (50% from Group A, and 50% from Group B) were approved by the model.

Demographic parity

20 applicants (50% from **Group A**)
14 approvals (50% from **Group A**)

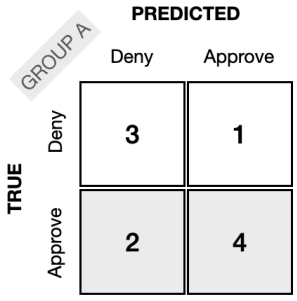


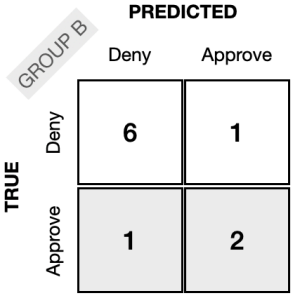


For equal opportunity fairness, the TPR for each group should be the same; in the example below, it is 66% in each case.

Equal opportunity

Group A: 66% true positive rate: 4/(4+2)
Group B: 66% true positive rate: 2/(1+2)

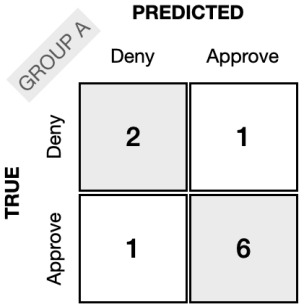


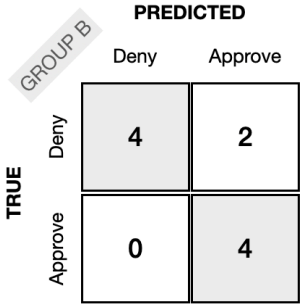


Next, we can see how the confusion matrices might look for equal accuracy fairness. For each group, the model was 80% accurate.

Equal accuracy

Group A: 80% accurate: (6+2)/10
Group B: 80% accurate: (4+4)/10





Note that group unaware fairness cannot be detected from the confusion matrix, and is more concerned with removing group membership information from the dataset.

Take the time now to study these toy examples, and use it to build your intuition for the differences between the different types of fairness. How does the example change if Group A has double the number of applicants of Group B?

Also note that none of the examples satisfy more than one type of fairness. For instance, the demographic parity example does not satisfy equal accuracy or equal opportunity. Take the time to verify this now. In practice, it is not possible to optimize a model for more than one type of fairness: to read more about this, explore the Impossibility Theorem of Machine Fairness. *So which fairness criterion should you select, if you can only satisfy one?* As with most ethical questions, the correct answer is usually not straightforward, and picking a criterion should be a long conversation involving everyone on your team.

When working with a real project, the data will be much, much larger. In this case, confusion matrices are still a useful tool for analyzing model performance. One important thing to note, however, is that real-world models typically cannot be expected to satisfy any fairness definition perfectly. For instance, if "demographic parity" is chosen as the fairness metric, where the goal is for a model to select 50% men, it may be the case that the final model ultimately selects some percentage close to, but not exactly 50% (like 48% or 53%).

Reference

- Explore different types of fairness with an interactive tool.
- You can read more about equal opportunity in this blog post.
- Analyze ML fairness with this walkthrough of the What-If Tool, created by the People and AI Research (PAIR) team at Google. This tool allows you to quickly amend an ML model, once you've picked the fairness criterion that is best for your use case.

Model Cards

Introduction

A **model card** is a short document that provides key information about a machine learning model. Model cards increase transparency by communicating information about trained models to broad audiences.

Model cards

Though AI systems are playing increasingly important roles in every industry, few people understand how these systems work. AI researchers are exploring many ways to communicate key information about models to inform people who use AI systems, people who are affected by AI systems and others.

Model cards - introduced in a 2019 paper - are one way for teams to communicate key information about their AI system to a broad audience. This information generally includes intended uses for the model, how the model works, and how the model performs in different situations.

You can think of model cards as similar to the nutritional labels that you find on packaged foods.

Examples of model cards

Before we continue, it might be useful to briefly skim some examples of model cards.

- Salesforce's model cards
- Open AI's model card for GPT-3
- Google Cloud's example model cards

Who is the audience of your model card?

A model card should strike a balance between being easy-to-understand and communicating important technical information. When writing a model card, you should consider your audience: the groups of people who are most likely to read your model card. These groups will vary according to the AI system's purpose.

For example, a model card for an AI system that helps medical professionals interpret x-rays to better diagnose musculoskeletal injuries is likely to be read by medical professionals, scientists,

patients, researchers, policymakers and developers of similar AI systems. The model card may therefore assume some knowledge of health care and of AI systems.

What sections should a model card contain?

Per the original paper, a model card should have the following nine sections. Note that different organizations may add, subtract or rearrange model card sections according to their needs (and you may have noticed this in some of the examples above).

As you read about the different sections, you're encouraged to review the two example model cards from the original paper. Before proceeding, open each of these model card examples in a new window:

- Model Card - Smiling Detection in Images
- Model Card - Toxicity in Text

1. Model Details

- Include background information, such as developer and model version.

2. Intended Use

- What use cases are in scope?
- Who are your intended users?
- What use cases are out of scope?

3. Factors

- What factors affect the impact of the model? For example, the smiling detection model's results vary by demographic factors like age, gender or ethnicity, environmental factors like lighting or rain and instrumentation like camera type.

4. Metrics

- What metrics are you using to measure the performance of the model? Why did you pick those metrics?
    - For **classification systems** – in which the output is a class label – potential error types include false positive rate, false negative rate, false discovery rate, and false omission rate. The relative importance of each of these depends on the use case.
    - For **score-based analyses** – in which the output is a score or price – consider reporting model performance across groups.

5. Evaluation Data

- Which datasets did you use to evaluate model performance? Provide the datasets if you can.
- Why did you choose these datasets for evaluation?
- Are the datasets representative of typical use cases, anticipated test cases and/or challenging cases?

6. Training Data

- Which data was the model trained on?

7. Quantitative Analyses

- How did the model perform on the metrics you chose? Break down performance by important factors and their intersections. For example, in the smiling detection example, performance is broken down by age (eg, young, old), gender (eg, female, male), and then both (eg, old-female, old-male, young-female, young-male).

8. Ethical Considerations

- Describe ethical considerations related to the model, such as sensitive data used to train the model, whether the model has implications for human life, health, or safety, how risk was mitigated, and what harms may be present in model usage.

9. Caveats and Recommendations

- Add anything important that you have not covered elsewhere in the model card.

How can you use model cards in your organization?

The use of detailed model cards can often be challenging because an organization may not want to reveal its processes, proprietary data or trade secrets. In such cases, the developer team should think about how model cards can be useful and empowering, without including sensitive information.

Some teams use other formats - such as FactSheets - to collect and log ML model information.

Answers

1. Solution: These are all good ways to start!
2. Solution: Yes, it would. People would generally agree that the goal is desirable, especially since the AI system will be working with pathologists rather than in their place. AI can help people with repetitive tasks and AI systems have proven effective in similar medical image recognition use cases. That said, it is important to follow current industry best practices and to be thorough in the rest of the design process, including in analyzing harms and in considering how medical practitioners will actually interact with the product in a medical setting.
3. Solution: One potential harm is that the AI system could be biased against certain groups, flagging, delaying or denying their legitimate transactions at higher rates than those of other groups. The bank can reduce these harms by selecting data carefully, identifying and mitigating potential bias (see Lessons 3 and 4), not operationalizing the system until potential bias is addressed and ensuring appropriate and continuous human oversight of the system once it is operational.
4. Solution: The correct answer is: Build a non-AI prototype quickly and start testing it with a diverse group of potential users. Iterating on a non-AI prototype is easier, faster and less expensive than iterating on an AI prototype. Iterating on a non-AI prototype also provides early information on user expectations, interactions and needs. This information should inform the eventual design of AI prototypes.
5. Solution: The social media company should ask customers how they would want to challenge a determination. It could be by easily accessing a challenge form on which a user can describe why their message does not contain misinformation, requesting further review by a human reviewer, requesting an explanation of why the content was flagged or a combination of these and other means.

6. Solution: All of these are great ways to improve safety.