Welcome to “Introduction to Responsible AI”.

01 AI & Responsibility

02 Google's AI principles

03 Responsible AI practices

04 Case study: Google Flights

This Module consists of 4 lessons.

You will learn to ...

01	Describe the importance of Responsible AI
02	List Google's AI principles
03	Describe the best practices for responsible AI development
04	Identify how responsible AI principles can be applied during product development

Today, you will learn how to:

- Describe the importance of Responsible AI.
- List Google's AI principles.
- Describe the best practices for responsible AI development.
- Identify how responsible AI principles can be applied during product development.

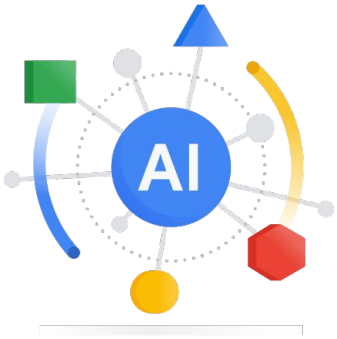
01 AI & Responsibility

02 Google's AI principles

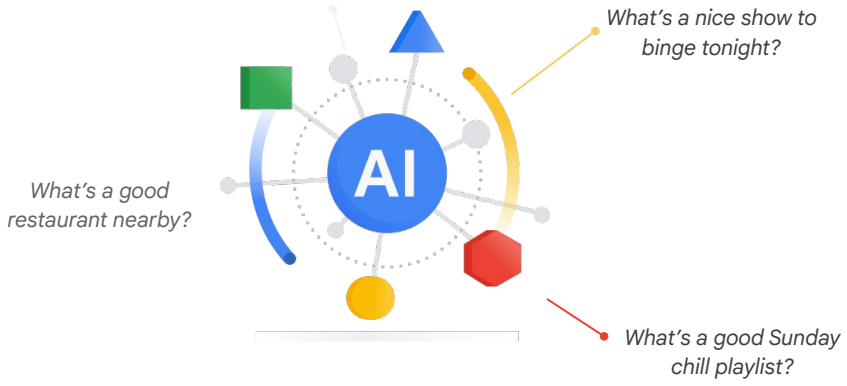
03 Responsible AI practices

04 Case study: Google Flights

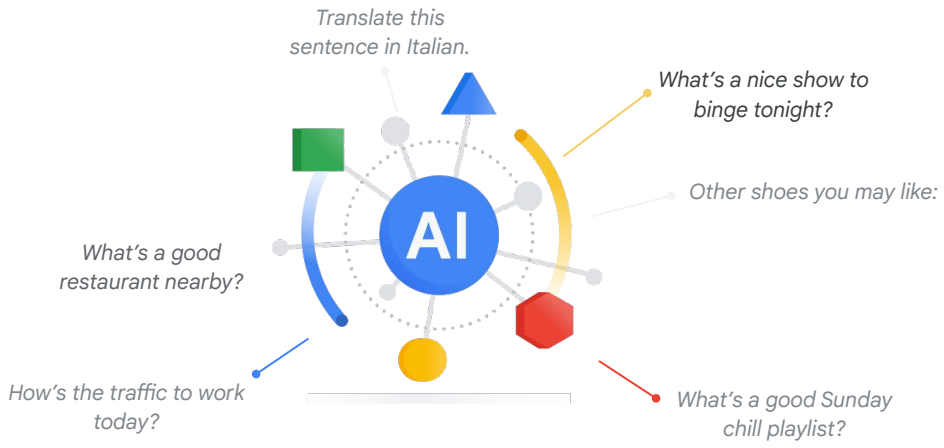
Let's start with AI and responsibility.



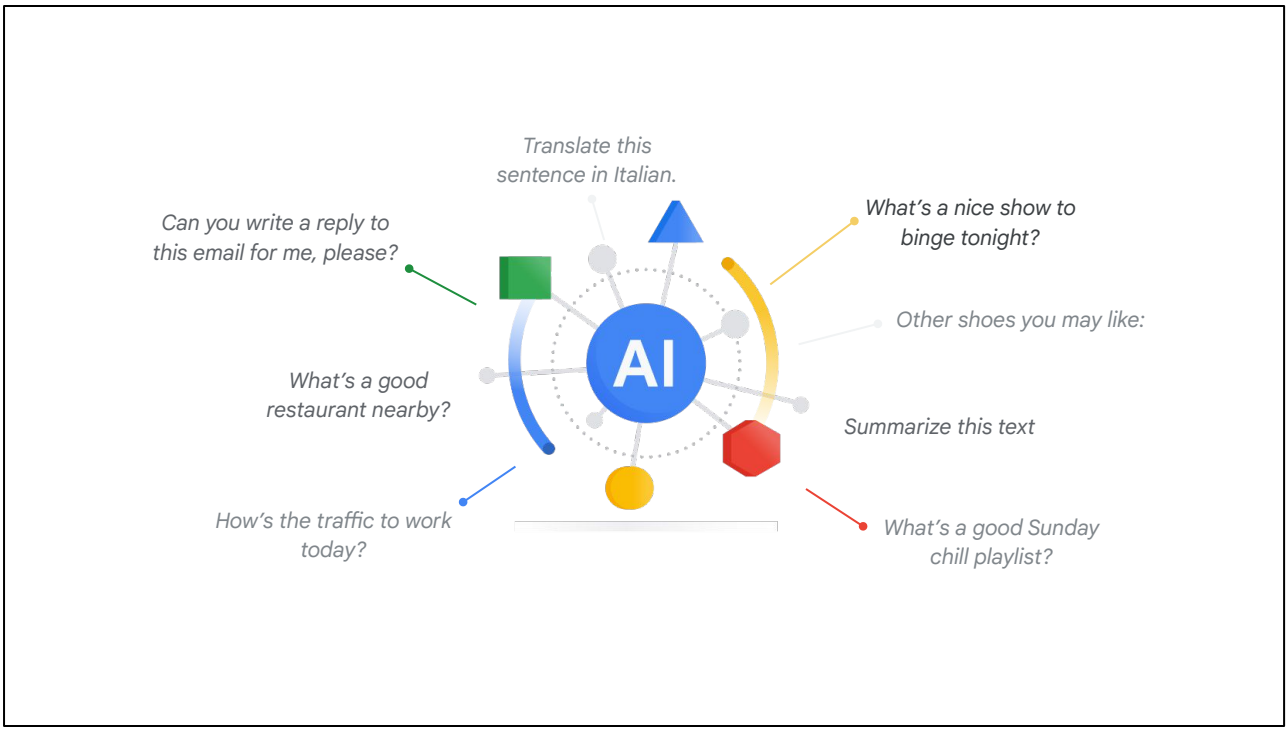
Artificial intelligence, or AI, is everywhere (you may have even heard the phrase “generative AI”).



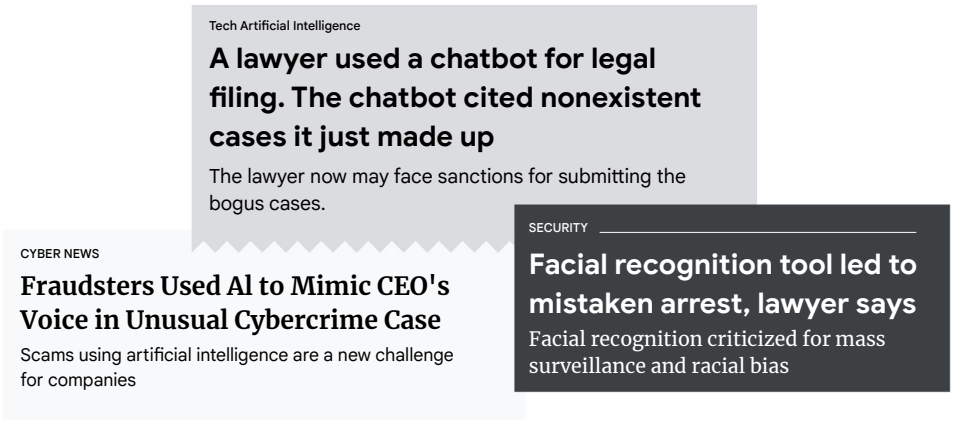
Most of us already have daily interactions with artificial intelligence,



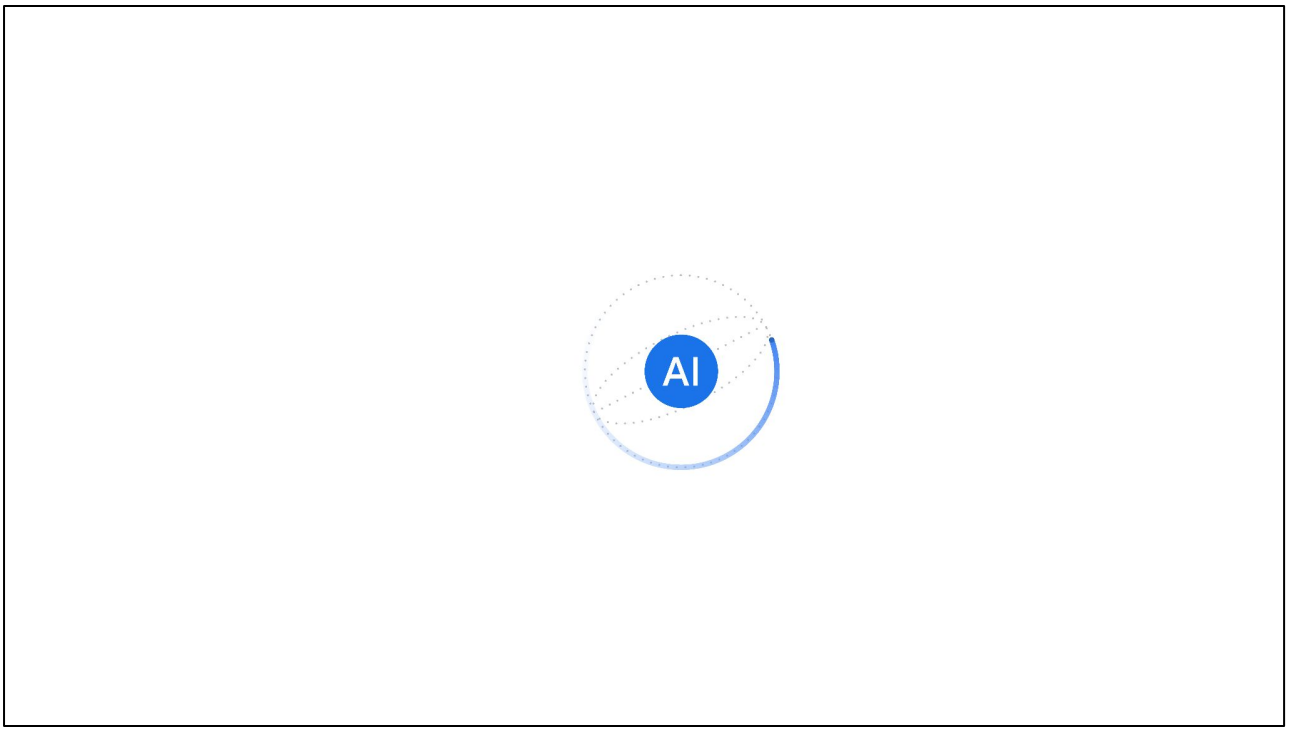
from predictions for traffic and weather, to recommendations for TV shows.



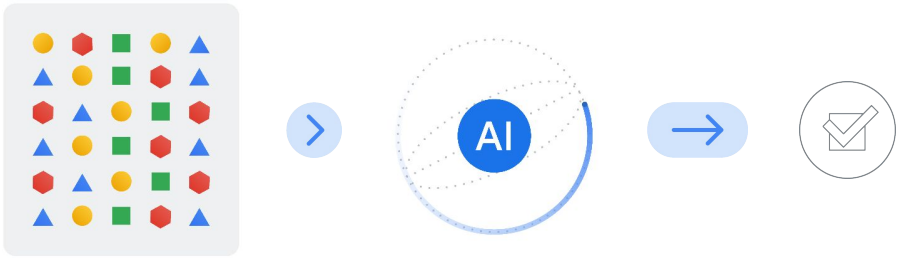
AI is becoming more common and new AI systems are continuously being developed at an extraordinary pace.



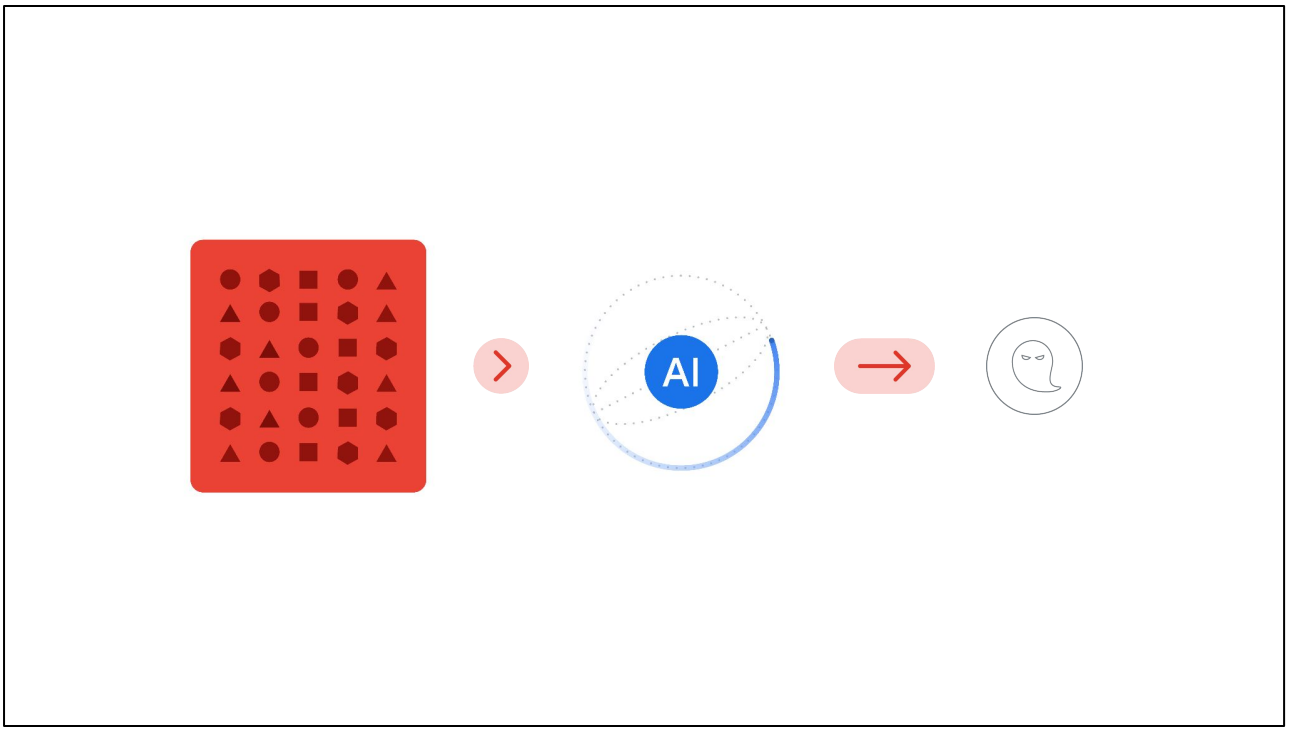
Still, AI is not infallible.
No AI model has 100% accuracy and AI systems are complex. When AI systems are used in real-world contexts, they can fail to behave in expected ways, which reduces their realized benefit.
Failures in behavior can come from model misuse (like the case of fraudsters who use AI to mimic a CEO's voice).



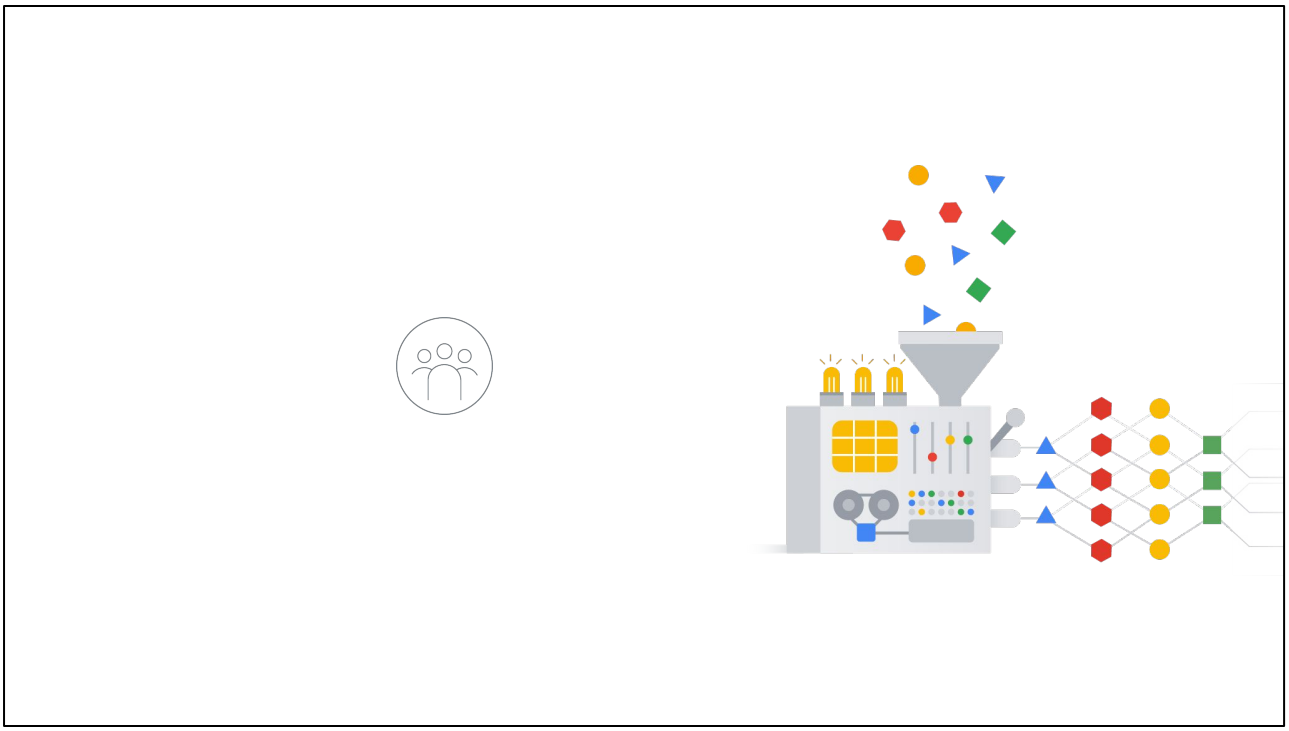
This is because AI models are often “underspecified”:



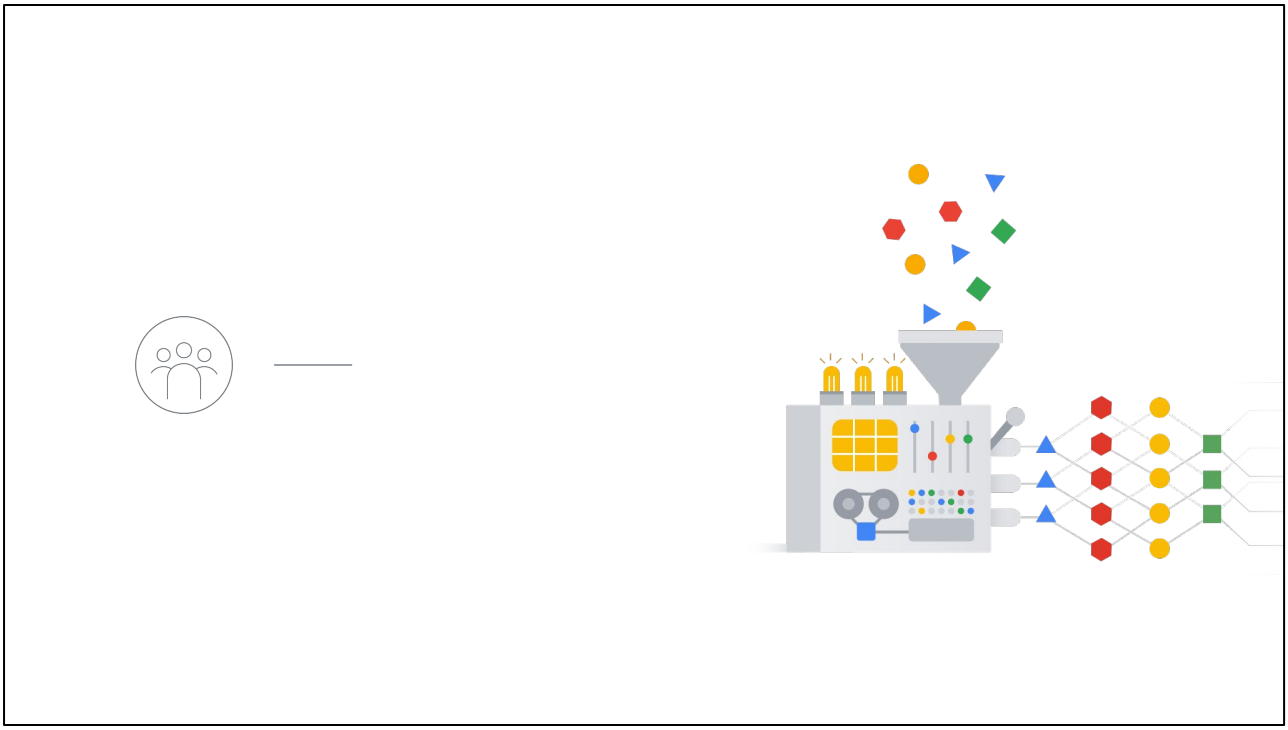
they perform well in the situation in which they are trained,



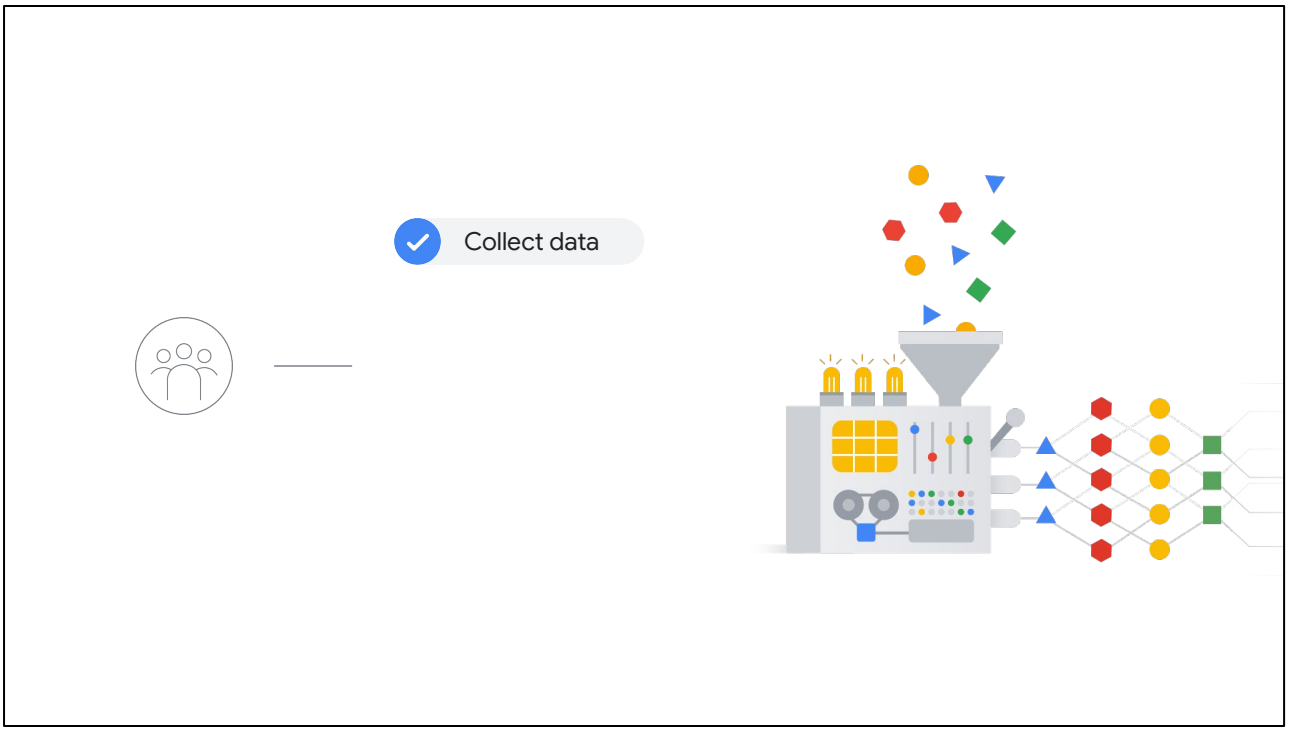
but might not be robust or fair in new situations.



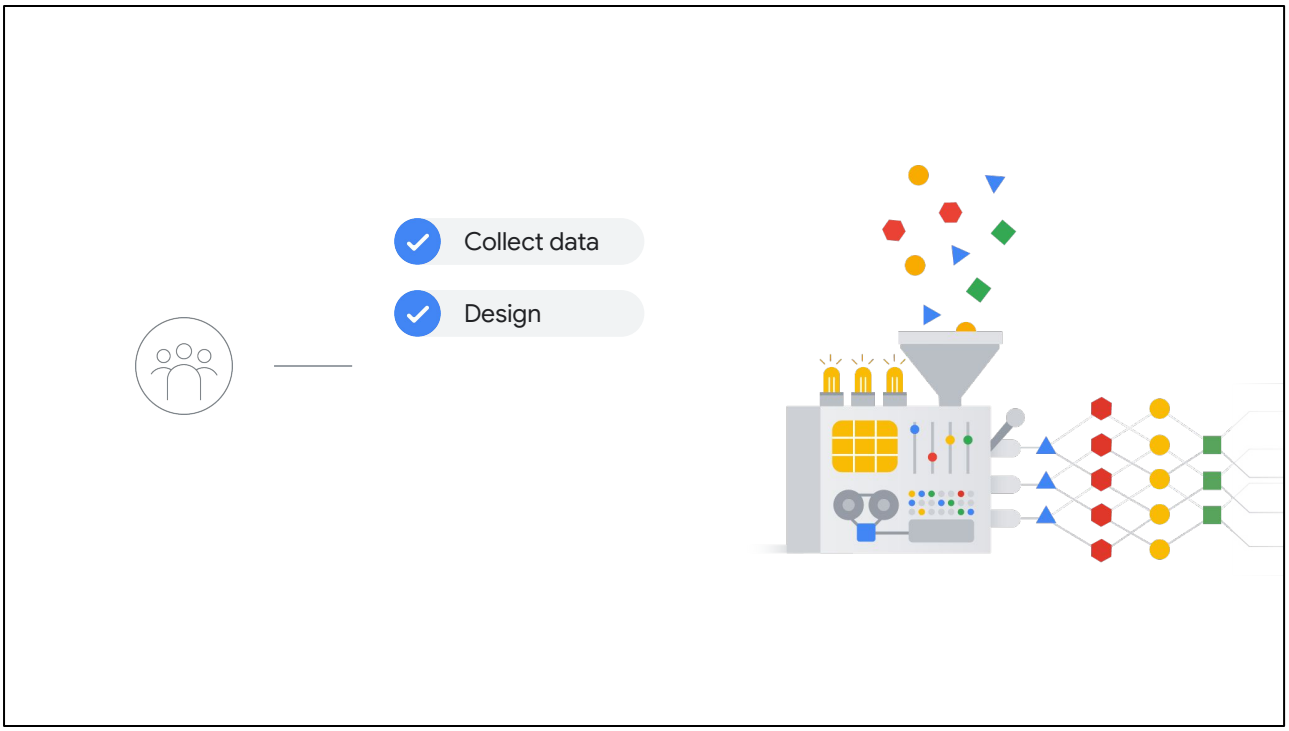
There is a common misconception with artificial intelligence that machines play the central decision-making role. In reality, it’s people who make decisions.



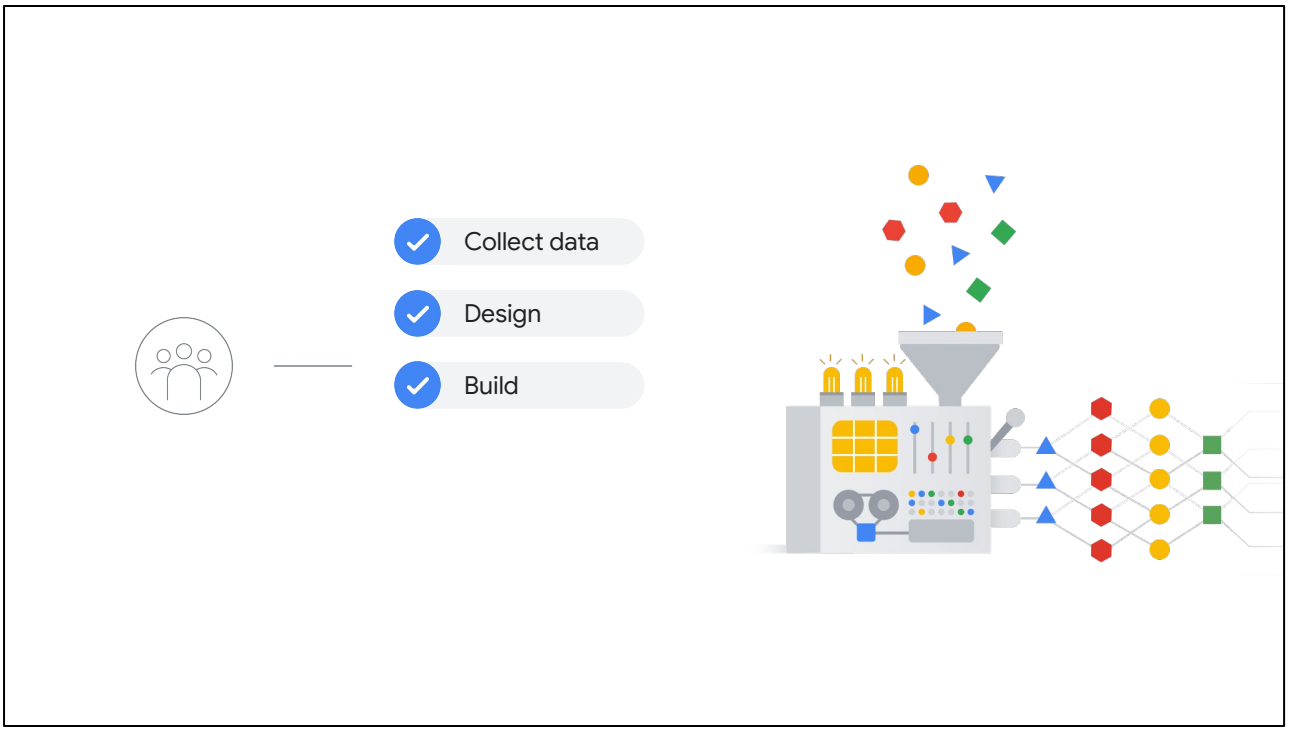
People are involved in each aspect of AI development.



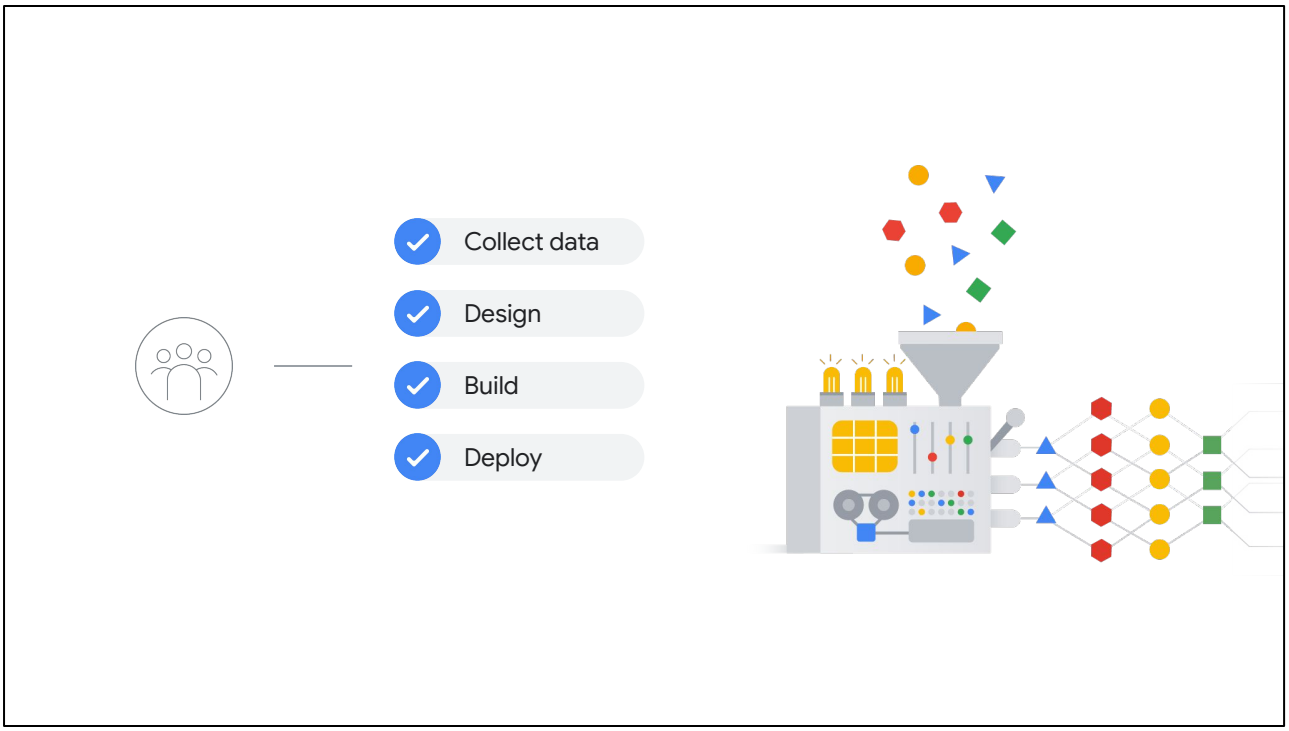
They collect or create the data that the model is trained on.



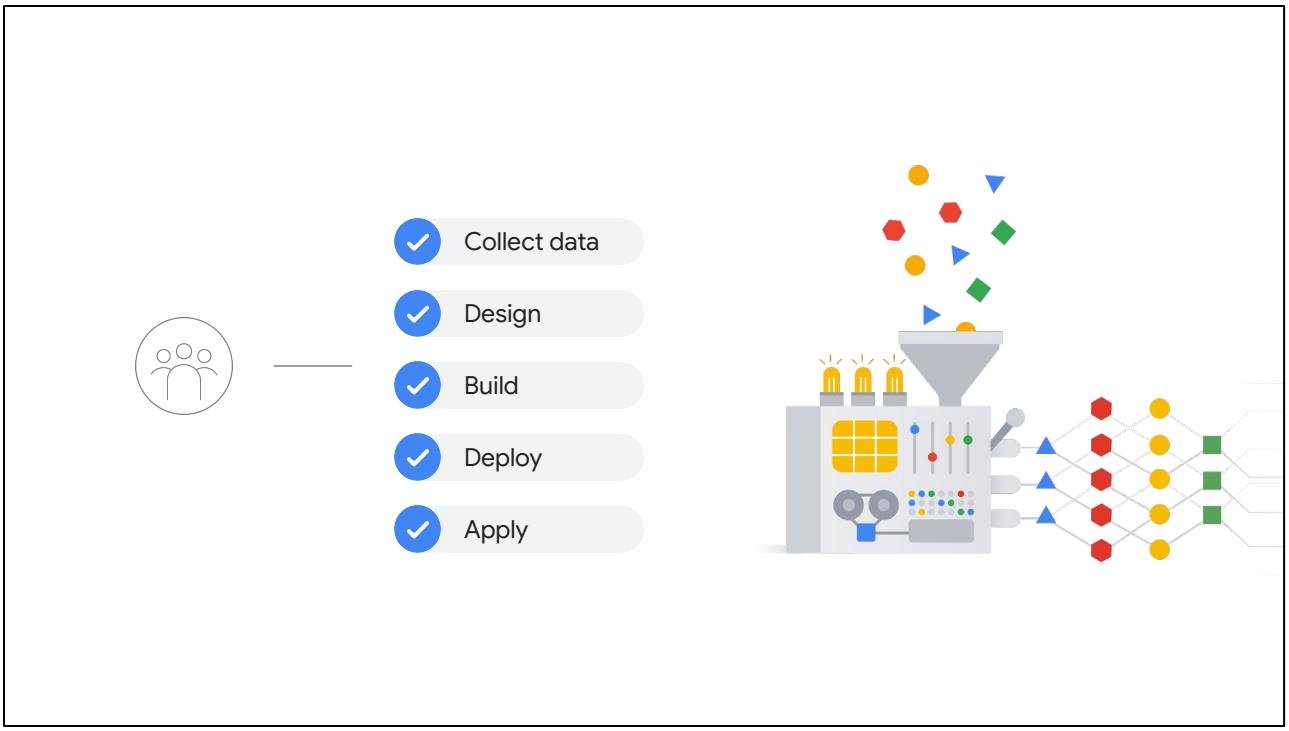
They design



and build these models.

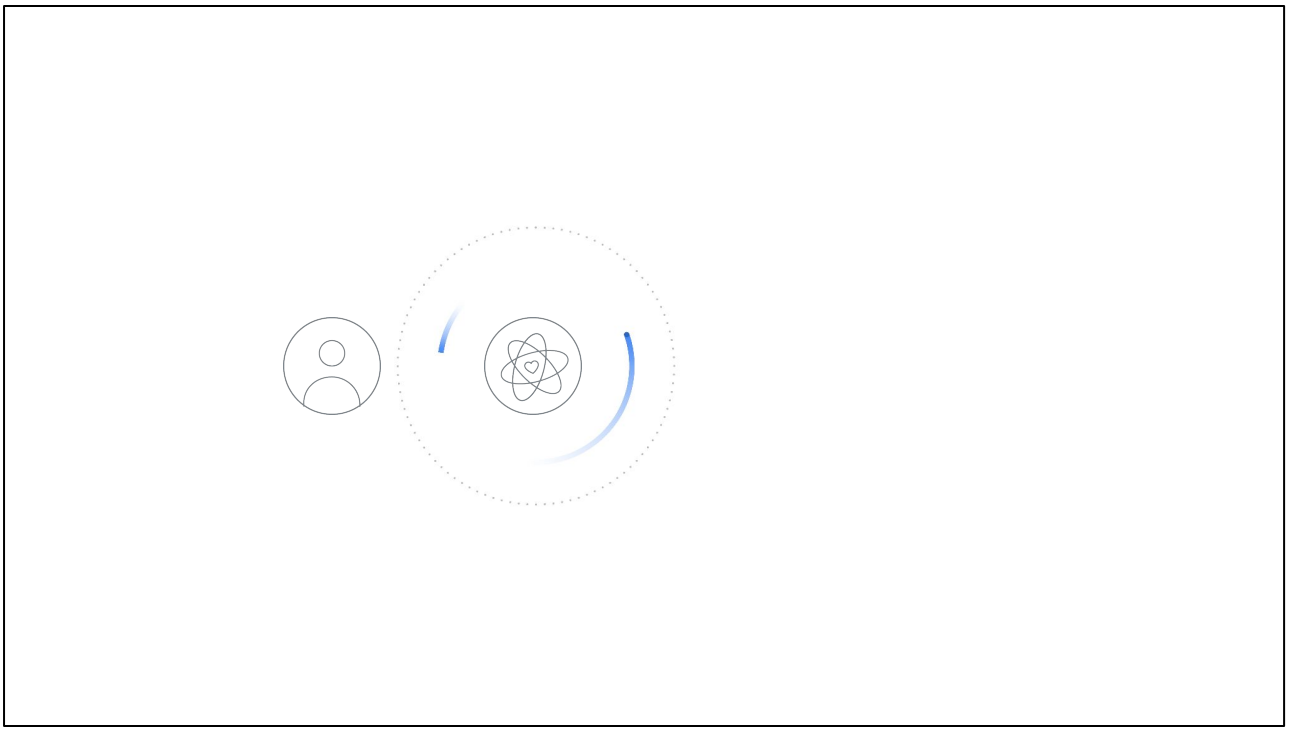


They control the deployment of the AI

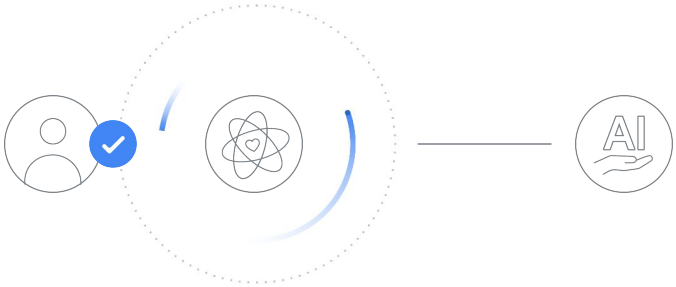


and how it is applied in a given context.

Essentially, human decisions are threaded throughout our technology products.



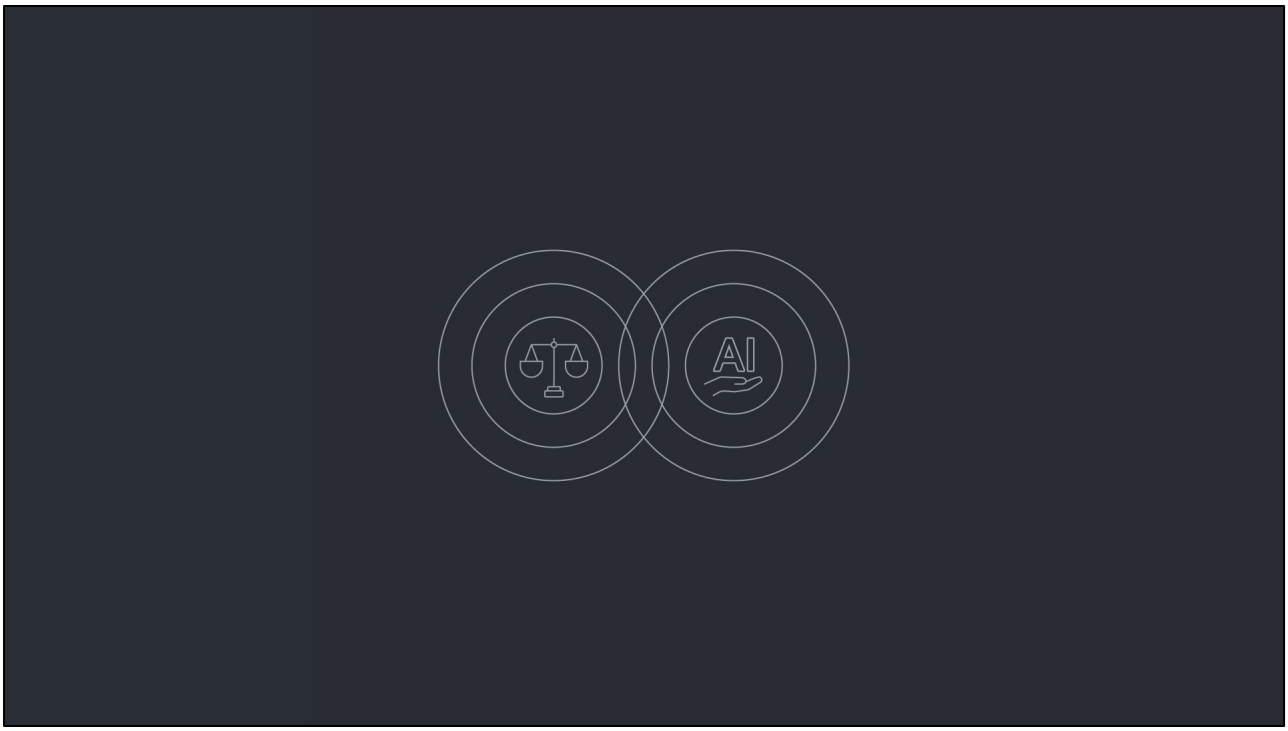
And every time a person makes a decision,



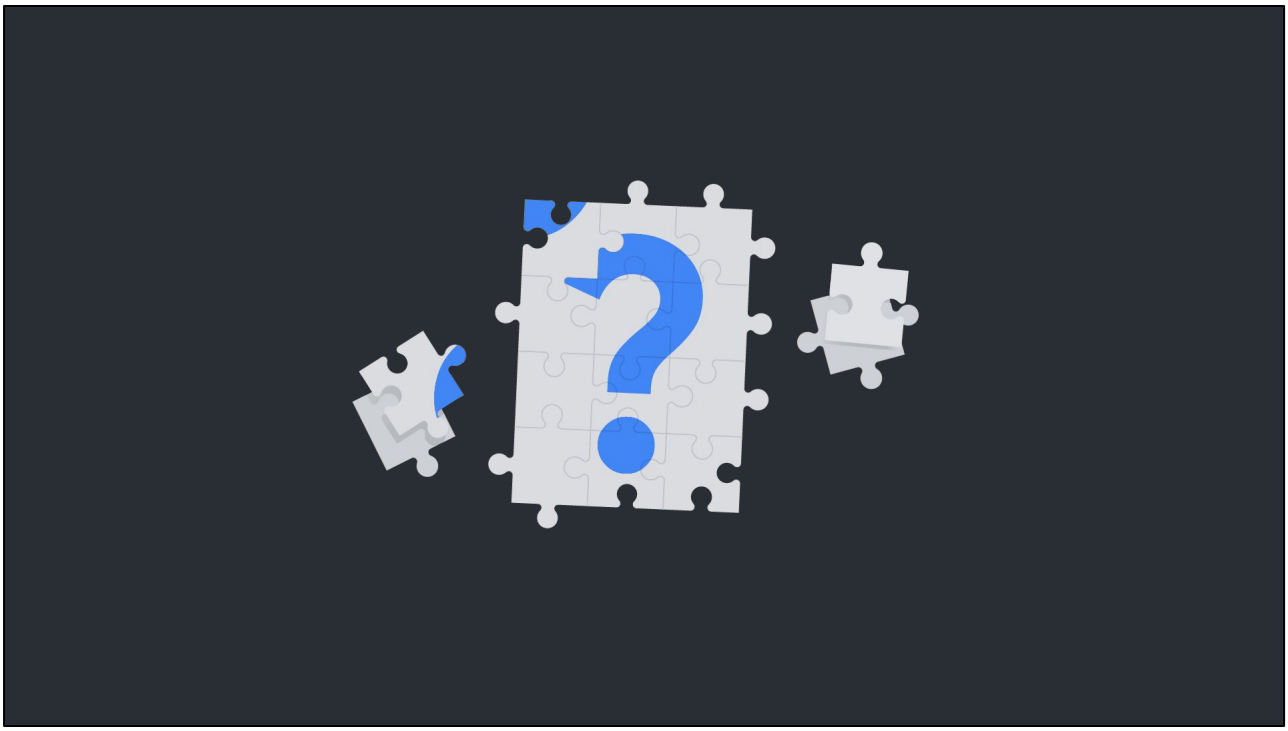
they are actually making a choice based on their own values. Whether it's the decision to use generative AI to solve a problem, as opposed to other methods, or anywhere throughout the machine learning lifecycle, that person introduces their own set of values.

Every decision point requires consideration and evaluation to ensure that choices have been made responsibly.

This means that every decision point requires consideration and evaluation to ensure that choices have been made responsibly from concept through deployment and maintenance.



Because there’s potential to affect many areas of society, it's important to develop AI technologies with ethics in mind.



When we’re talking about new AI technologies and ethics, there's a lot of work in progress on laws, policies, regulations, etc. Unfortunately, there is not a wide global consensus on what these ethics are yet. What we have learned is we can’t wait for society to catch up and codify the rules. We have a voice in shaping what those norms will be.

It should be noted that ethics is different from laws and policies. Law draws insights from ethics, and ethics inform policy but most ethical norms are not codified.

So we need to define: what is ethical behavior?
To put it simply, ethics can be explained in these three ways:

01 What we **ought** to do.

- Ethics is what we ought to do. (Not the same as what is actually done, or what most people say or think should be done.)

01 What we **ought** to do.

02 What others can rightly blame us for **not doing**.

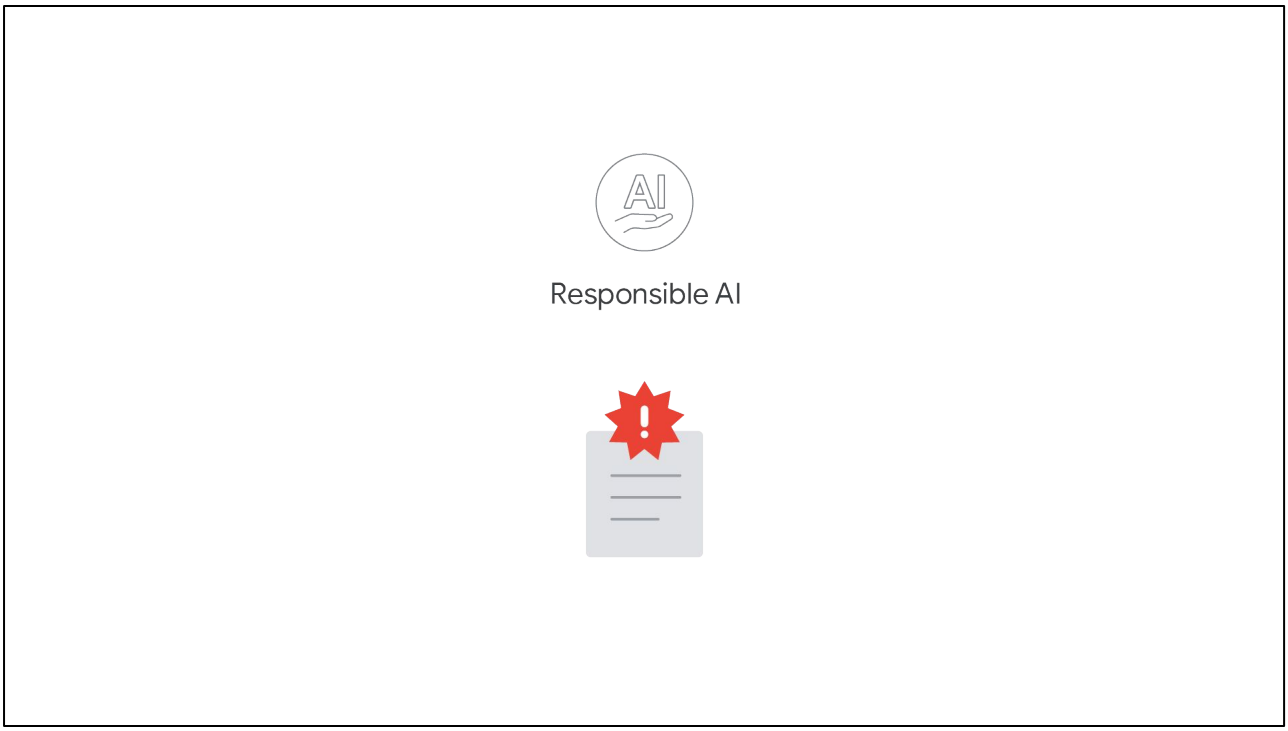
- Ethics is what others can rightly blame us for not doing (even if we suffer no actual punishment).

01 What we **ought** to do.

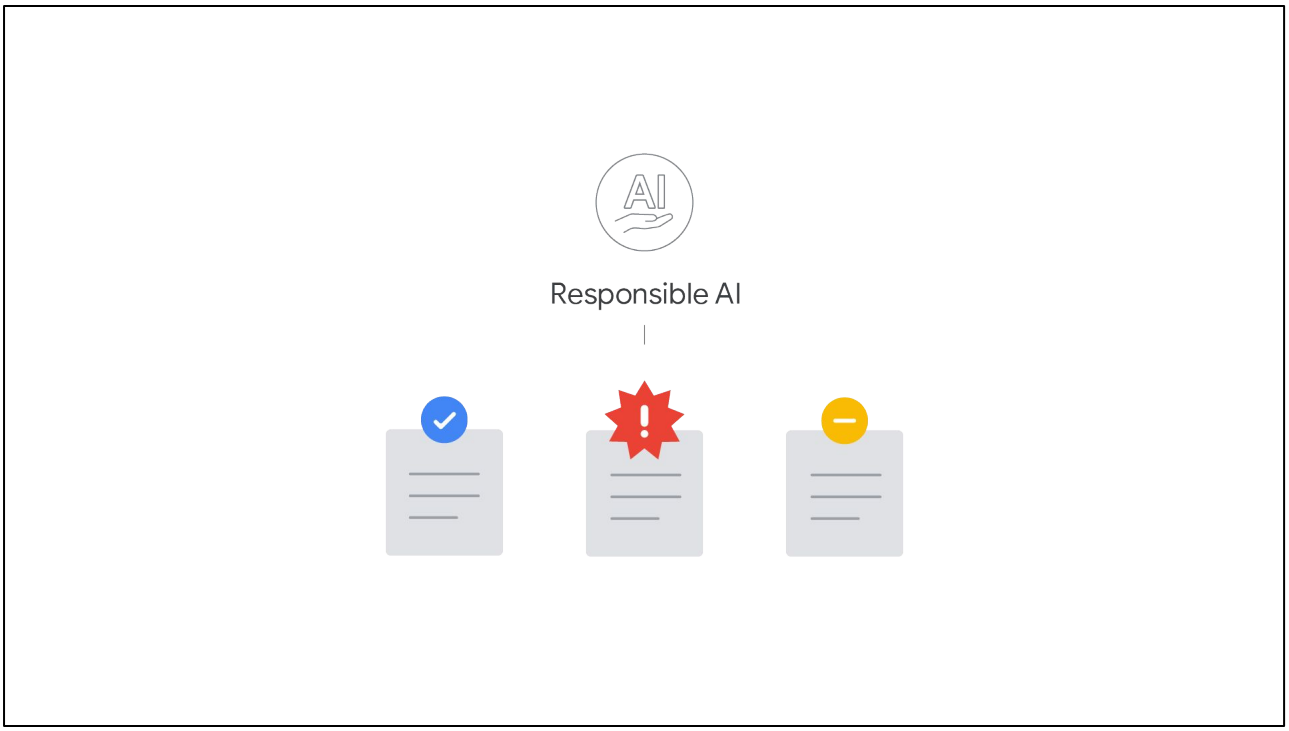
02 What others can rightly blame us for **not doing**.

03 What sustains our flourishing **together** in human society.

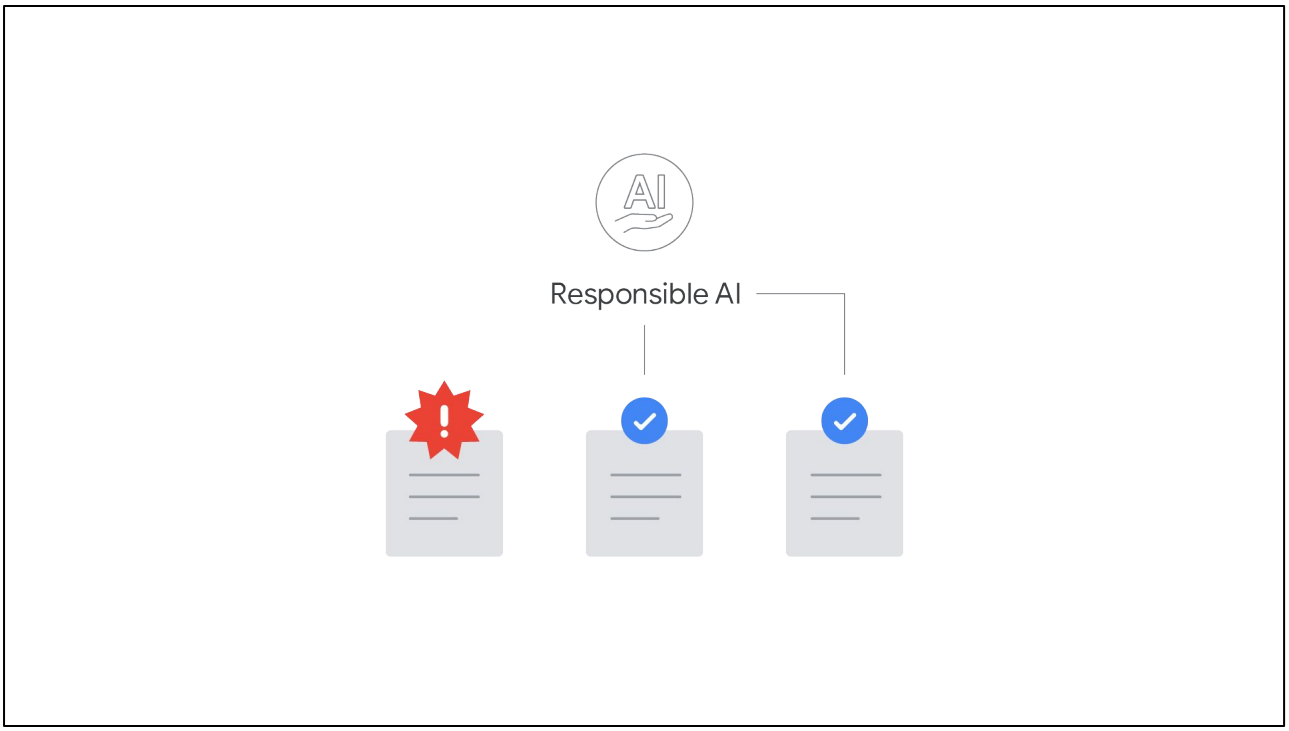
- Ethics is what sustains our flourishing together in human society (ethics is an evolved tool for living well as social creatures).



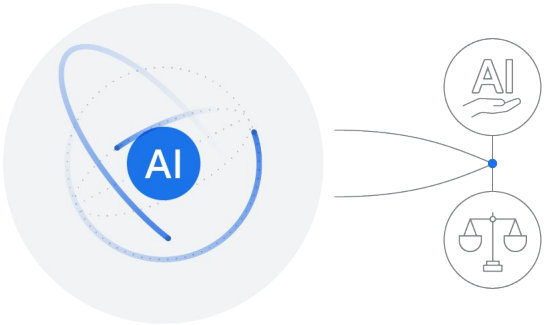
Responsible AI doesn’t mean to focus only on the obviously controversial use cases though.



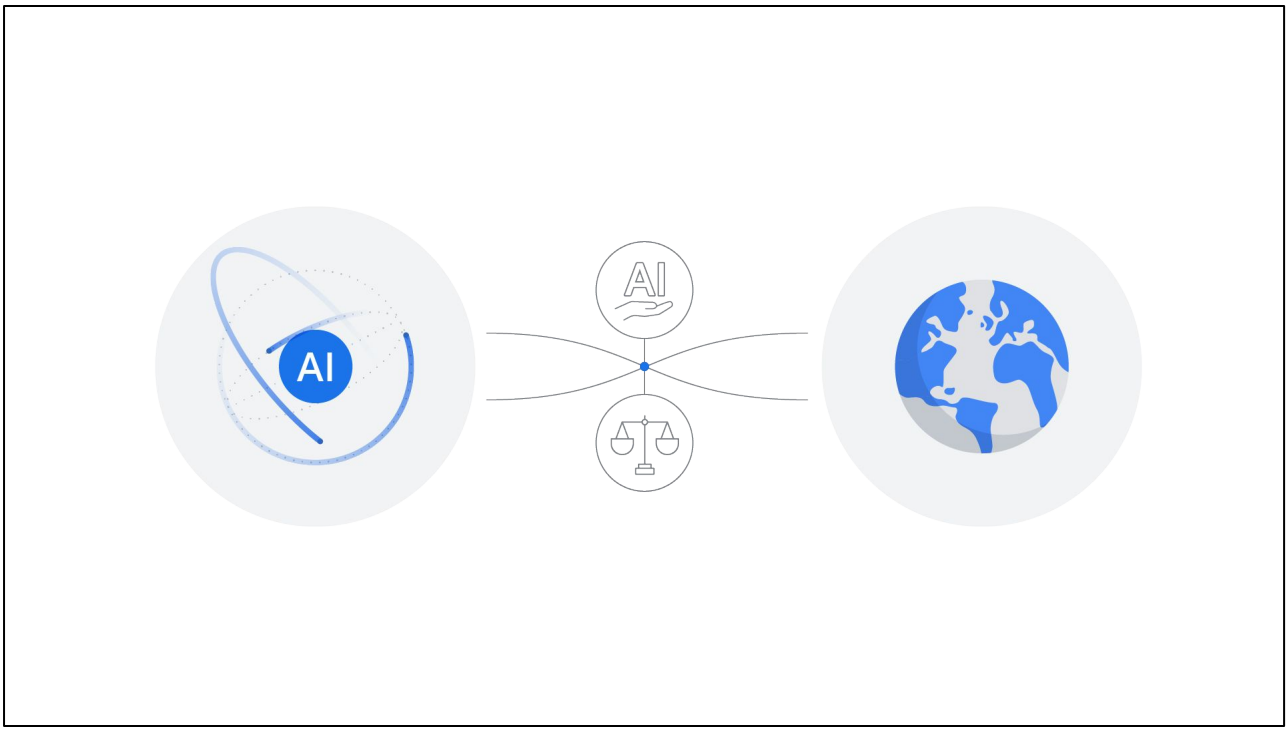
Without responsible AI practices, even seemingly innocuous or good intent AI use cases



could still cause ethical issues or unintended outcomes. They might not even be as beneficial as they could be.



Ethics and responsibility are important, not least because they represent the right thing to do,



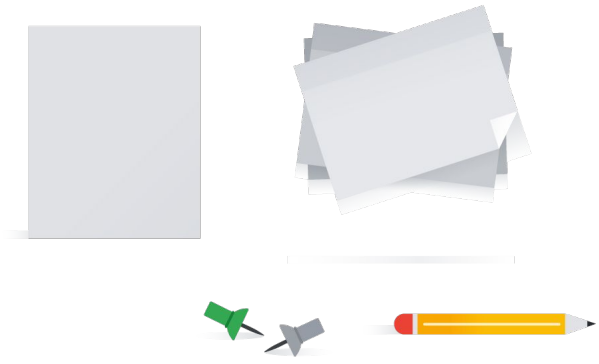
but also because they can guide AI design to be more beneficial for people's lives.

Responsible AI requires an understanding of the possible issues, limitations, or unintended consequences.

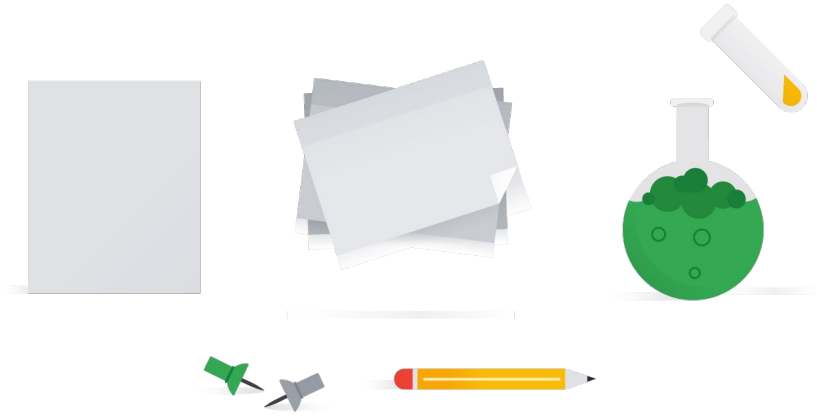
So what is responsible AI?
What we can say is that responsible AI requires an understanding of the possible issues, limitations, or unintended consequences. This understanding is aimed at developing AI ethically.

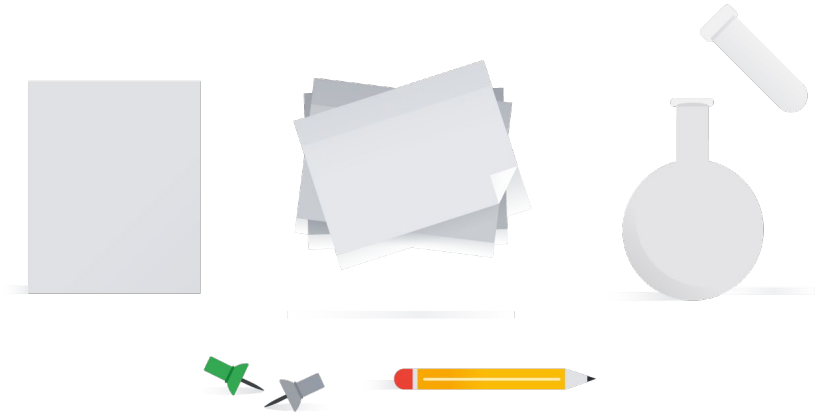
There is not a universal definition of “responsible AI,”

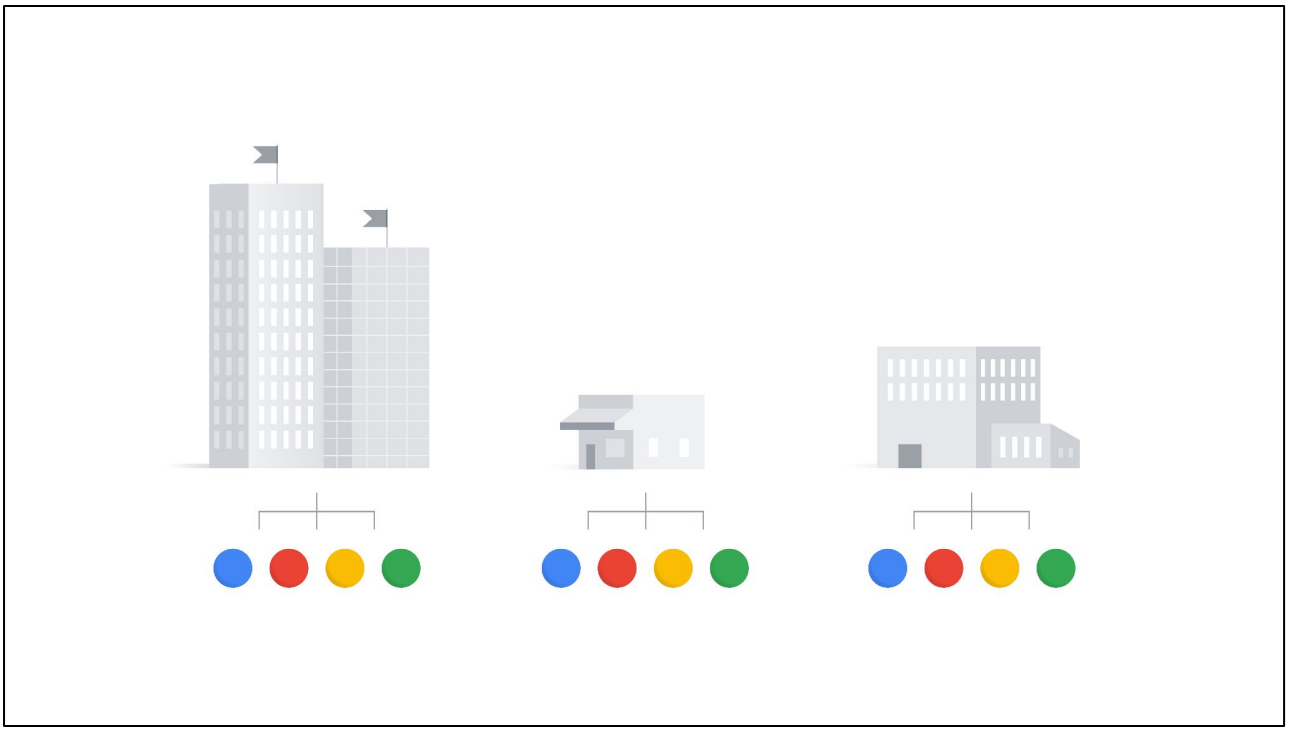
nor is there a simple checklist



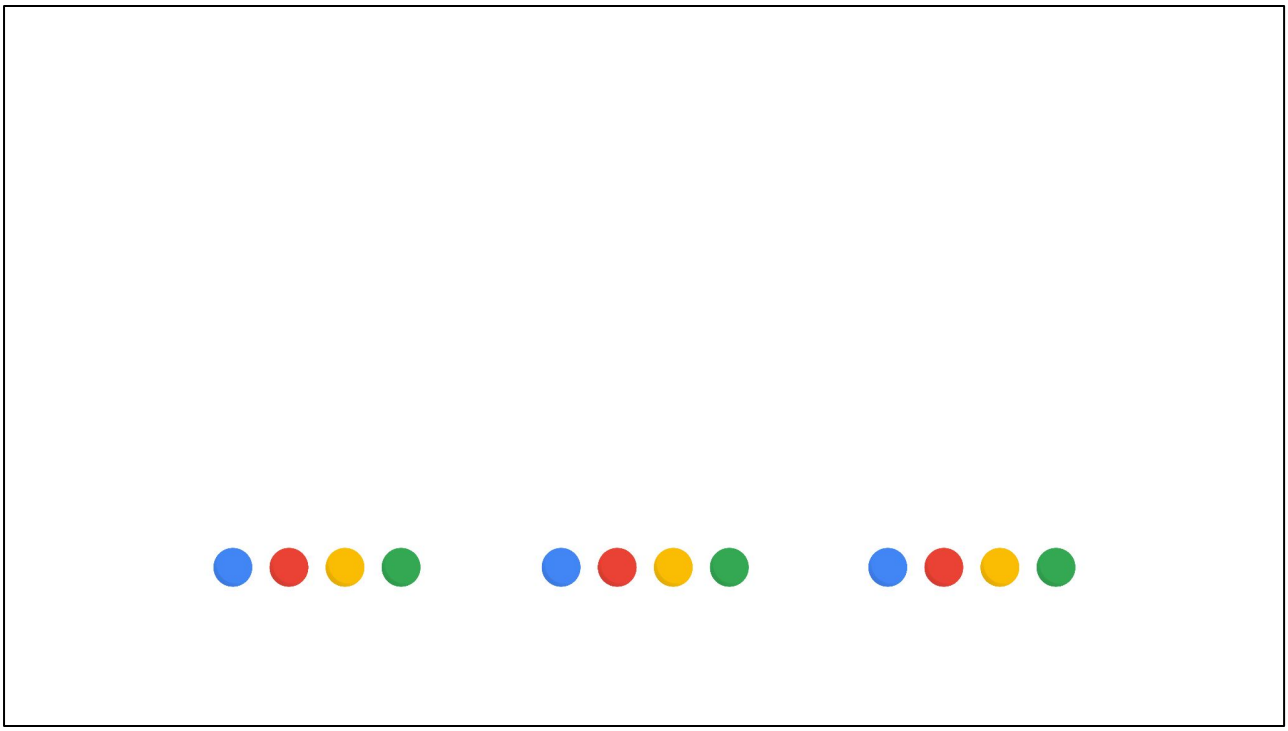
or formula that defines how responsible AI practices should be implemented.

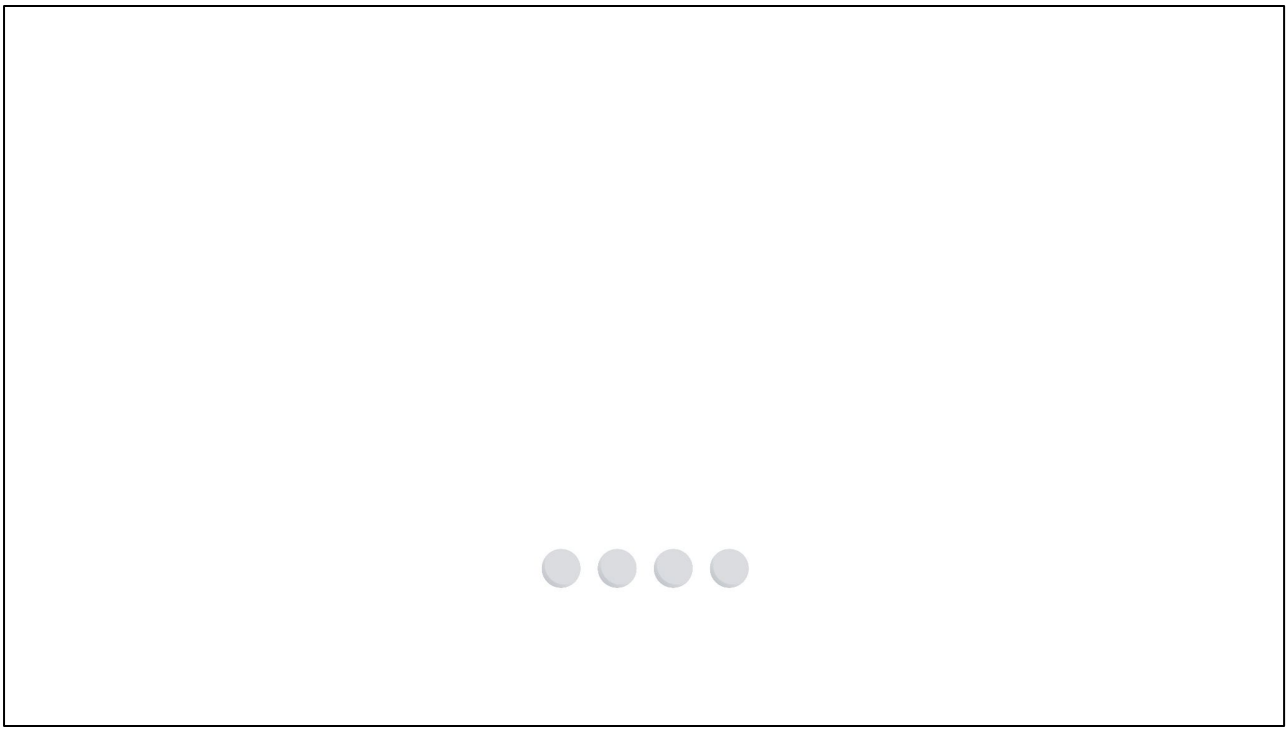


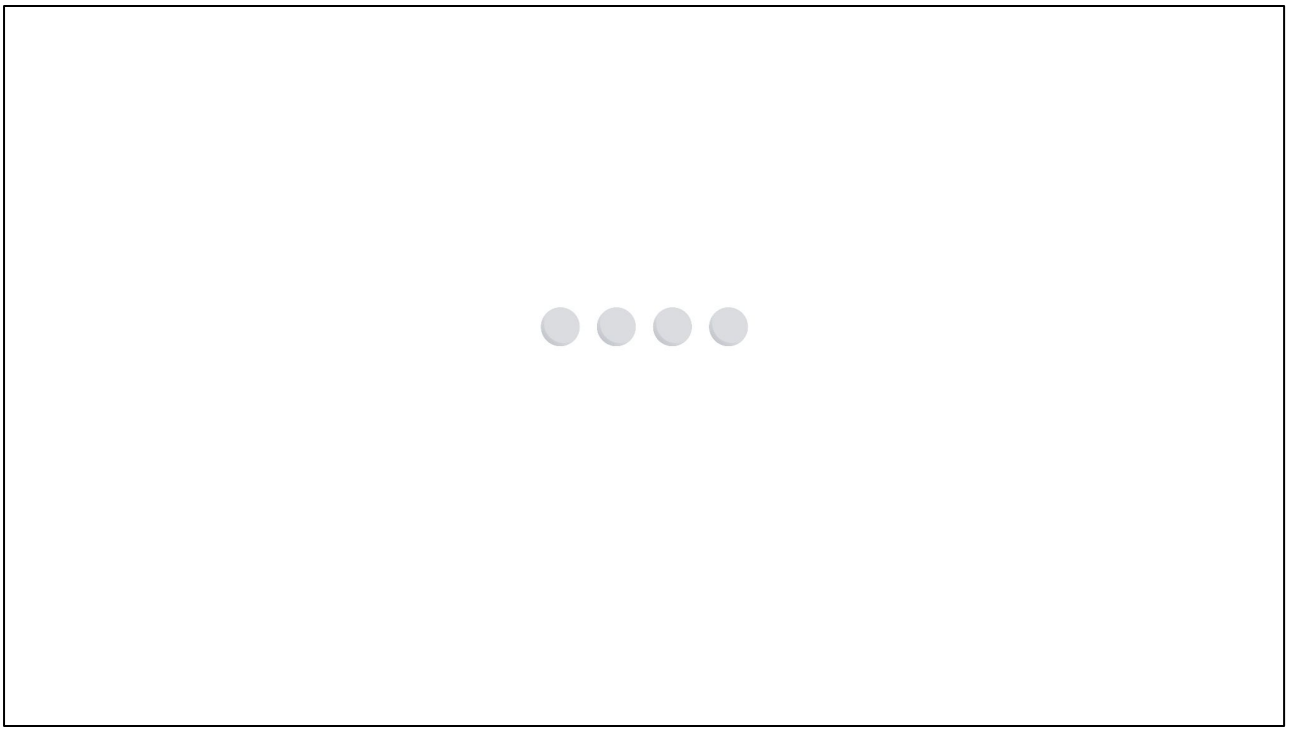




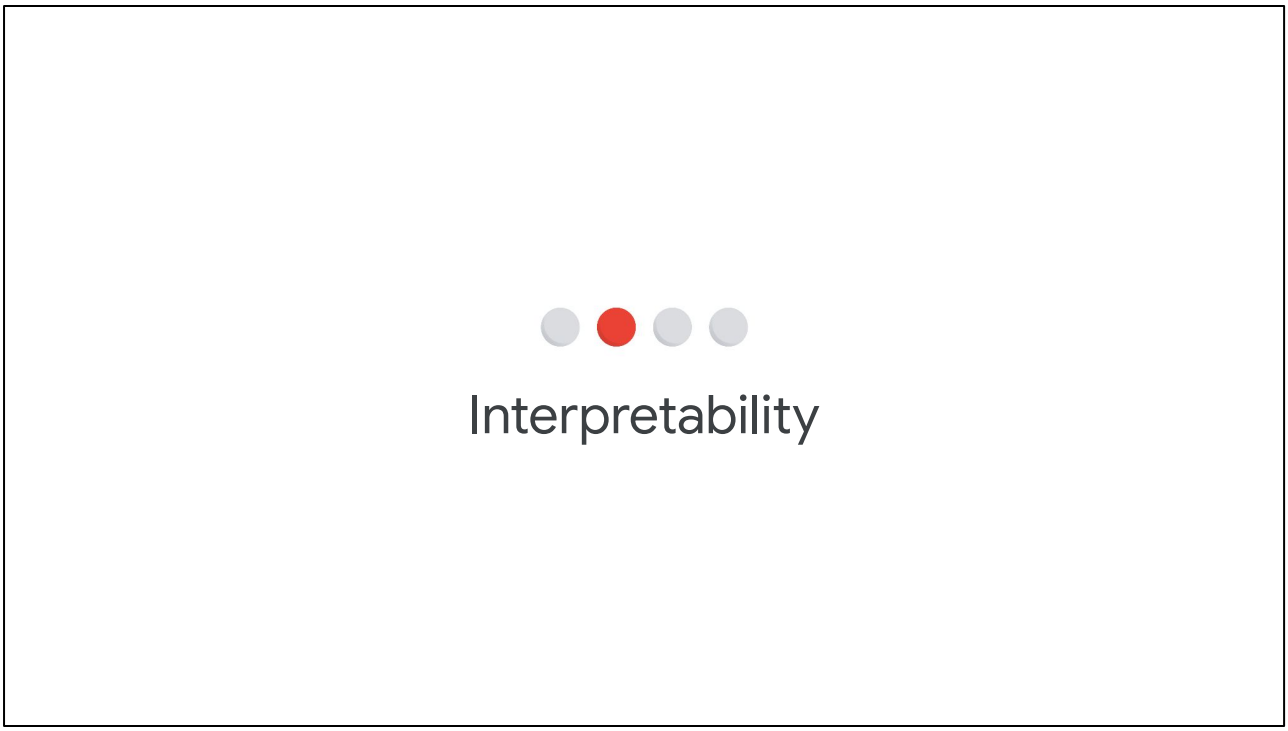
Instead, individuals and organizations are developing their own AI foundational principles that reflect their mission and values.
While these principles are unique to every organization, if you look for common themes, there is a consistent set of ideas across



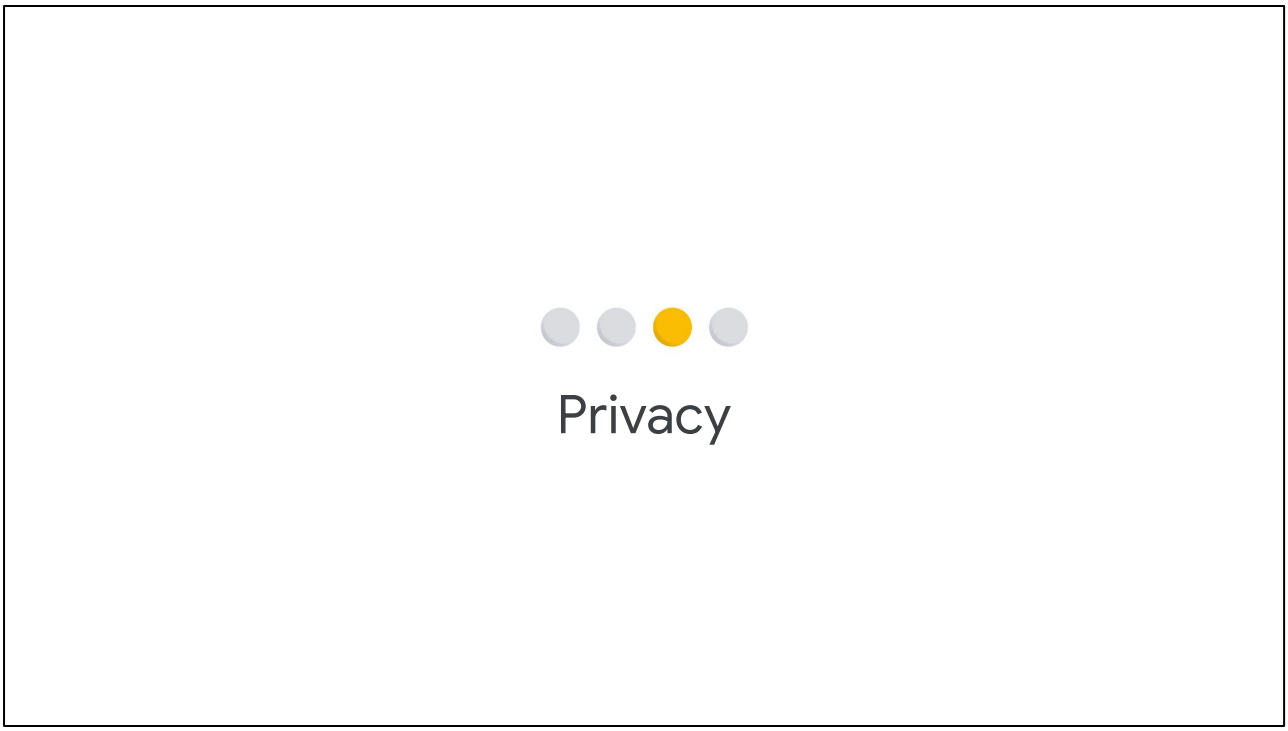




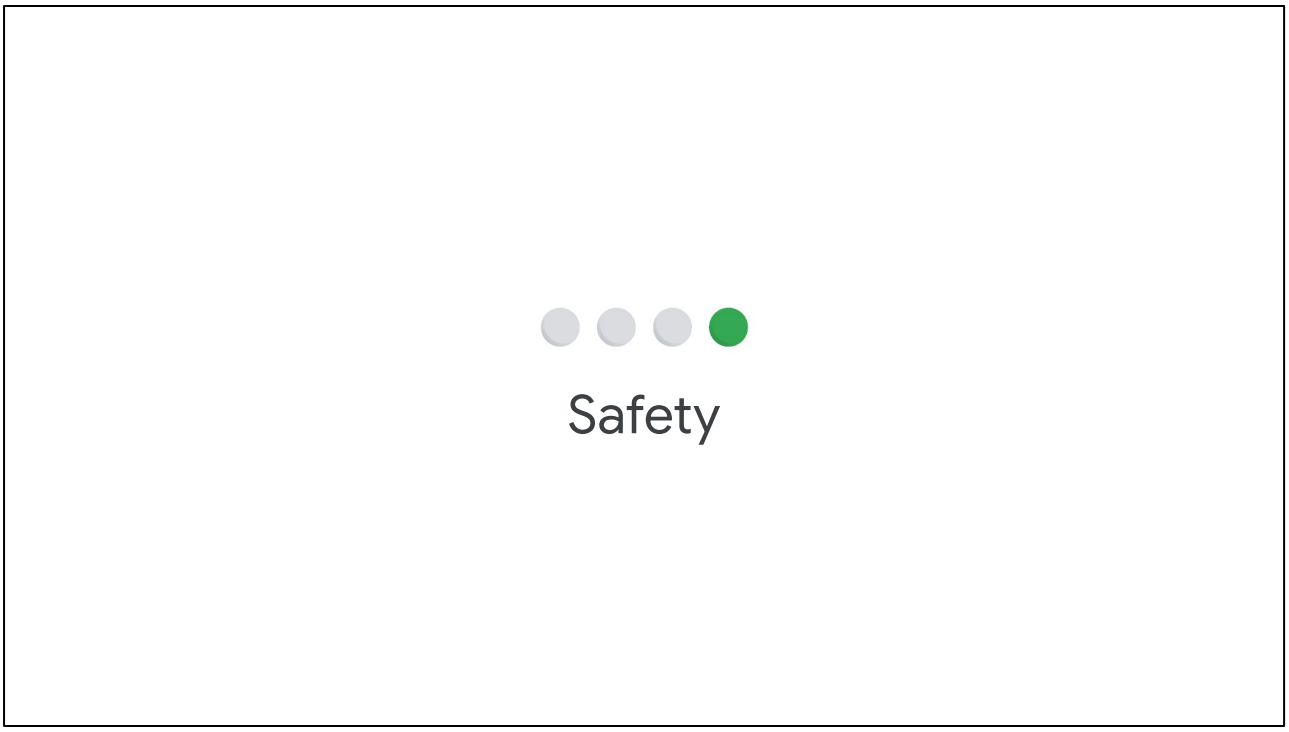
fairness



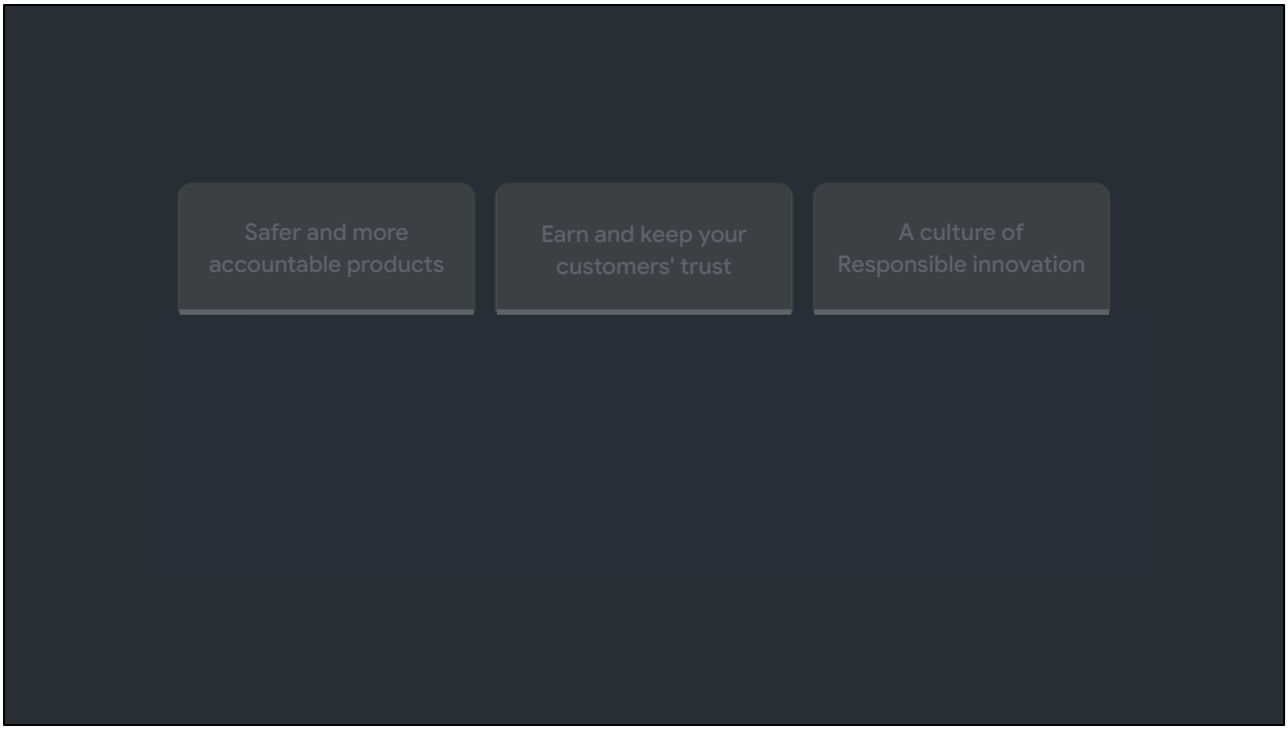
Interpretability



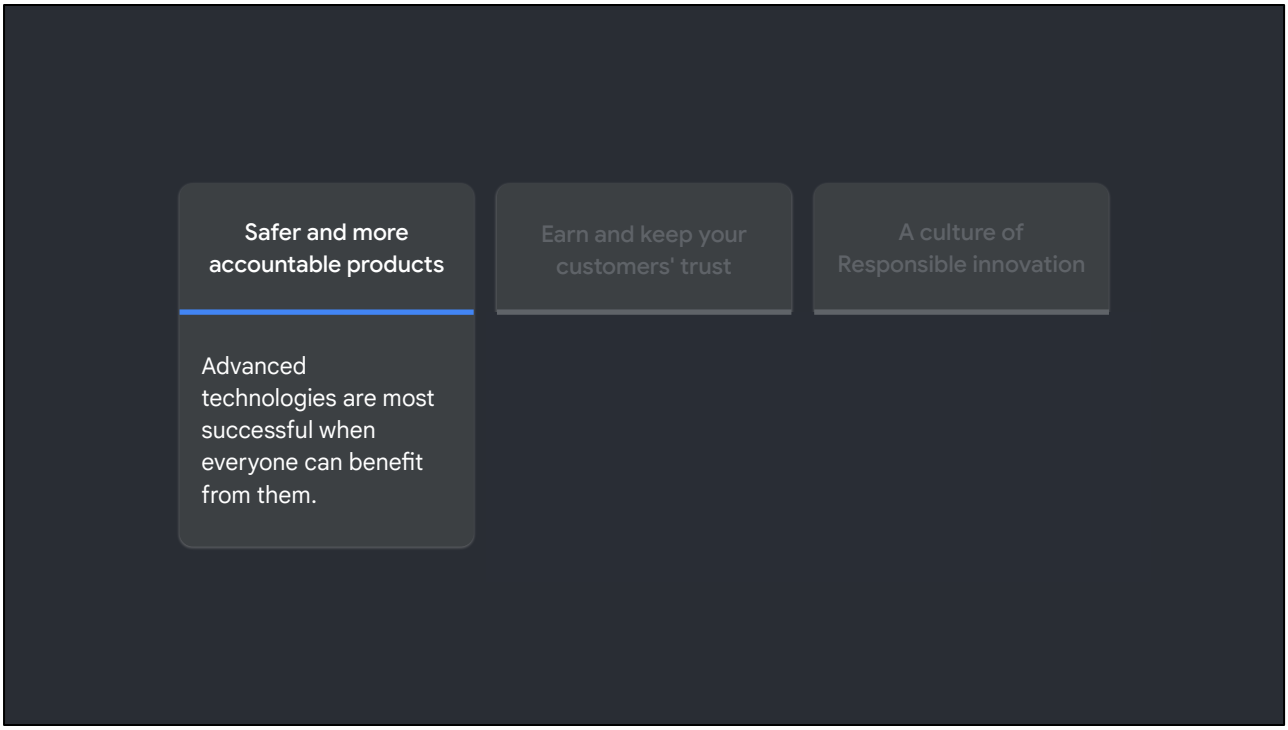
privacy



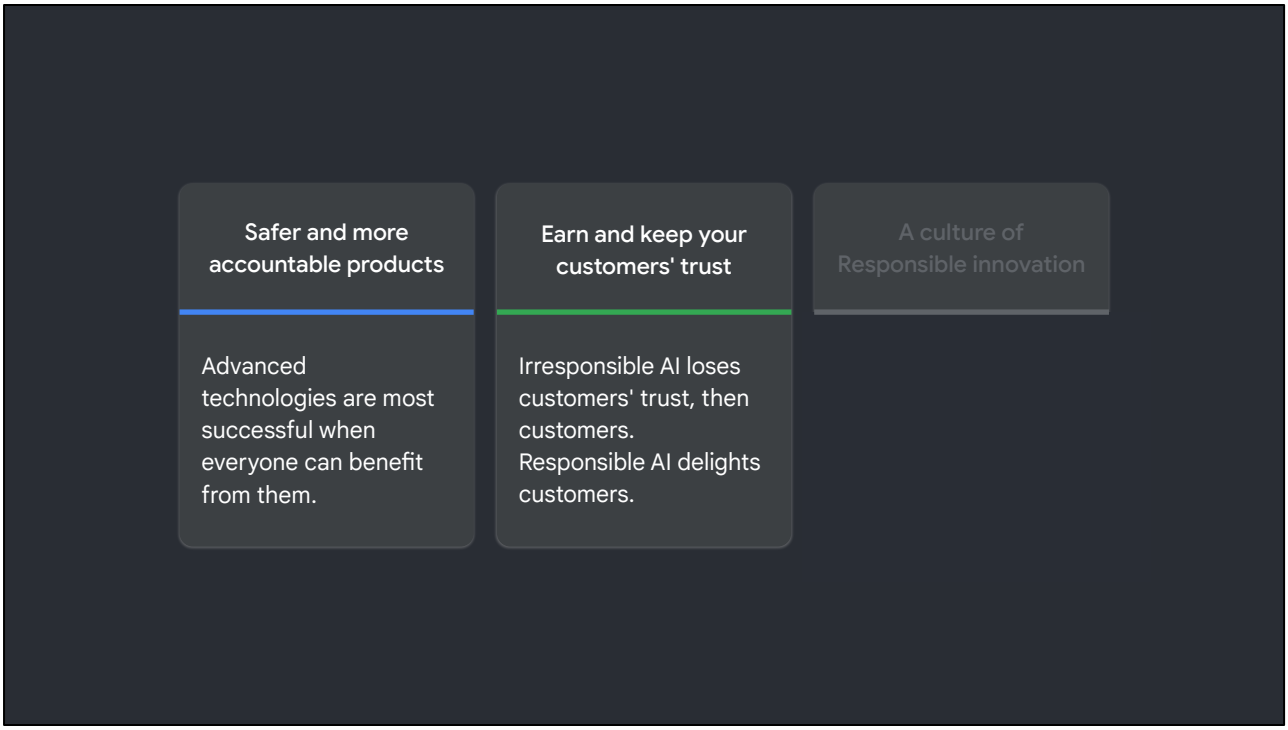
and safety.



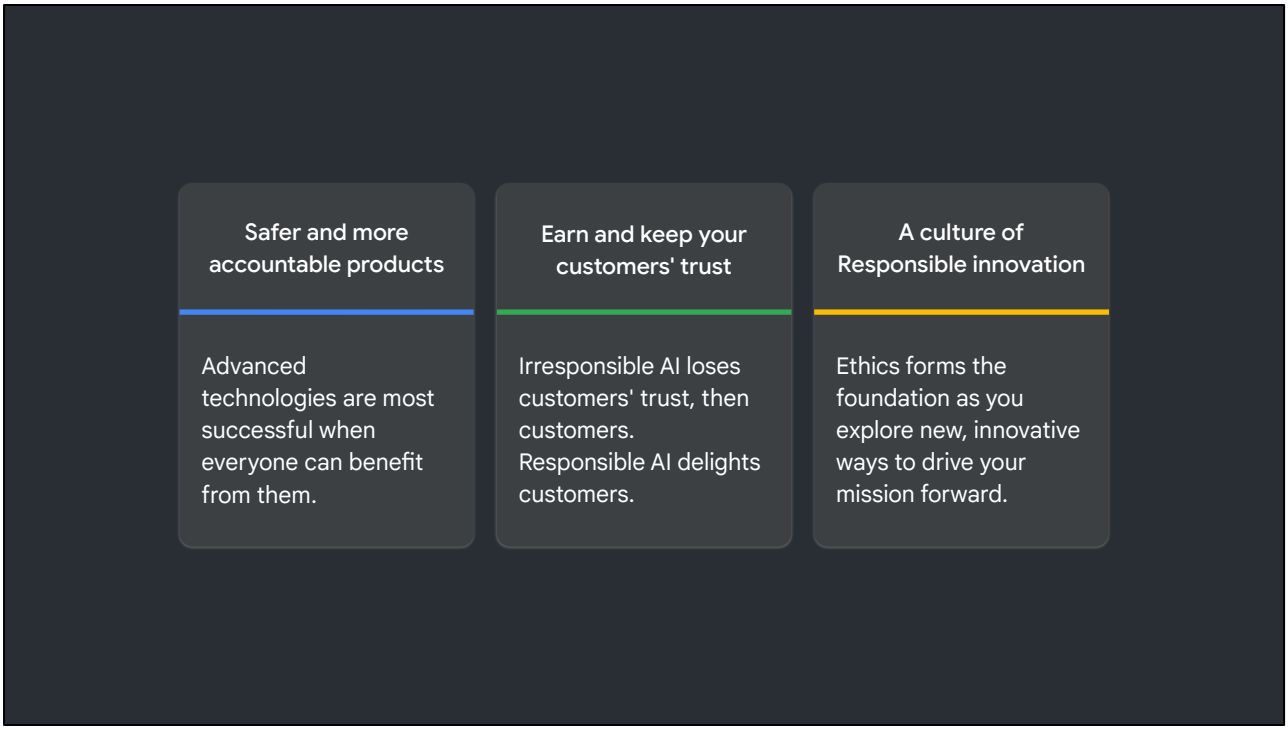
What we've found at Google is that following these foundational responsible AI principles leads to developing successful AI.
We learned that:



- Accountability ensures that your products are beneficial to everyone. Evaluating your AI systems, both when they perform as intended and when they don’t, is crucial to building accountable products.



- Building responsibility into any AI deployment makes better models and builds trust with your customers and your customers’ customers. If at any point that trust is broken, you run the risk of AI deployments being stalled or unsuccessful. At worst, it is harmful to the stakeholders that those products affect. Lack of trust in AI systems is a growing barrier to adoption in enterprise with more organizations selecting enterprise products based on AI commitments and practices.



- Ethical development drives innovation. Empowering AI decision-makers and developers to consider ethical considerations enables them to find new, innovative ways to drive your mission forward.

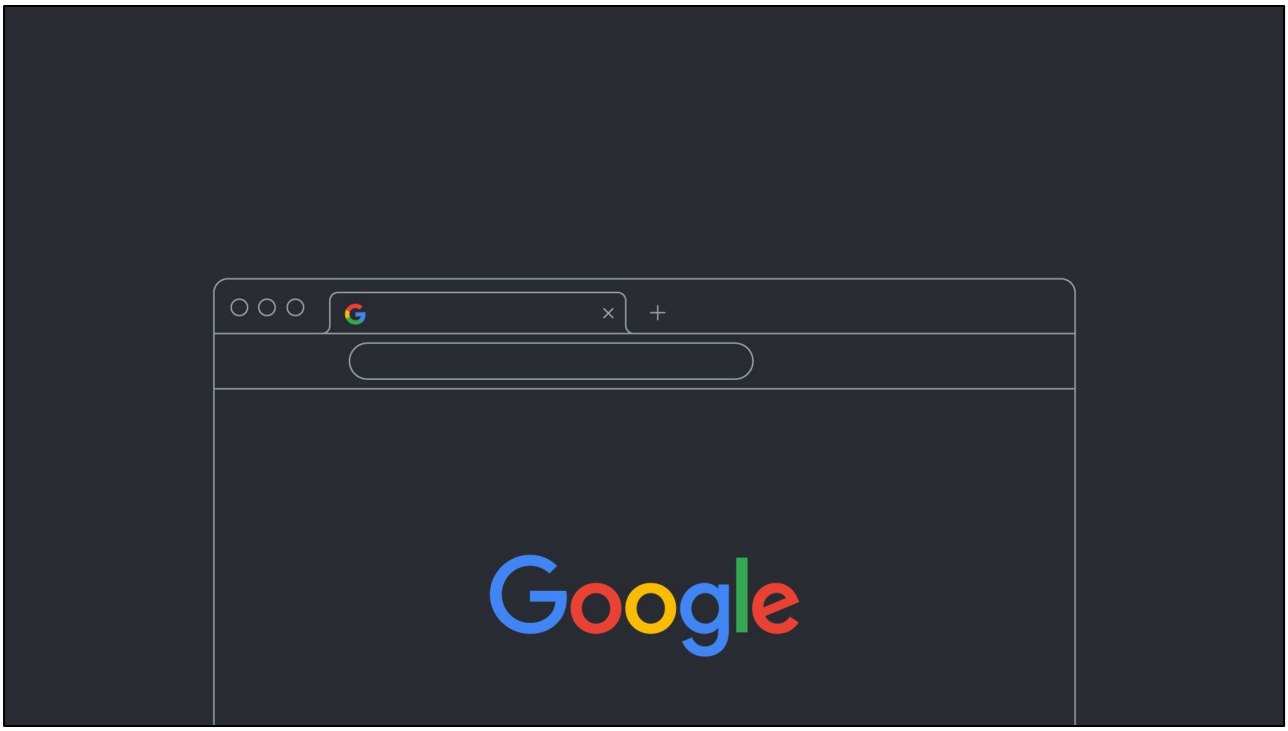
01 AI & Responsibility

02 Google's AI principles

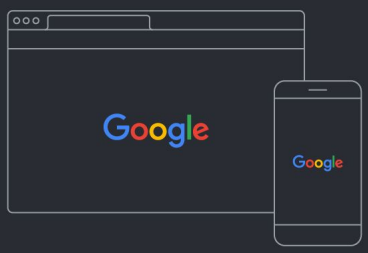
03 Responsible AI practices

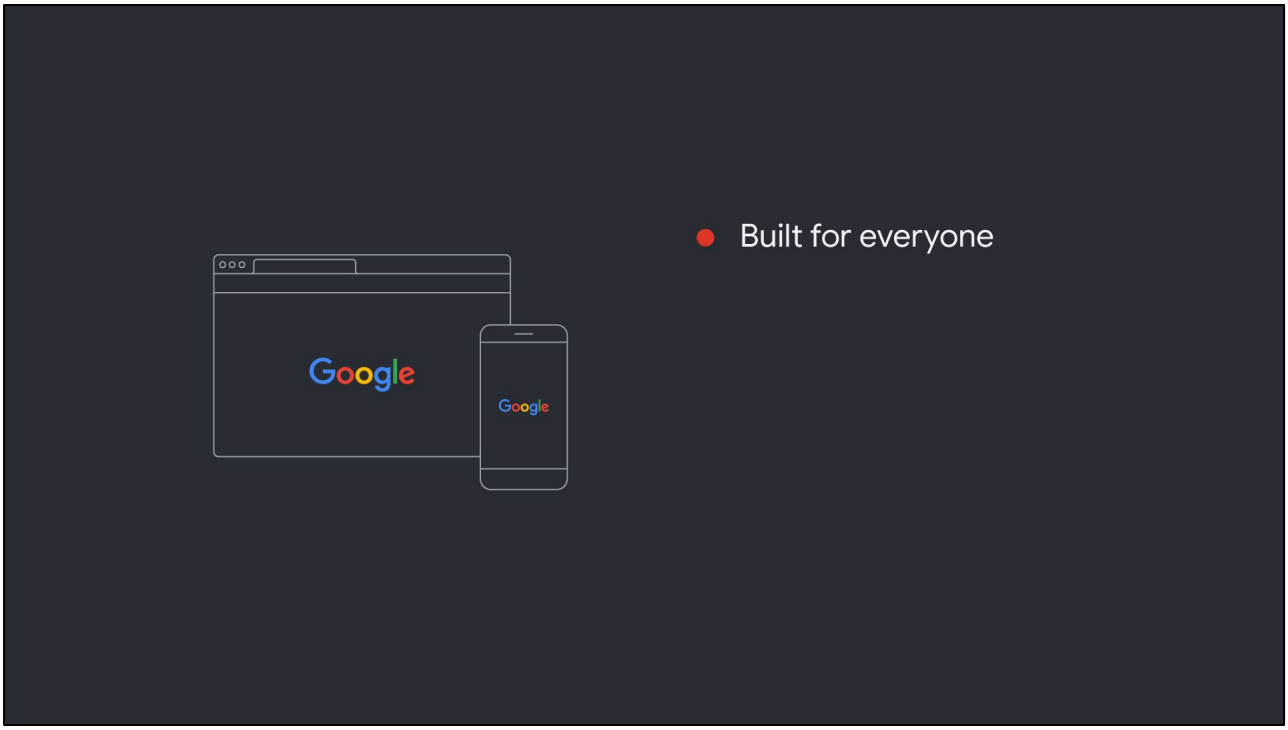
04 Case study: Google Flights

In 2018, Google announced its AI principles. These principles are concrete standards that actively govern our research and product development and affect our business decisions. We incorporated responsibility by design into our products, and even more importantly, our organization. Google has been constantly updating them since these principles were announced.

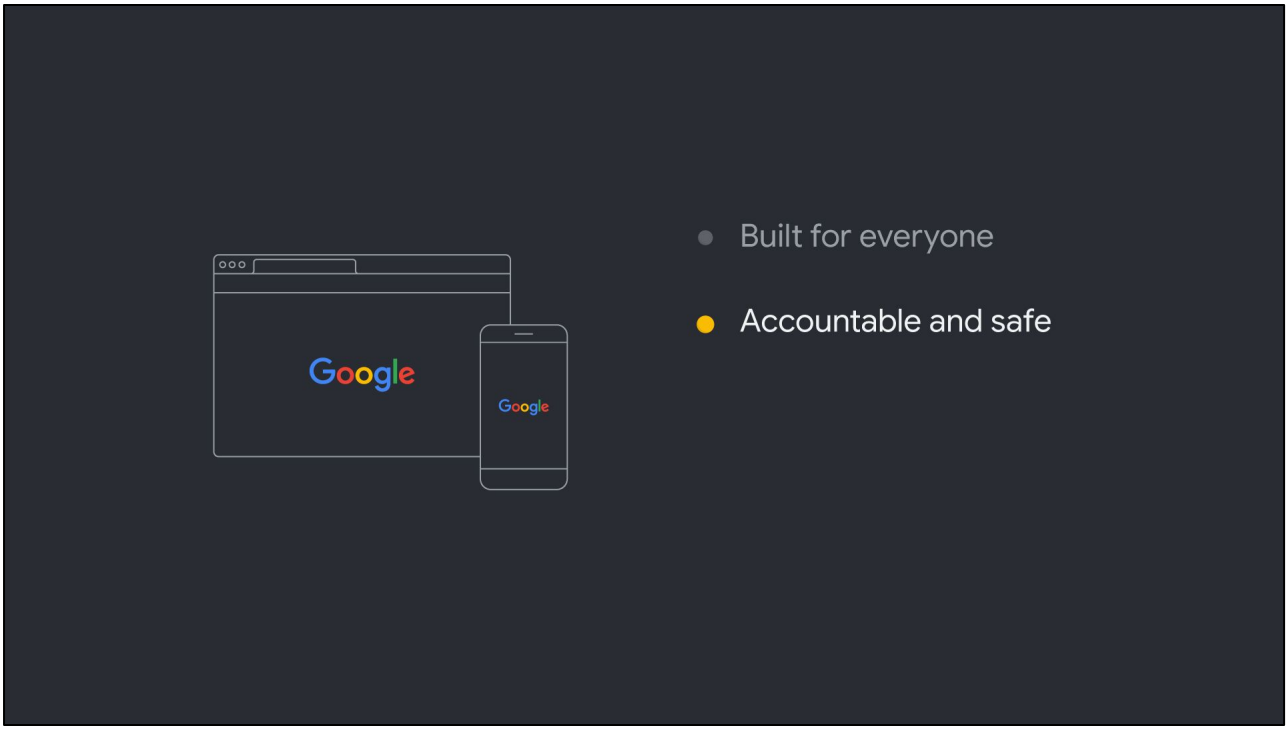


Our approach to responsible AI is rooted in a commitment to strive toward AI that

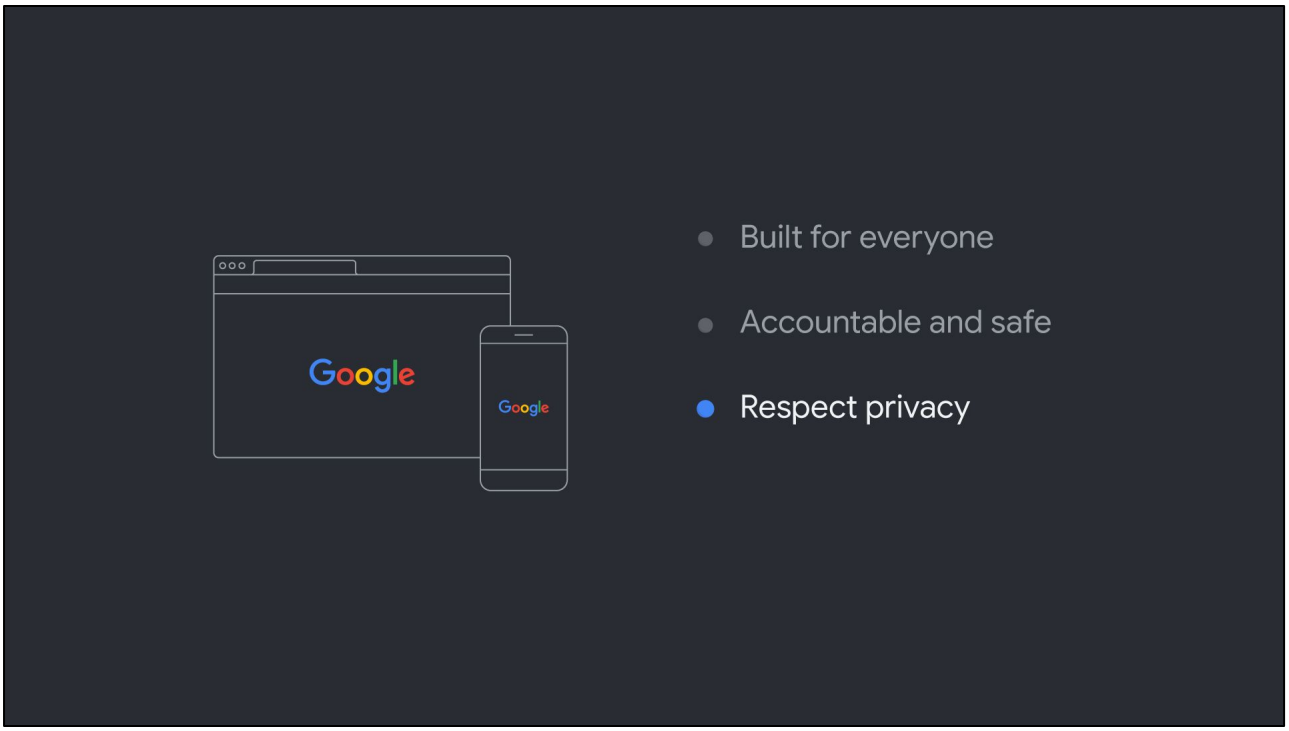

Google
Google



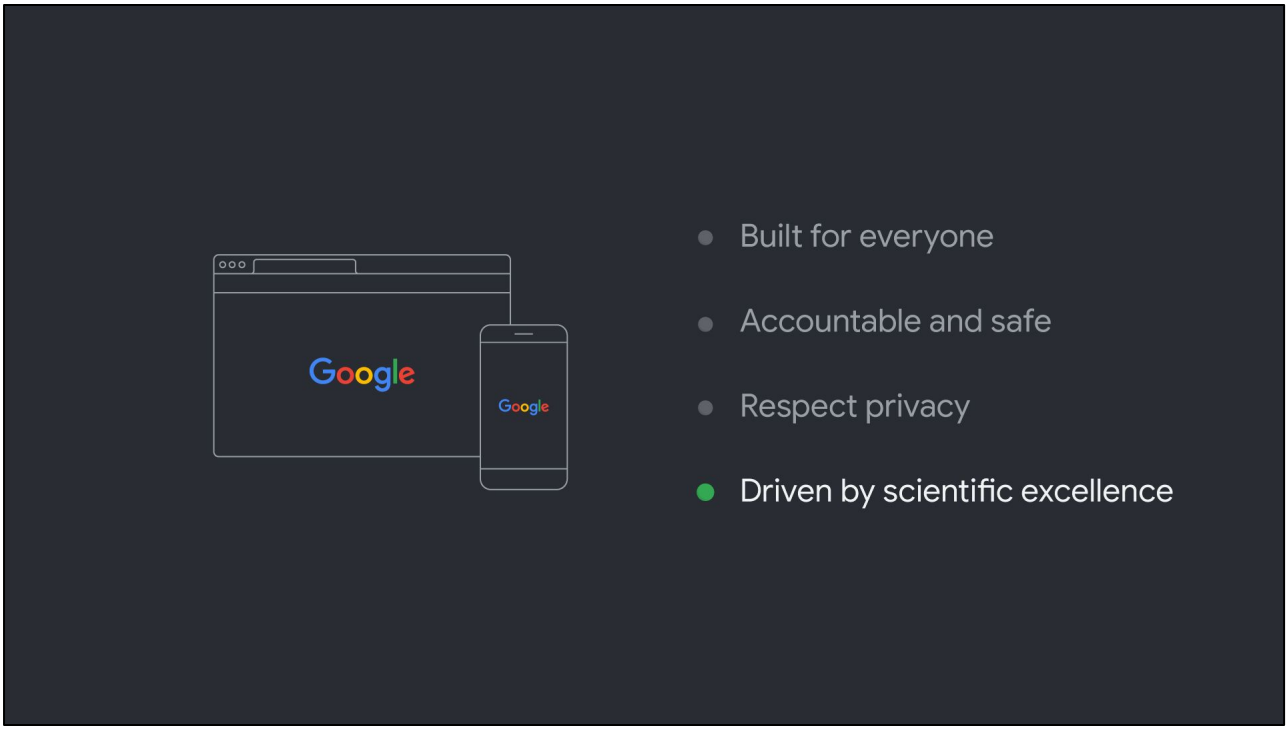
is built for everyone,



accountable and safe,

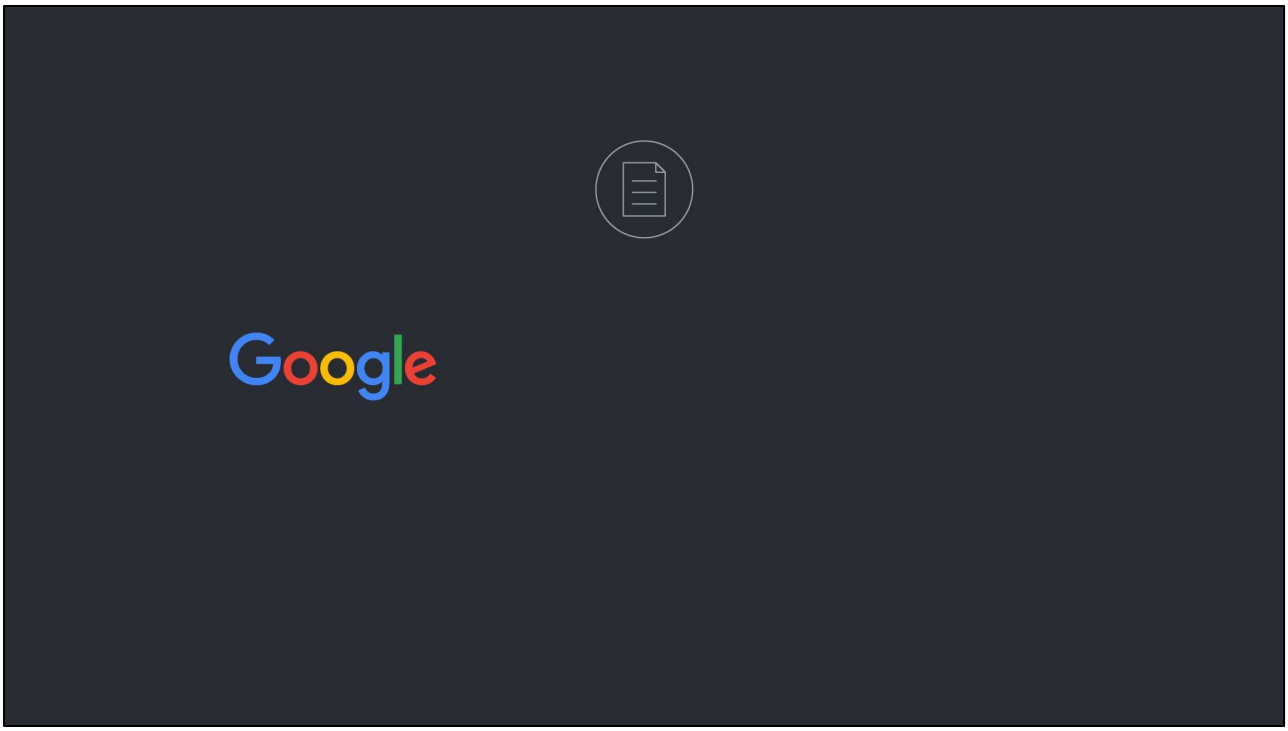


respects privacy,

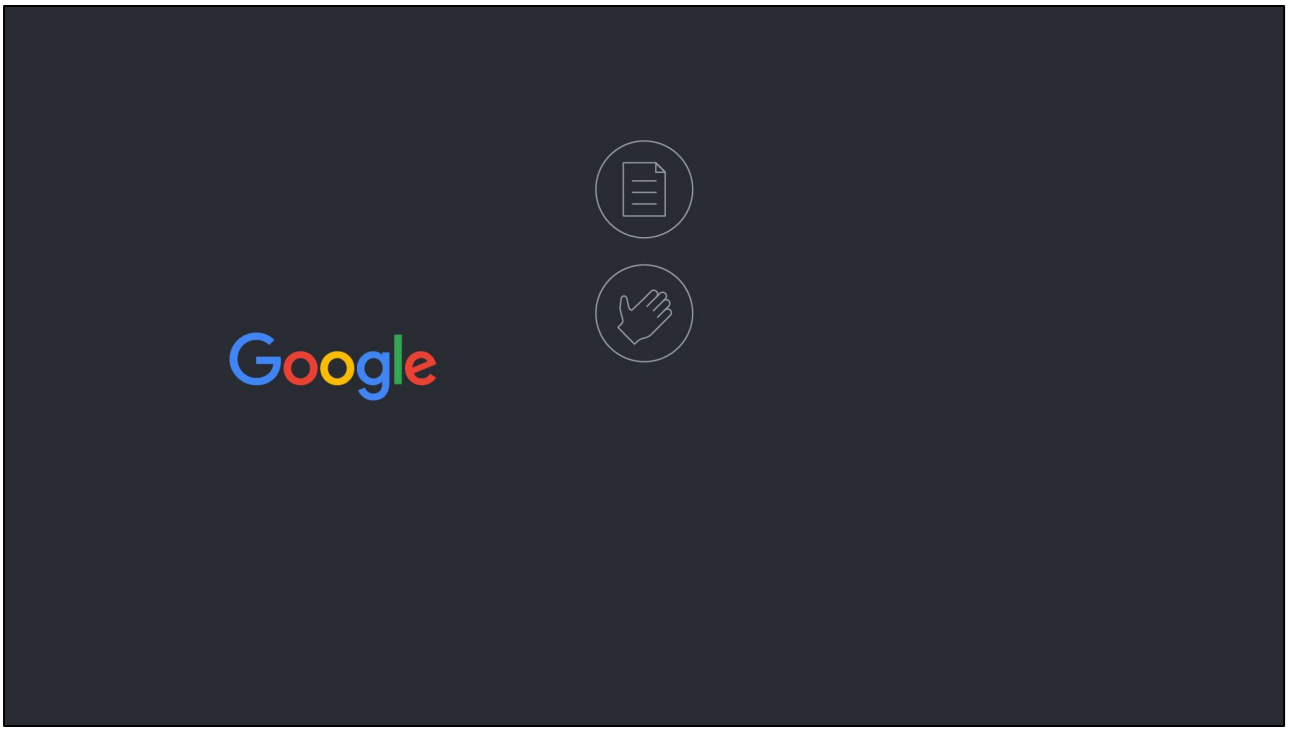


and driven by scientific excellence.

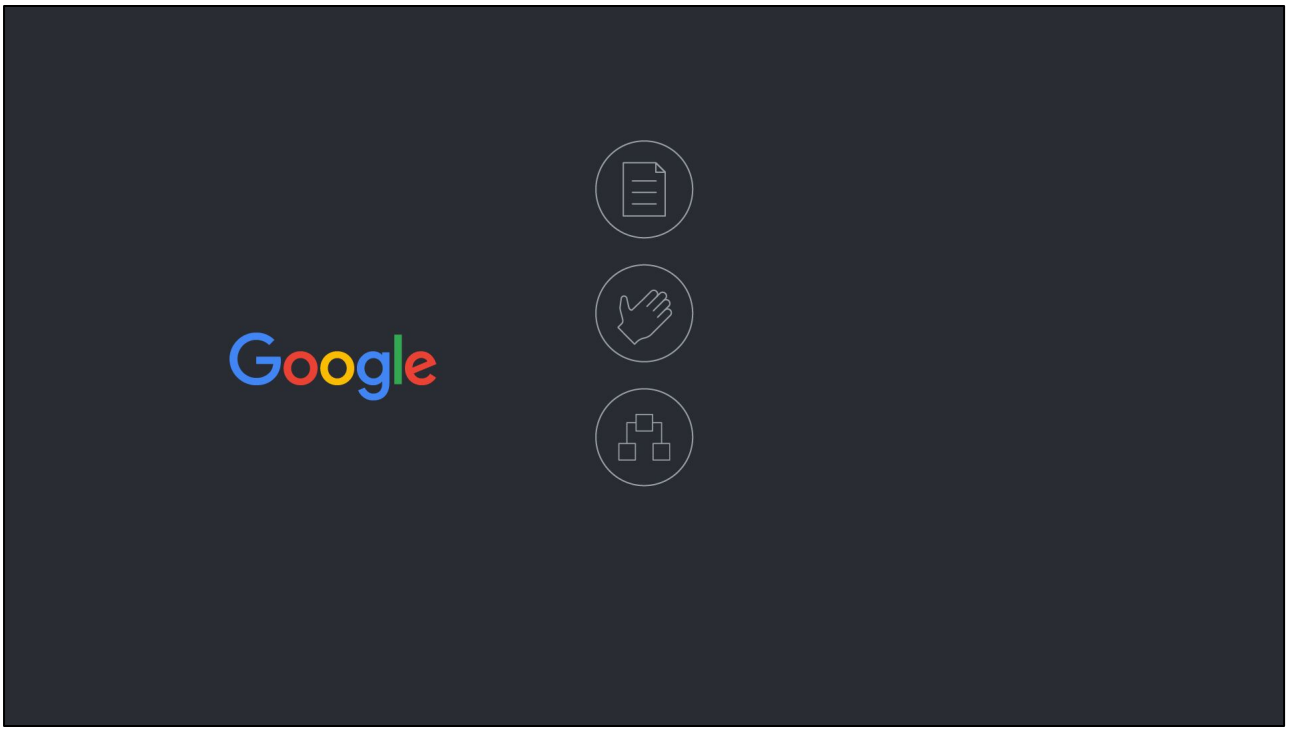
We’ve developed our own



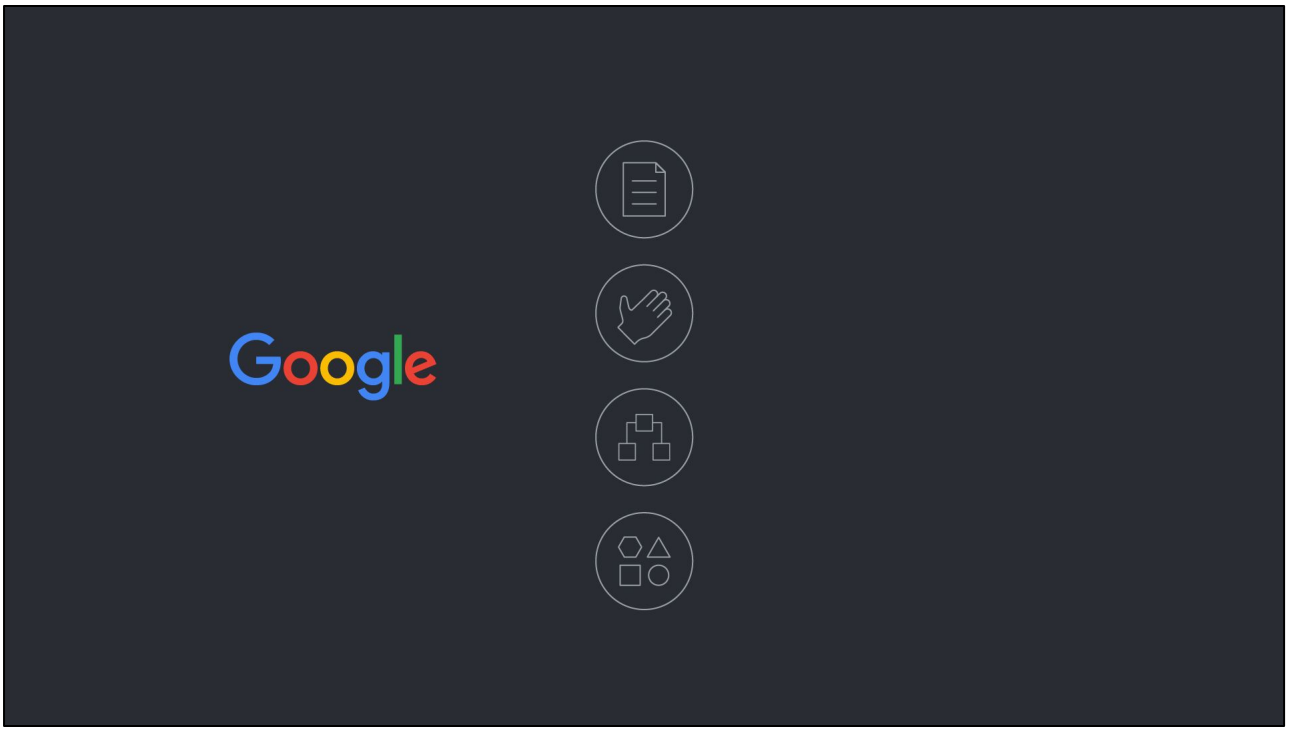
AI principles,



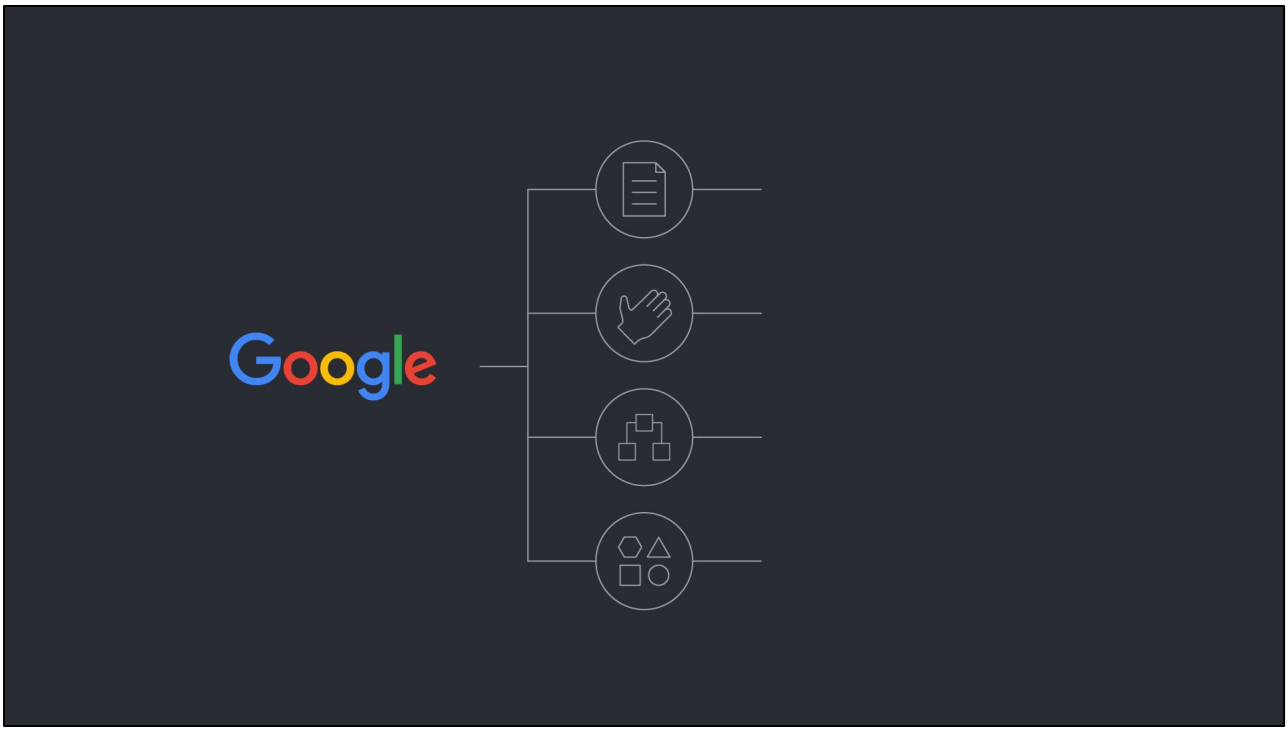
practices,



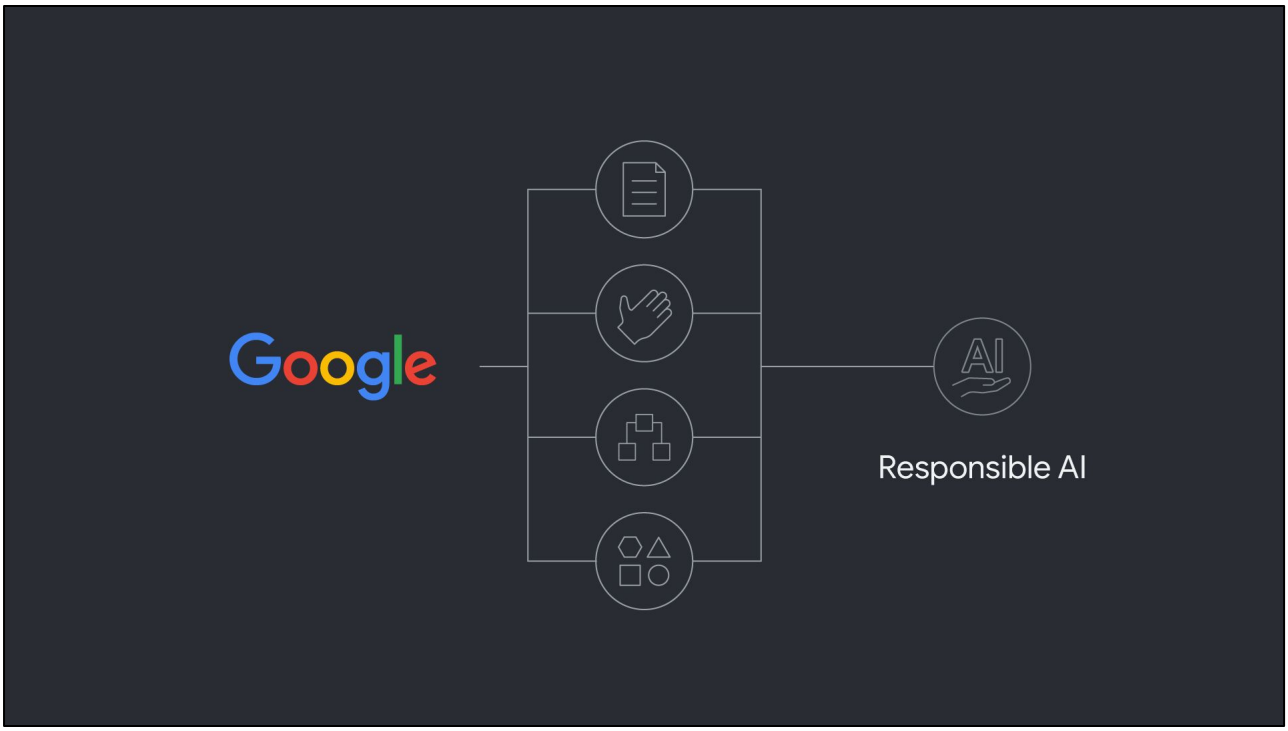
governance processes,



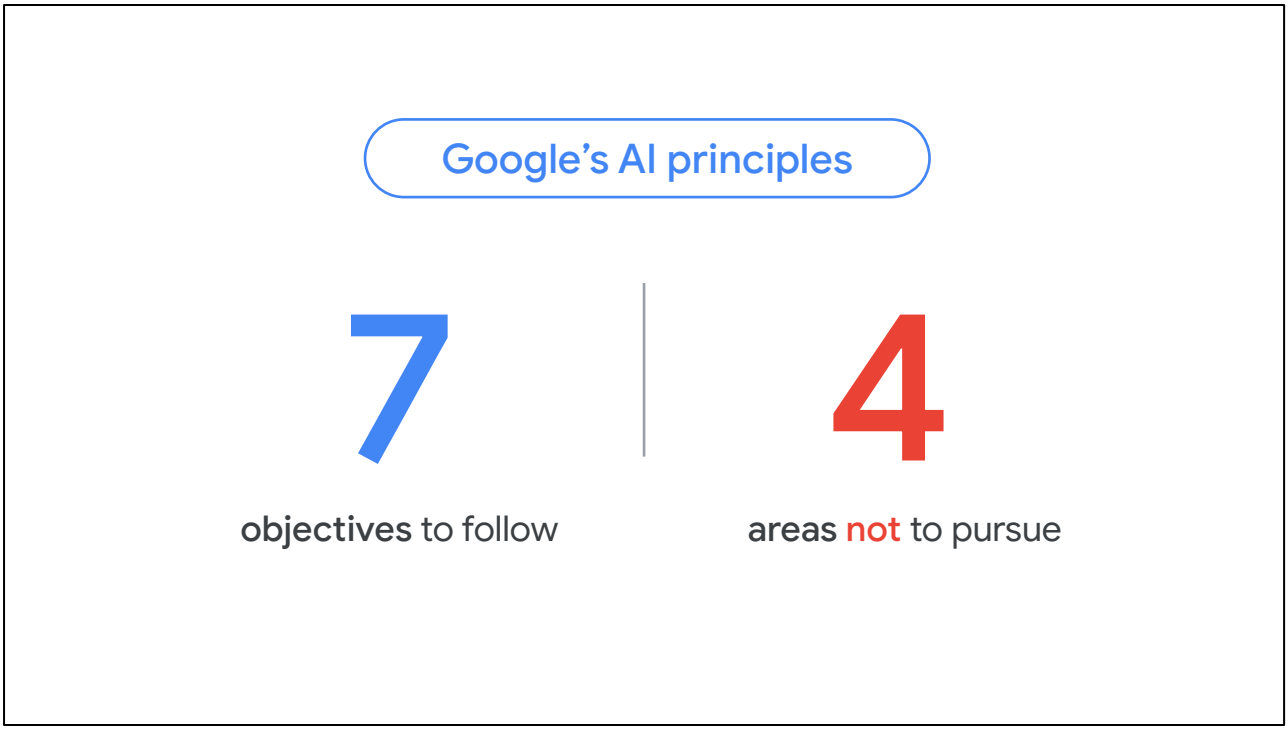
and tools



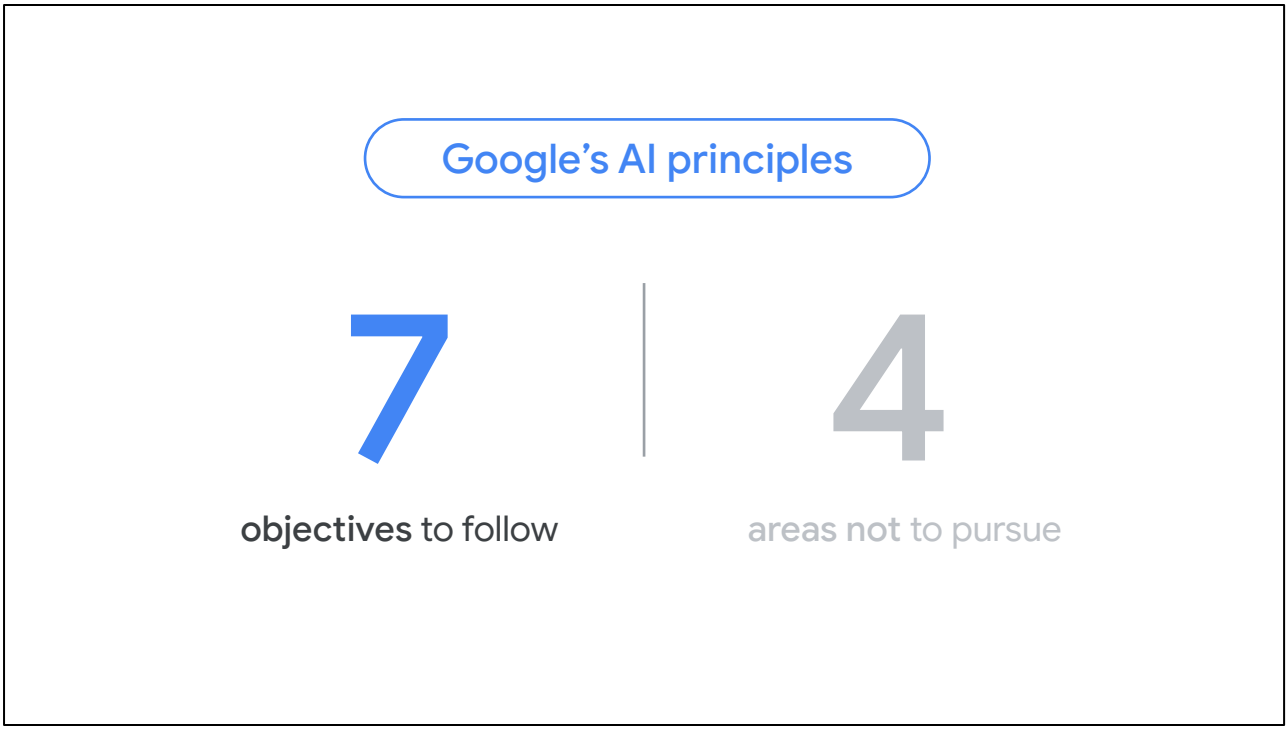
that together embody our values and guide our approach



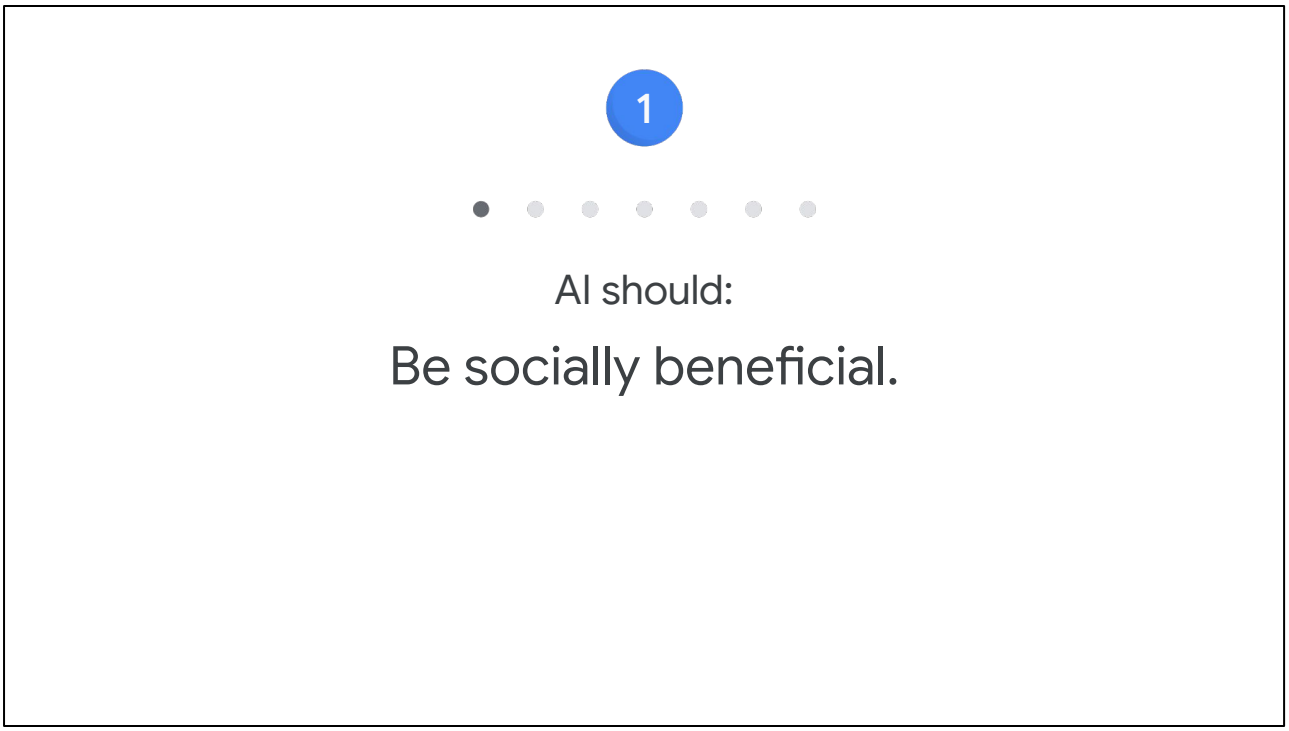
to responsible AI.



Google's AI principles describe our commitment to developing technology responsibly, and work to establish specific application areas we will not pursue.
Let's explore each of Google's 7 principles for AI applications, and 4 areas you should not pursue.

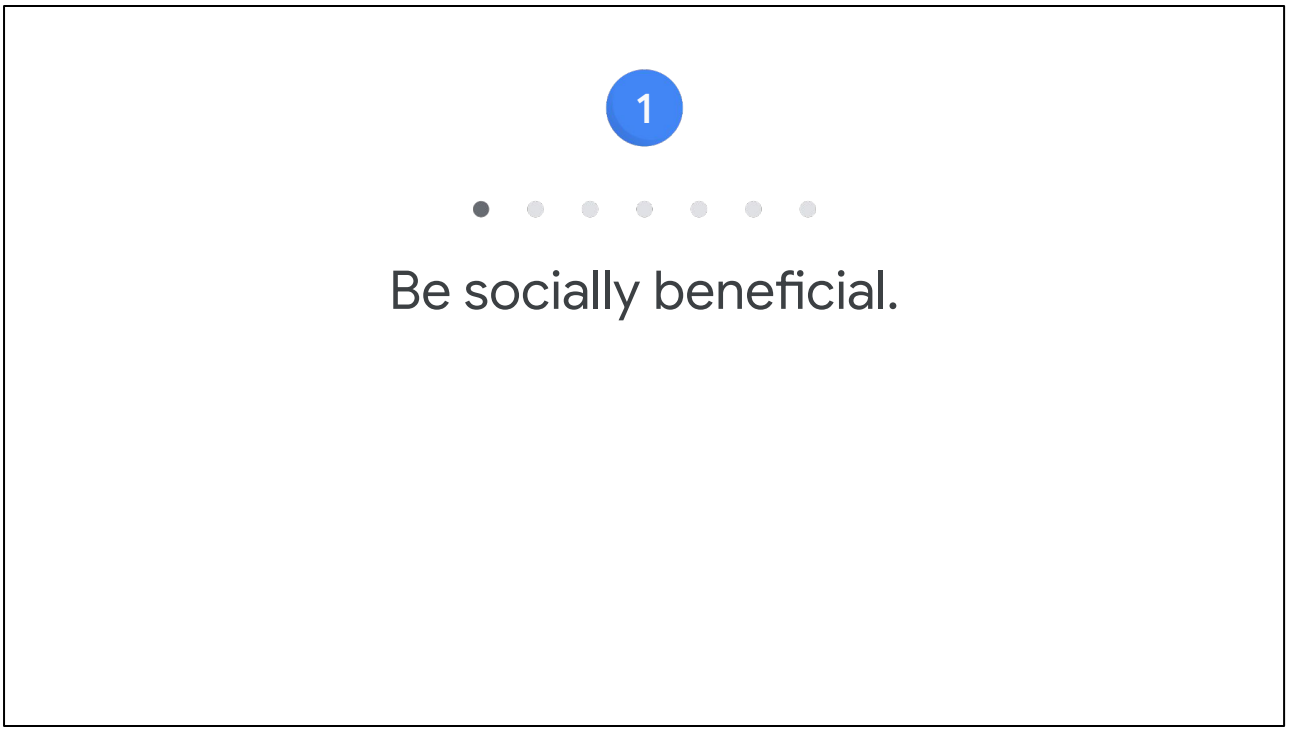


Let's start with the 7 principles for AI applications.

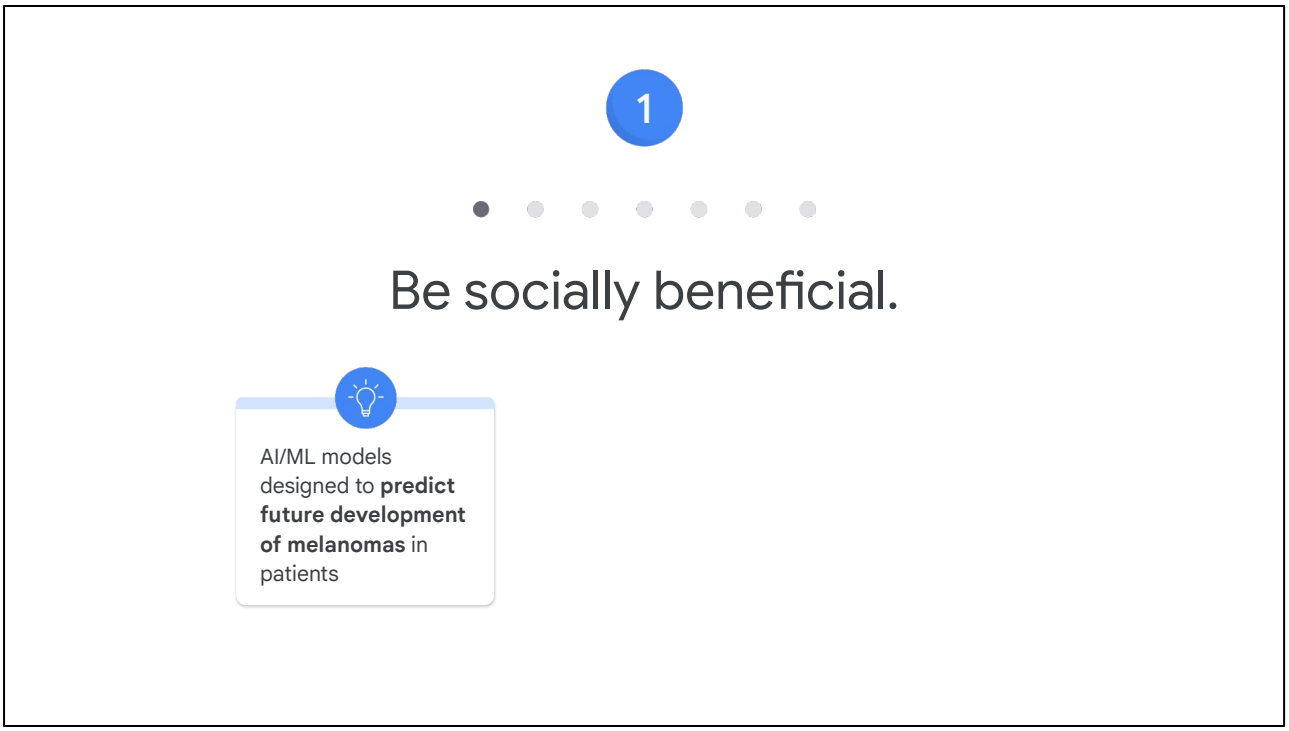


Google's first principle states that AI should be socially beneficial. For any project, we should:

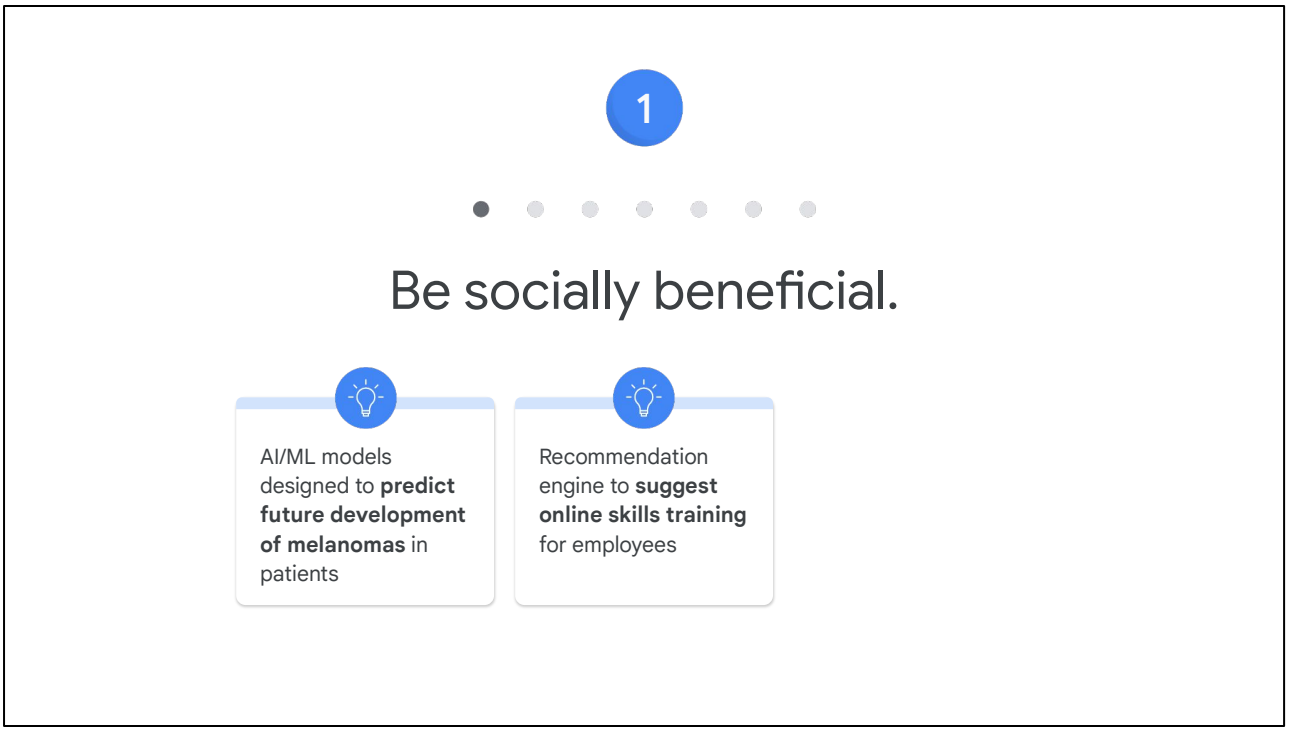
- Consider a broad range of social and economic factors, and proceed only where we believe that the overall likely benefits substantially exceed the foreseeable risks and downsides.
- Make high-quality and accurate information readily available using AI, while continuing to respect cultural, social, and legal norms in the countries where we operate.
- Evaluate when to make our technologies available on a non-commercial basis.



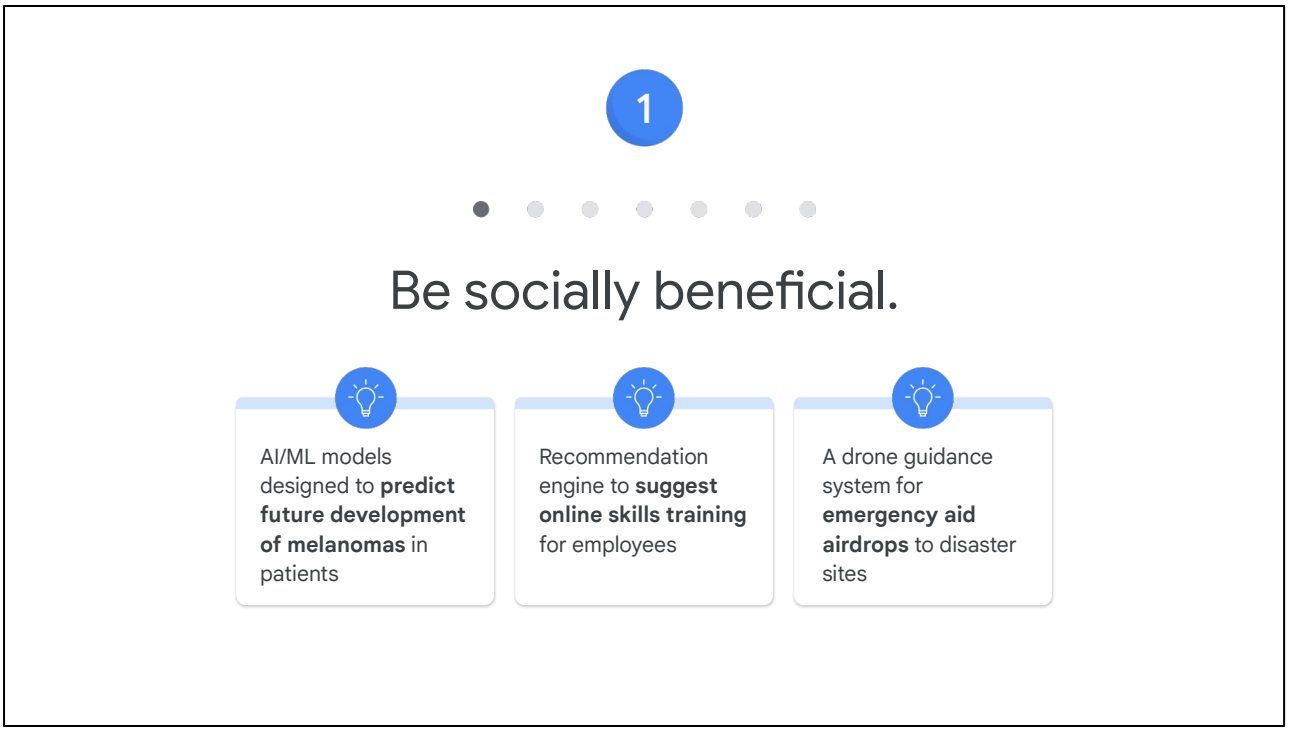
Some examples of the first principle and socially beneficial AI applications could include:



- AI/ML models designed to predict future development of skin cancer in patients.

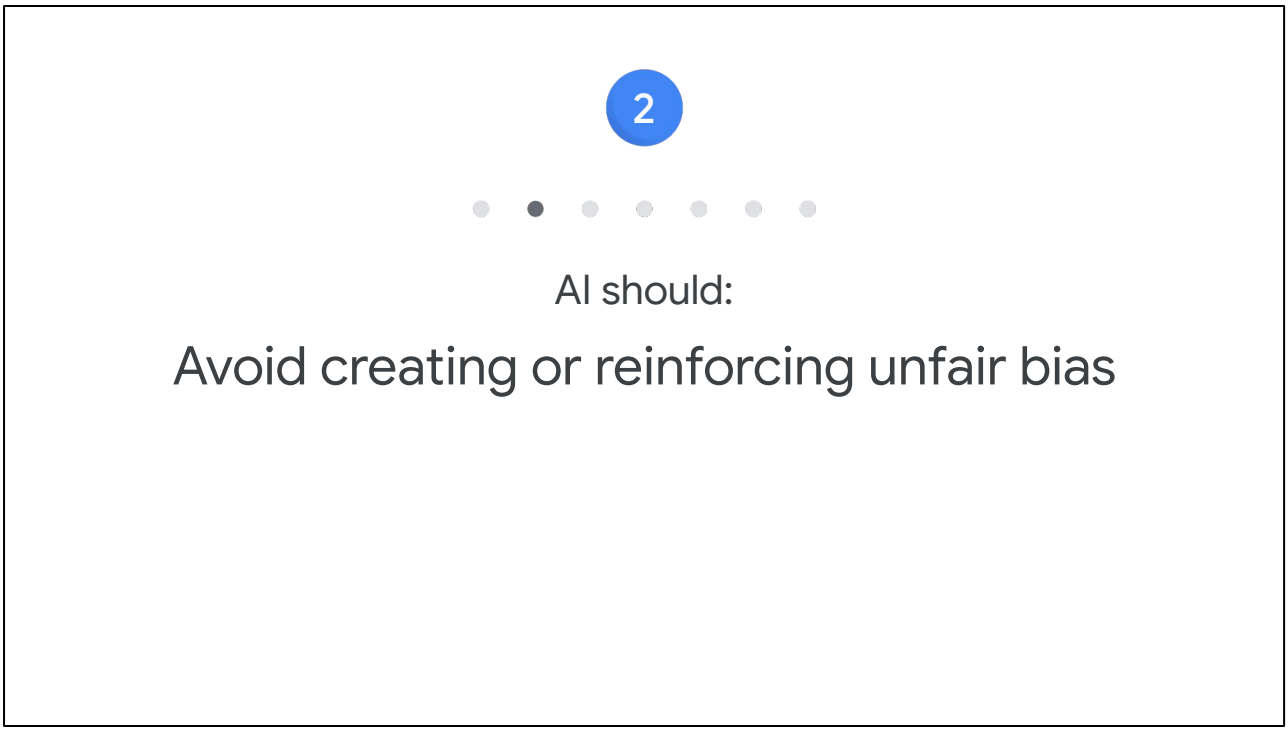


- A recommendation engine to suggest online skills training for retail employees.



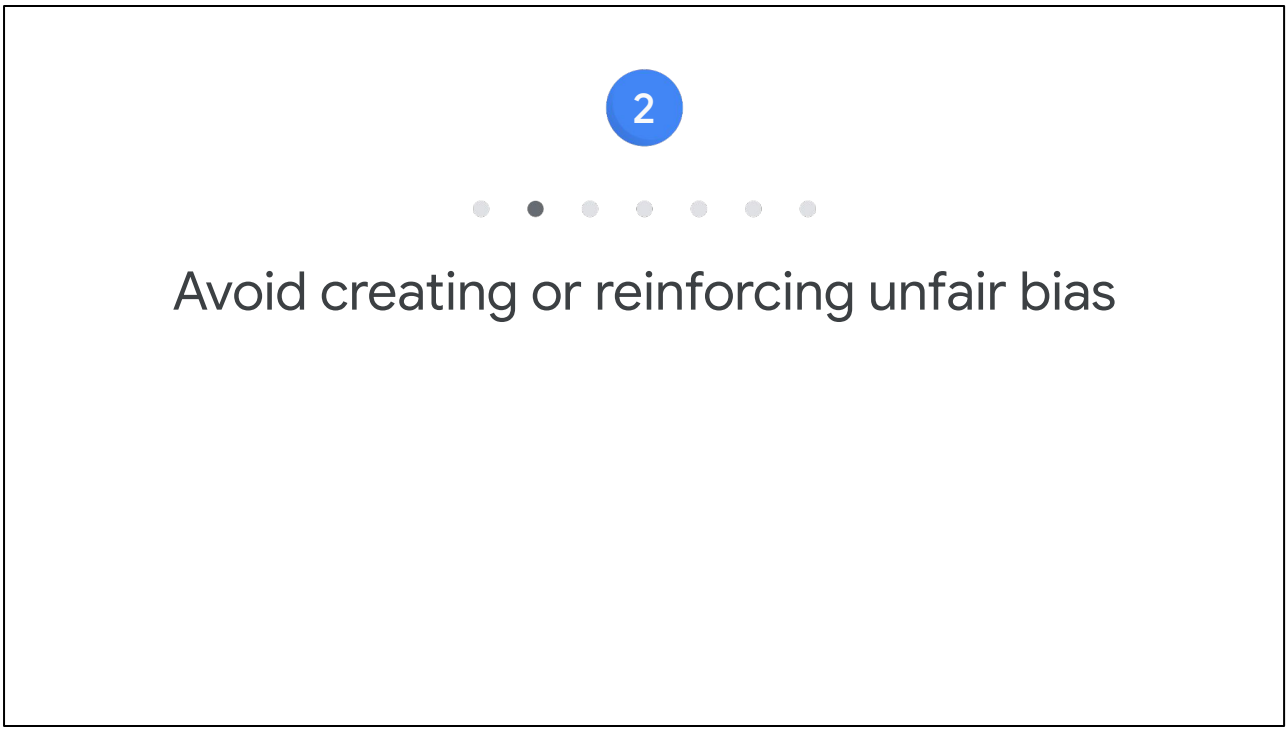
- Or a drone guidance system for emergency aid airdrops to disaster sites.

However, no AI application, no matter how well-intentioned, is inherently "absolutely beneficial." Its impact depends entirely on how responsibly we design and deploy it.

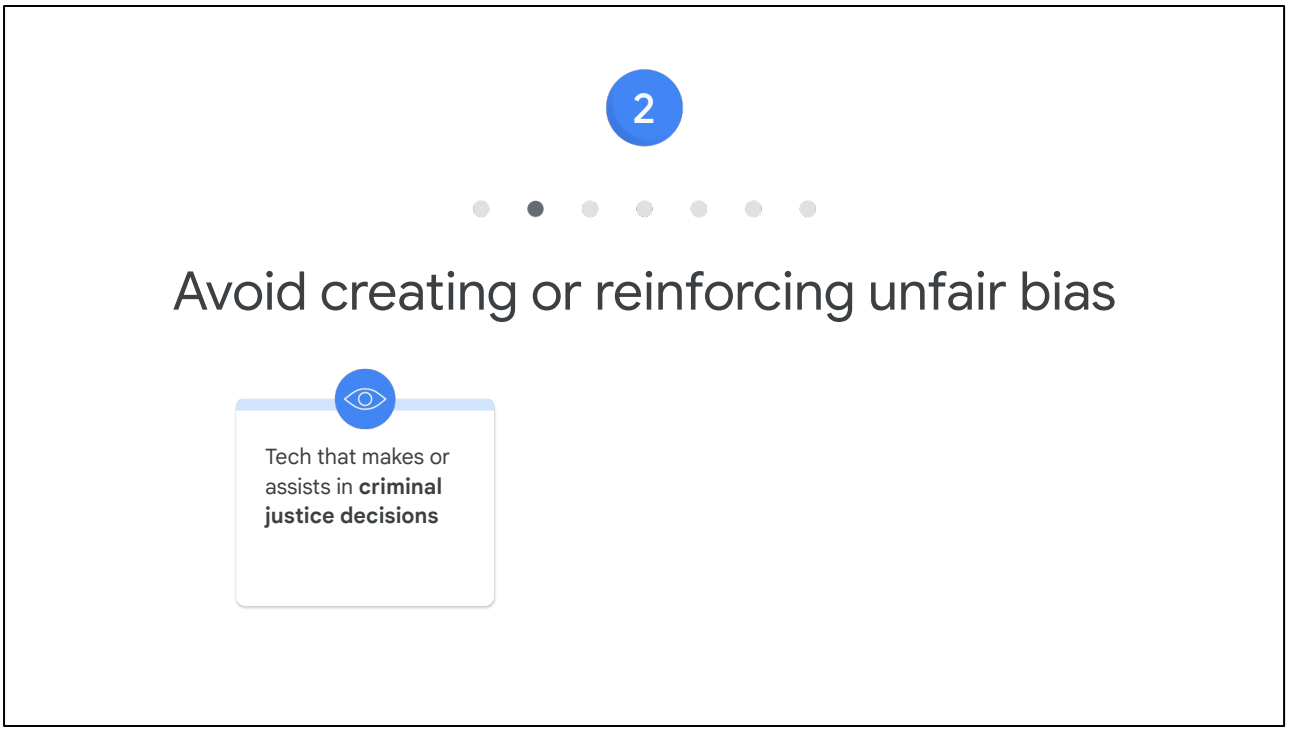


Google's second principle states: AI should avoid creating or reinforcing unfair bias and unjust effects on people, particularly those related to sensitive characteristics such as race, ethnicity, gender, ability, and political or religious belief.

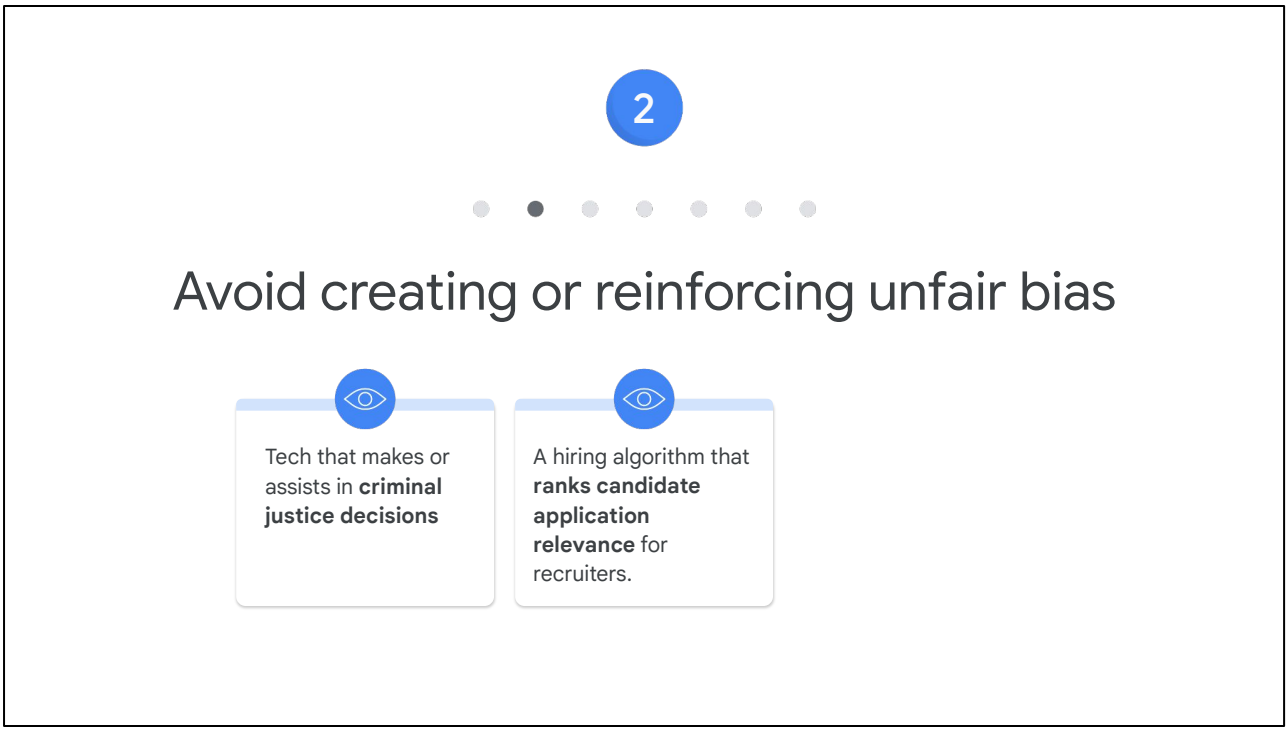
Distinguishing fair from unfair biases is not always simple, and differs across cultures and societies.



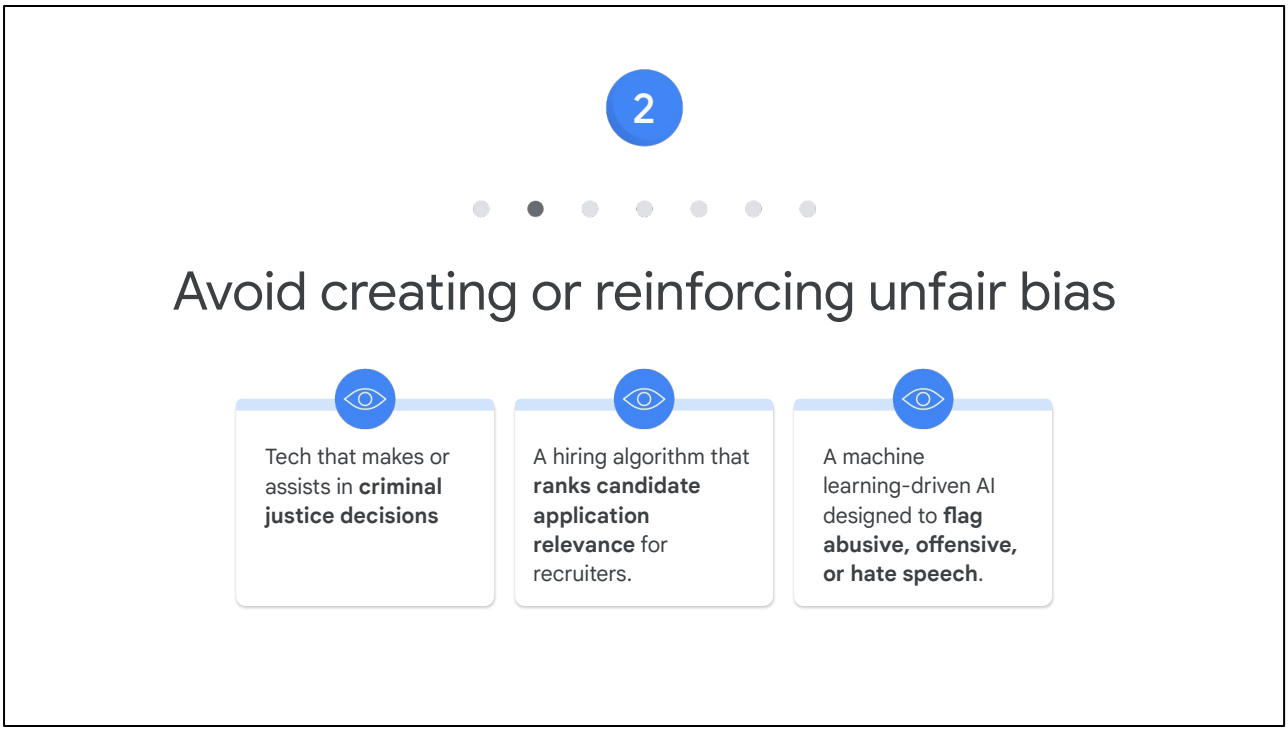
Here are a few examples of the second principle where it's important to avoid applications that create or reinforce unfair bias. They include:



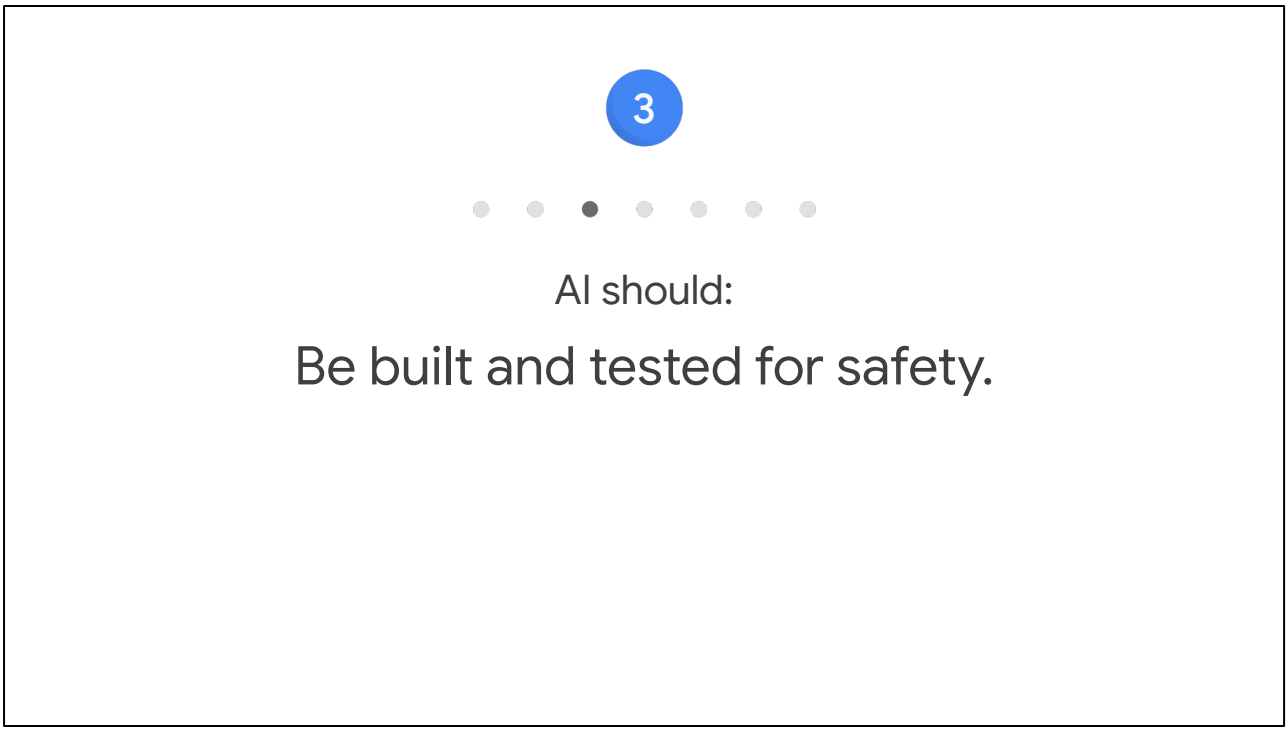
- Tech that makes or assists in criminal justice decisions.



- A hiring algorithm that ranks candidate application relevance for recruiters.

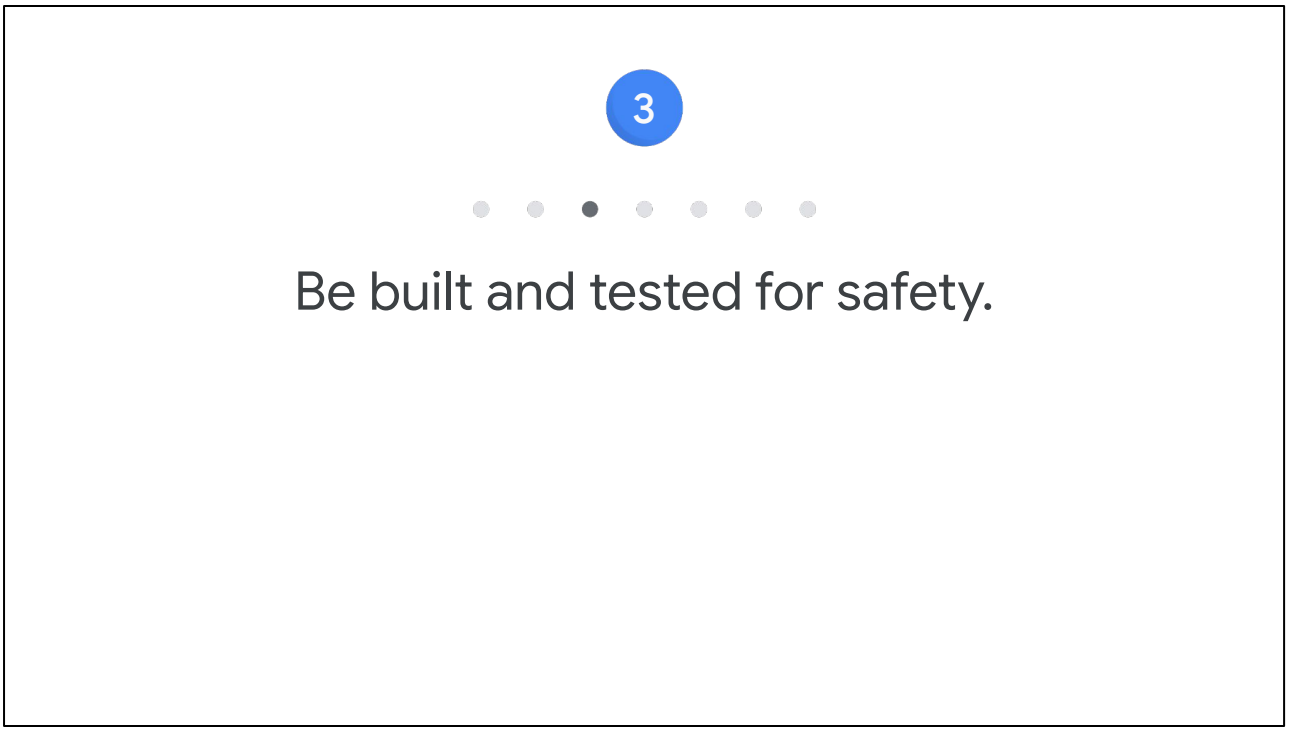


- A machine learning-driven AI designed to flag abusive, offensive, or hate speech.

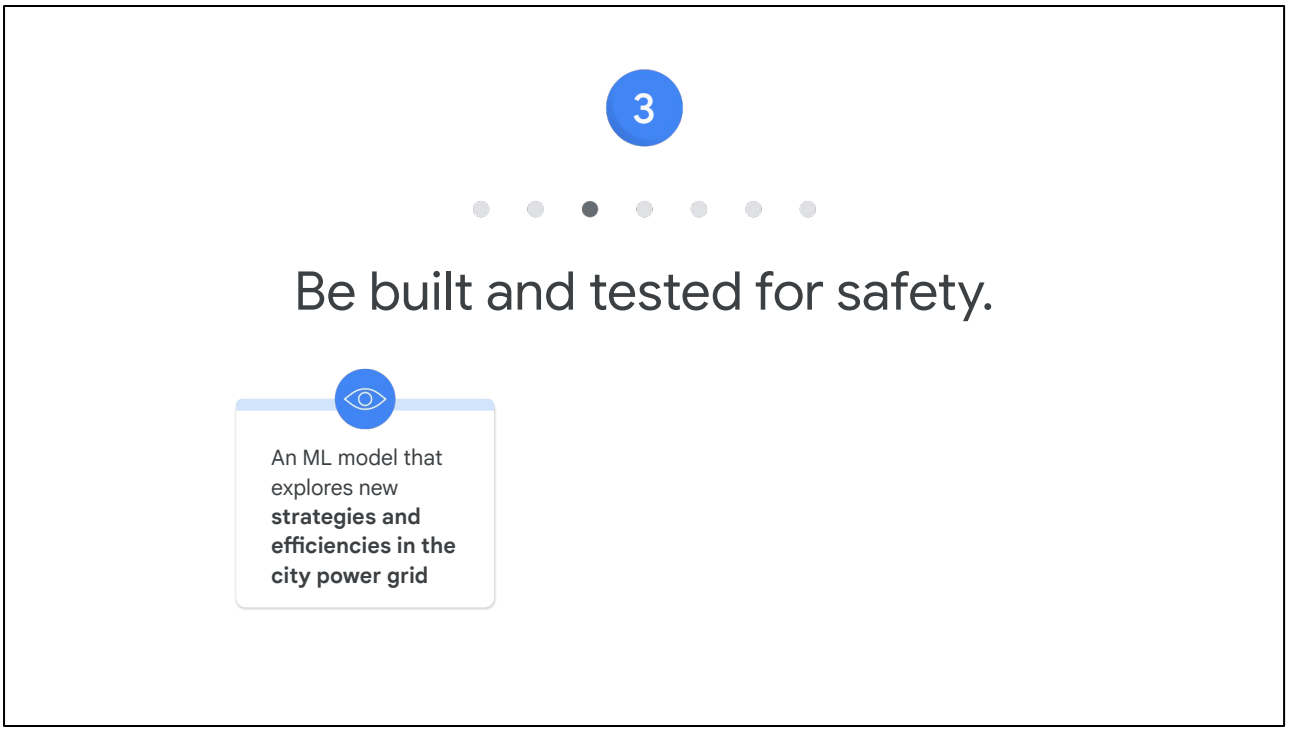


Google's third principle states: AI should be built and tested for safety. For any project, we should:

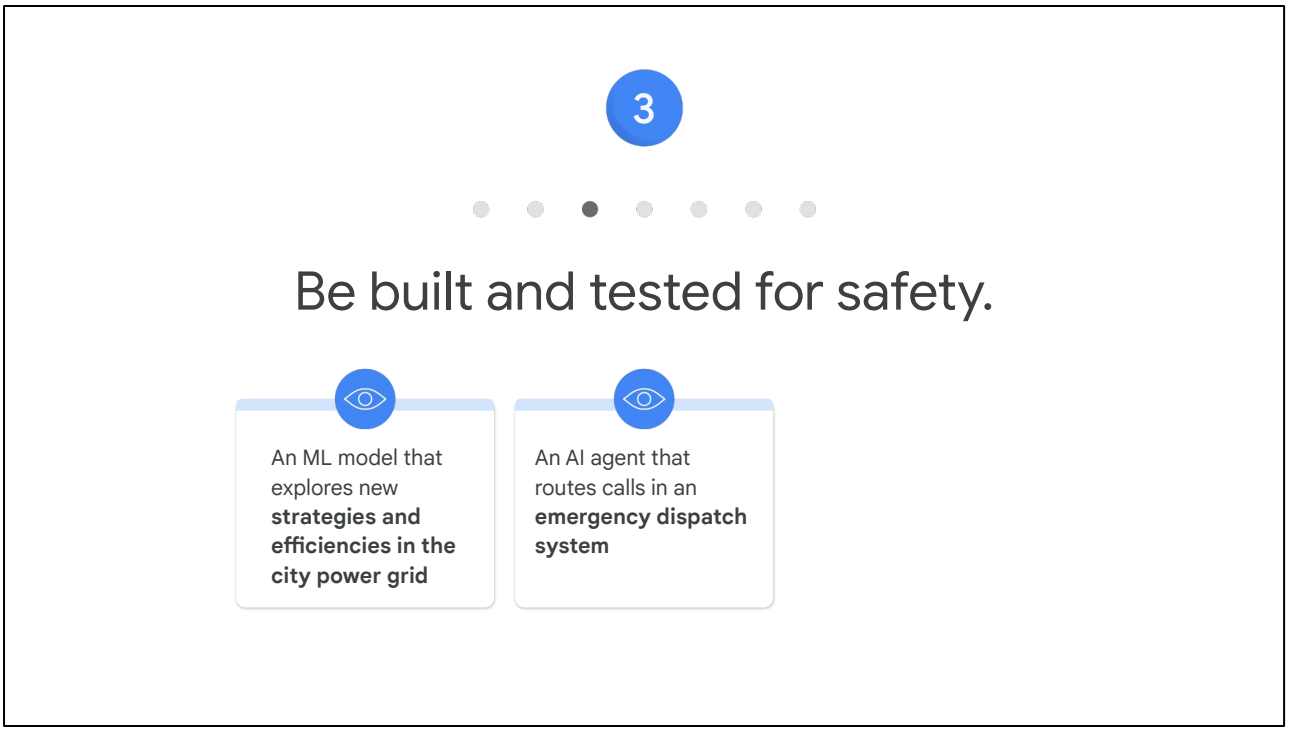
- Avoid unintended results that create risks of harm.
- Be appropriately cautious.



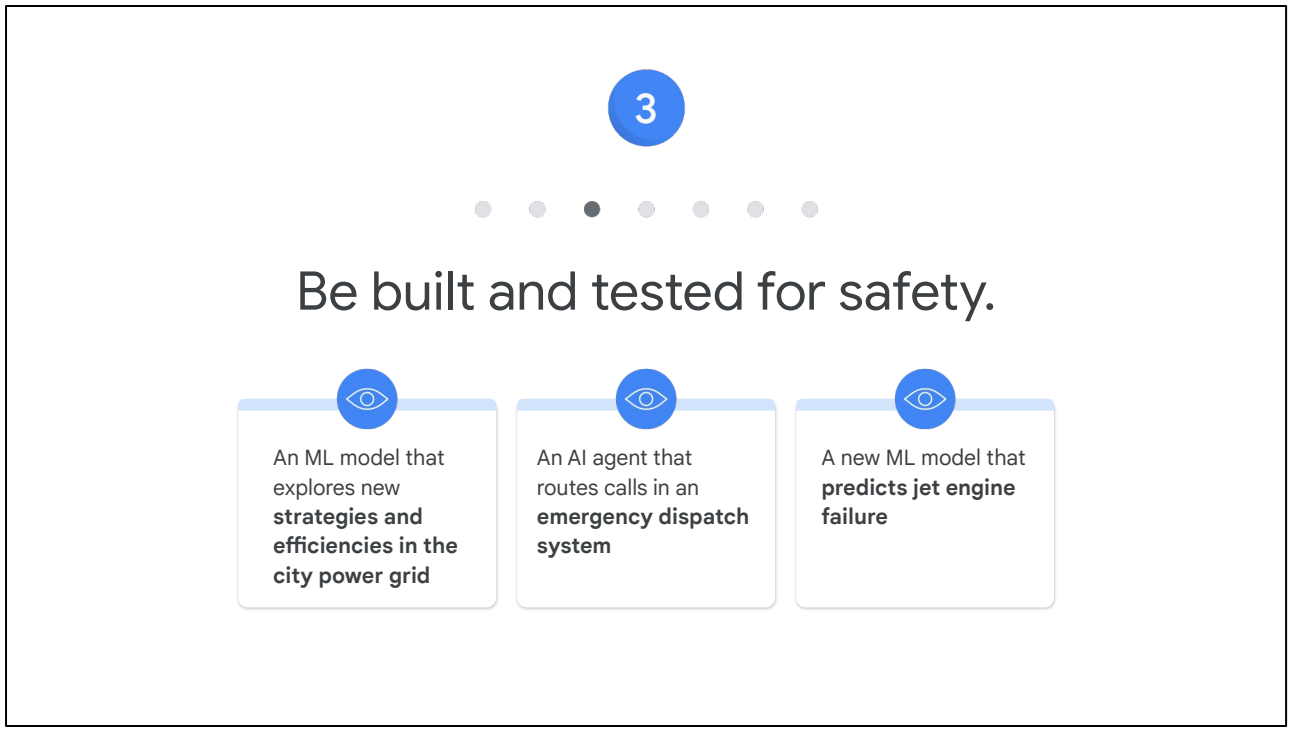
Some use case examples where the third principle is especially important include:



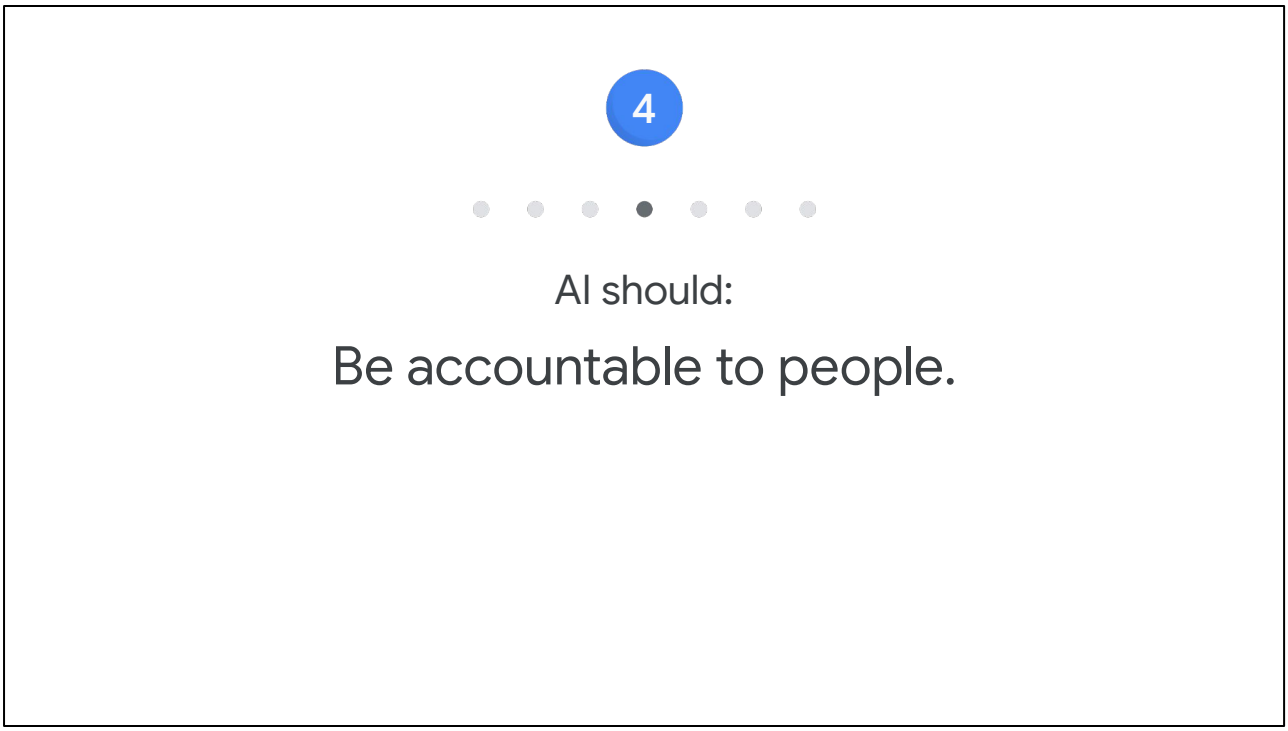
- An ML model that explores new strategies and efficiencies in the city power grid.



- An AI agent that routes calls in an emergency dispatch system.

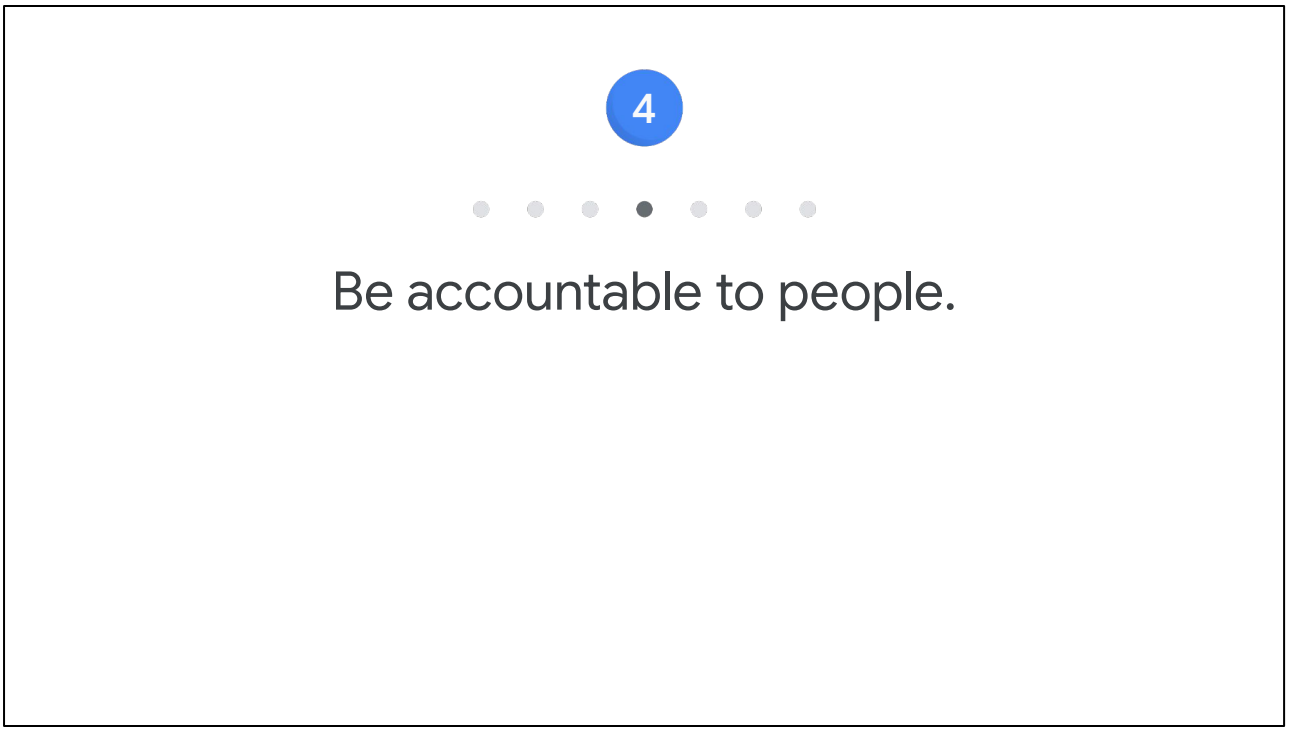


- A new ML model that predicts jet engine failure.

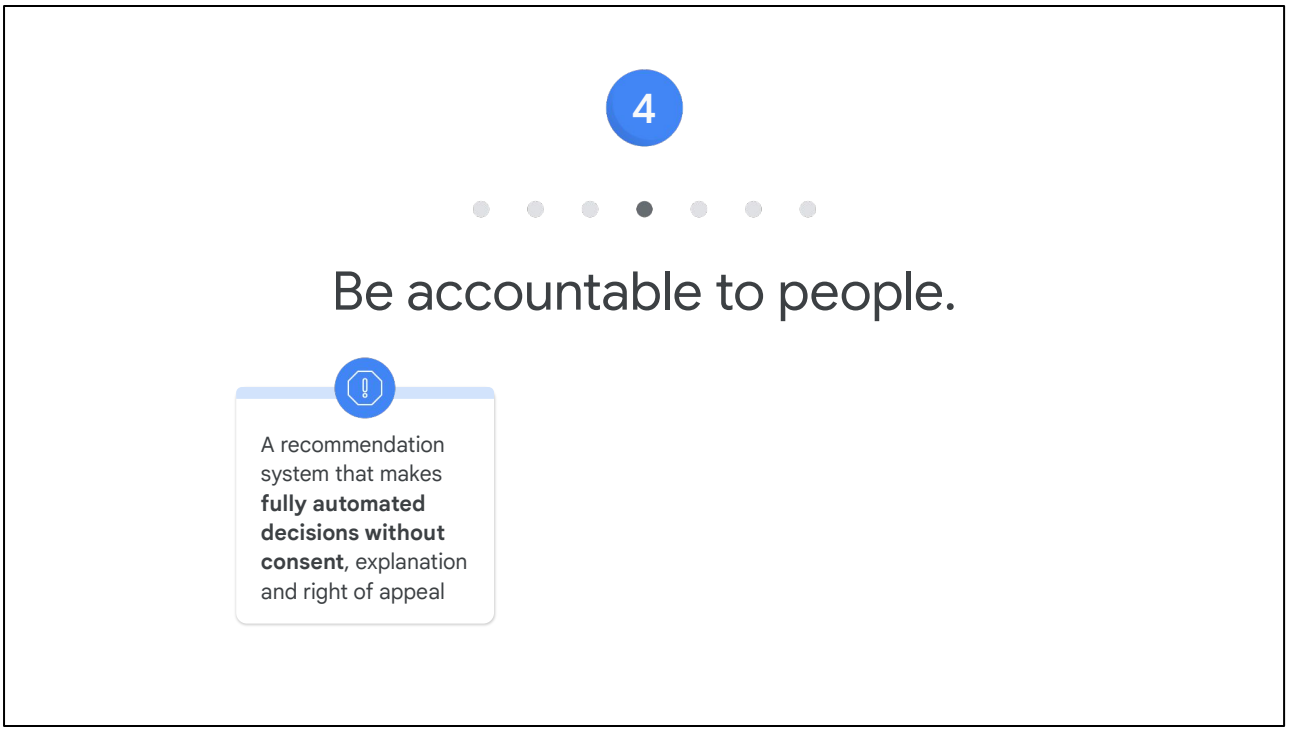


Google's fourth principle states: AI should be accountable to people. For any project, we should:

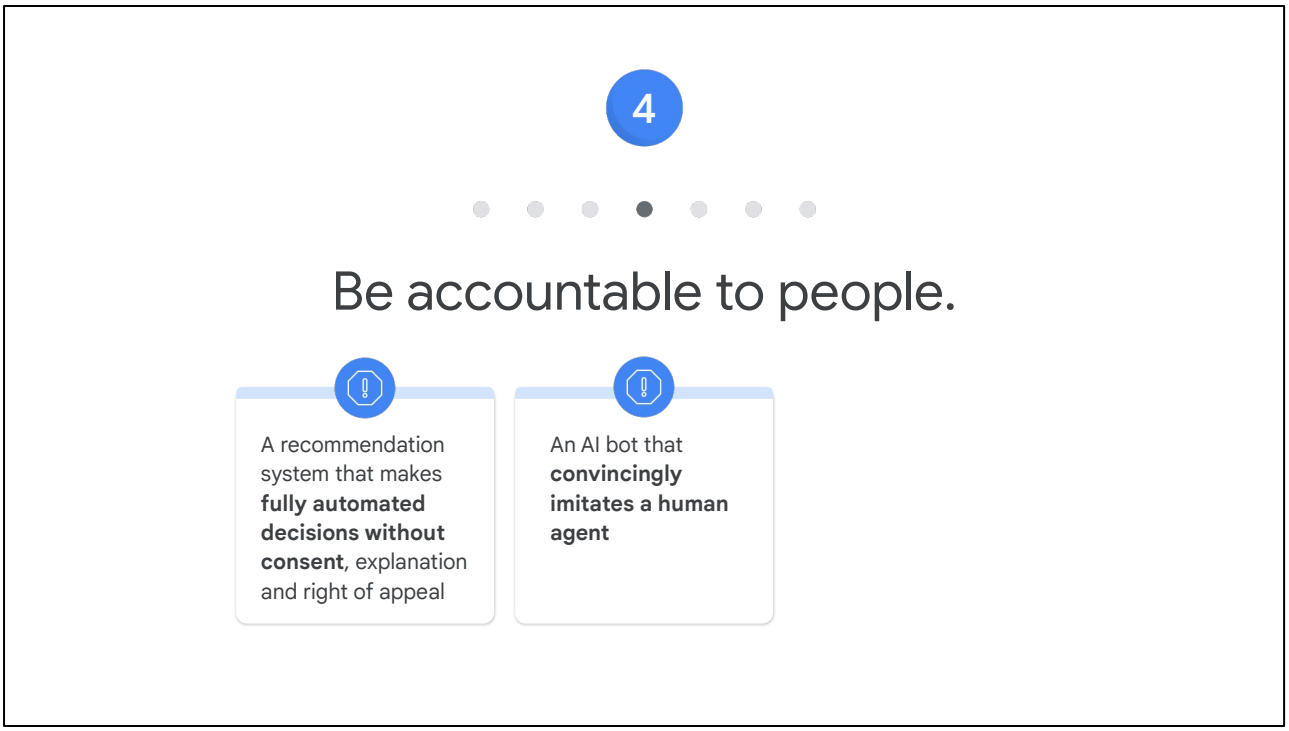
- Provide appropriate opportunities for feedback, relevant explanations, and appeal.
- Introduce appropriate human direction and control.



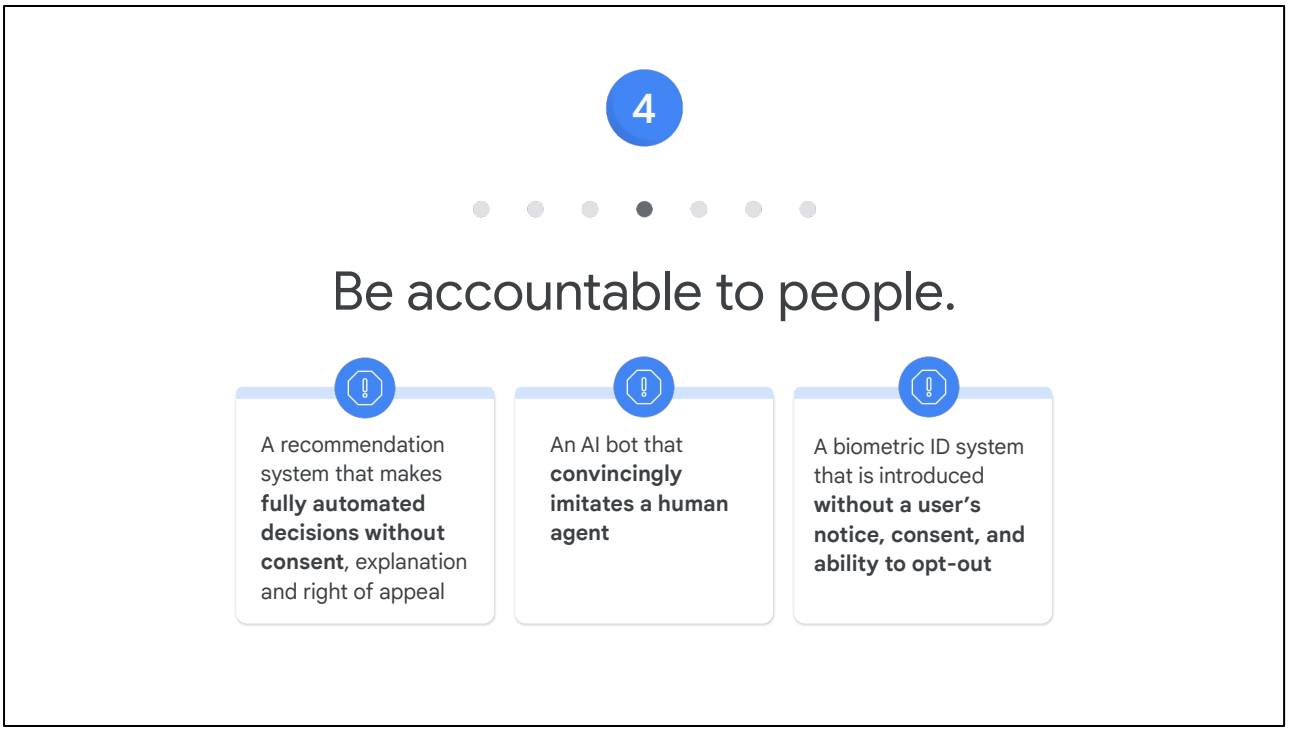
There have been examples of the fourth principle applications which have failed to be accountable to people. These include:



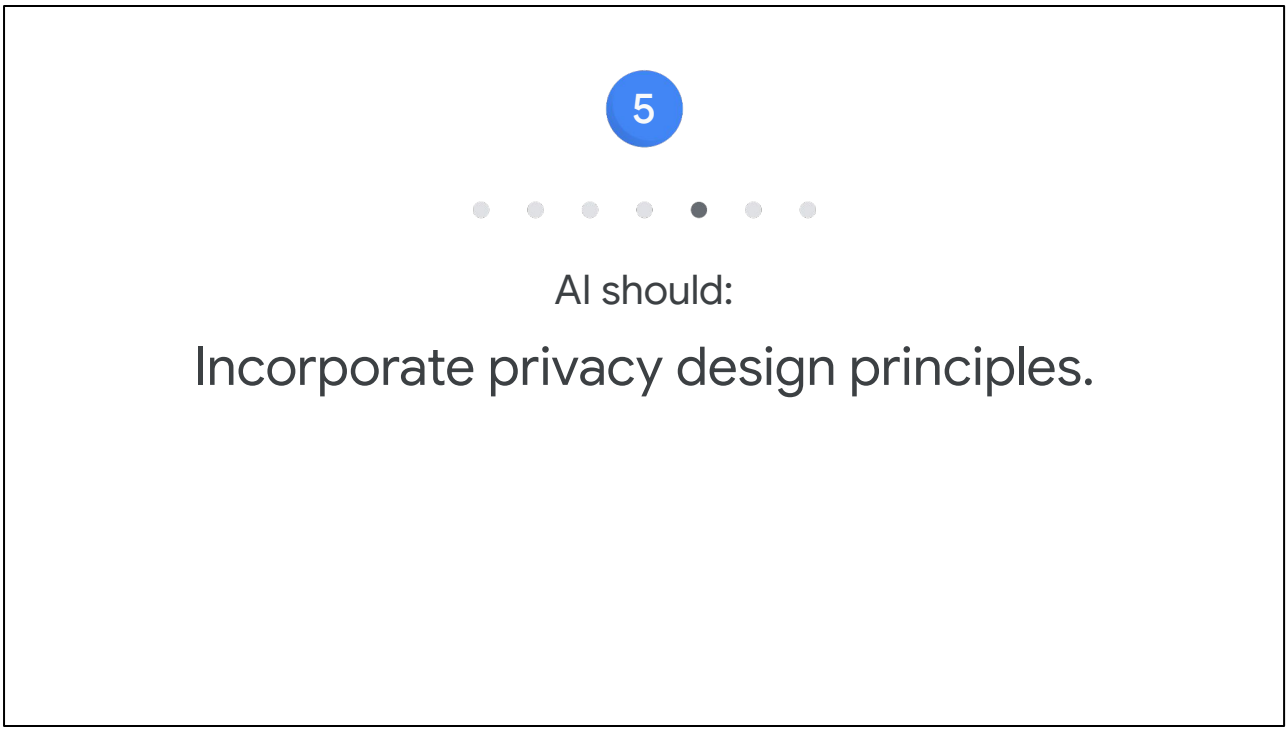
- A recommendation system that makes fully automated decisions without consent, explanation and right of appeal, such as credit and insurance decisions.



- An AI bot that convincingly imitates a human agent.

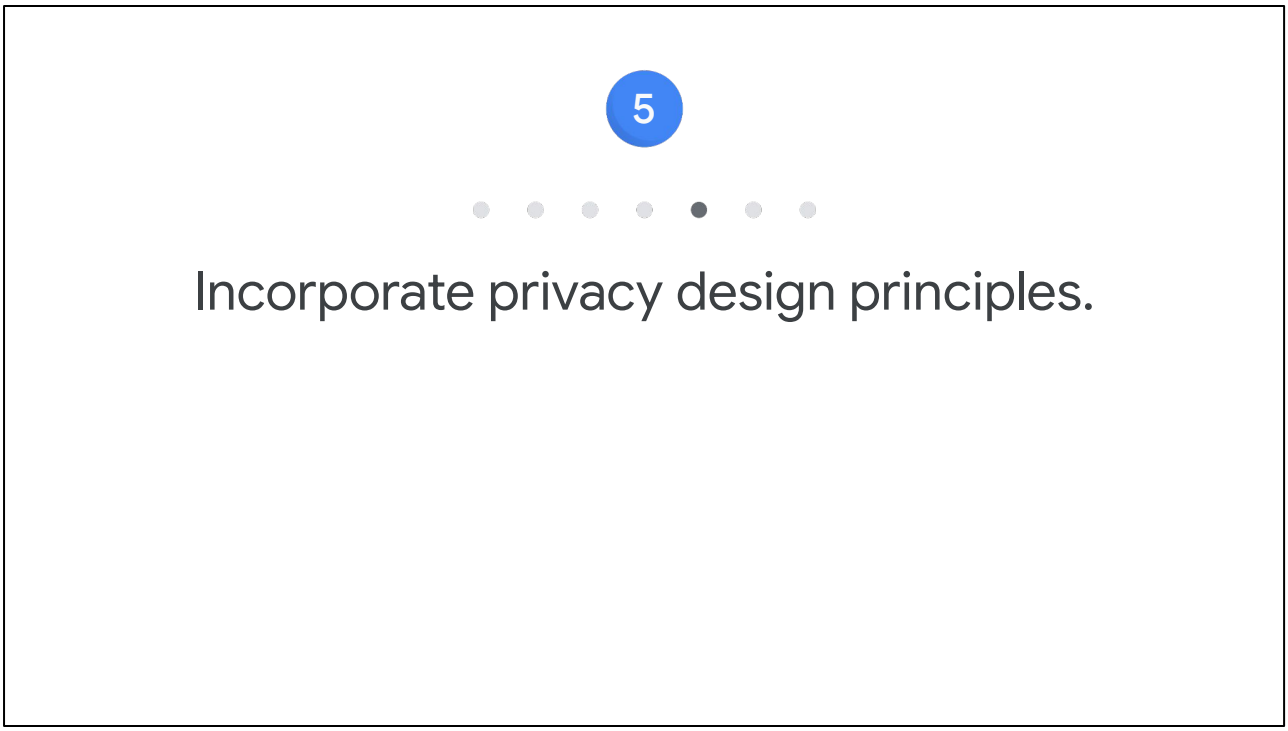


- A biometric ID system that is introduced without a user's notice, consent, and ability to opt-out.

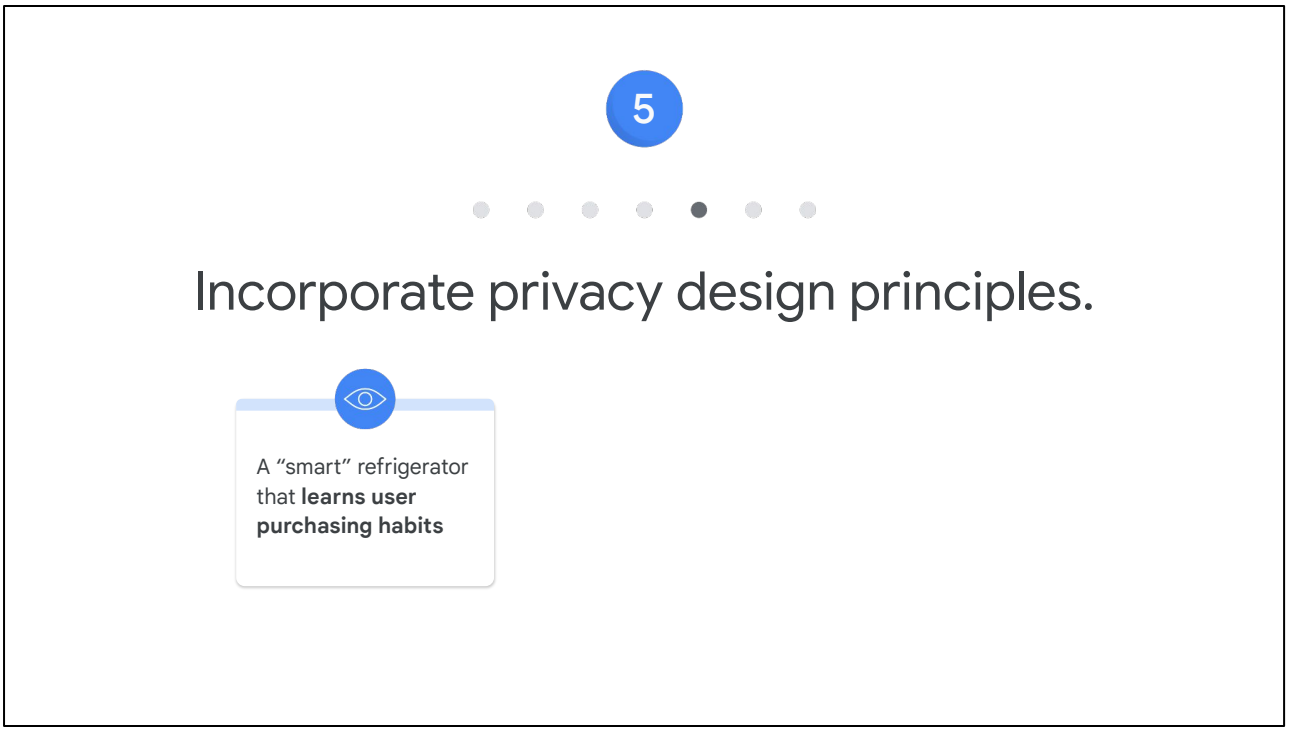


Google's fifth principle states: AI should incorporate privacy design principles. For any project, we should:

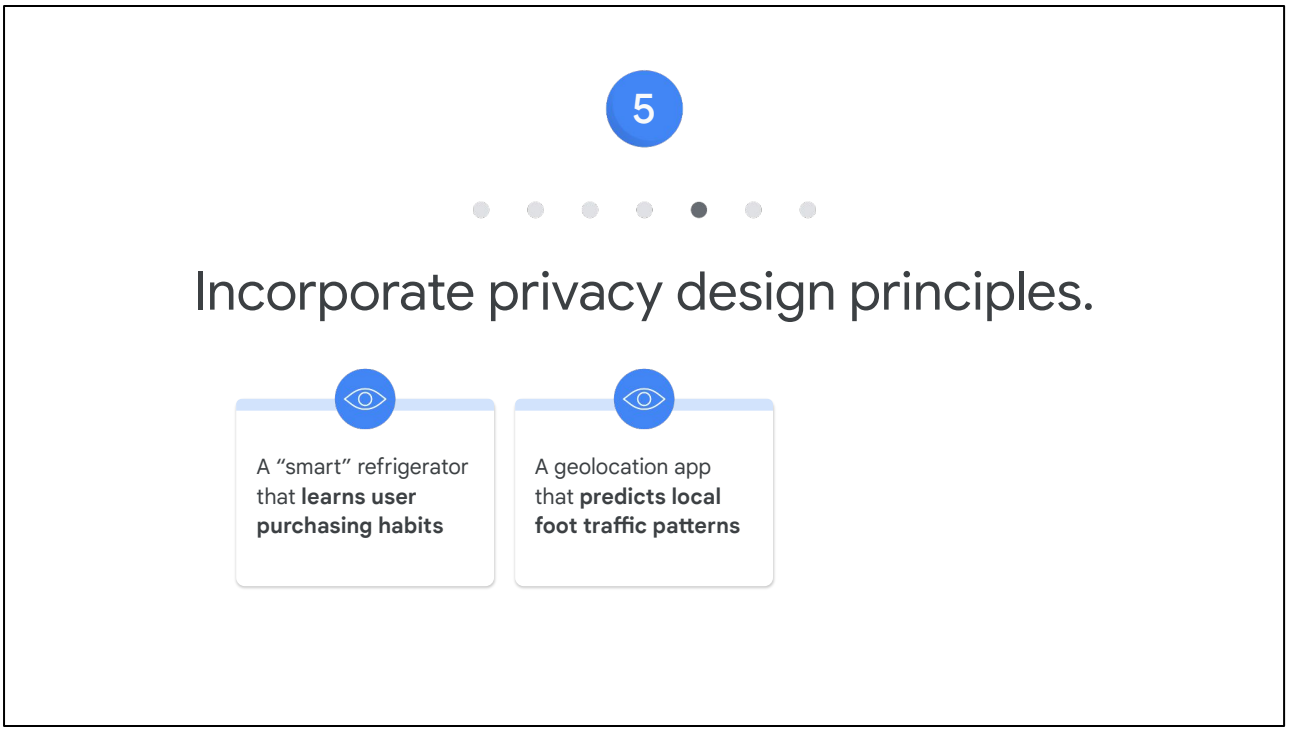
- Give opportunities for notice and consent.
- Encourage architectures that have privacy safeguards.
- Provide appropriate transparency and control over the use of data



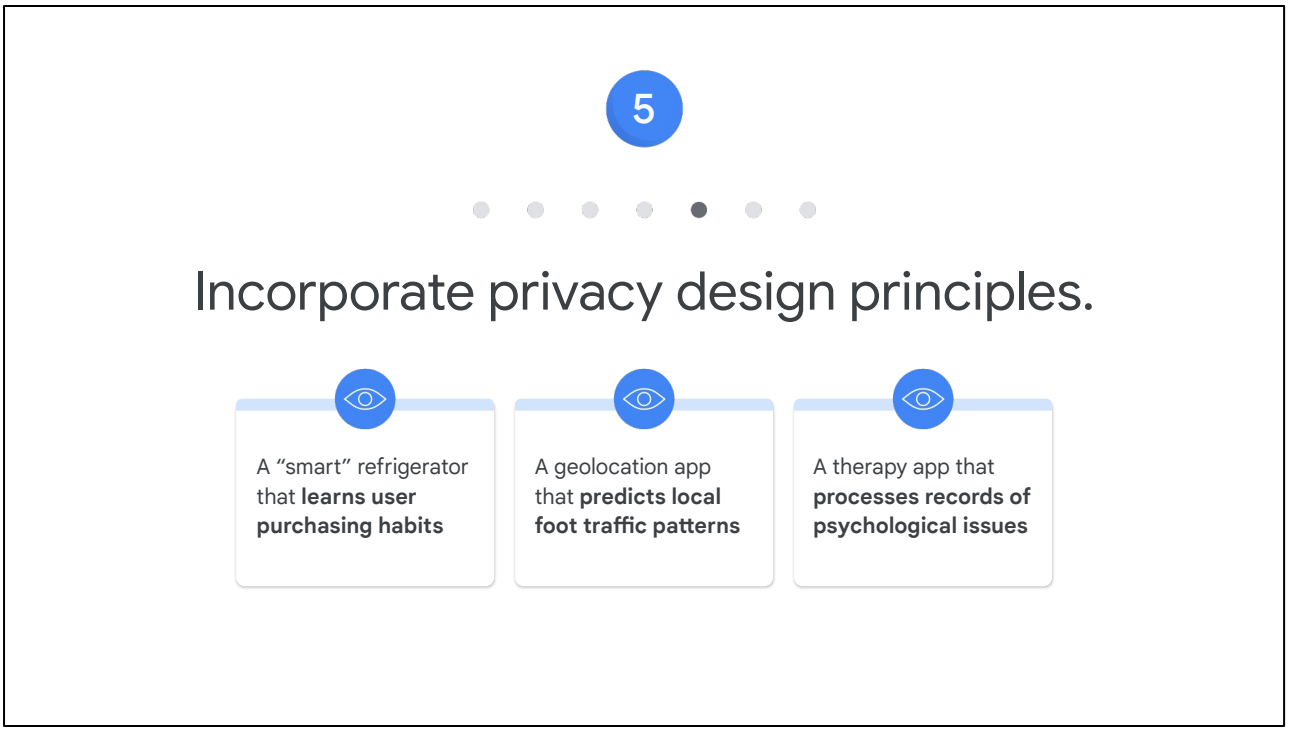
Here are some examples of fifth principle applications where incorporating privacy design is especially important.



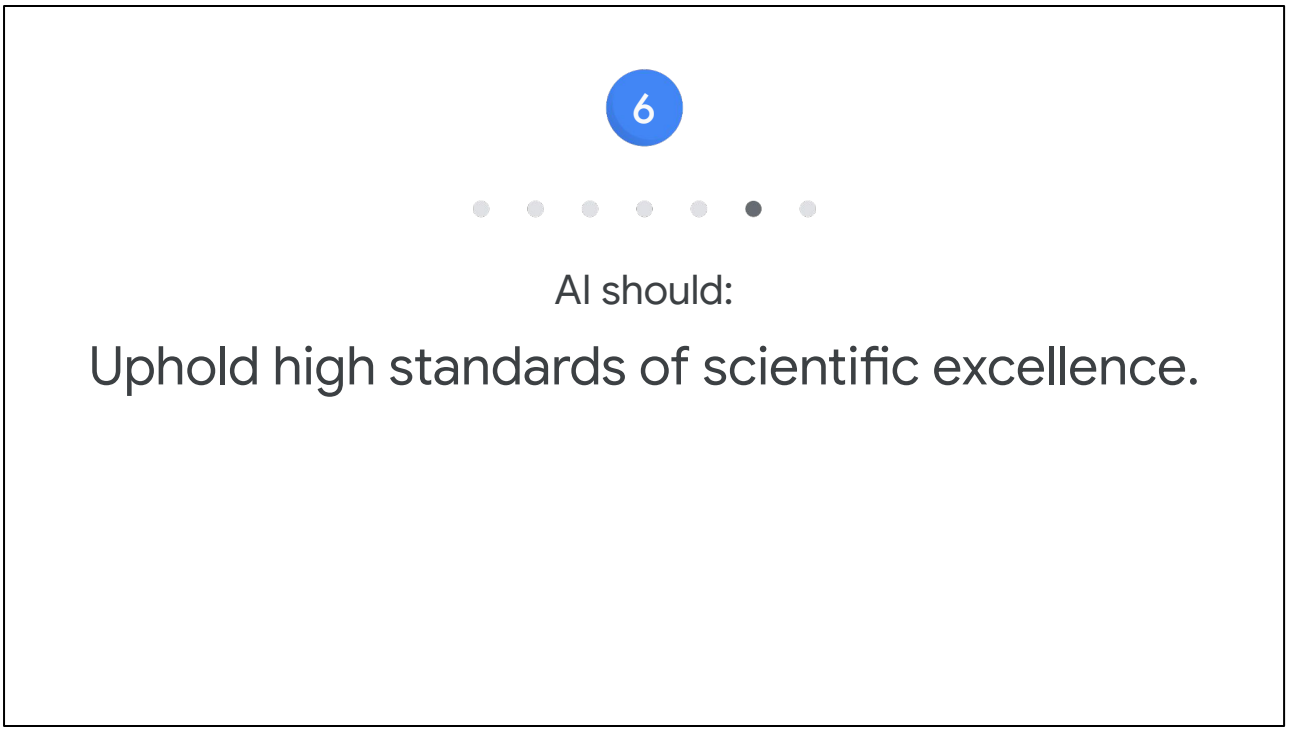
- A “smart” refrigerator that learns user purchasing habits.



- A geolocation app that predicts local foot traffic patterns.

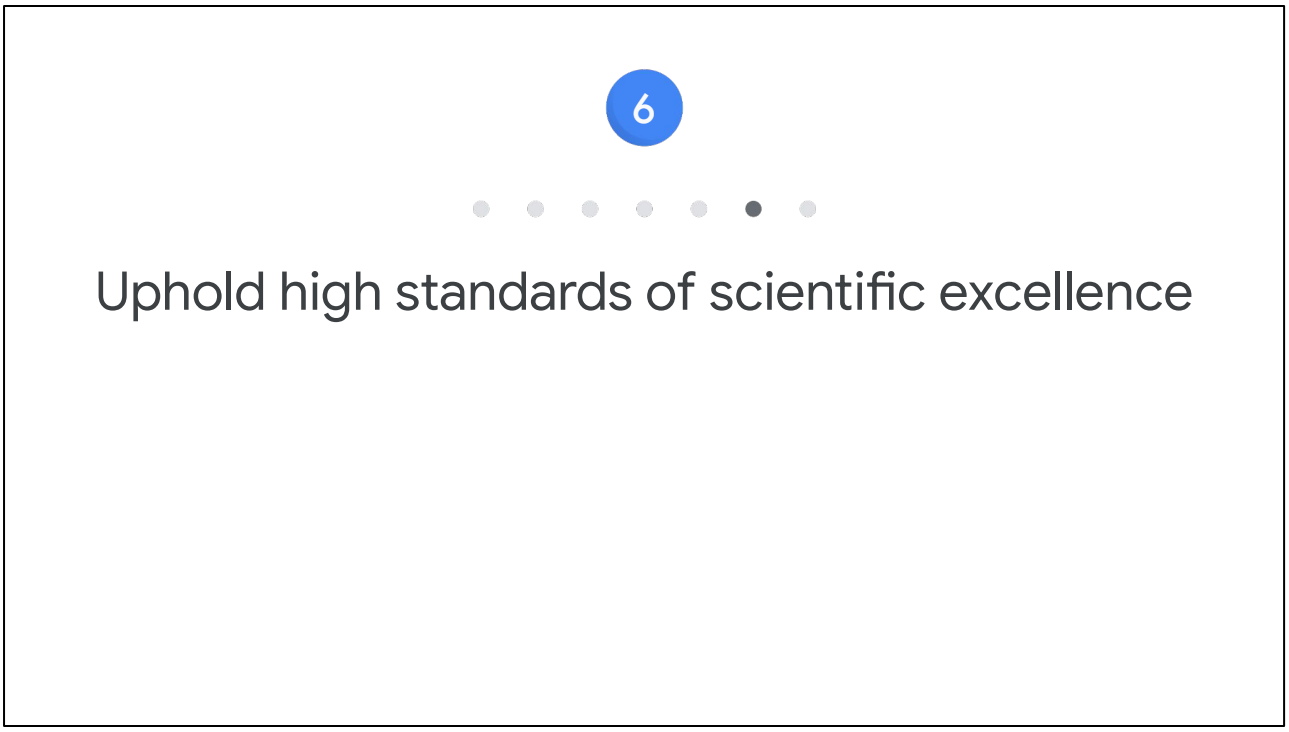


- A therapy app that processes records of psychological issues.

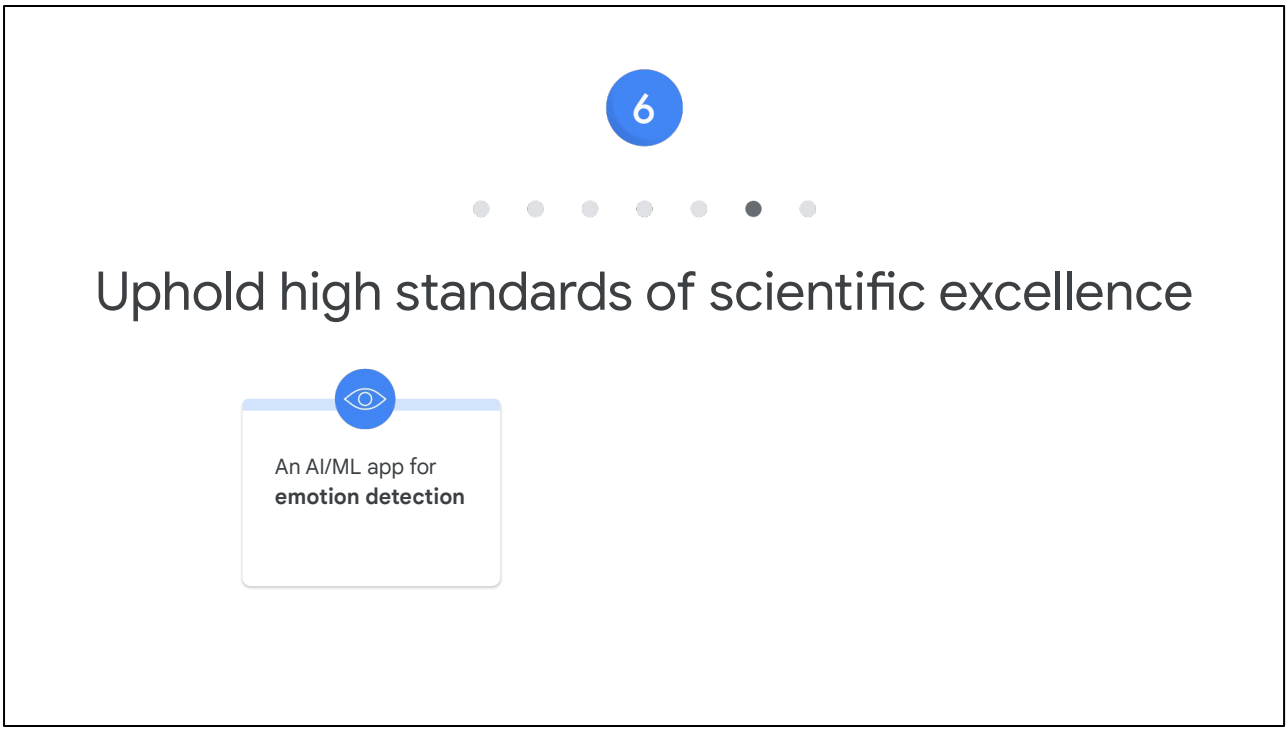


Google's sixth principle: AI should uphold high standards of scientific excellence. For any project, we should:

- Work with a range of stakeholders to promote thoughtful leadership in this area, drawing on scientifically rigorous and multidisciplinary approaches.
- Responsibly share AI knowledge by publishing educational materials, best practices, and research that enable more people to develop useful AI applications.



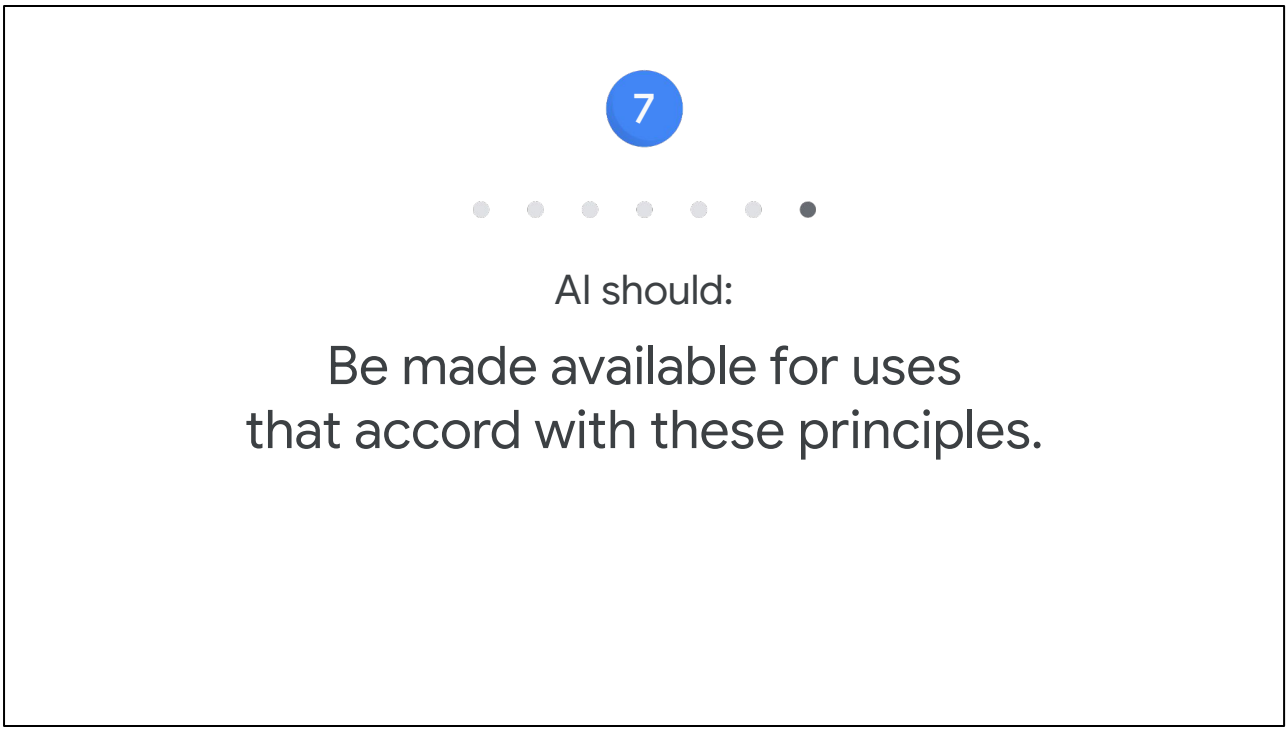
Here are some examples of sixth principle applications that demonstrate the importance to uphold high standards of scientific excellence.



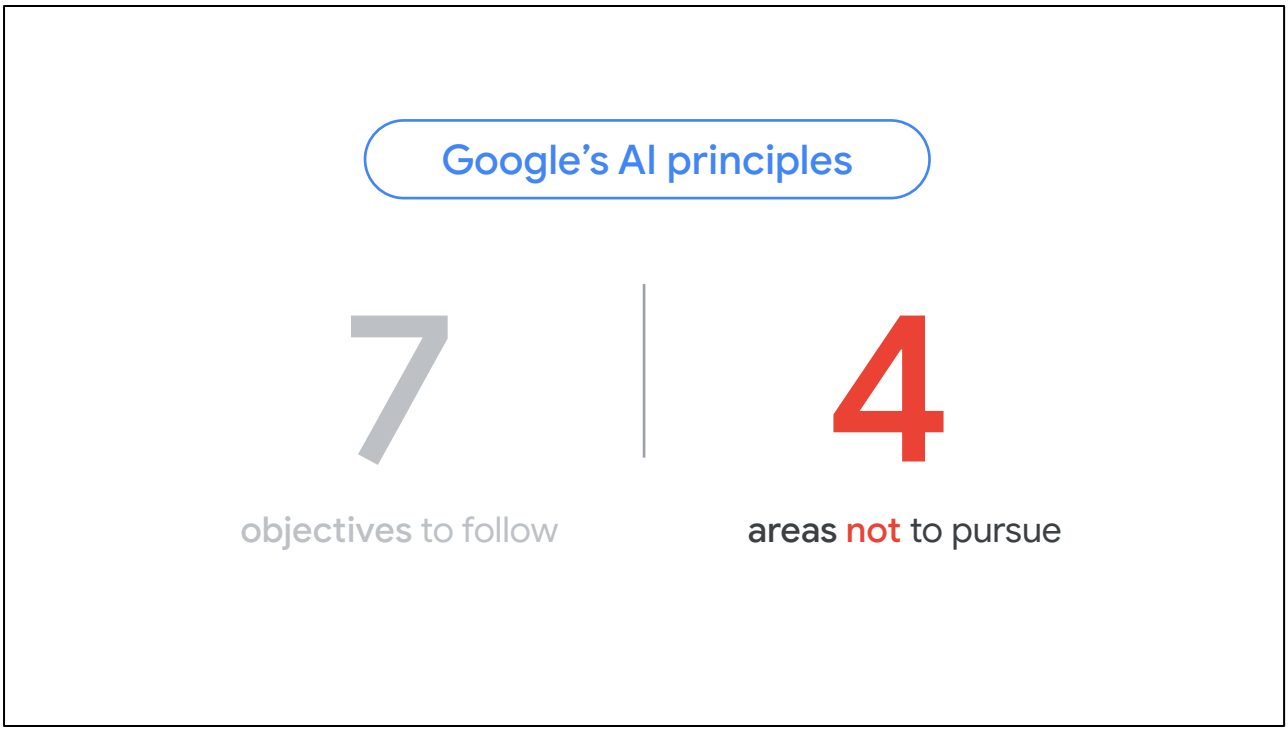
- An AI/ML app for emotion detection.

- An AI/ML app that detects signs of clinical depression.

- An AI/ML tool that advances deepfake detection.



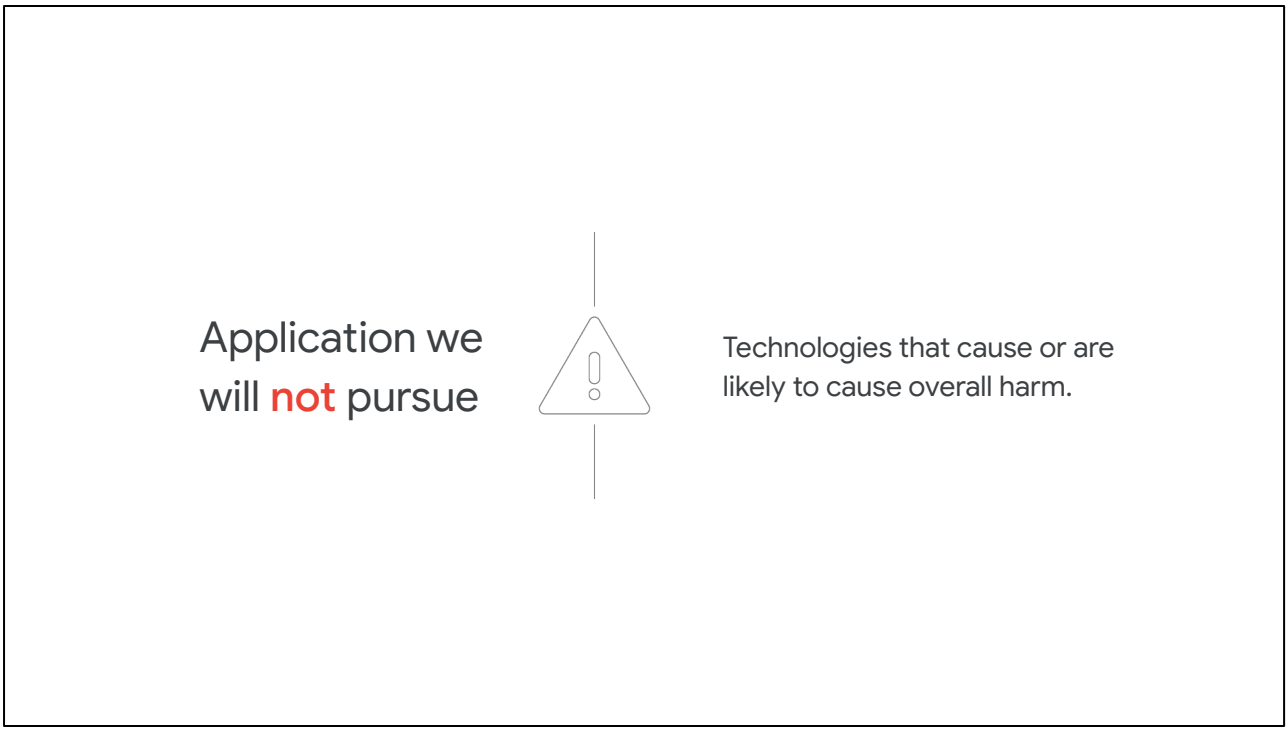
Google's seventh and last principle states: AI should be made available for uses that accord with these principles. Many technologies have multiple uses, so we will work to limit potentially harmful or abusive applications.



In addition to these seven principles, Google also has a commitment as part of its AI Principles to NOT design or deploy AI in four application areas.

Application we will not pursue

The first area is:



technologies that cause or are likely to cause overall harm. Where there is a material risk of harm, Google will proceed only where they believe that the benefits substantially outweigh the risks, and will incorporate appropriate safety constraints.

Application we will not pursue

The second area which Google will not pursue is:

Application we will not pursue

Weapons or other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people.

Weapons or other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people.

Application we will not pursue

The third area which Google will not pursue is:

Application we will not pursue

Technologies that gather or use information for surveillance that violates internationally accepted norms.

Technologies that gather or use information for surveillance that violates internationally accepted norms.

Application we will not pursue

The fourth and last area which Google will not pursue is:

Application we will **not** pursue

Technologies where the purpose contravenes widely accepted principles of international law and human rights.

Technologies where the purpose contravenes widely accepted principles of international law and human rights.

01 AI & Responsibility

02 Google's AI principles

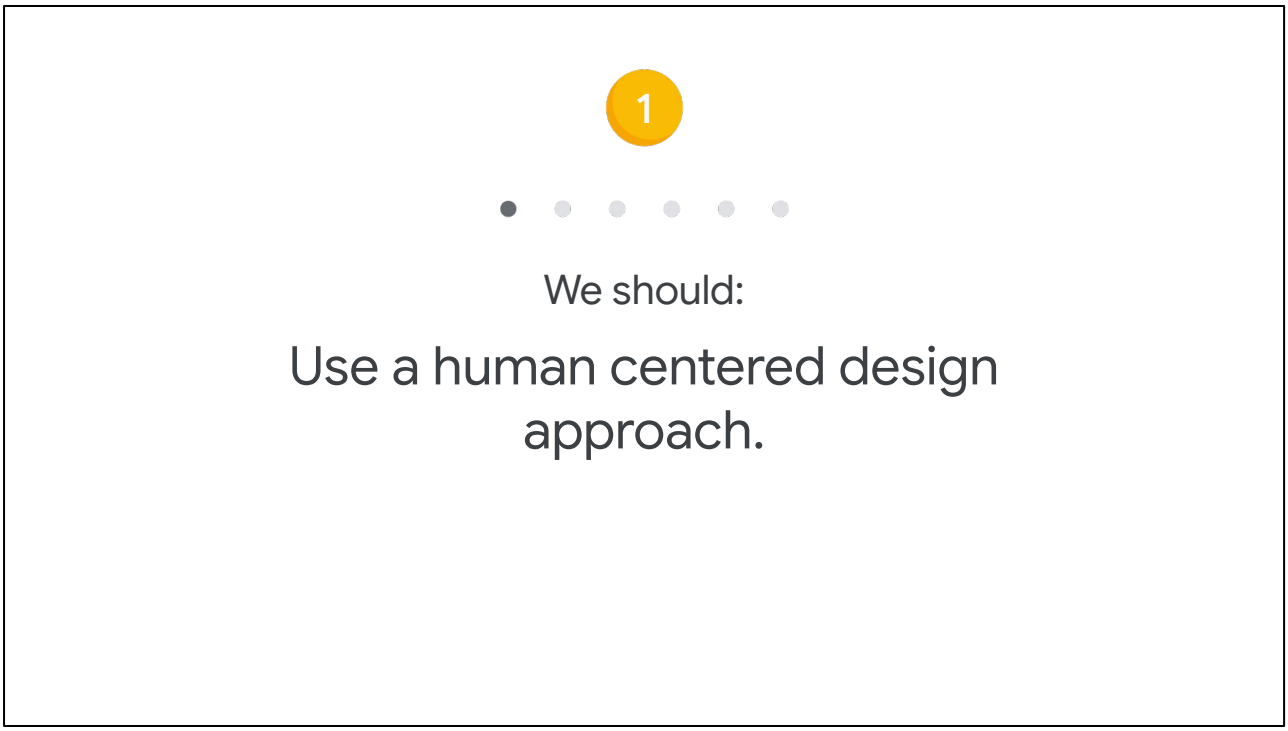
03 Responsible AI practices

04 Case study: Google Flights

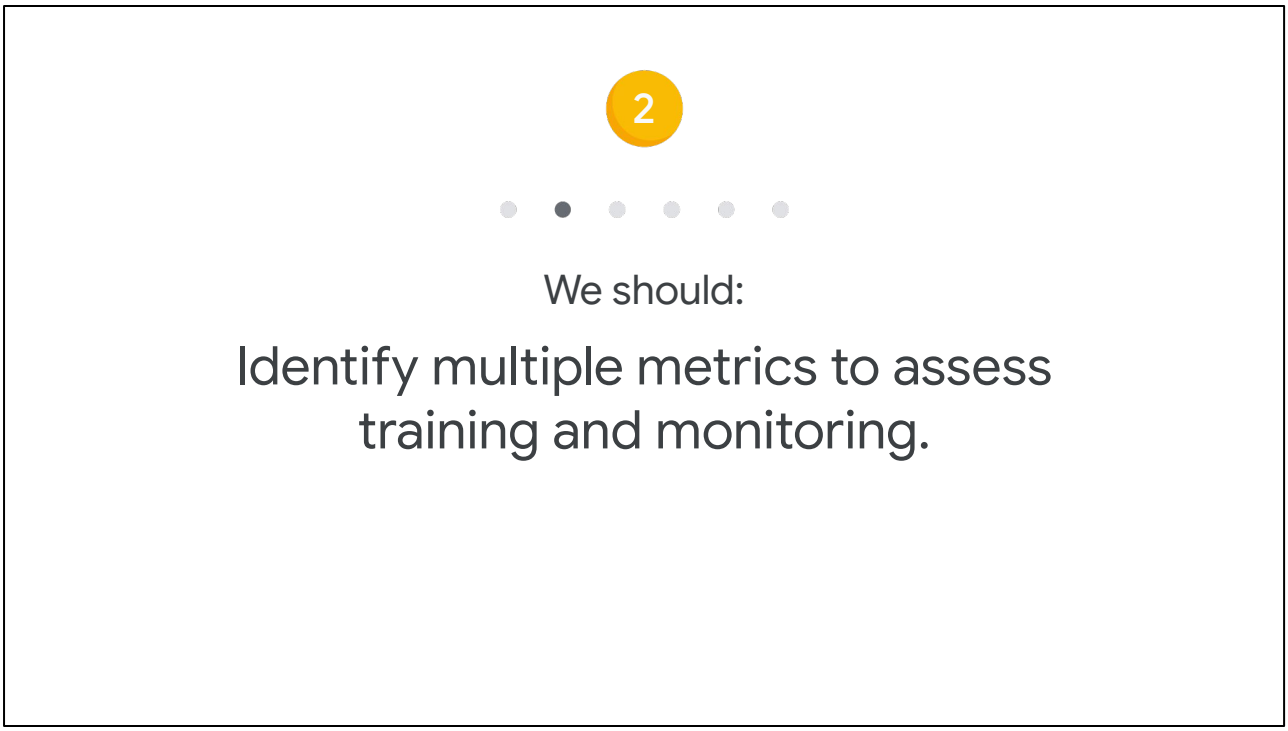
Now, let's look at recommended practices for Responsible AI.

AI is software, and so general best practices for software systems should always be followed when designing AI systems.
There are also various considerations unique to machine learning for us to discover.

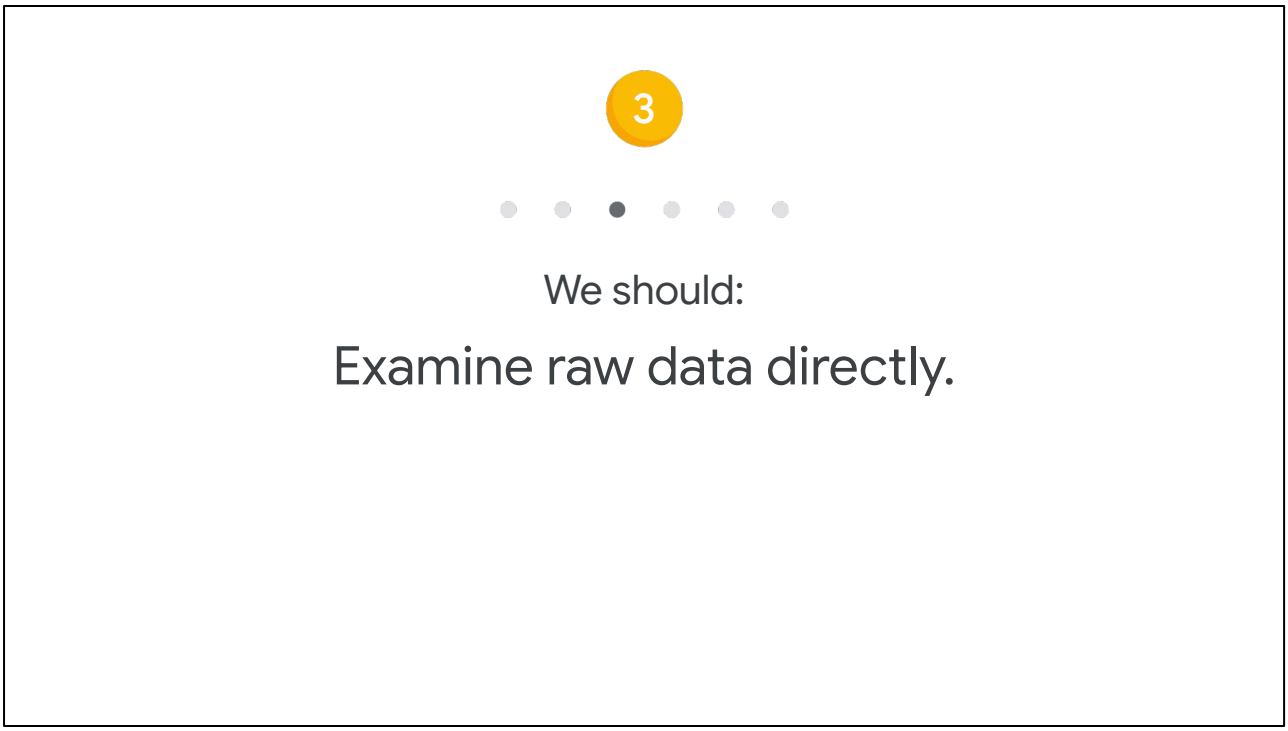
Here are 6 recommended practices for developing AI with responsible AI principles in mind.



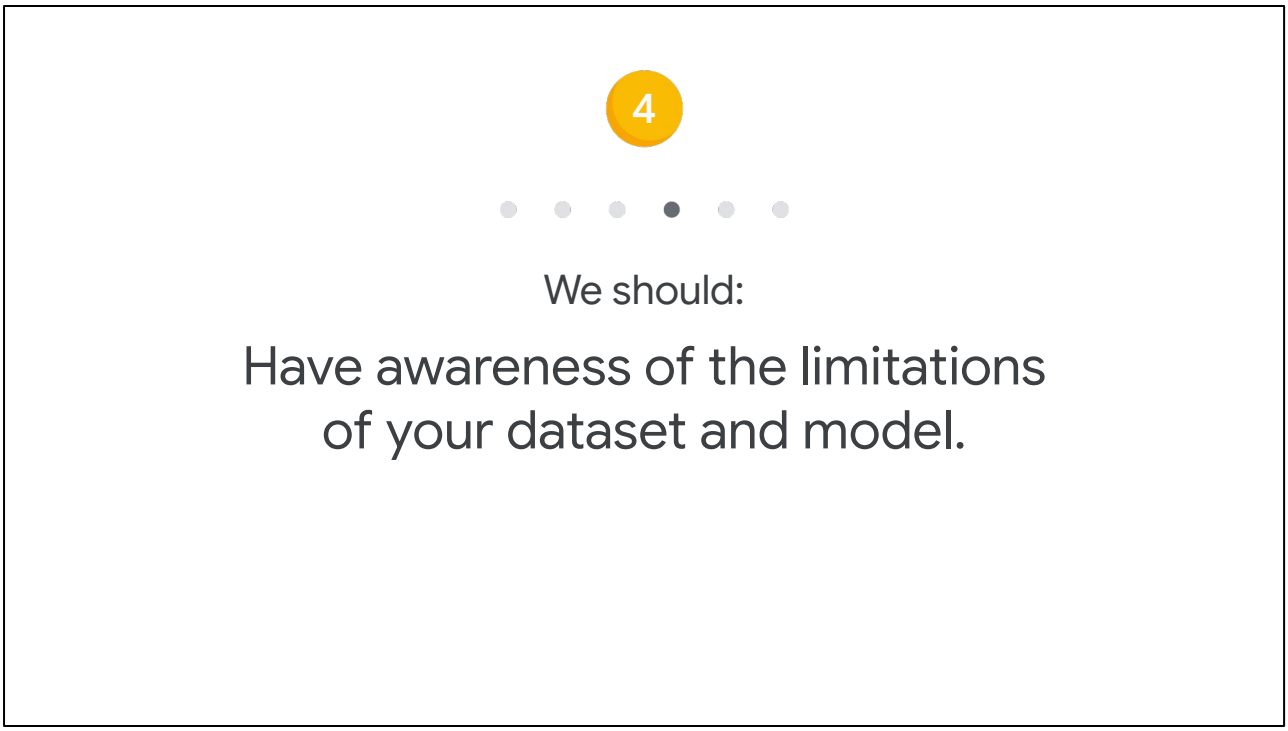
- Use a human centered design approach.



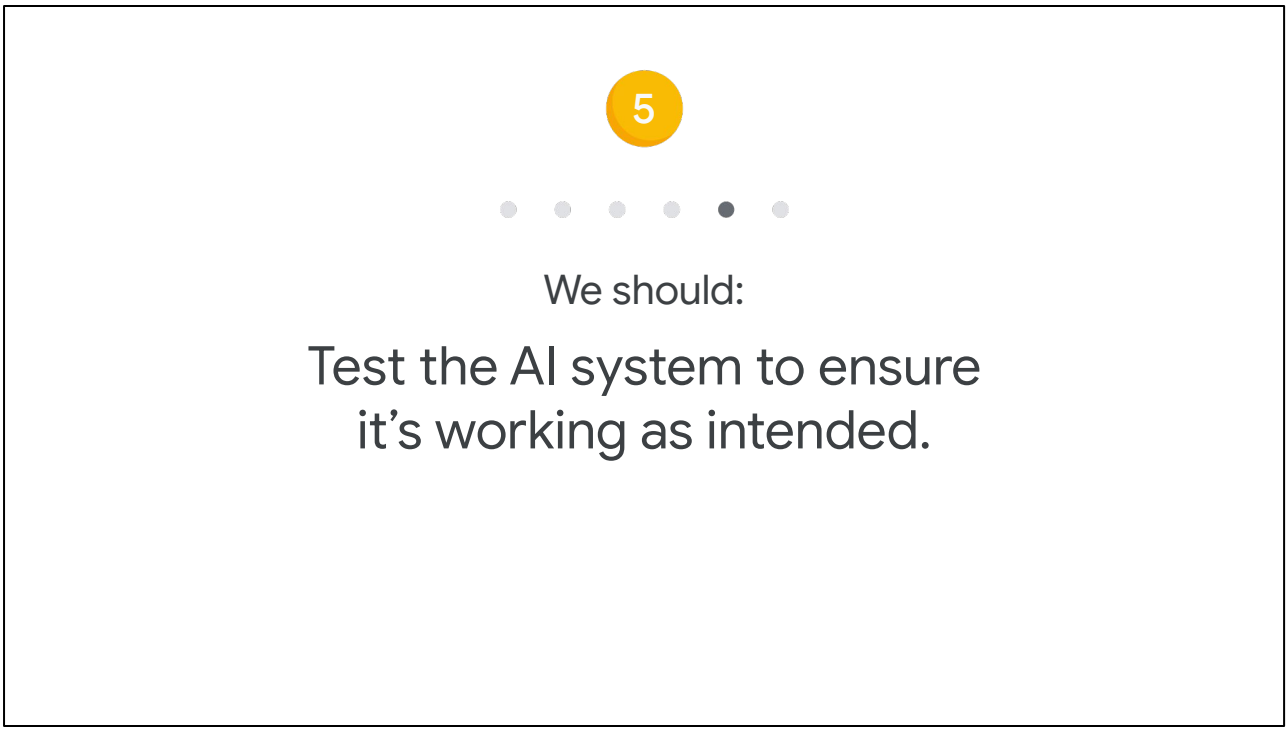
- Identify multiple metrics to assess training and monitoring.



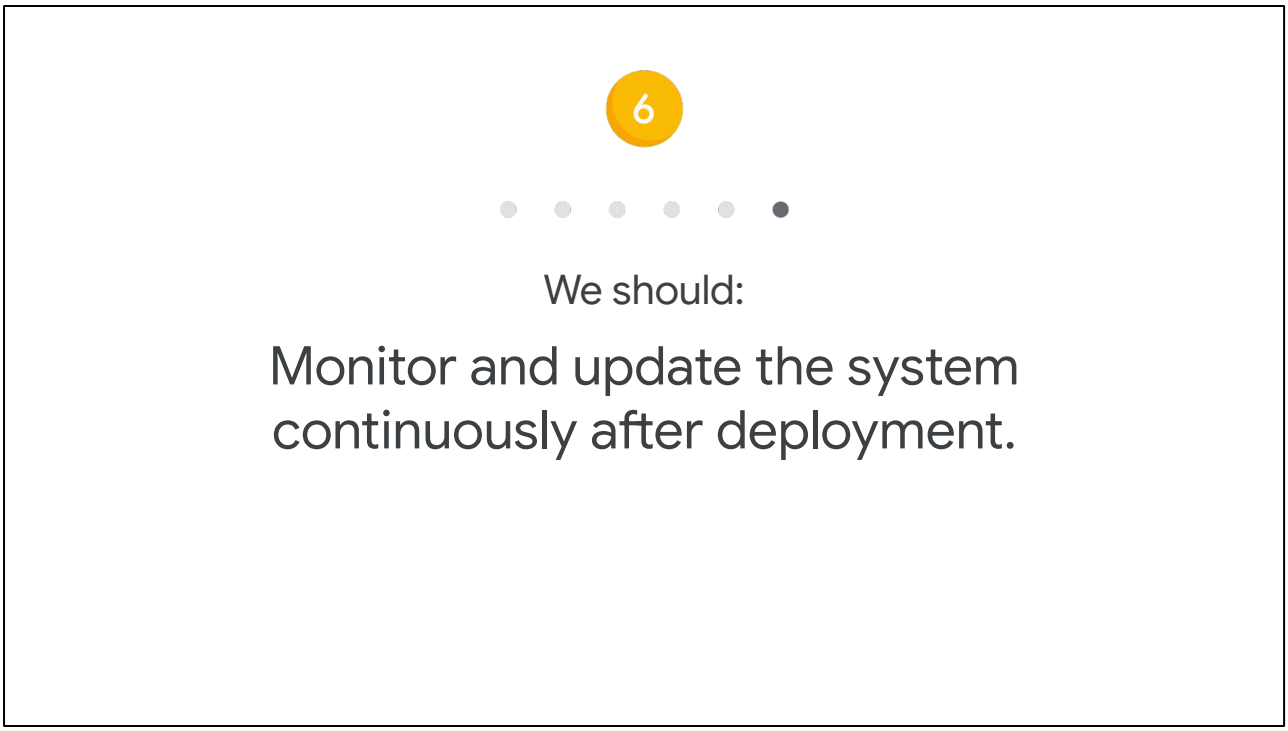
- When possible, examine raw data directly.



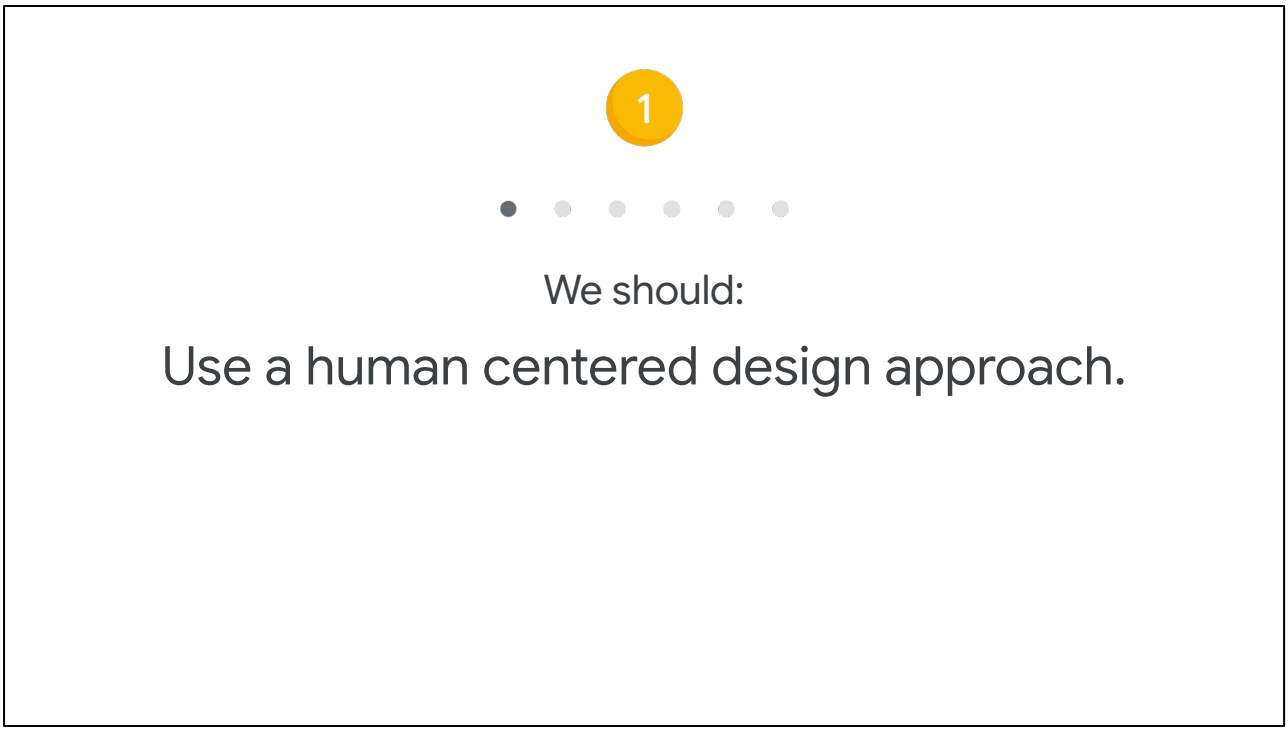
- Have awareness of the limitations of your dataset and model.



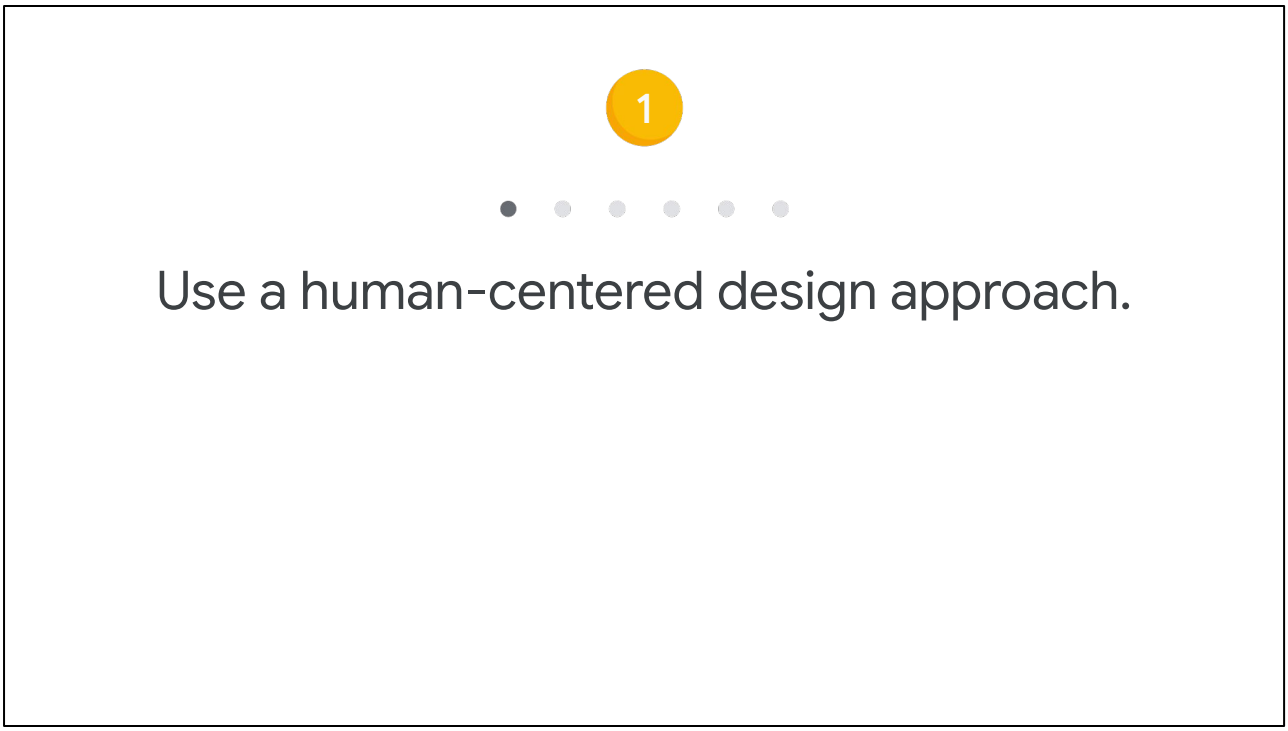
- Test the AI system to ensure it’s working as intended.



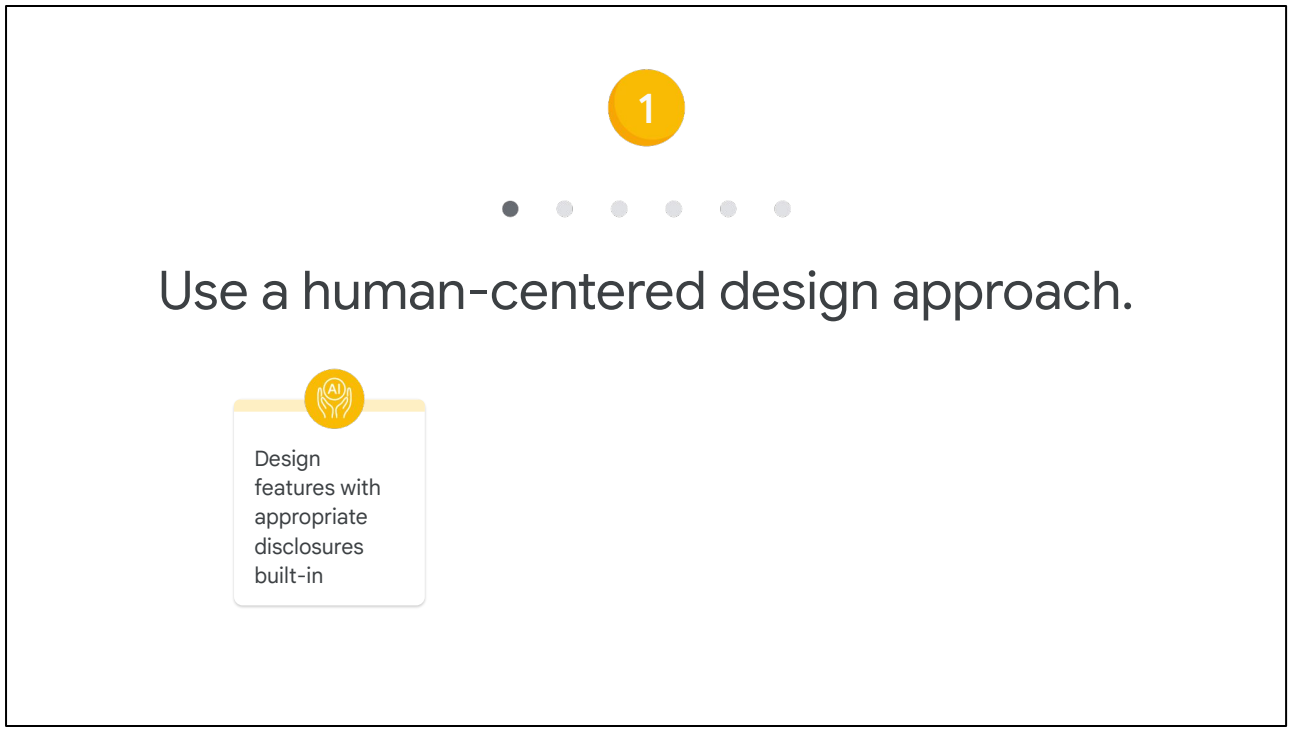
- Monitor and update the system continuously after deployment.



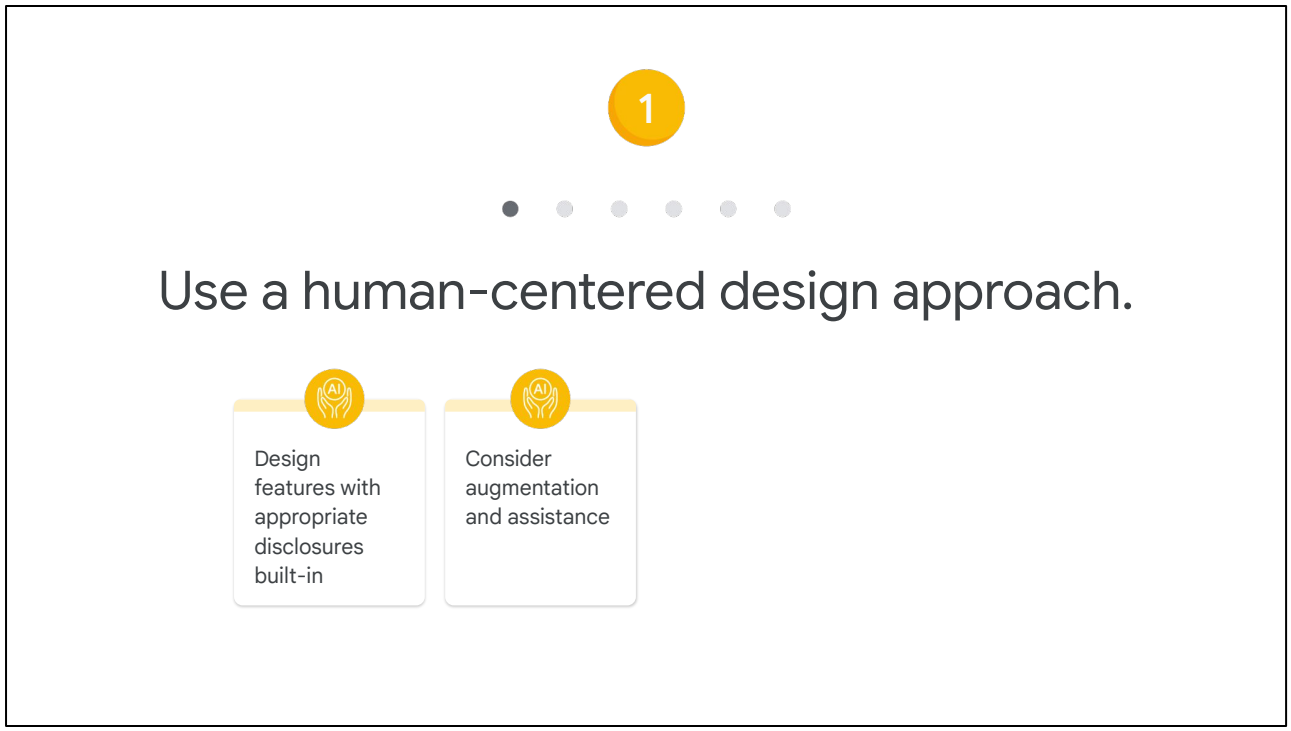
Let’s discuss the first recommended practice: Use a human-centered design approach.



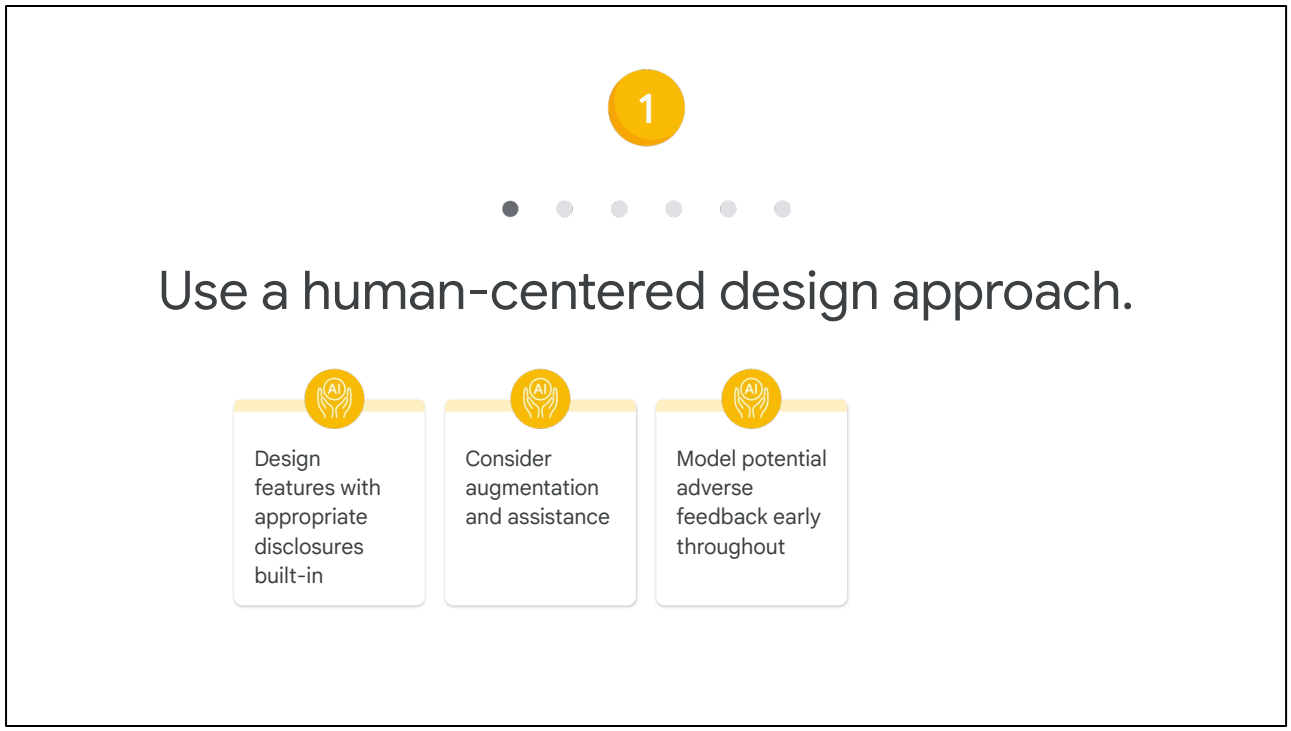
The way that actual users experience your system is essential to assessing the true impact of its predictions, recommendations, and decisions.
To use a human-centered design approach, you should:



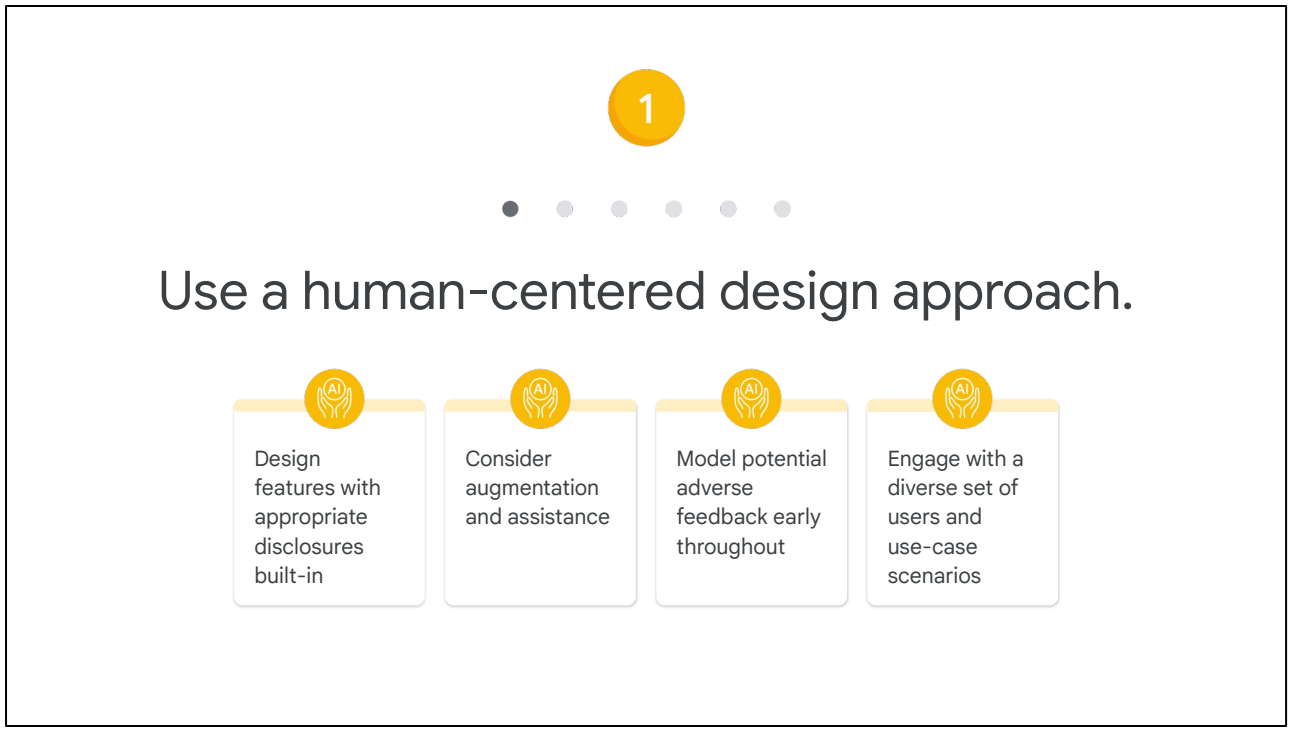
Design features with appropriate disclosures built-in: clarity and control are crucial to a good user experience.



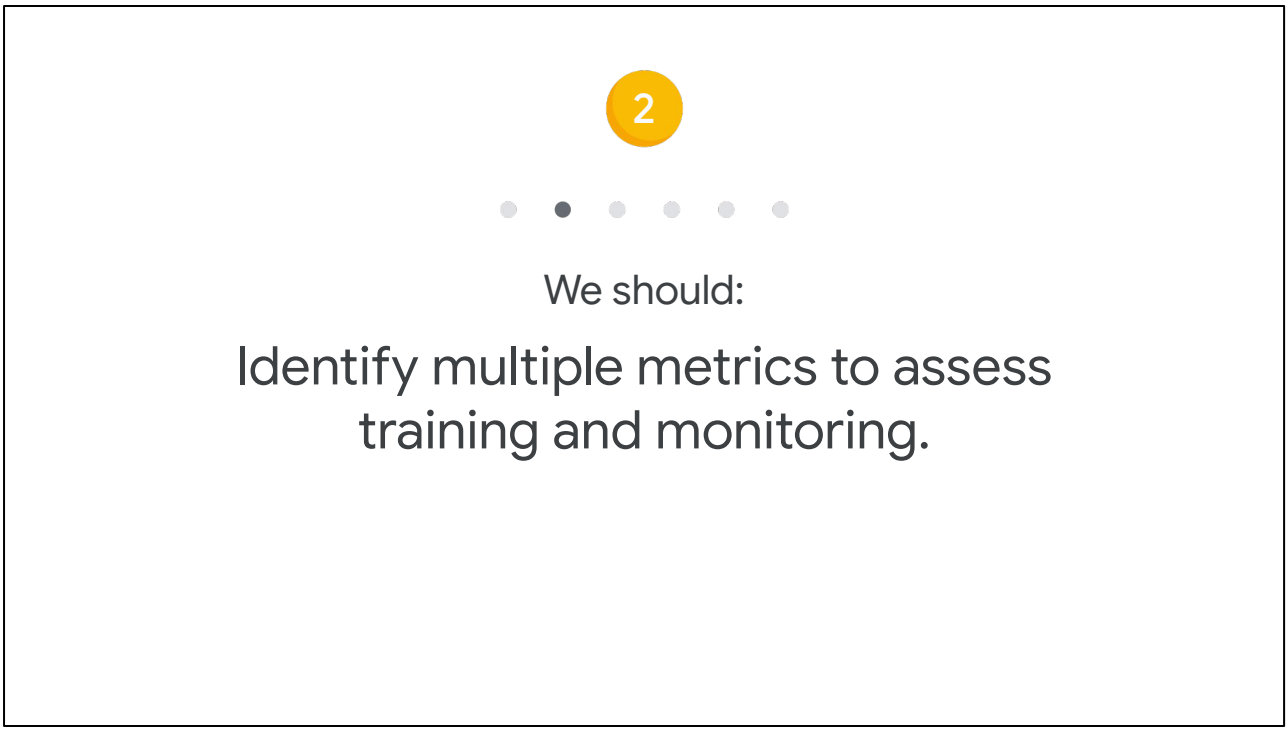
Consider augmentation and assistance. In some cases, it might be optimal for your system to suggest a few options rather than one to the user. Technically, it's much more difficult to achieve good precision at one answer versus precision at a few answers.



Model potential adverse feedback early in the design process, followed by specific live testing and iteration for a small fraction of traffic before full deployment.

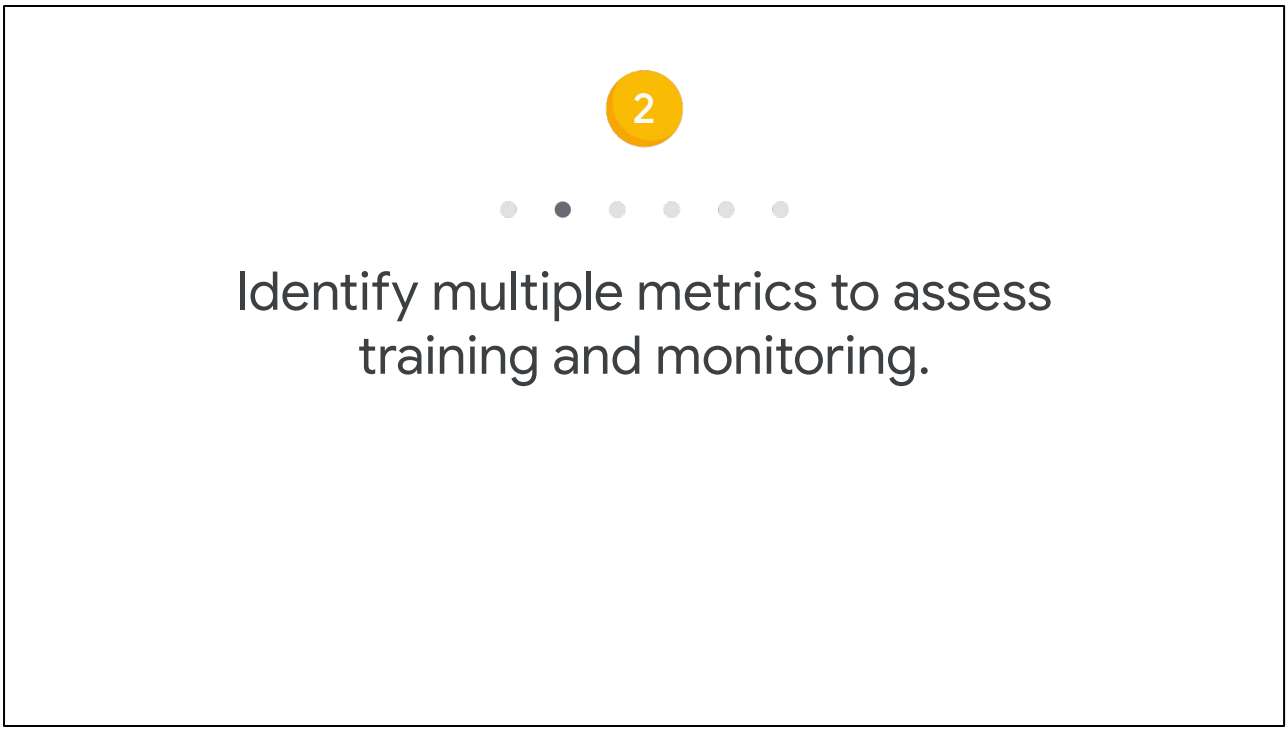


Engage with a diverse set of users and use-case scenarios, and incorporate feedback before and throughout project development. This builds a rich variety of user perspectives into the project and increases the number of people who benefit from the technology.

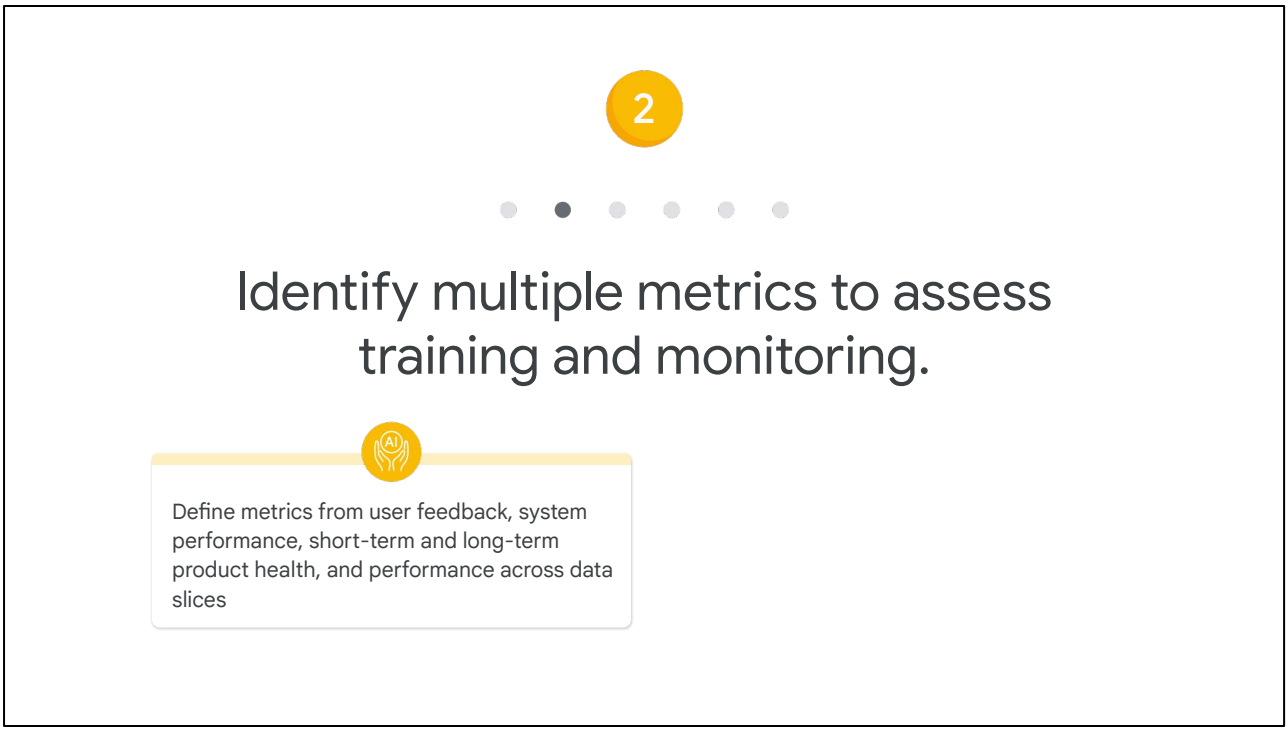


The next practice is to identify multiple metrics to assess training and monitoring.

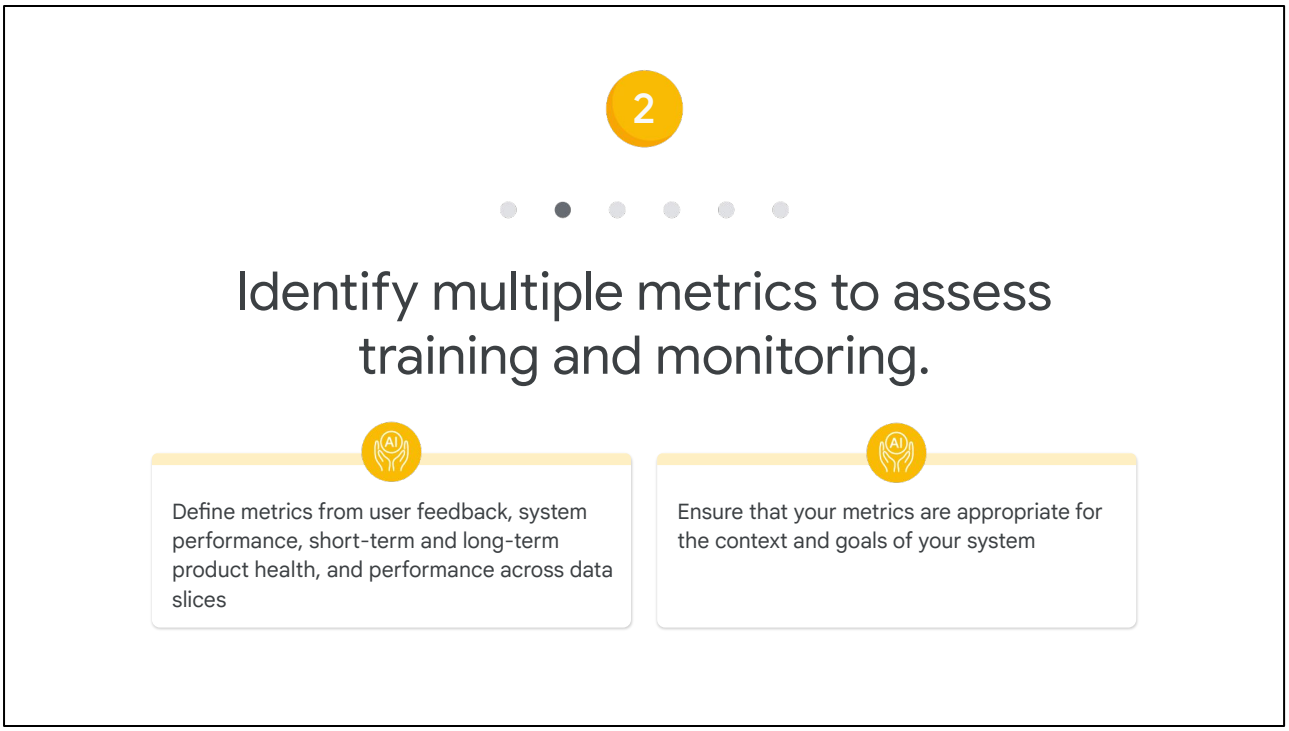
The use of several metrics instead of a single one will help you to understand tradeoffs between different kinds of errors and experiences.



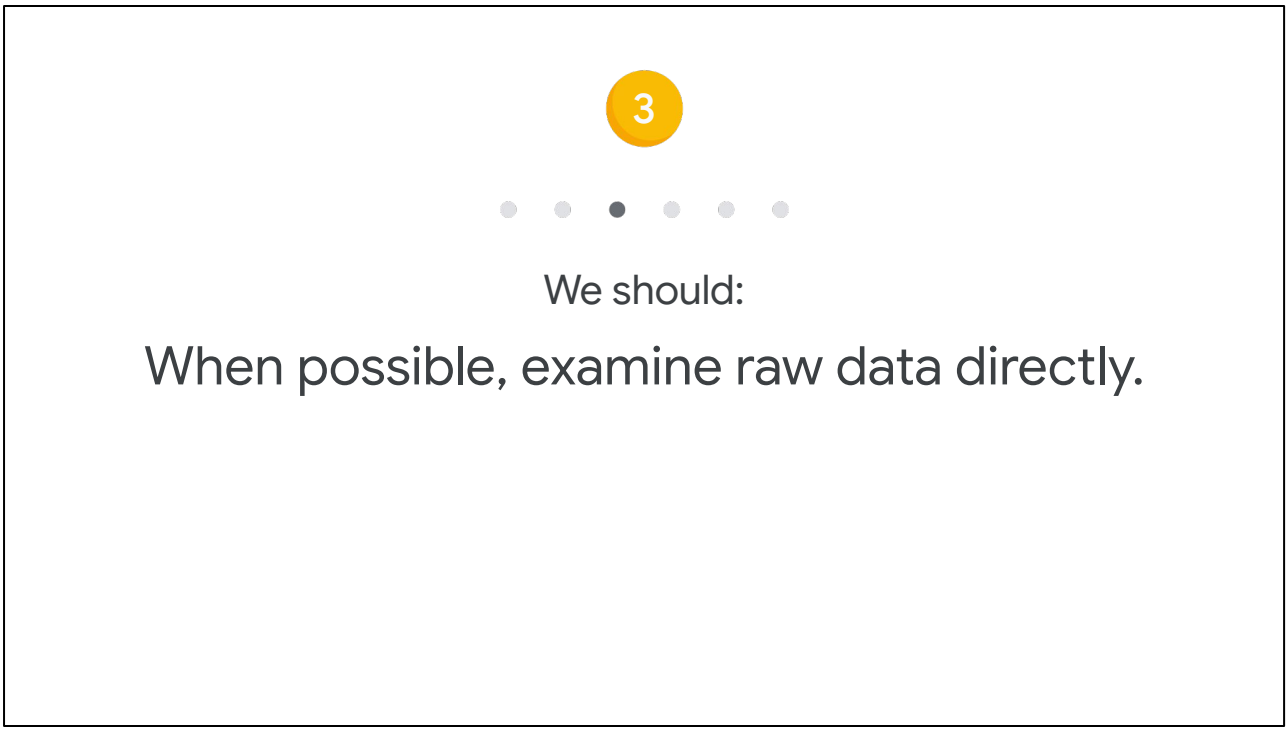
This means you should:



Define metrics including feedback from user surveys, quantities that track overall system performance, and short and long-term product health (for example, click-through rate and customer lifetime value, respectively), and performance sliced across different subgroups.

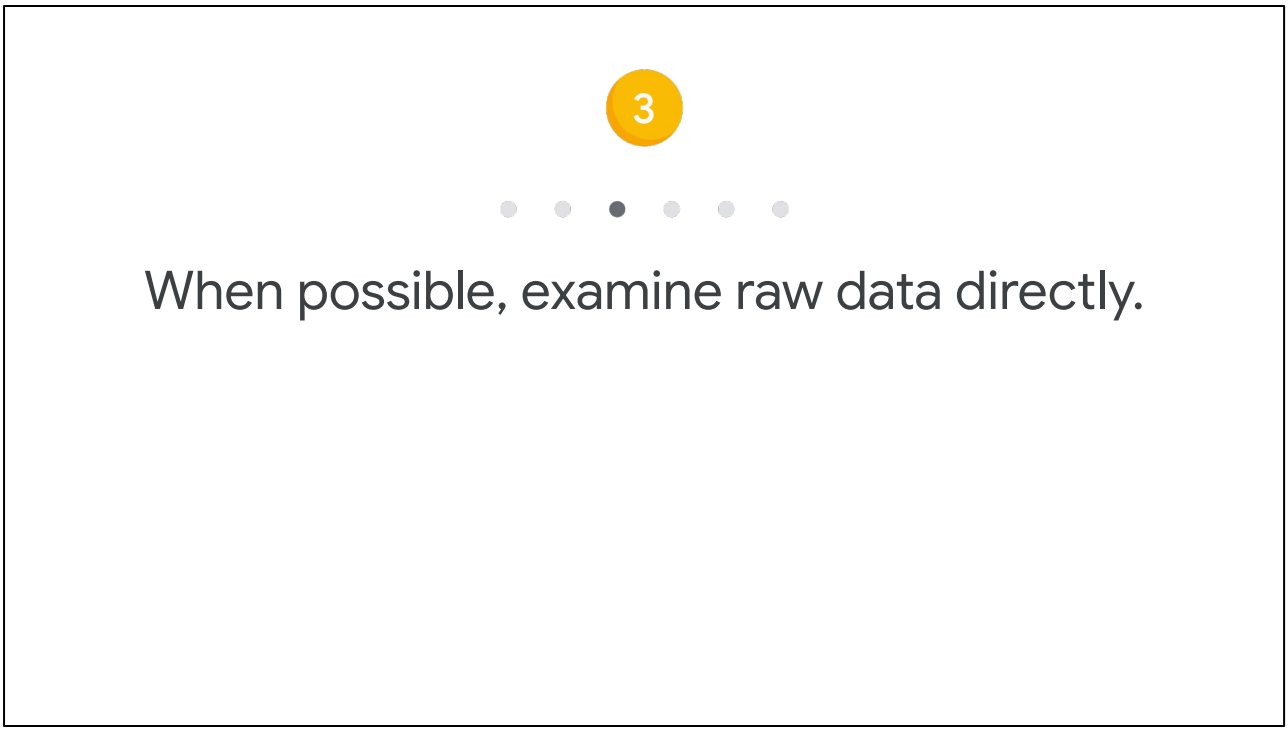


Ensure that your metrics are appropriate for the context and goals of your system. For example, a fire alarm system should have high recall, even if that means the occasional false alarm.

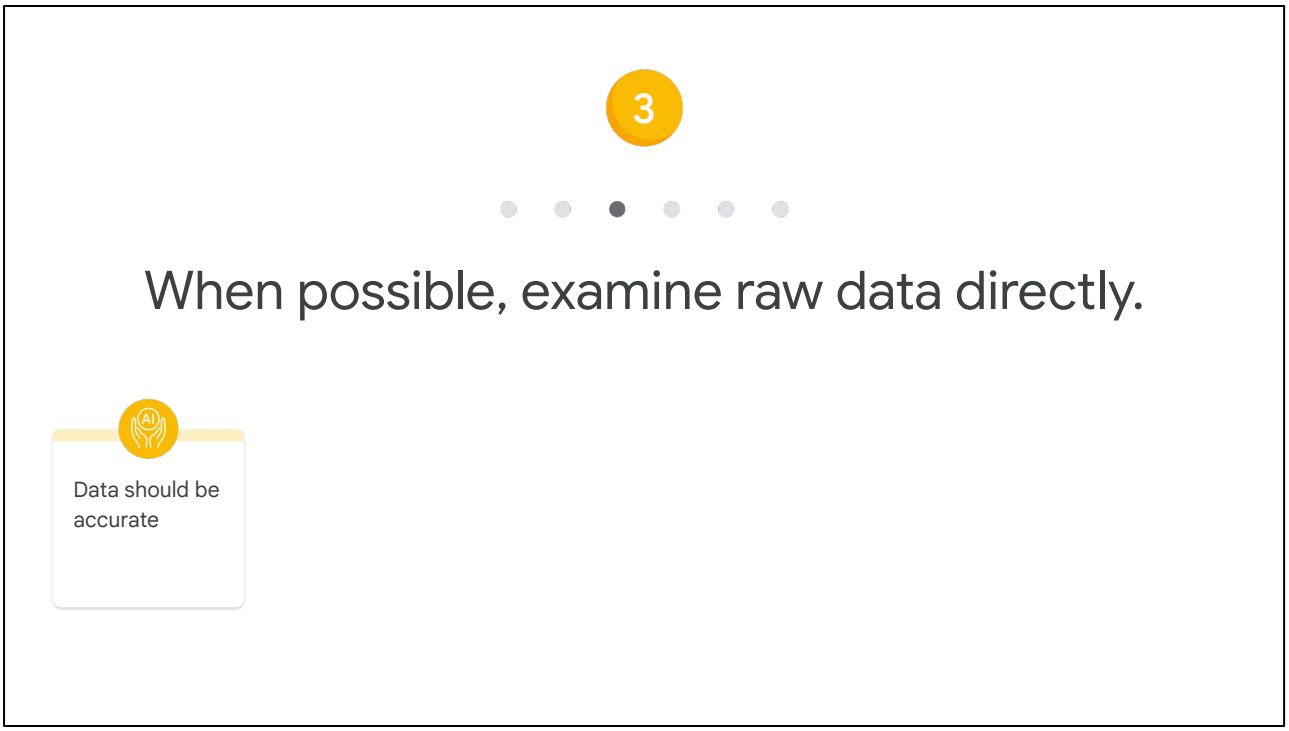


Next: the practice of directly examining your raw data.

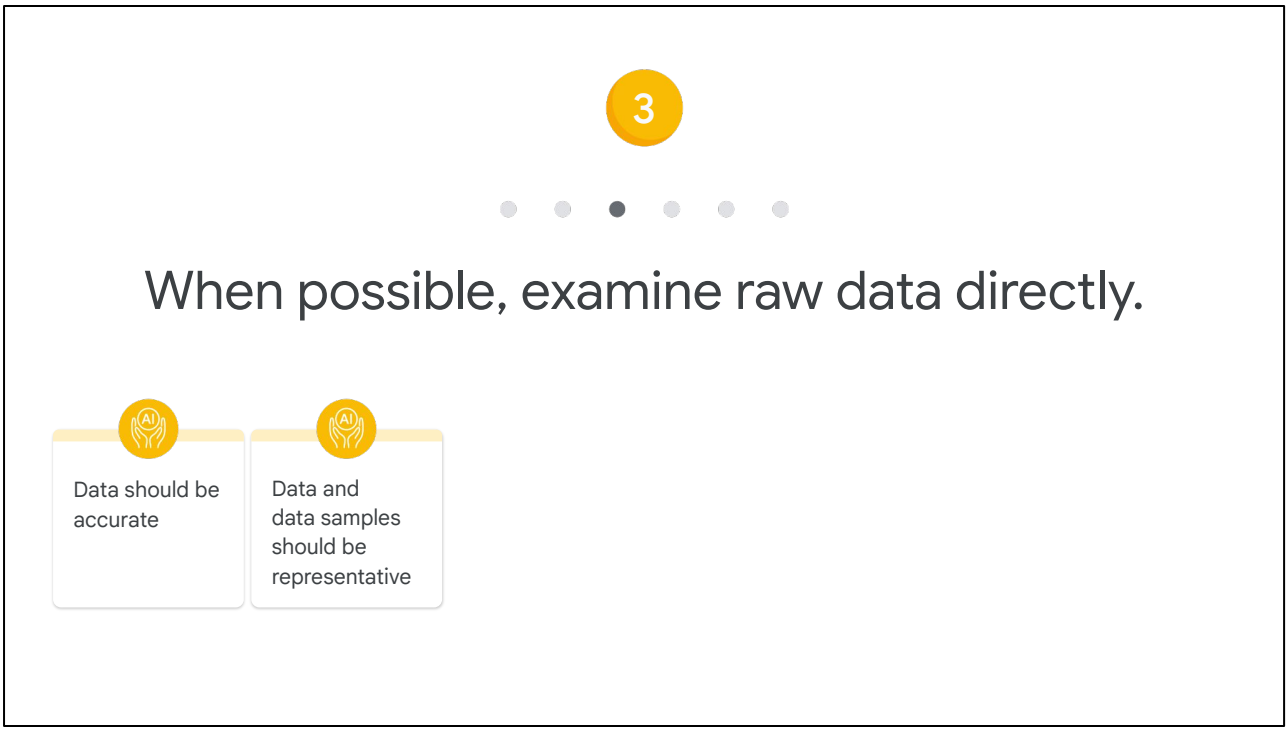
Machine learning models will reflect the data they are trained on, so analyze your raw data carefully to ensure you understand it. In cases where this is not possible, like with sensitive raw data, understand your input data as much as possible while respecting privacy. For example, by computing aggregate, anonymized summaries.



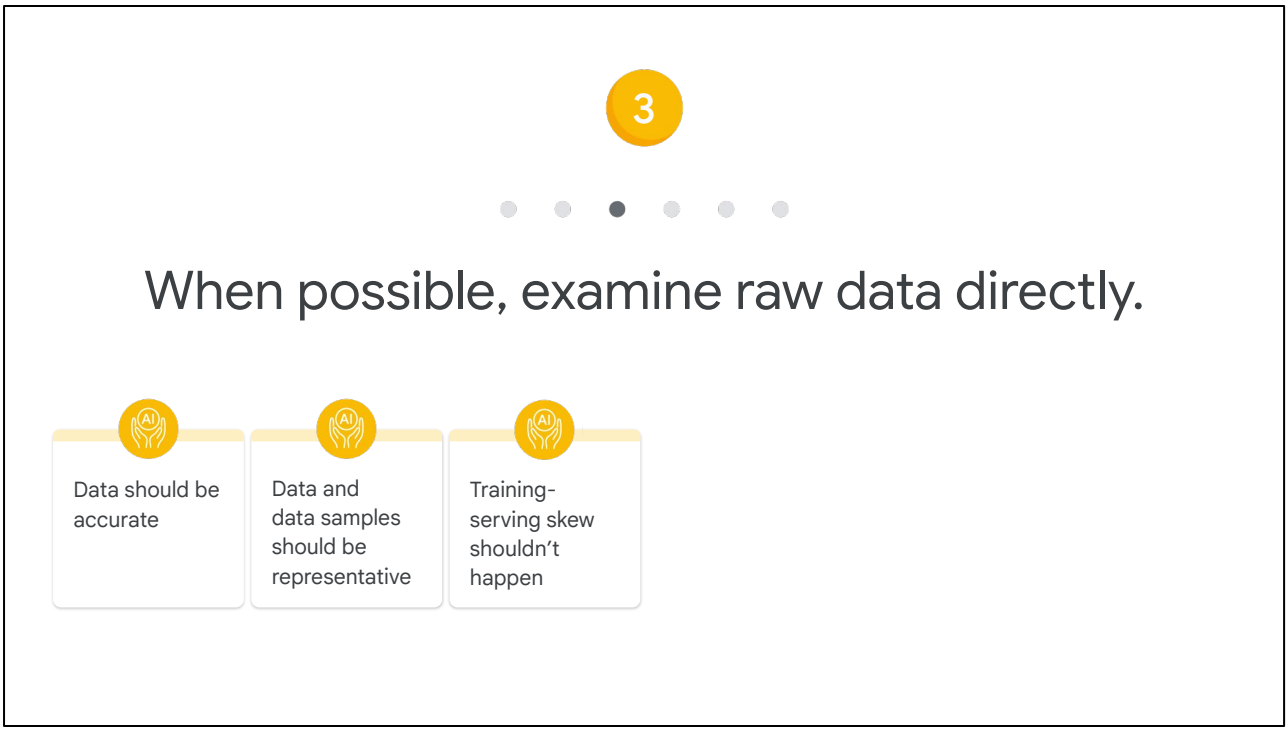
This means that data should be accurate.



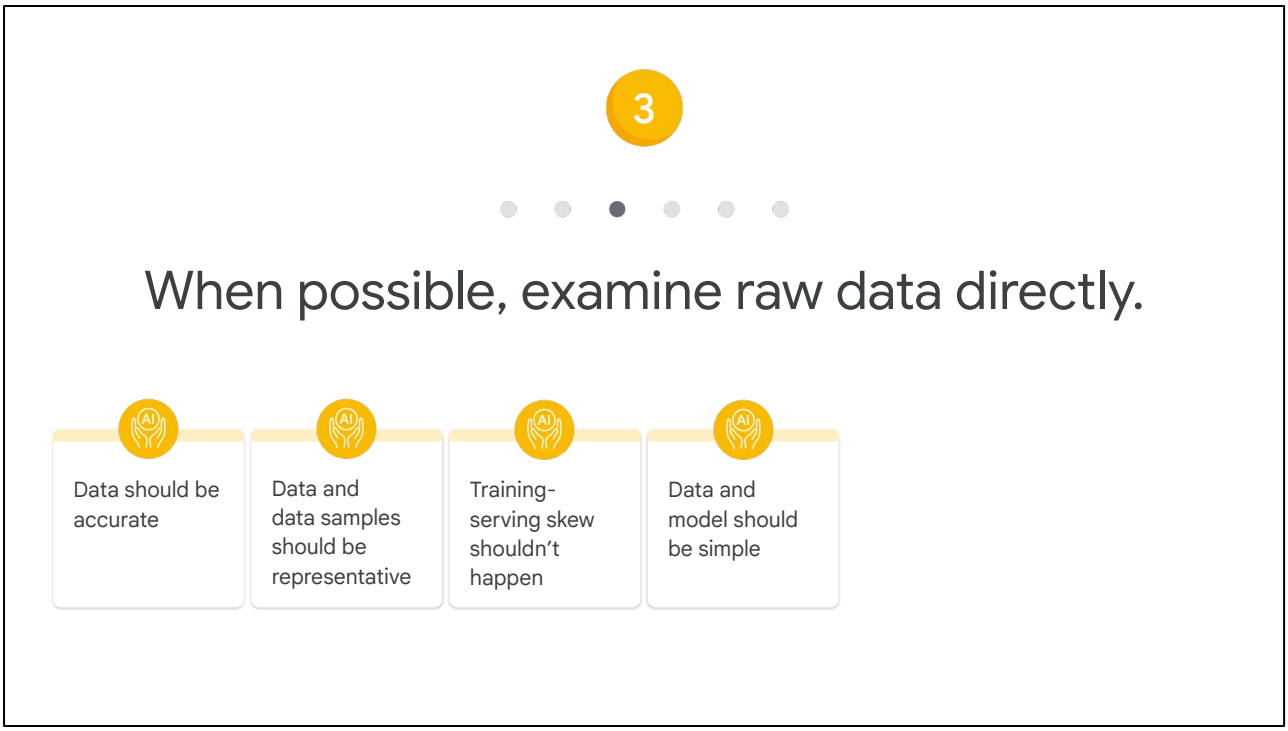
Ask yourself: does my data contain any mistakes (for example, missing values, incorrect labels)?



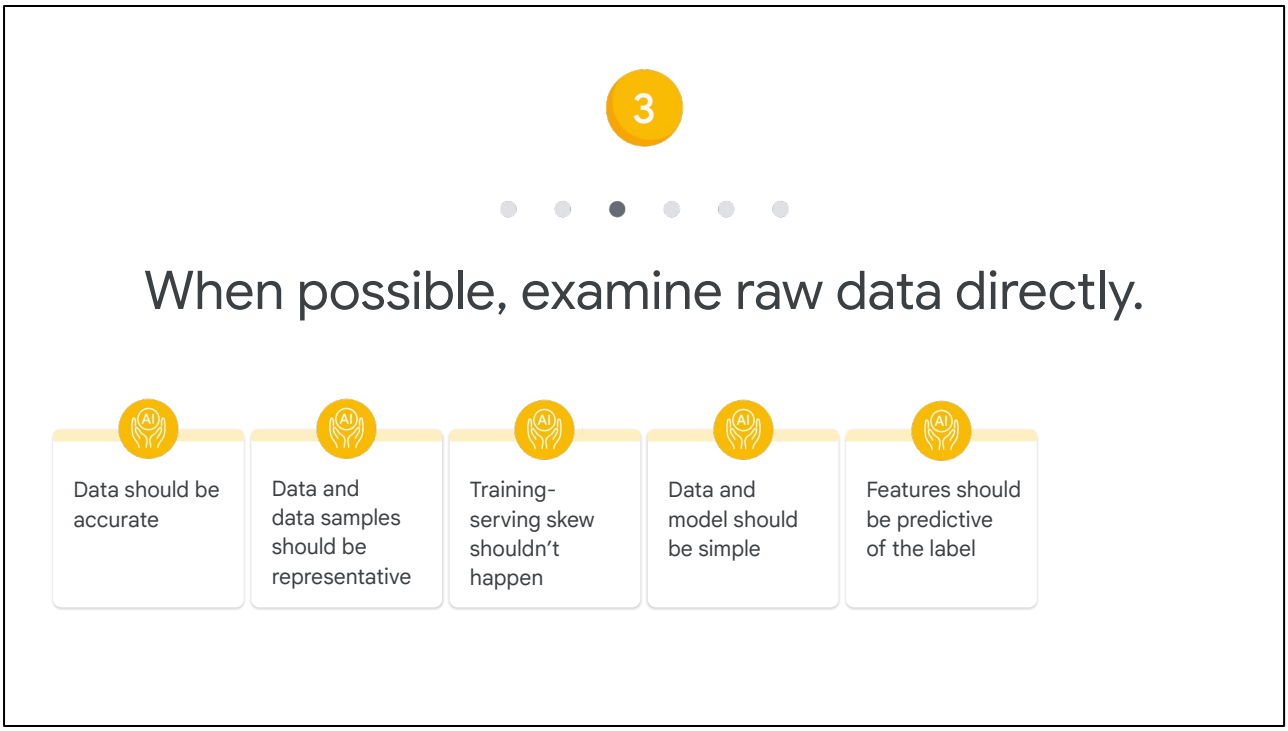
Data and data samples should be representative.
Ask yourself: Is my data sampled in a way that represents my users and the real-world setting?



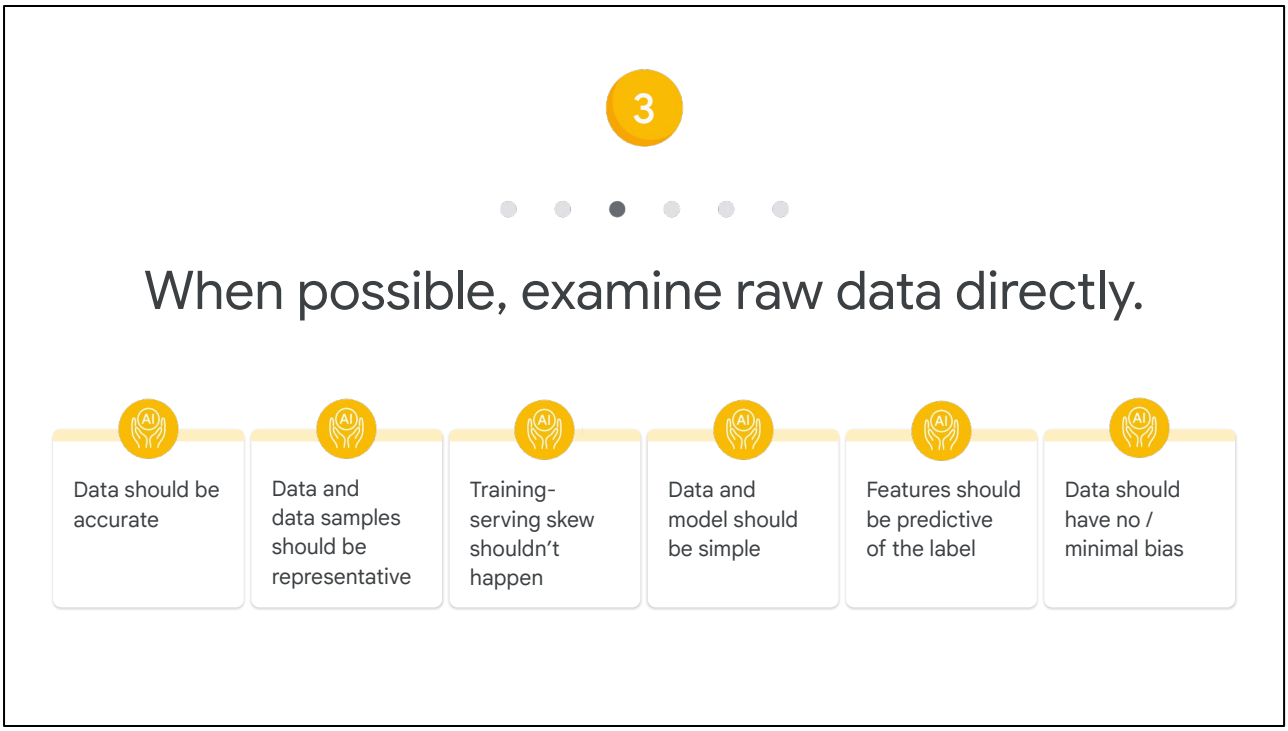
Training-serving skew shouldn't happen.
The difference between performance during training and performance during serving, is a persistent challenge.
During training, you may need to adjust your training data or objective function. During evaluation, continue to ensure that evaluation data is as representative as possible of the deployed setting.



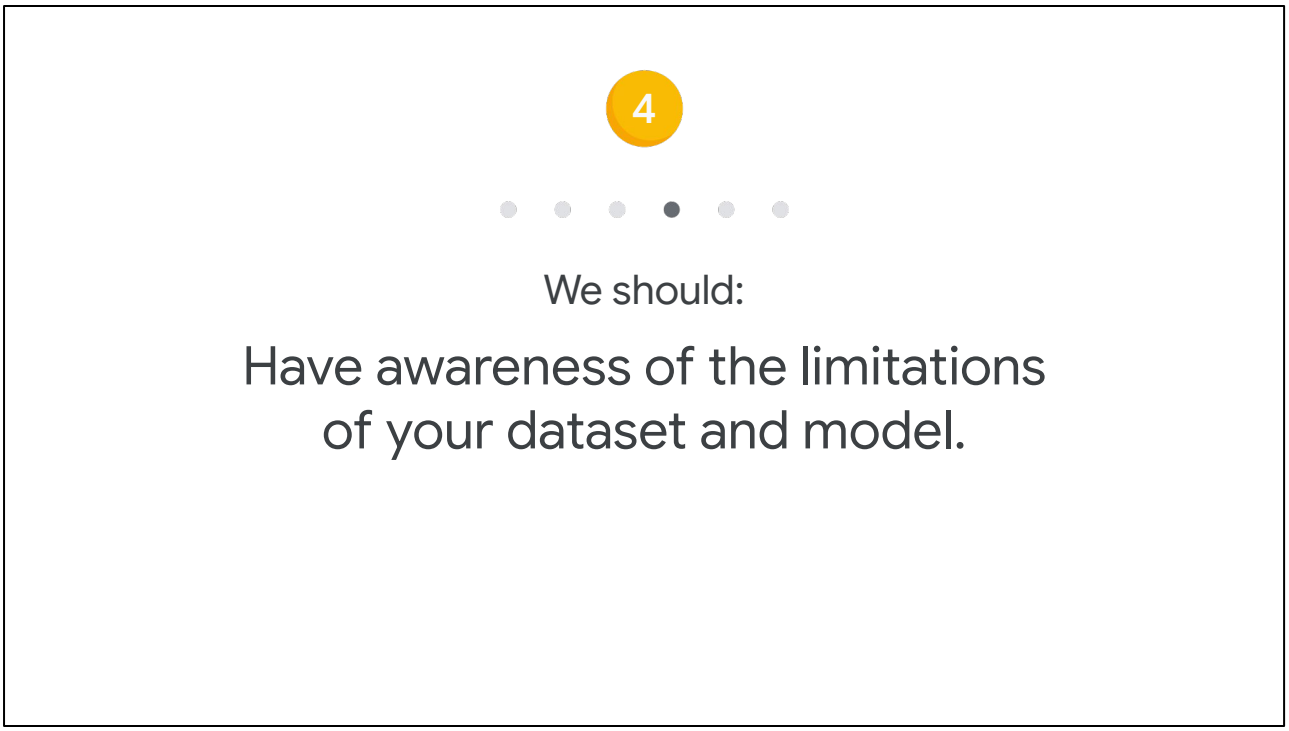
Data and model should be simple.
Ask yourself: Are any features in my model redundant or unnecessary?
Is my model unnecessarily complex?



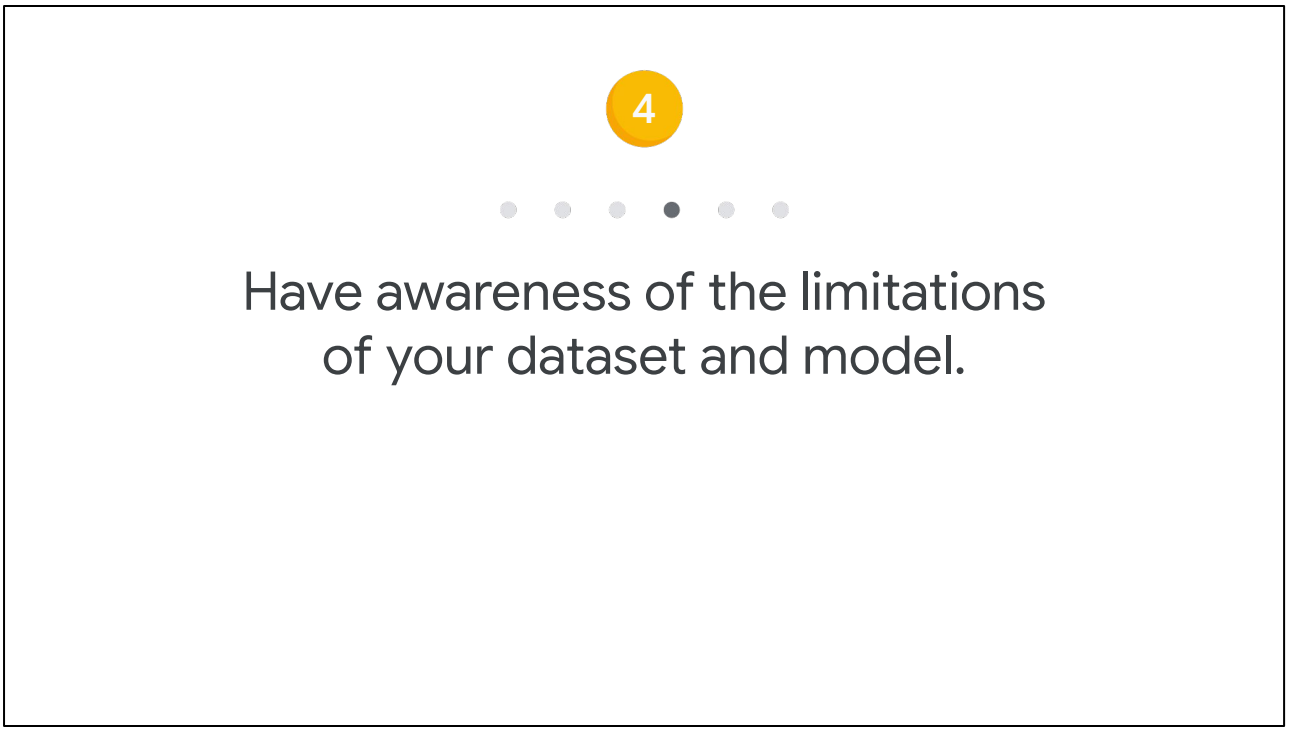
For supervised systems, consider the relationship between the data labels you have, and the items you're trying to predict.



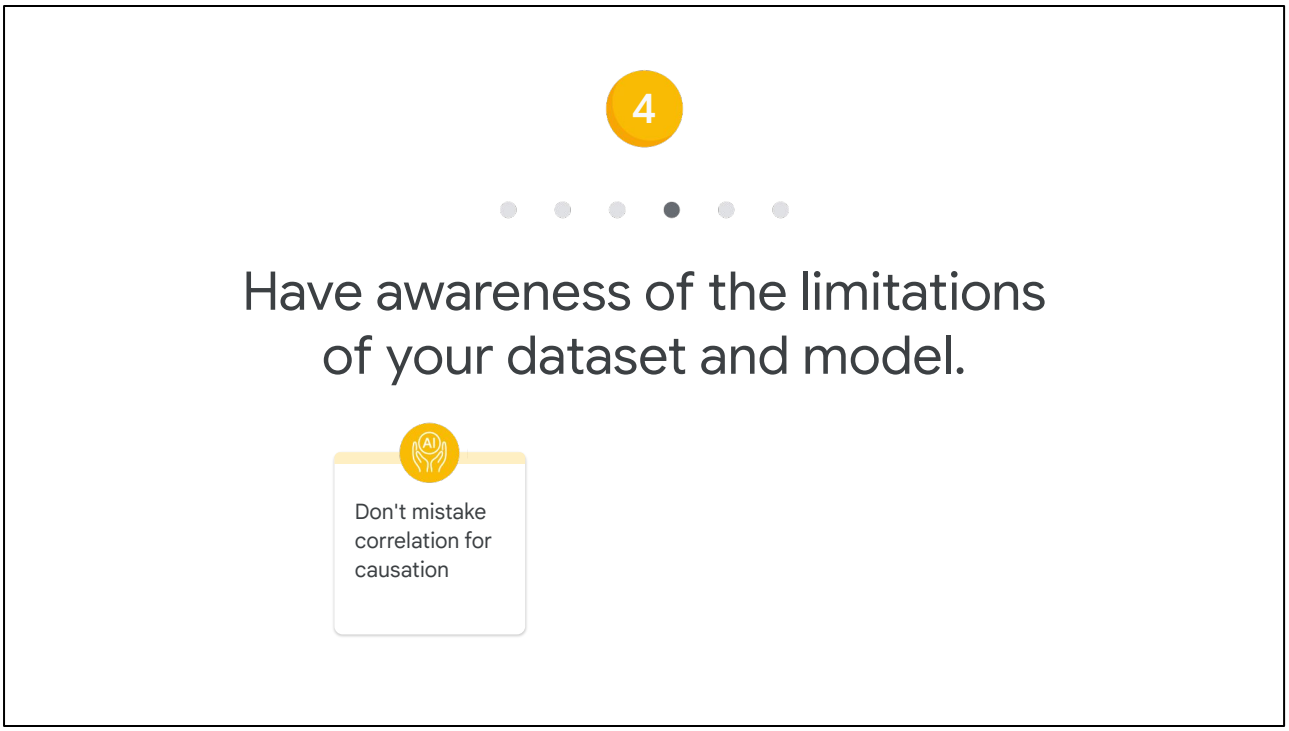
Bias should be minimized in your data. First, you want to ensure that your data is fairly representative of the entire population.



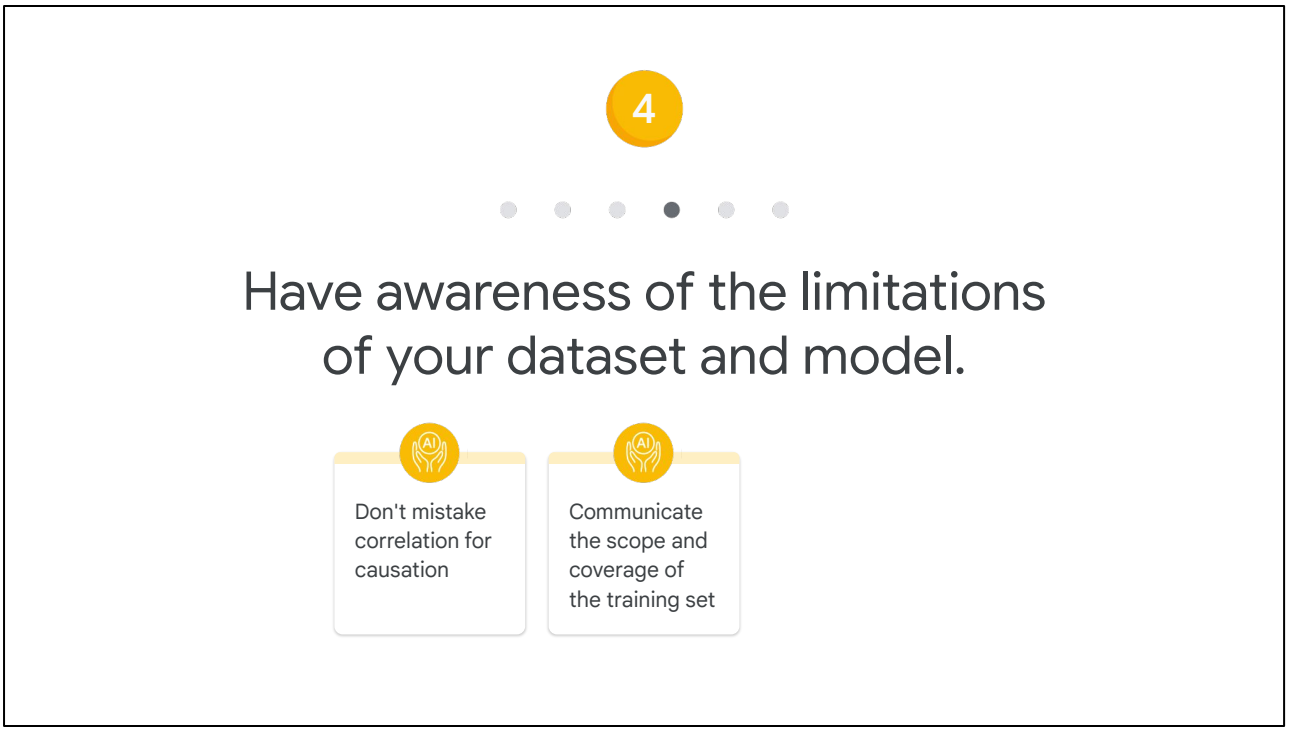
When you develop AI, you should have awareness of the limitations of your dataset and model.



This means you should:

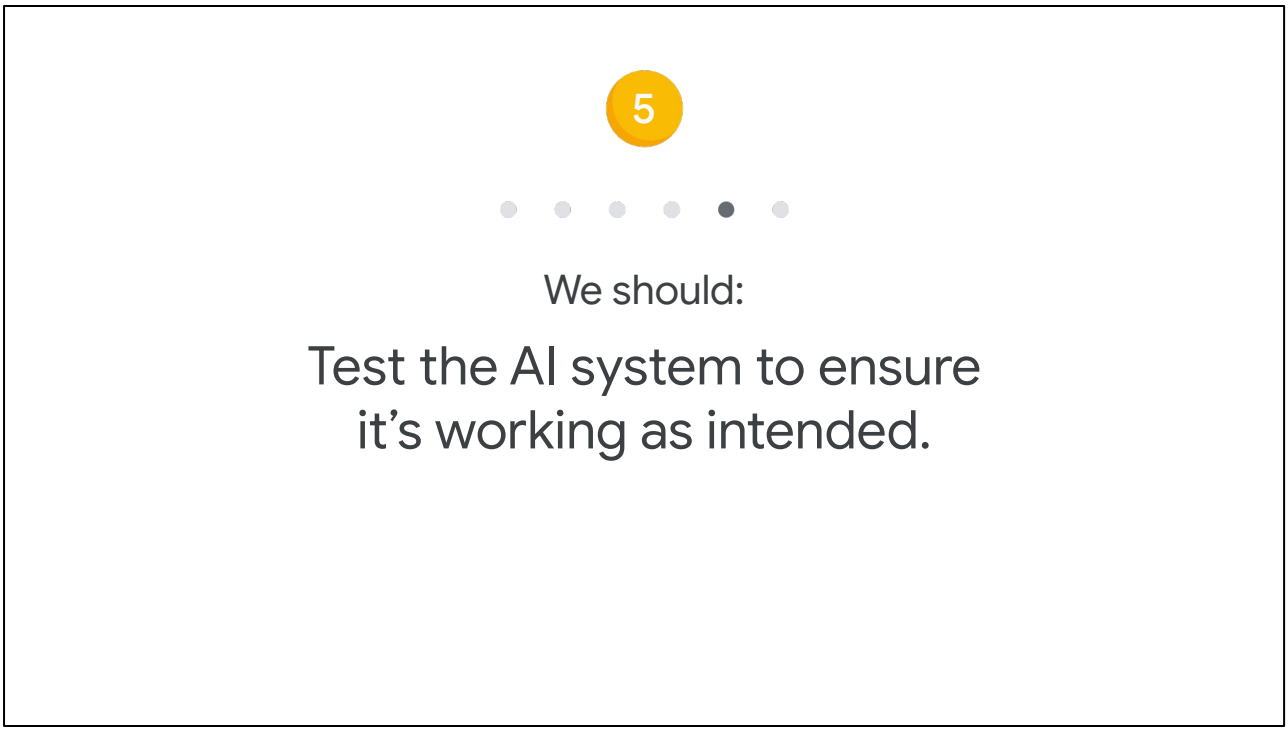


Not mistake correlation for causation.
For example, your model might learn that people who buy basketball shoes are taller on average, but this does not mean that a user who buys basketball shoes will become taller as a result.

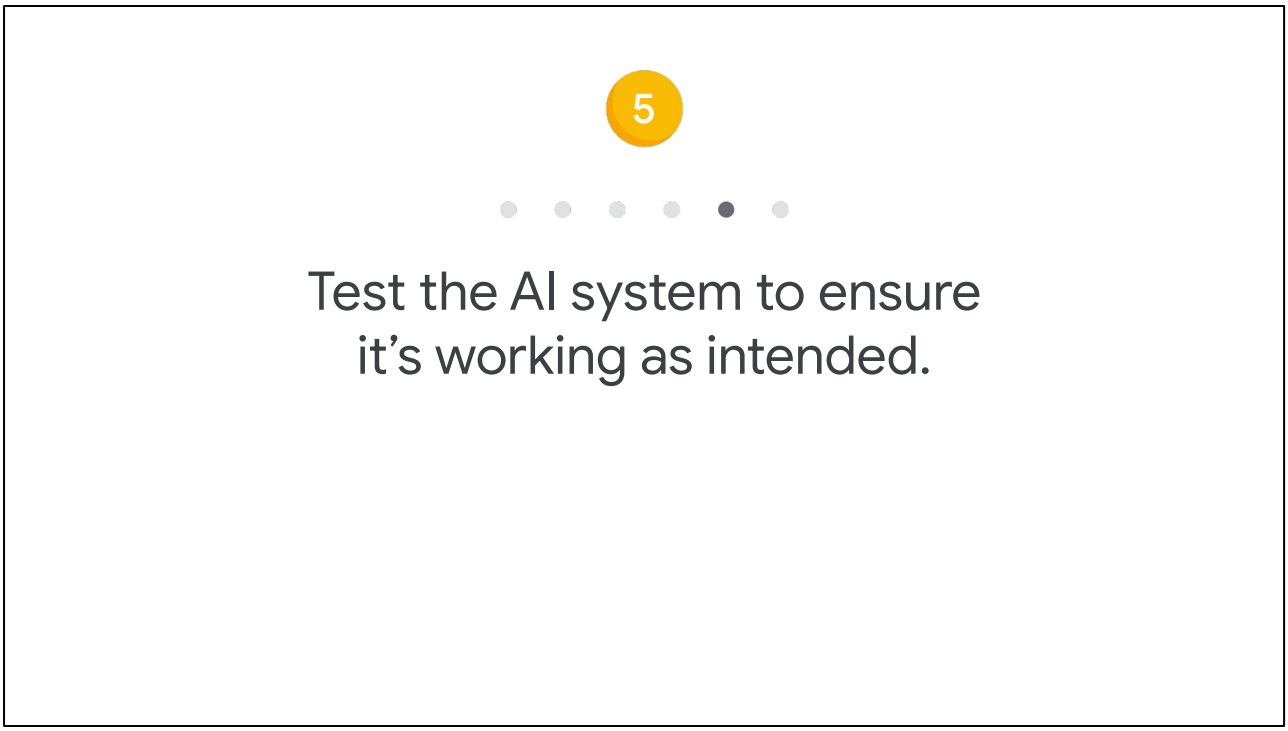


Communicate the scope of the training set.
For example, a shoe detector trained with stock photos can work best with stock photos but has limited capability when tested with user-generated mobile device photos.

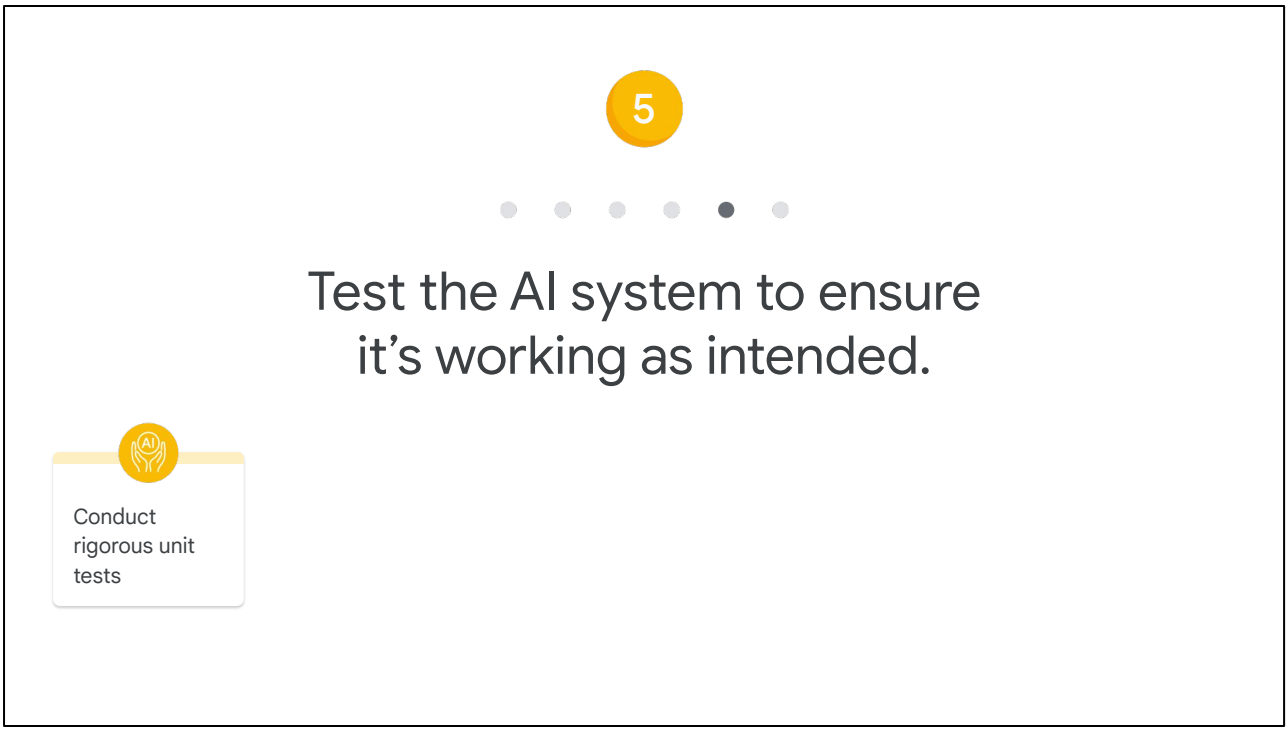
Communicate limitations to users where possible.
For example, an app that uses machine learning to recognize specific bird species might communicate that the model was trained on a small set of images from a specific region of the world.
By better educating the user, you might also improve the feedback provided from users about your feature or application.



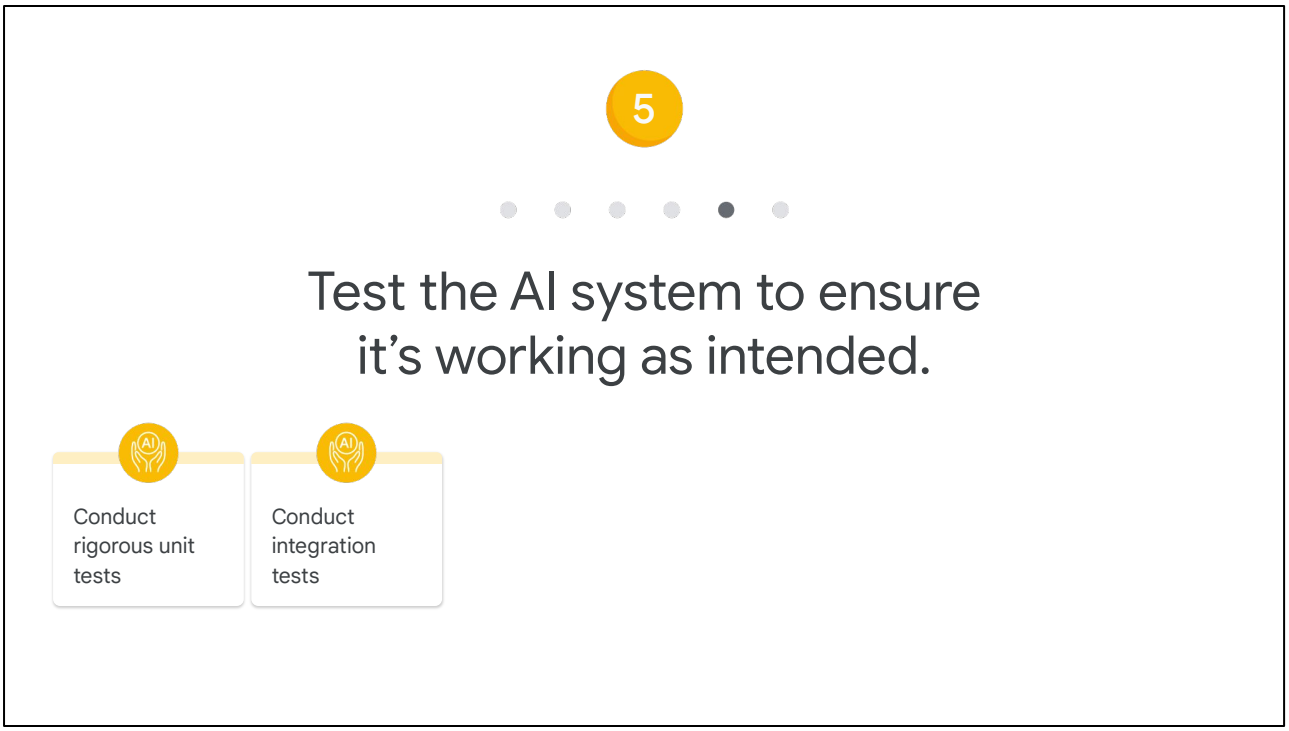
Learn from software engineering best test practices and quality engineering to ensure that the AI system is working as intended and can be trusted.



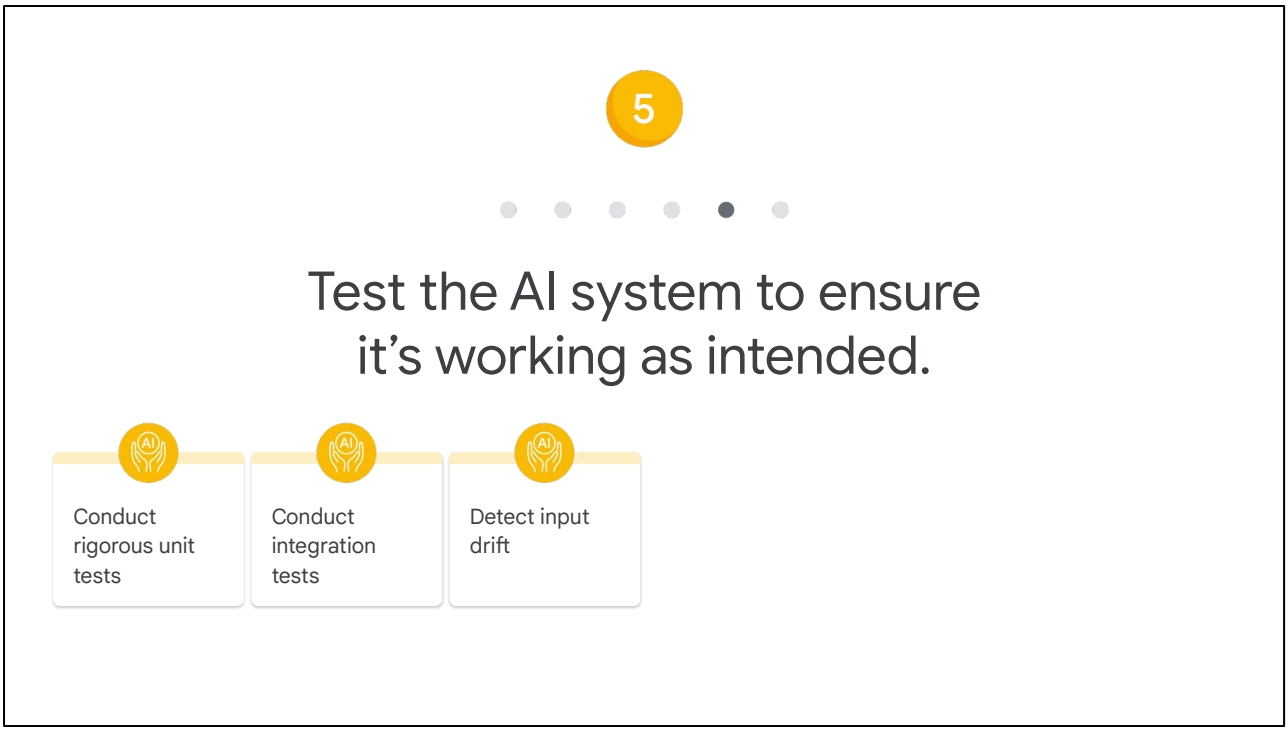
This means you should:



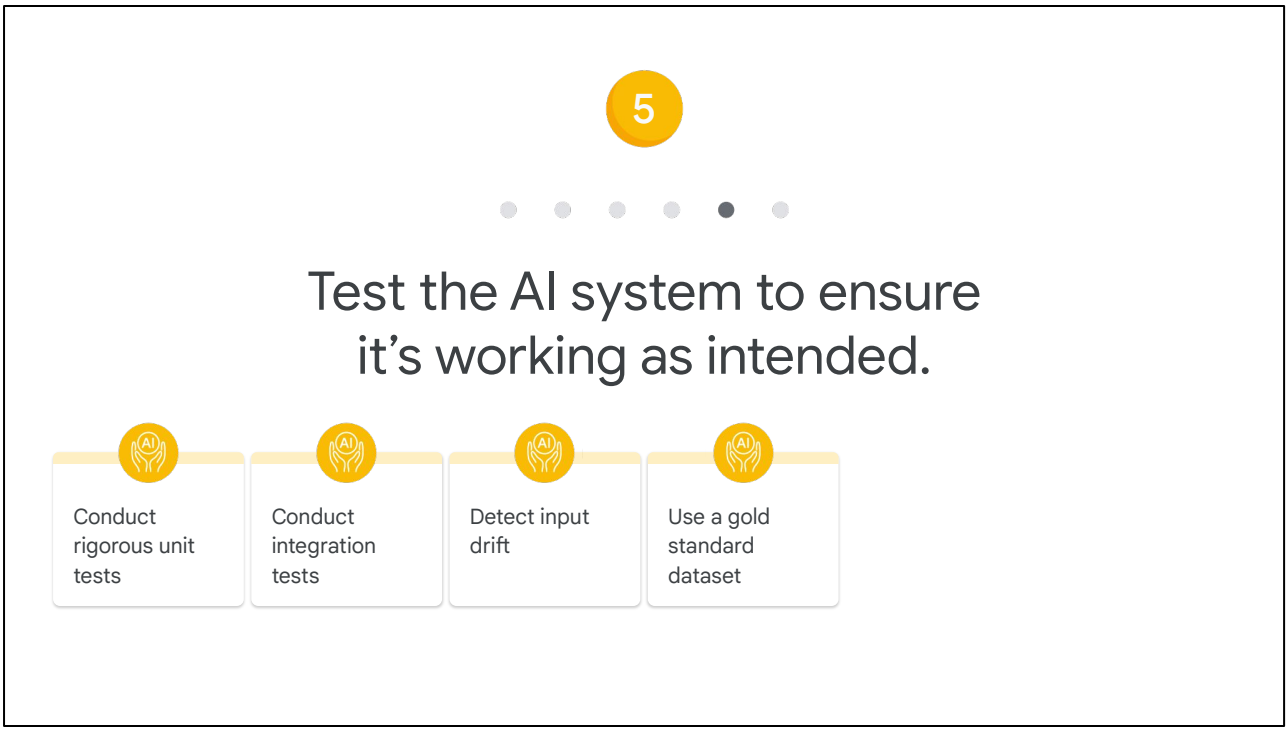
- Conduct rigorous unit tests to test each component of the system in isolation.



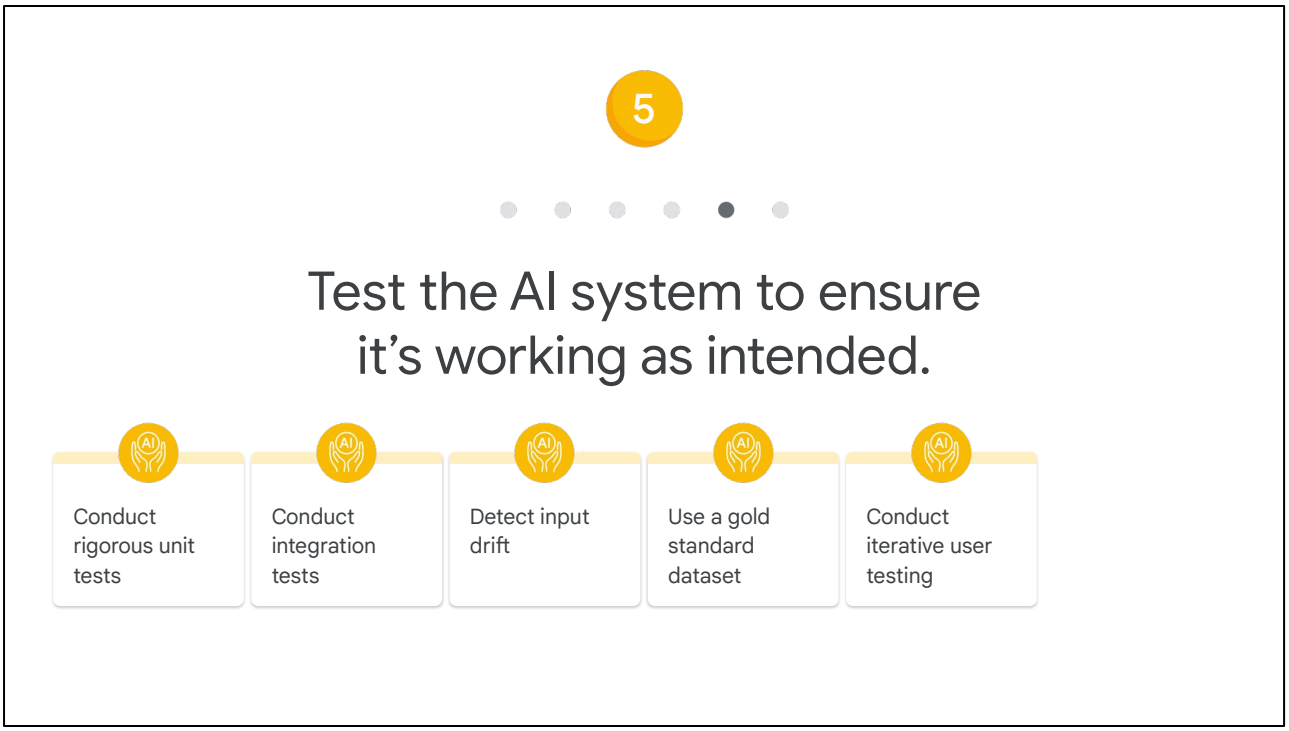
- Conduct integration tests to understand how individual ML components interact with other parts of the overall system.



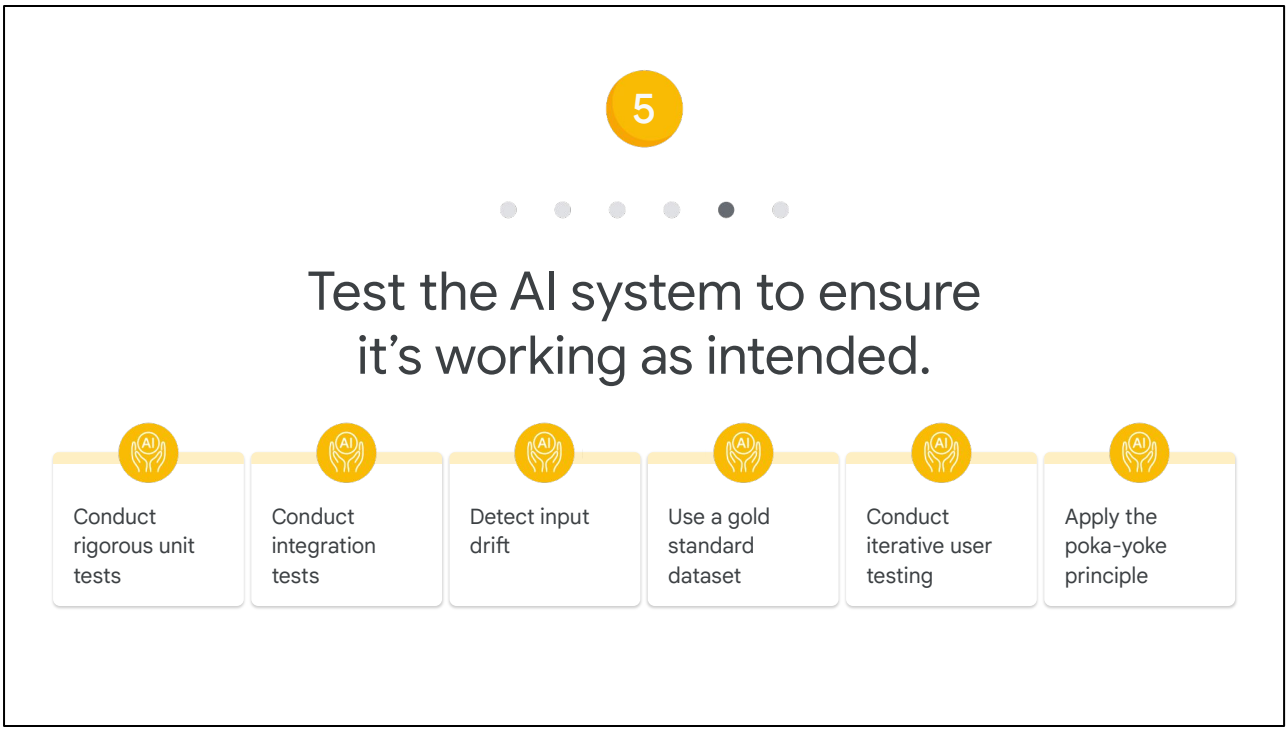
- Proactively detect input drift by testing that data distributions are not changing in unexpected ways.



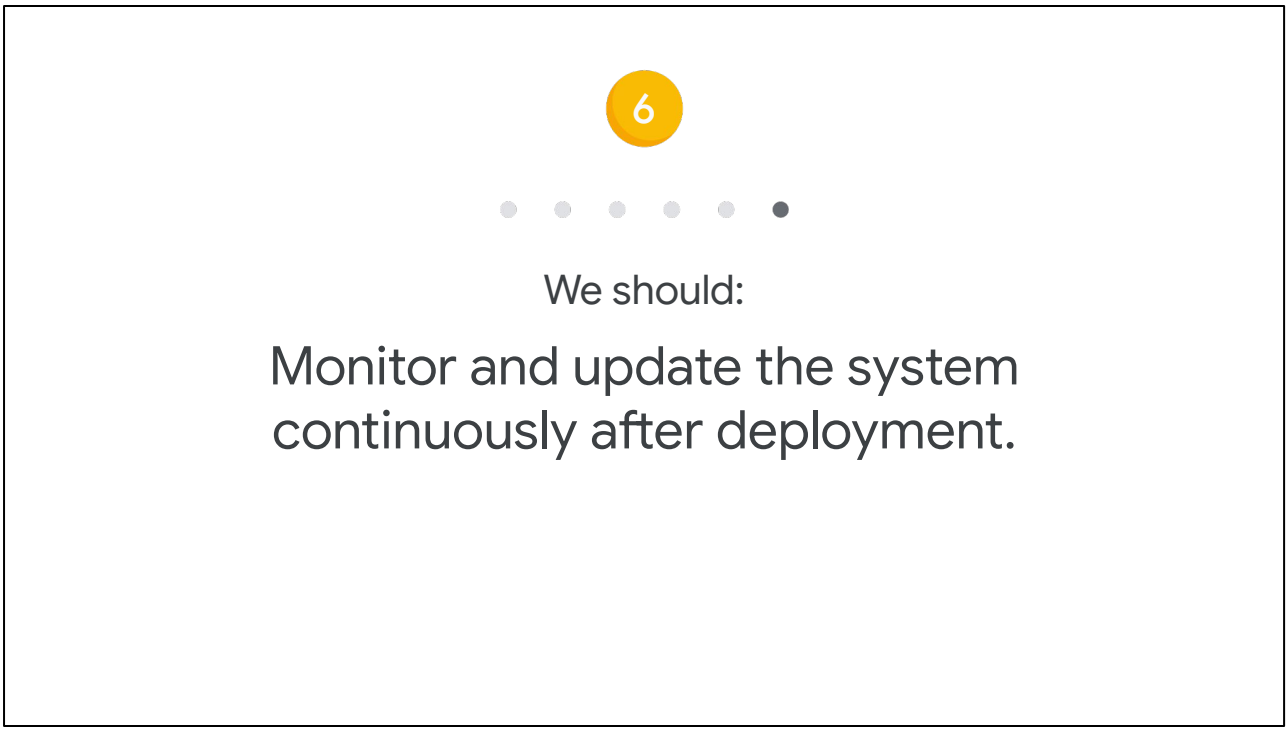
- Use a gold standard dataset to test the system and ensure that it continues to behave as expected by updating it regularly.



- Conduct iterative user testing to incorporate a diverse set of users' needs in the development cycles.

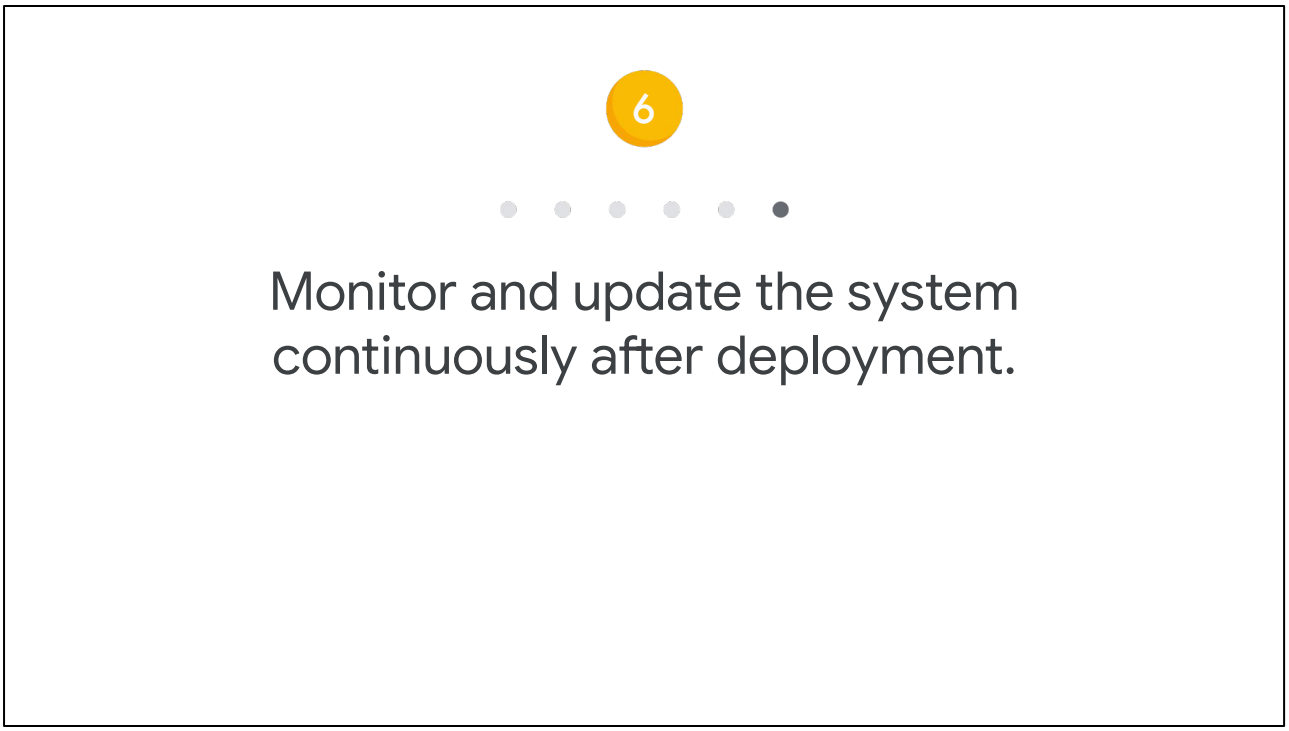


- Apply the quality engineering principle of poka-yoke. The principle pushes you to build quality checks into a system, so that unintended failures either cannot happen or they trigger an immediate response. For example, if an important feature is unexpectedly missing, the AI system won't output a prediction.

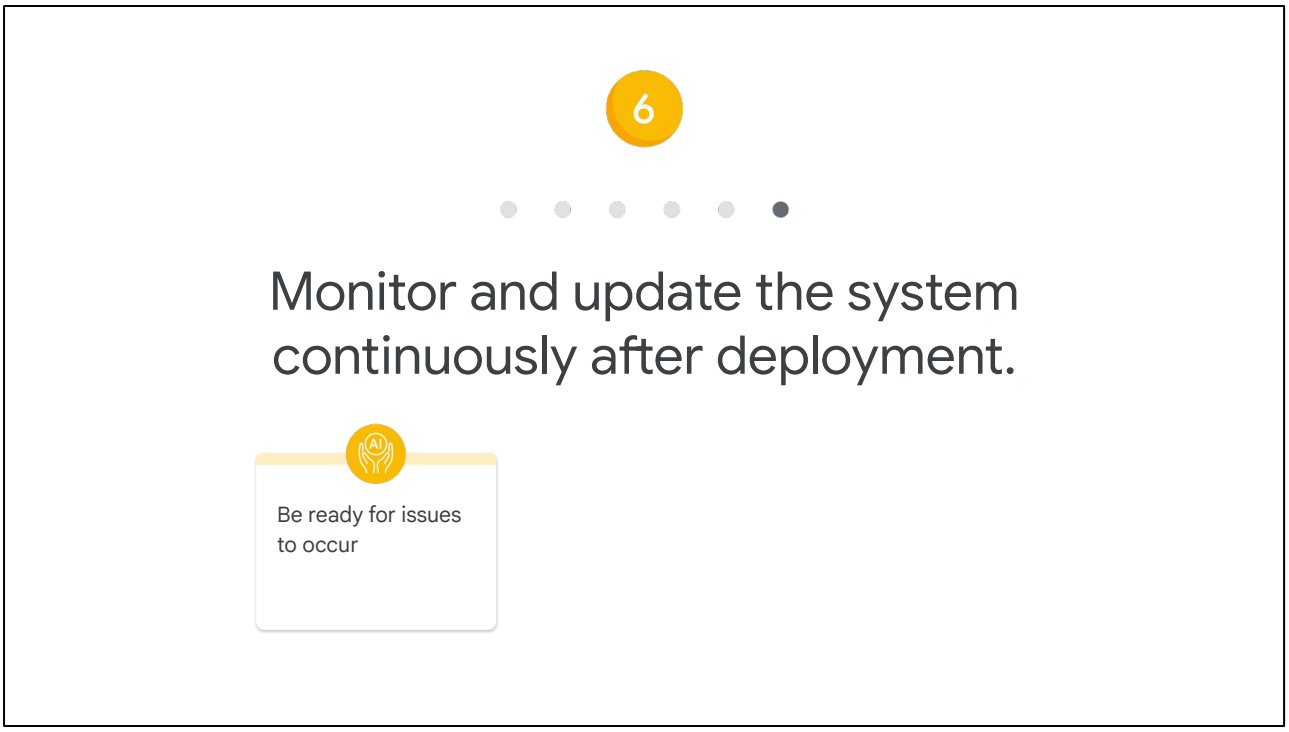


Lastly, you should monitor and update the system continuously after deployment.

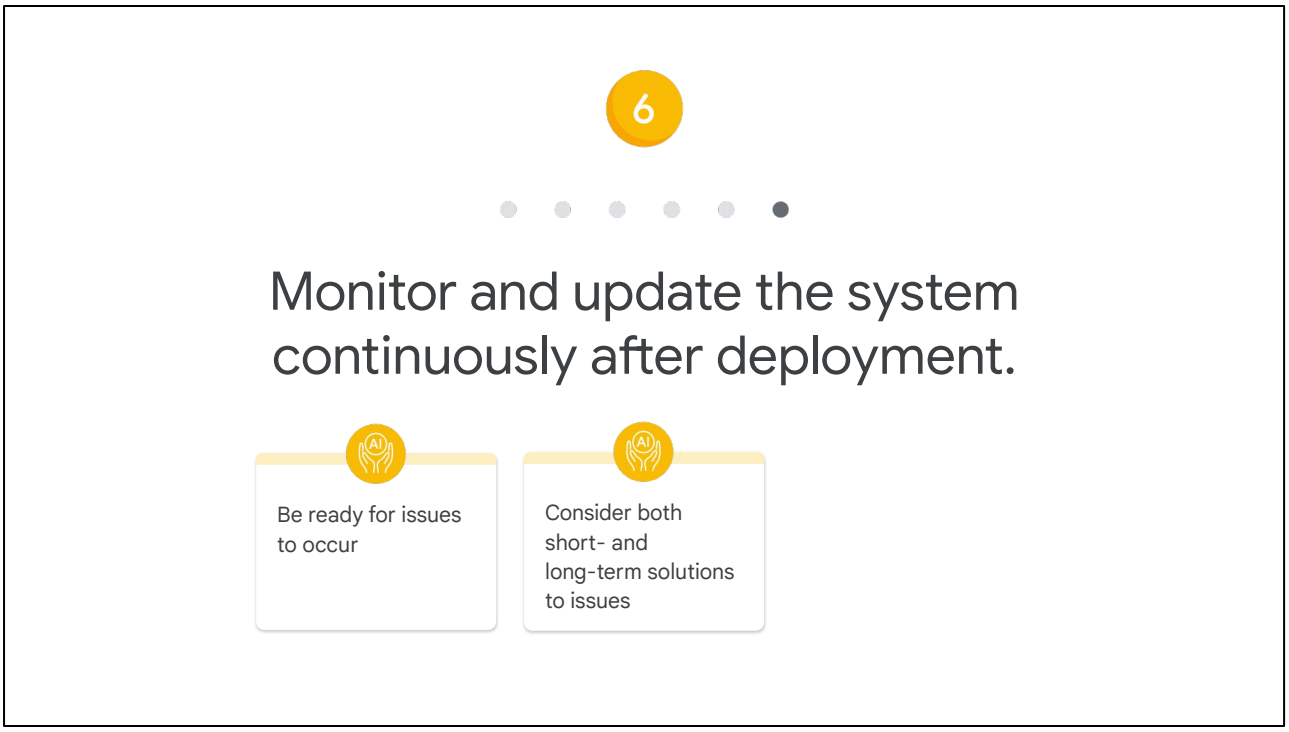
Continued monitoring ensures that your model takes real-world performance and user feedback (like, happiness tracking surveys, HEART framework) into account.



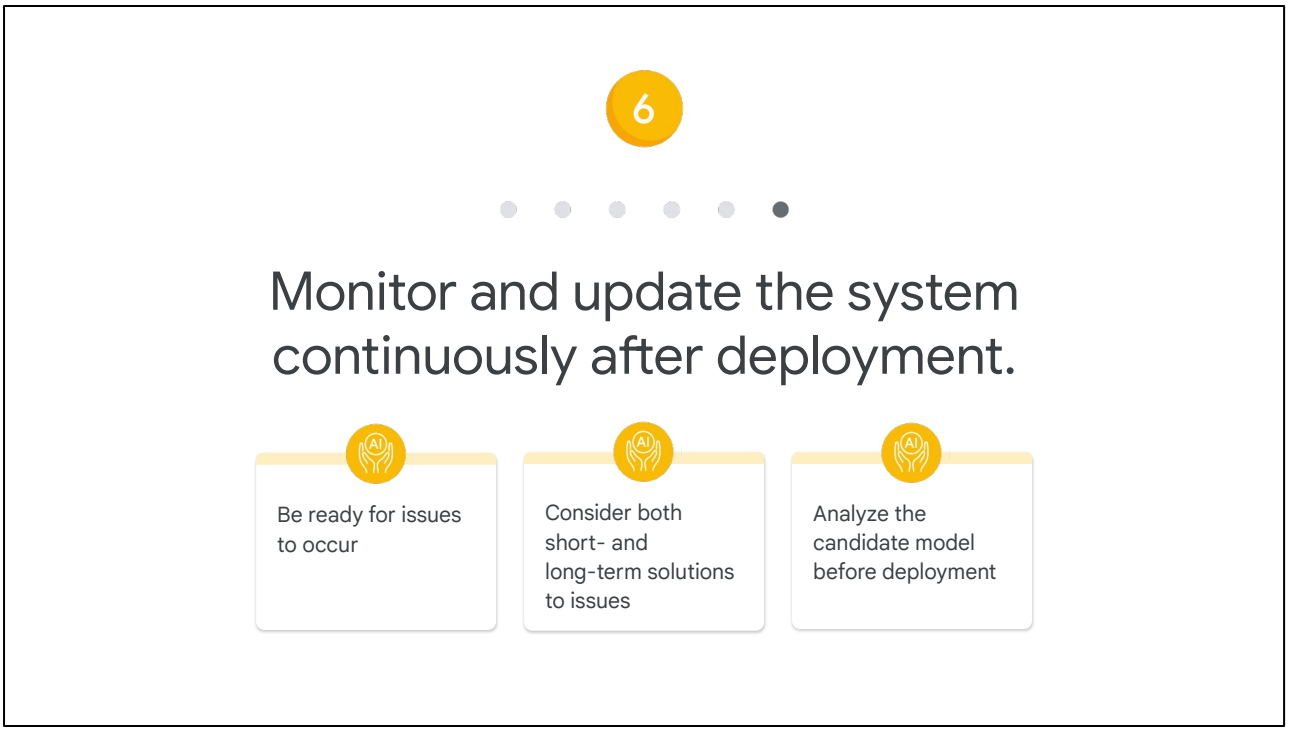
This means that you should:



- Be ready for issues to occur: any model of the world is imperfect almost by definition. Build time into your product roadmap to let you address issues.



- Consider both short and long-term solutions to issues. Balance short-term simple fixes with longer-term learned solutions.



- Analyze the candidate model before deployment. Specifically, how it differs from the deployed model, and how the update will affect the overall system quality and user experience.

01 AI & Responsibility

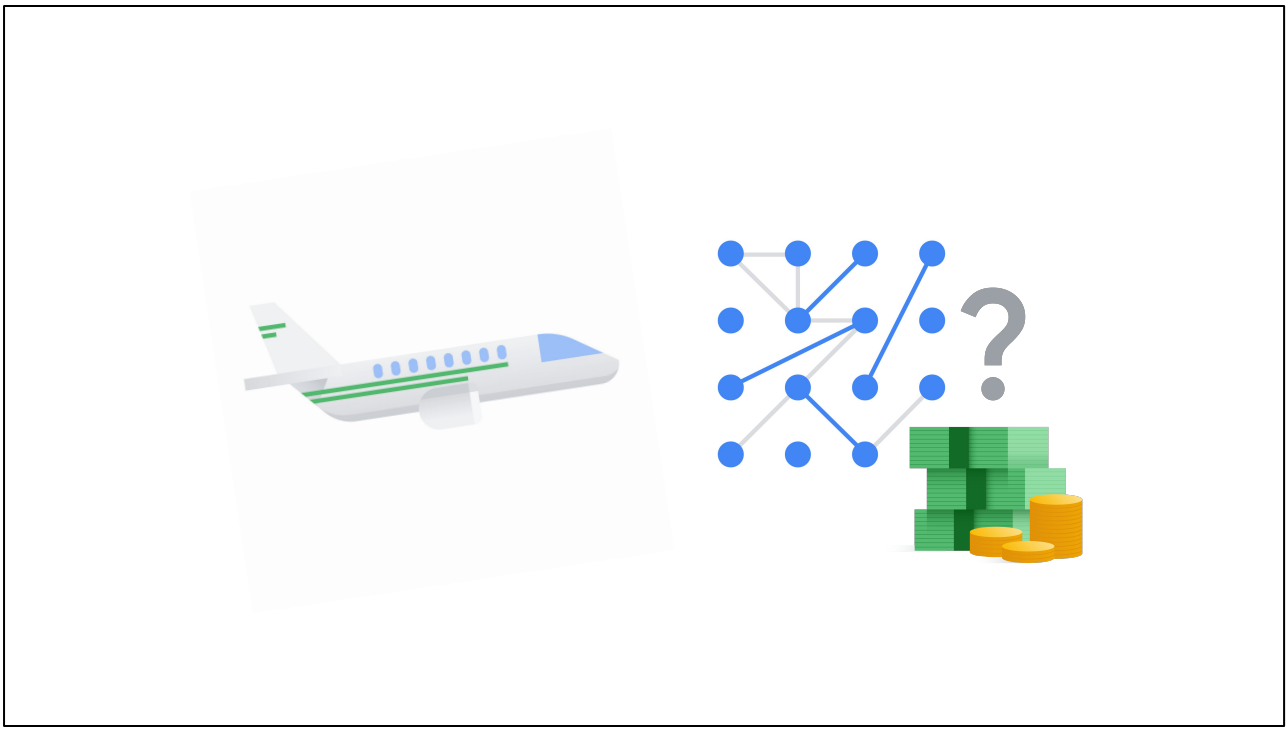
02 Google's AI principles

03 Responsible AI practices

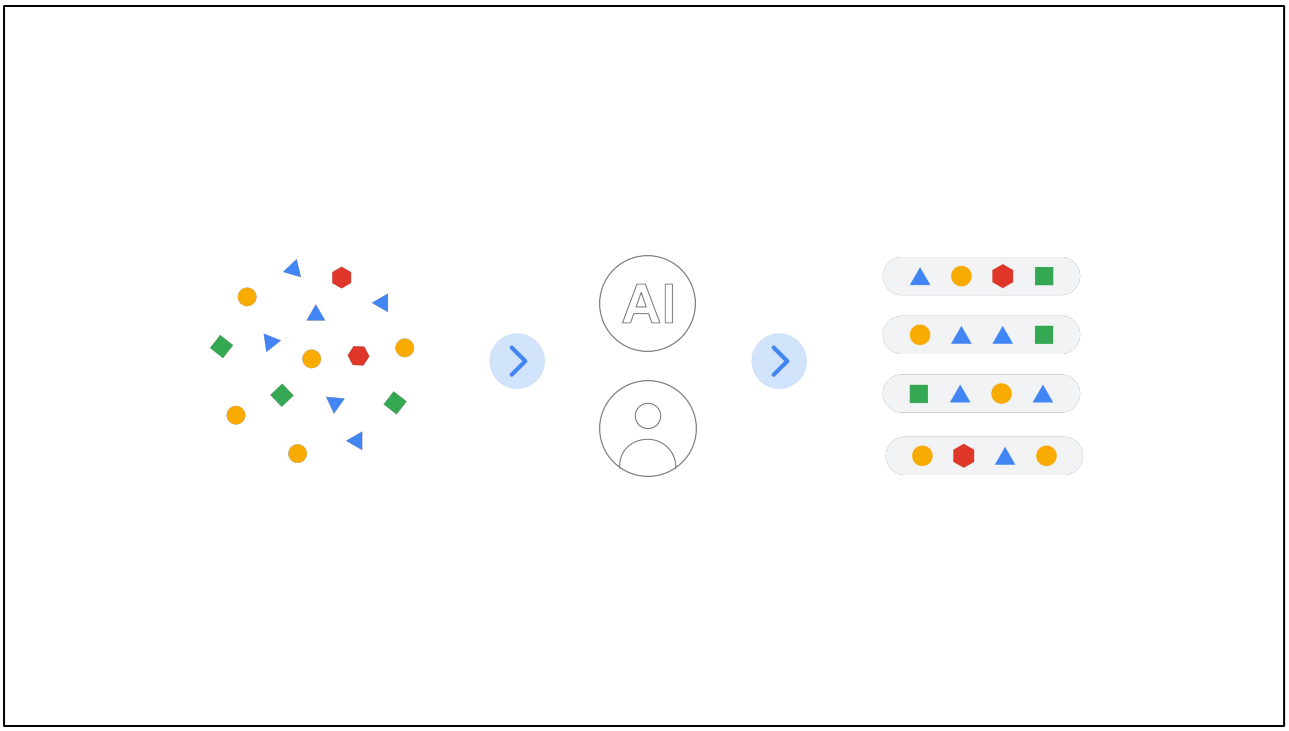
04 Case study: Google Flights

Now, let's look at a case study on Google Flights.

Imagine the scene: You were busy at work and suddenly realized you forgot to book the flight for your upcoming trip.

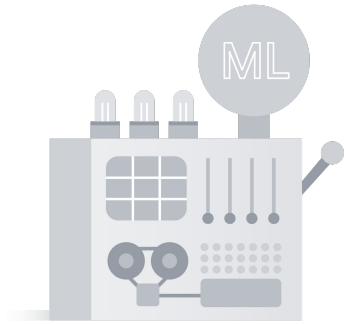


Purchasing flight tickets can be difficult. Identifying patterns in flight prices is challenging due to changes in flight prices, inconsistent pricing across sites, and sometimes, even pricing breakdowns are hard to understand.



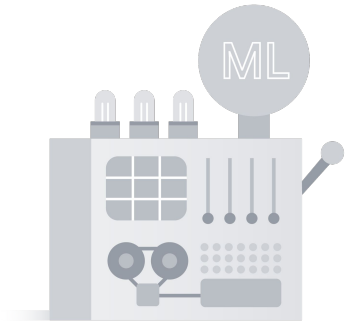
Google Flights believes that by putting some of the data and smart AI in the hands of users, it would help them demystify what they need to pay for a certain flight at a certain time. This would potentially save users time, stress, and money.

Machine learning predictions can't be 100% right all the time.



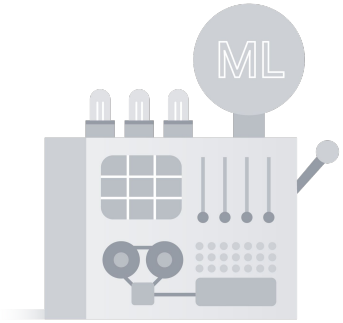
One of the challenges with predictions, and a common theme across AI-driven products, is that machine learning predictions can't be 100% right all the time for two reasons:

- Machine learning predictions can't be 100% right all the time.
  - The predictions are specific to certain flights to certain places at specific times.



First, the predictions are specific to certain flights to certain places at specific times, and

- Machine learning predictions can't be 100% right all the time.
  - The predictions are specific to certain flights to certain places at specific times.
  - For some places, we don't have enough pricing data to provide an accurate prediction of whether the price is fair or not.



second, we don't have enough pricing data to provide an accurate prediction of whether the price is fair or not.

As they thought through how to build the tool, they felt it was important to help users make informed and better decisions by explaining where this data was coming from and what it relates to. Therefore, users were allowed to:

Assess price ‘goodness’ today and in the future

- Assess price “goodness” today and in the future.

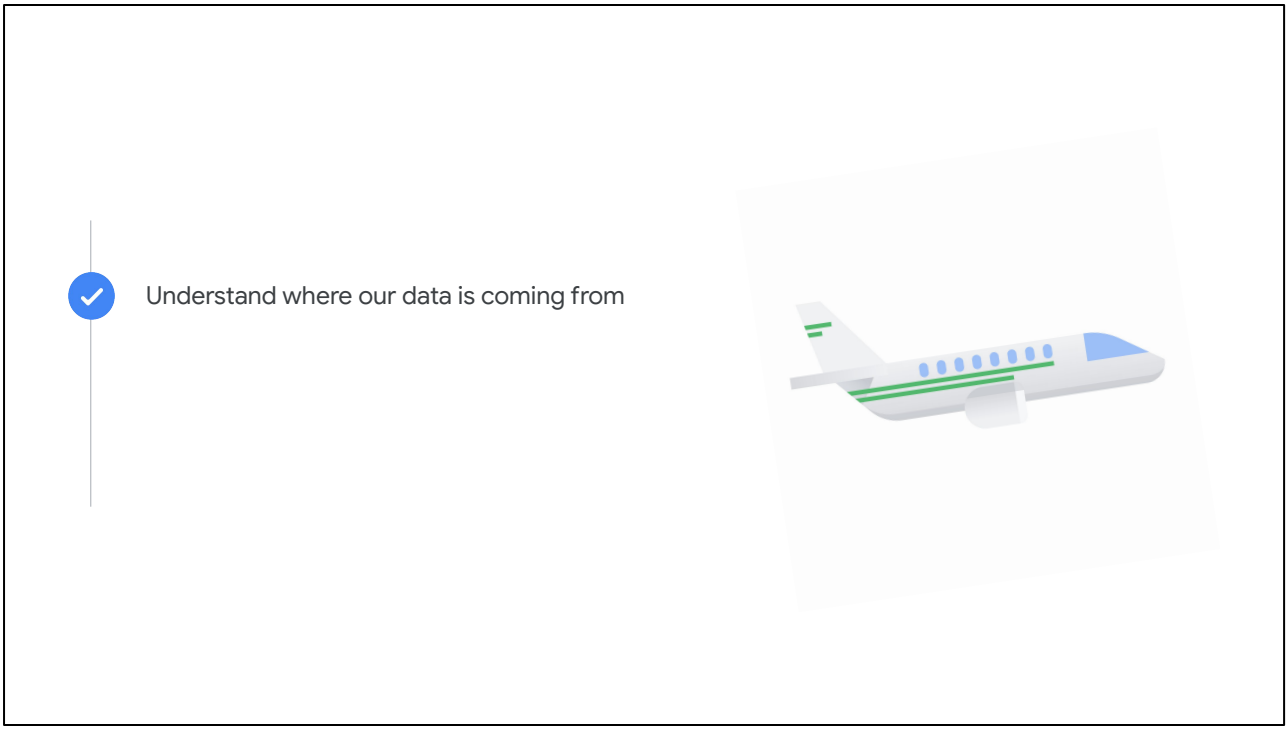
- Assess price ‘goodness’ today and in the future
- Track the model’s predictions and check them

- Track the model’s predictions and check them.

- Assess price ‘goodness’ today and in the future
- Track the model’s predictions and check them
- Make confident decisions about when to book

- Make confident decisions about when to book.

While at the same time, ensuring that users:



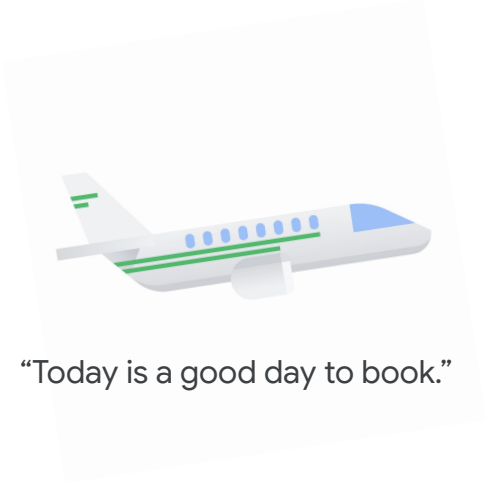
- Understand where our data is coming from.

- Understand where our data is coming from
- View the general trends in flight pricing

- View the general trends in flight pricing.

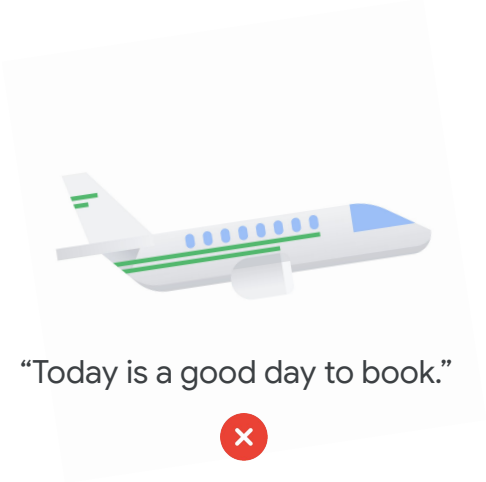
- Understand where our data is coming from
- View the general trends in flight pricing
- Have reasonable expectations for the correctness of our predictions

- Have reasonable expectations for the correctness of our predictions



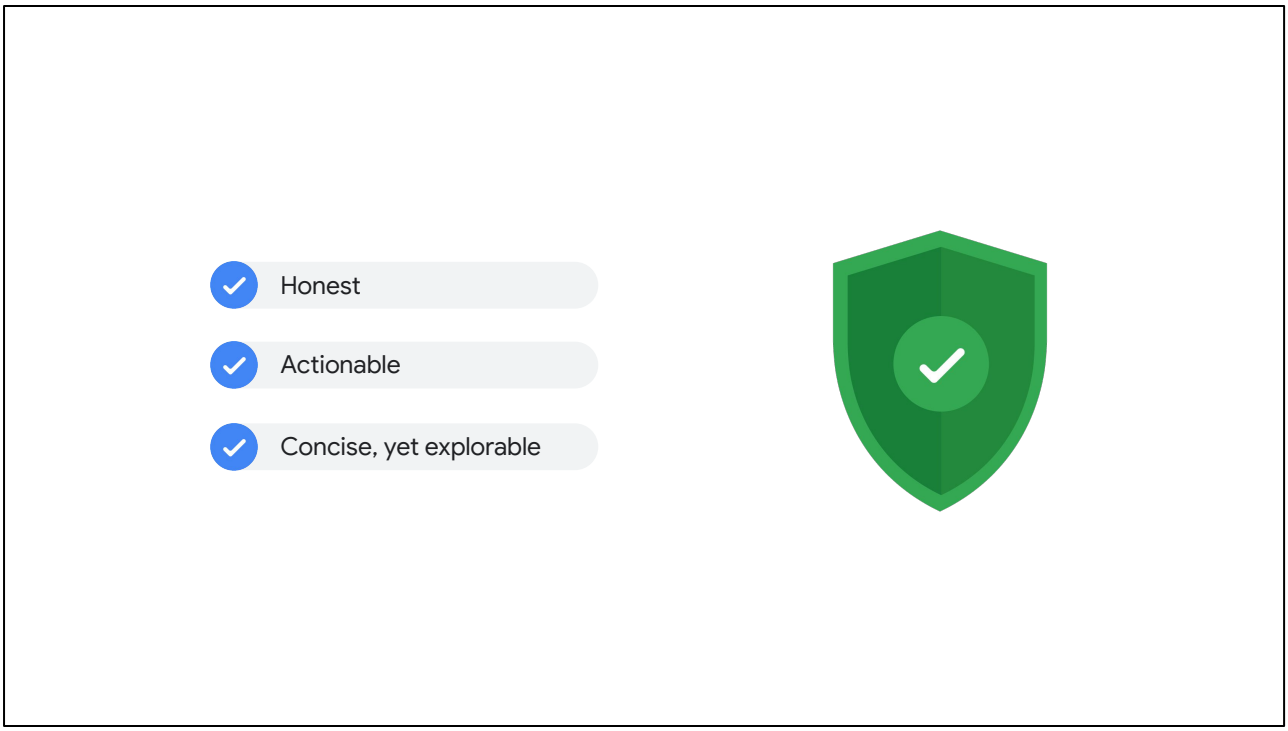
Google Flights started designing a new tool to help users understand whether the prices for a given flight are currently high,
low, or typical, and to help users learn market trends for similar trips.

At one point, they considered a pivot to a more direct approach: hiding the complicated calculations in the background,
and giving users a simple conclusion, such as, “Today is a good day to book.”



But during testing, users expressed that this felt salesy, was upsetting, and didn’t feel trustworthy.

That was not an option, because they wanted a tool that only worked if people trusted it.



To start setting some guide rails for further iterations, the team created three design principles for price intelligence. Any Google Flight price insights surfaced to users would have to be:

- Honest
- Actionable, and
- Concise, yet explorable.

Accurate, engaging output explanations to include:

So how to explain the machine learning model output to users in a way that is actionable and compelling, but also accurate?

Through the following ways:

Accurate, engaging output explanations to include:

A price “goodness” indicator.

- A price ‘goodness’ indicator, with corresponding descriptions of ‘high’, ‘typical’, or ‘low’

Accurate, engaging output explanations to include:

- A price “goodness” indicator.
- A single-line explanation of the usual price.

- A single-line explanation of the usual price for a trip like the one the user is planning

Accurate, engaging output explanations to include:

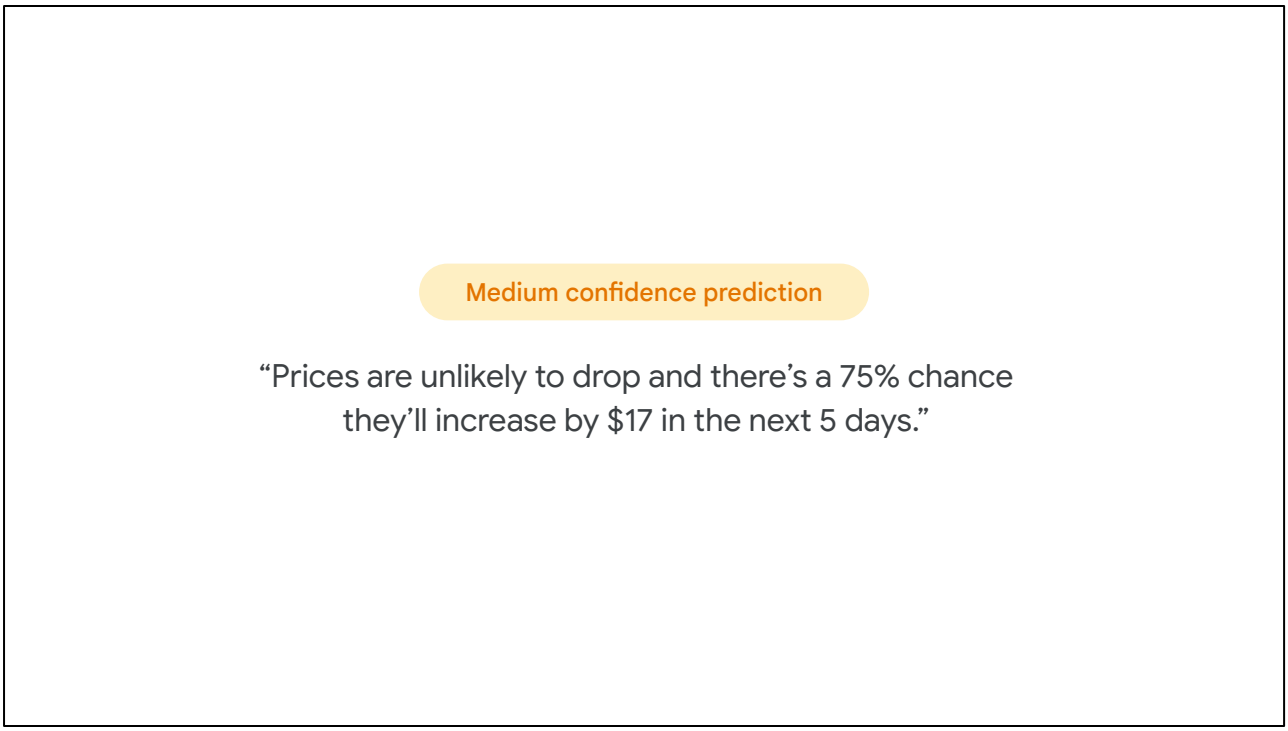
- A price “goodness” indicator.
- A single-line explanation of the usual price.
- Prediction text.

- Prediction text stating whether prices are likely to go up or down

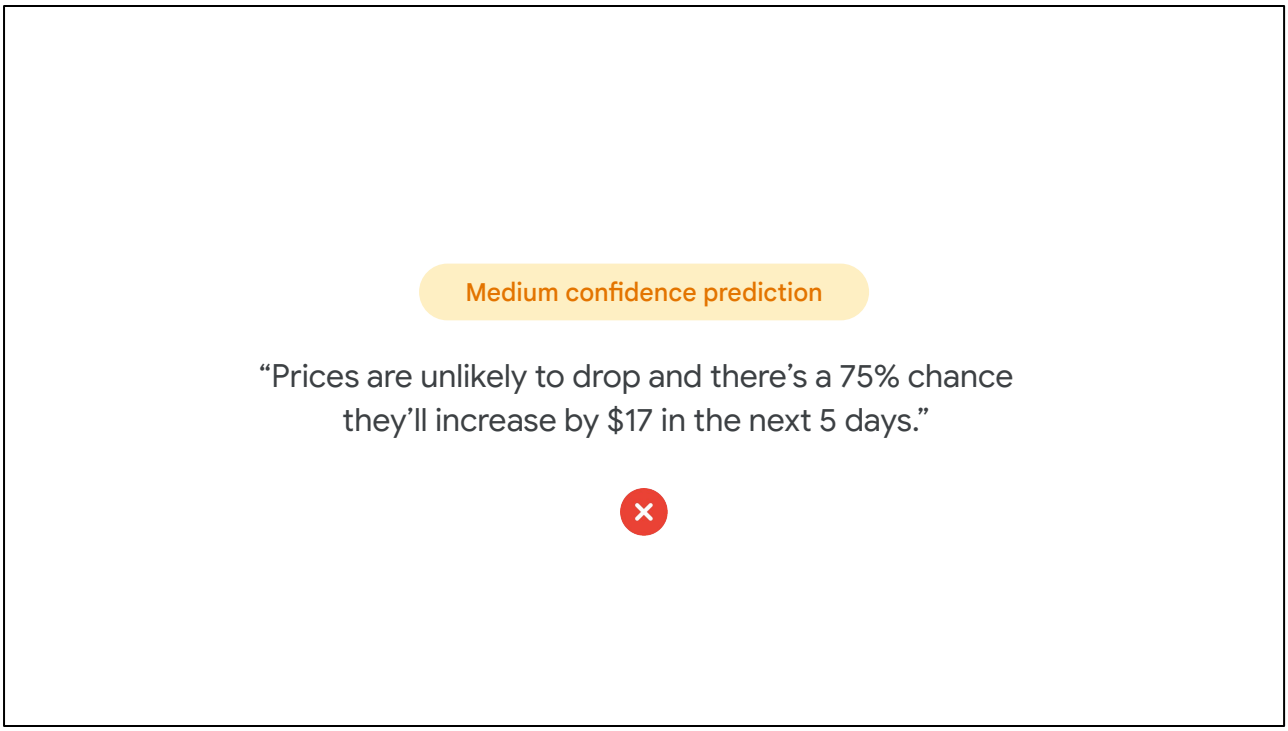
Accurate, engaging output explanations to include:

- A price “goodness” indicator.
- A single-line explanation of the usual price.
- Prediction text.
- An info icon to provide more explanation.

- An info icon that opens an explanation bubble with text explaining which data sources were used to compute the insight.



At first, the tool showed the likelihood that a price would go up or down in a very specific way.
For example, a “medium” confidence prediction could say “Prices are unlikely to drop and there’s a 75% chance they’ll increase by $17 in the next 5 days.”



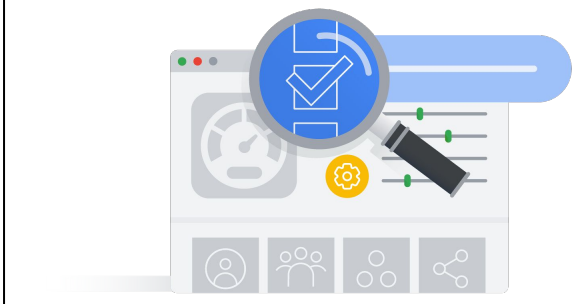
However, this was too much information for a user to process in decision-making.

Display confidence ratings of 90% or higher

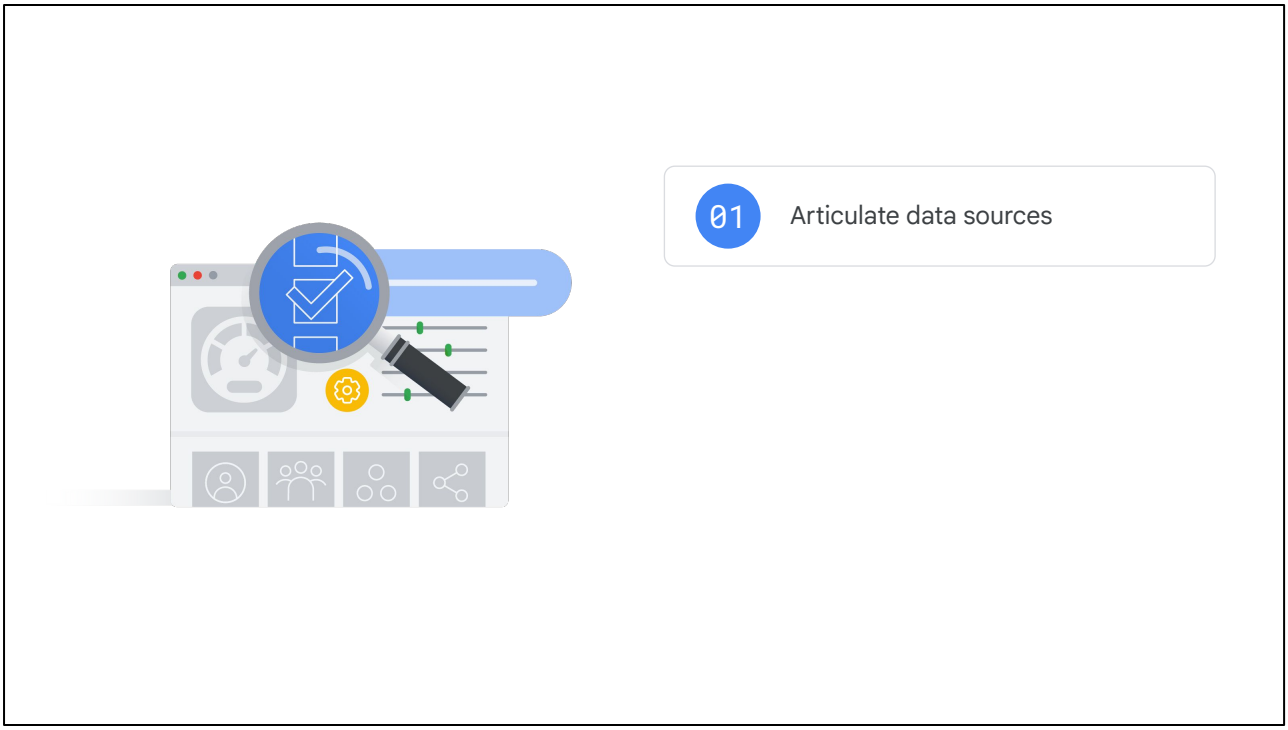
Because “medium” confidence predictions were confusing and not actionable, the decision was made to use a confidence rating of 90% or higher.

- Display confidence ratings of 90% or higher
- Simplify the wording

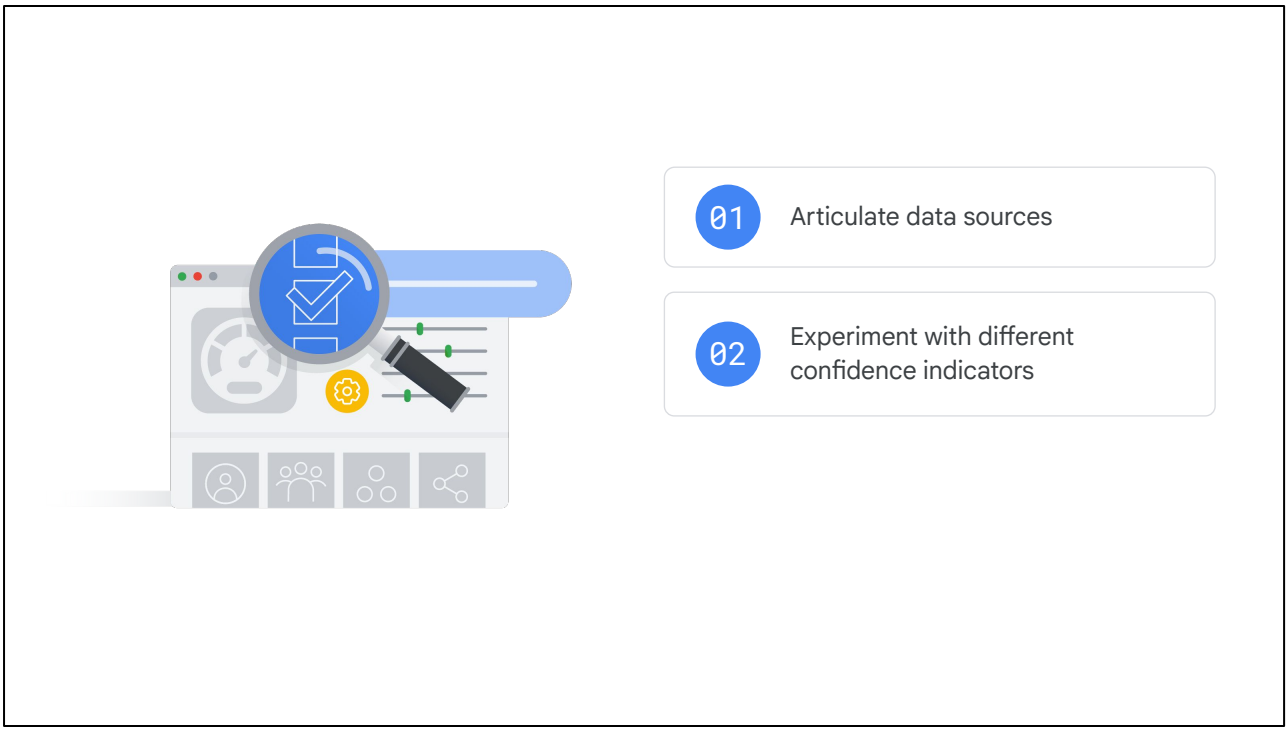
Wording was also changed to say “likely to go up” or “not likely to go down”.



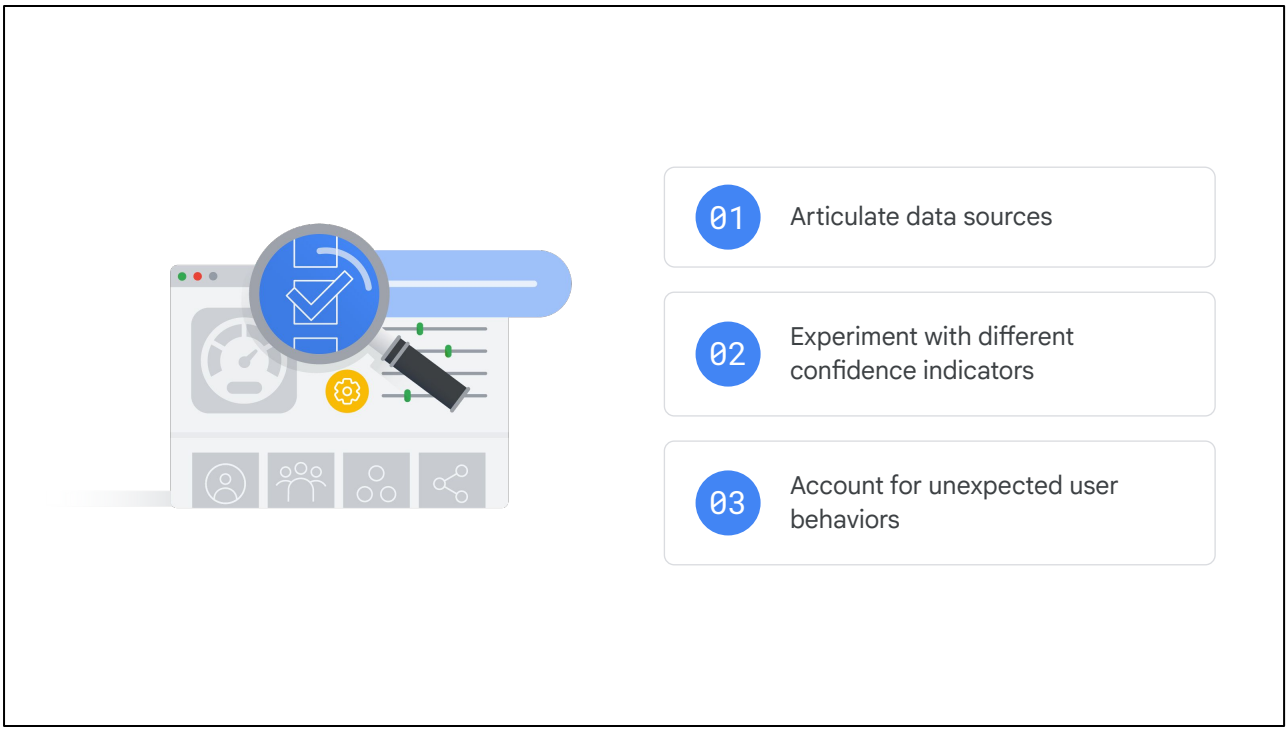
As a result, three strategies were most helpful in the design of price insights in Flights:



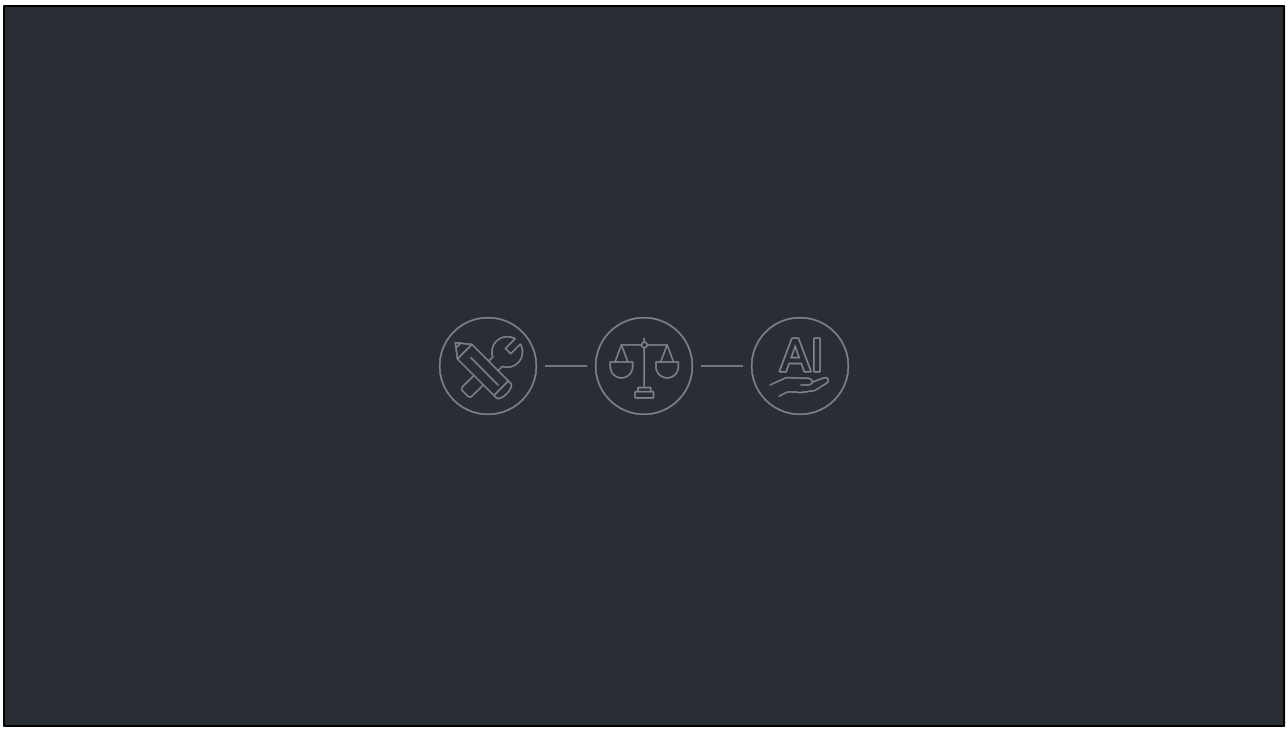
1. Articulate data sources: Telling the user what data was being used in the AI's prediction helped the product team avoid contextual surprises and privacy suspicion, and helped the user know when to apply their own judgment.



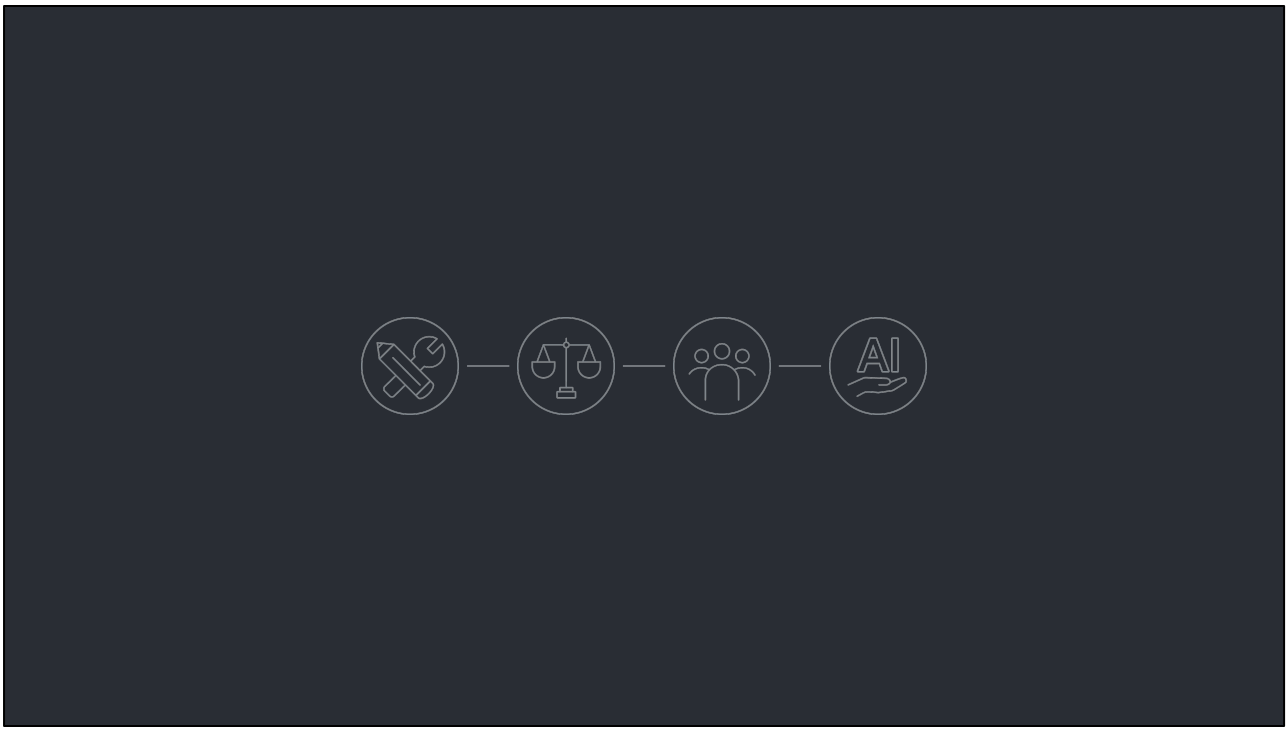
2. Experiment with different confidence indicators: Showing model confidence in categorical buckets and visual graphs helped give users relevant information about flight prices in a way that was easy for them to understand, and



3. Account for unexpected user behaviors: Conducting user research early and frequently helped anticipate any unintended consequences of detailed explanations. This helped the product team change its communications approach and therefore, bolster user trust.



In summary, building machine learning products can be challenging. As you build, applying responsible AI principles throughout your project can profoundly influence a user's trust in both your system, and the system’s machine learning’s usefulness in decision-making.



We all have a role in how responsible AI is applied. As you saw in this case study, whatever stage in the AI process you are involved with, from design to deployment or application, your decisions have an impact. Remember, it's important that you too have a defined and repeatable process for using AI responsibly.