# Explainability + Trust

Explaining predictions, recommendations, and other AI output to users is critical for building trust. This chapter covers:

How much should the user trust the AI system?

When should we provide explanations?

What should we do if we can’t show why the AI made a given prediction?

How should we show users the confidence associated with an AI prediction?

Want to drive discussions, speed iteration, and avoid pitfalls? Use the worksheet.

# What’s new when working with AI

Because AI-powered systems are based on probability and uncertainty, the right level of explanation is key to helping users understand how the system works. Once users have clear mental models of the system’s capabilities and limits, they can understand how and when to trust it to help accomplish their goals. In short, explainability and trust are inherently linked.

In this chapter, we’ll discuss considerations for how and when to explain what your AI does, what data it uses to make decisions, and the confidence level of your model’s output.

Key considerations for explaining AI systems:

① **Help users calibrate their trust**. Because AI products are based on statistics and probability, the user shouldn’t trust the system completely. Rather, based on system explanations, the user should know when to trust the system’s predictions and when to apply their own judgement.

② **Plan for trust calibration throughout the product experience**. Establishing the right level of trust takes time. AI can change and adapt over time, and so will the user’s relationship with the product.

③ **Optimize for understanding**. In some cases, there may be no explicit, comprehensive explanation for the output of a complex algorithm. Even the developers of the AI may not know precisely how it works. In other cases, the reasoning behind a prediction may be knowable, but difficult to explain to users in terms they will understand.

④ **Manage influence on user decisions**. AI systems often generate output that the user needs to act on. If, when, and how the system calculates and shows confidence levels can be critical in informing the user's decision making and calibrating their trust.

## Identify what goes into user trust

Trust is the willingness to take a risk based on the expectation of a benefit.

The following factors contribute to user trust:

- **Ability** is a product's competence to get the job done. Does the product address the user's need, and has it improved their experience? Strive for a product that provides meaningful value that is easy to recognize.
- **Reliability** indicates how consistently your product delivers on its abilities. Is it meeting quality standards according to the expectations set with the user? Only launch if you can meet the bar that you've set and described transparently to the user.
- **Benevolence** is the belief that the trusted party wants to do good for the user. What does the user get out of their relationship with your product, and what do you get out of it? Be honest and up-front about this.

Take the example of an app that can identify plants. A user will calibrate their trust in the app based on their understanding of how well the app can recognize a safe vs. non-safe plant from a photo they've just taken while on a hike in nature, how consistently the app works during different seasons and in different lighting conditions, and finally, how helpful the app is at keeping them safe from plants that they are allergic to.

The process to earn user trust is slow, and it’ll require proper calibration of the user’s expectations and understanding of what the product can and can’t do. In this chapter, you’ll see guidance on when and how to help users set the right level of trust in an AI product.

# ① Help users calibrate their trust

Users shouldn’t implicitly trust your AI system in all circumstances, but rather calibrate their trust correctly. There are many research examples of “algorithm aversion”, where people are suspicious of software systems. Researchers have also found cases of people over-trusting an AI system to do something that it can’t. Ideally, users have the appropriate level of trust given what the system can and cannot do.

For example, indicating that a prediction could be wrong may cause the user to trust that particular prediction less. However, in the long term, users may come to use or rely on your product or company more, because they’re less likely to over-trust your system and be disappointed.

## Articulate data sources

Every AI prediction is based on data, so data sources have to be part of your explanations. However, remember that there may be legal, fairness, and ethical considerations for collecting and communicating about data sources used in AI. We cover those in more detail in the chapter on Data Collection + Evaluation.

Sometimes users can be surprised by their own information when they see it in a new context. These moments often occur when someone sees their data used in a way that appears as if it isn’t private or when they see data they didn’t know the system had access to, both of which can erode trust. To avoid this, explain to users where their data is coming from and how it is being used by the AI system.

Equally important, telling users what data the model is using can help them know when they have a critical piece of information that the system does not. This knowledge can help the user avoid over-trusting the system in certain situations.

For example, say you're installing an AI-powered navigation app, and you click to accept all terms and conditions, which includes the ability for the navigation app to access data from your calendar app. Later, the navigation app alerts you to leave your home in 5 minutes in order to be on time for an appointment. If you didn't read, realize, or remember that you allowed the navigation app to access your appointment information, then this could be very surprising.

Your trust in the app's capabilities depends on your expectations for how it should work and how alerts like these are worded. For instance, you could become suspicious of the app's data sources; or, you could over-trust that it has complete access to all your schedule information. Neither of these outcomes are the right level of trust. One way to avoid this is to explain the connected data source — how the navigation app knows about the appointment — as part of the notification, and to provide the option to opt out of that kind of data sharing in the future. In fact, regulations in some countries may require such specific, contextual explanations and data controls.

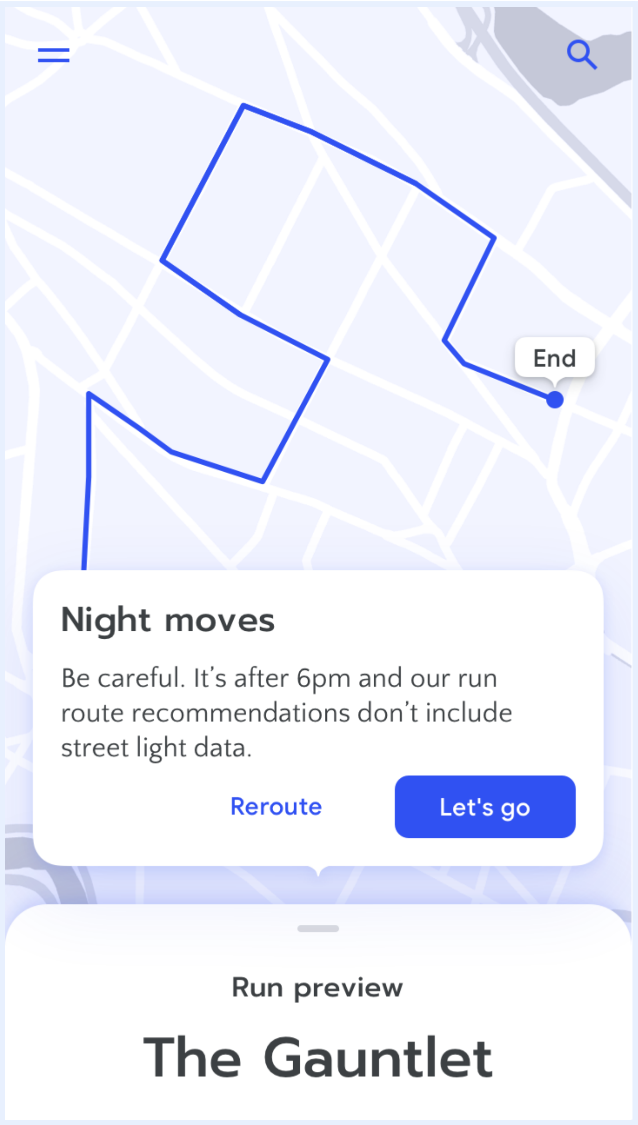
## Key concept

Whenever possible, the AI system should explain the following aspects about data use:

- **Scope**. Show an overview of the data being collected about an individual user, and which aspects of their data are being used for what purpose.
- **Reach**. Explain whether the system is personalized to one user or device, or if it is using aggregated data across all users.

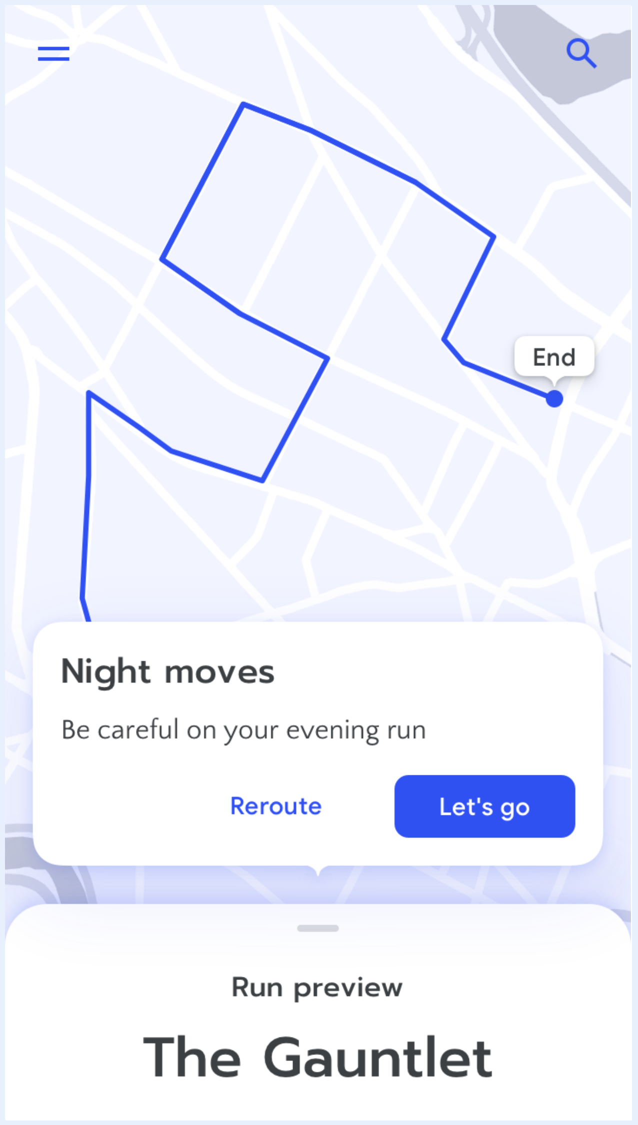
- **Removal**. Tell users whether they can remove or reset some of the data being used.

Apply the concepts from this section in Exercise 1 in the worksheet.



Night moves

Be careful. It's after 6pm and our run route recommendations don't include street light data.



Night moves

Be careful on your evening run

**Aim for**
Tell the user when a lack of data might mean they'll need to use their own judgment.
Learn more

**Avoid**
Don't be afraid to admit when a lack of data could affect the quality of the AI recommendations.

## Tie explanations to user actions

People learn faster when they can see a response to their actions right away, because then it’s easier to identify cause and effect. This means the perfect time to show explanations is in response to a user’s action. If the user takes an action and the AI system doesn’t respond, or responds in an unexpected way, an explanation can go a long way in building or recovering a user’s trust. On the other hand, when the system is working well, responding to users’ actions is a great time to tell the user what they can do to help the system continue to be reliable.

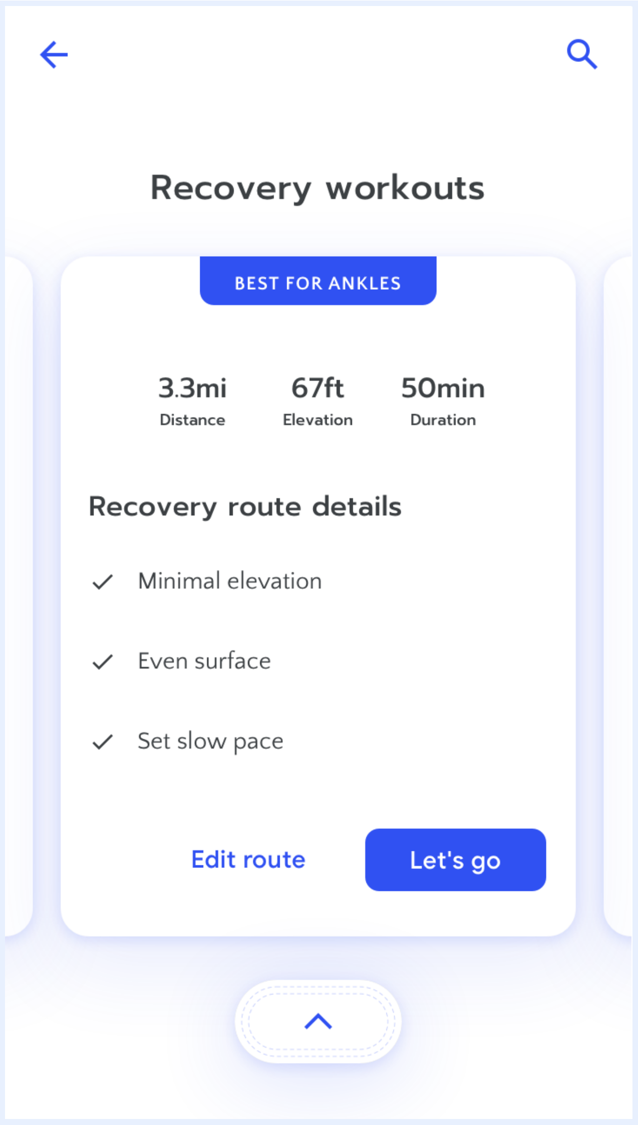
For example, let’s say a user taps on the “recommendations for me” section of an AI-powered restaurant reservation app. They only see recommendations for Italian restaurants, which they rarely visit, so they’re a bit disappointed and less trusting that the app can make relevant, personalized recommendations. If, however, the app’s recommendations include an explanation that the system only recommends restaurants within a one-block area, and the user is standing in the heart of Little Italy in New York City, then trust is likely to be maintained. The user can see how their actions — in this case asking for recommendations in a specific location — affects the system.
Just as you might build trust in another person through back and forth interactions that reveal their strengths and weaknesses, the user’s relationship with an AI system can evolve in the same way.

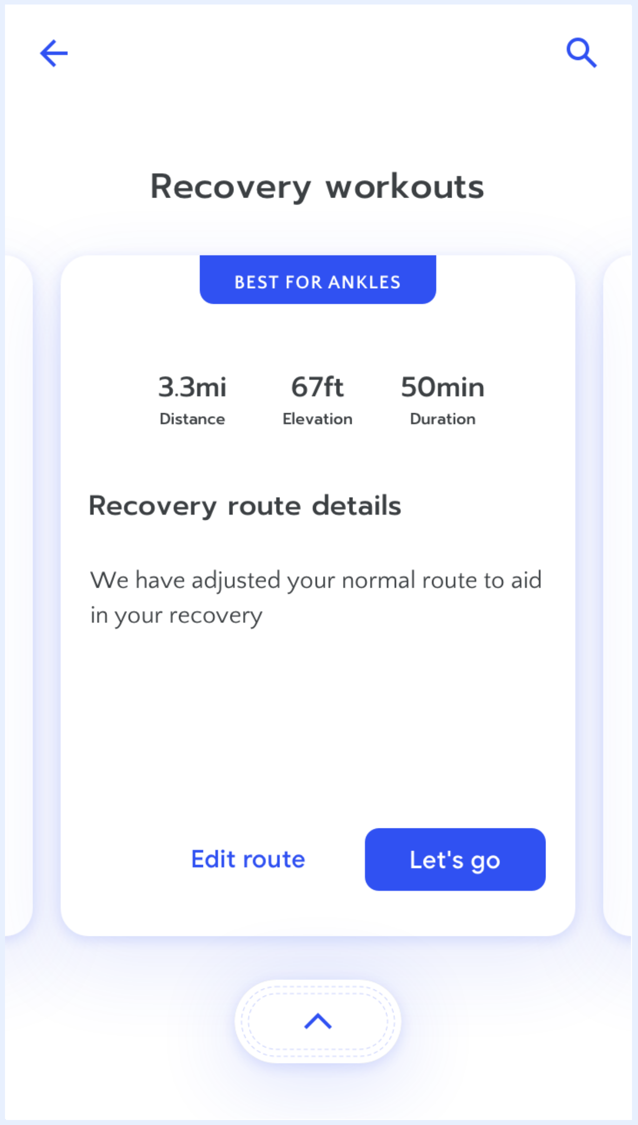
When it’s harder to tie explanations directly to user actions, you could use multi-modal design to show explanations. For example, if someone is using an assistant app with both visual and voice interfaces, you could leave out the explanation in the voice output but include it in the visual interface for the user to see when they have time.

## Account for situational stakes

You can use explanations to encourage users to trust an output more or less depending on the situation and potential consequences. It’s important to consider the risks of a user trusting a false positive, false negative, or a prediction that’s off by a certain percent.

For example, for an AI-powered navigation app, it may not be necessary to explain how arrival time is calculated for a daily commute. However, if someone is trying to catch a flight (higher stakes and less frequent than a commute) they may need to cross-check the timing of the recommended route. In that case, the system could prompt them with an explanation of its limitations. For example, if a user enters the local airport as their destination, let them know that traffic data only refreshes every hour.





We have adjusted your normal route to aid in your recovery

### Aim for

Give the user details about why a prediction was made in a high stakes scenario. Here, the user is exercising after an injury and needs confidence in the app's recommendation. Learn more

### Avoid

Don't say "what" without saying "why" in a high stakes scenario.

You can find detailed information about giving the user appropriate guidance in situations of failure or low-confidence predictions in the Errors + Graceful Failure chapter.

## Key concept

As a team, brainstorm what kinds of interactions, results, and corresponding explanations would decrease, maintain, or inflate trust in your AI system. These should fall somewhere along a trust spectrum of “No trust” to “Too much trust”.
Here are some examples from our running app:

- A user who has never run more than 3 miles at a time receives a recommendation for a marathon training series.
- A user takes the training recommendation to their personal trainer and their trainer agrees with the app’s suggestion.
- A user follows the app’s suggestion for a recovery run, but it’s too difficult for them to complete.

Apply the concepts from this section in Exercise 2 in the worksheet.

# ② Calibrate trust throughout the product experience

The process to build the right level of trust with users is slow and deliberate, and it starts even before users' first interaction with the product.
There are various opportunities to help users set their expectations of the product's abilities and limitations by providing explanations throughout, and outside of, the product experience:

- **Explain in-the-moment**. When appropriate, provide reasons for a given inference, recommendation, suggestion, etc.
- **Provide additional explanations in the product**. Leverage other in-product moments, such as onboarding, to explain AI systems.
- **Go beyond the product experience**. In-product information may not be sufficient, but you can support it with a variety of additional resources, such as marketing campaigns to raise awareness, and educational materials and literacy campaigns to develop mental models.

For example, you may onboard users to an app that identifies plants by explaining some basic, global properties of the model trained to classify plants, such as its known strengths and limitations, and what it's been designed to optimize for. If the model was trained using a dataset made up of 75% of its examples representing plants native to North America, and 25% of its examples coming from South America, this may help users in different locations calibrate their trust. You may also wish to explain to the user if recall was prioritized over precision, as this may mean more conservative recommendations that potentially flag certain harmless plants as unsafe. (For more information on the trade-offs between precision and recall, see the User Needs + Defining Success chapter.)

When the user tries to identify plants by giving the Plant Pal app a photo of a plant, they may then see an explanation that highlights which defining characteristics of the plant led the model to label it by type and safety.

In this section, you'll find guidance for building trust at specific phases of the product experience.

## Establish trust from the beginning

When considering using a product that relies on technology that is new to them, users may have certain concerns. They may wonder what the system can and can't do, how it works, and how they should interact with it. They may also wonder whether they can trust it, especially if they've had less favorable experiences with other products in the past.

To start establishing trust with users before they begin to use the product, some best practices include the following:

- **Communicate the product's capabilities and limitations clearly to set expectations, and do this early**. Focus on the benefit, not the technology. Users may have misconceptions about AI, trusting it too much or too little. For more guidance on how to onboard in stages, see the Mental Models chapter.

- **Highlight what's familiar**. Users look for familiar appearance and legibility, which can contribute to initial trust.
- **Contextualize recommendations with third-party sources**. When onboarding users to a new feature or product, you may benefit from pointing to third-party sources that they already trust to jump-start initial trust in your product.

For example, when a user opens up the running route recommendations page on the Run app for the first time, you can leverage a known and trusted third-party source by surfacing a short list of local running routes that are listed in their local municipality's parks and recreation website. The app may also display a note to the user that more personalized running route recommendations will appear as they start interacting with the app, by marking runs complete, and rating routes.

You'll find additional guidance that will help you calibrate users' trust early on in the Help users calibrate their trust section, earlier in this chapter.

## Grow trust early on

As you onboard users who are new to your product, they'll likely have a new set of concerns. They may want to understand which settings they can edit, especially those controlling privacy and security. They'll also want to understand how to interact with the system, and how they can expect the system to react to their input and feedback.
To build and calibrate users' trust as they get familiar with the AI product:

- **Communicate privacy and security settings on user data**. Explicitly share which data is shared, and which data isn't.

- **Make it easy to try the product first**. Onboarding users with a “sandbox” experience can allow them to explore and test a product with low initial commitment from them.
- **Engage users and give them some control as they get started**. Give them the ability to specify their preferences, make corrections when the system doesn’t behave as they expect, and give them opportunities to provide feedback. Setting the expectation that the system will learn as they teach it can help build trust.

For example, a user getting started with the Run app may wonder how the routes displayed to them were chosen. If you ask them for their geographic location as they’re setting up their account in the app, you’ll want to communicate clearly, and at the time you request the data, that it is the data that will be used to select routes in their vicinity, rather than data from a GPS, or another app. And once you’ve introduced the geographic location setting, make sure that the user can find it when they want to change it. This way, if they ever get tired of running routes in their neighborhood and want to try some further away, they can go back to the location setting and change it.

## Maintain trust

After a user has been onboarded, and they’ve specified their initial preferences and started teaching the product through their interactions with it, you’ll want to make sure to address some common concerns that may come up as the user-product relationship matures. Users will now want to know how to edit their settings, and they may wonder how the system will react to new needs and contexts.

To maintain trust with users as they continue using the product, you may choose to:

- **Progressively increase automation under user guidance**. Allow users to get used to the shift from human to machine control, and make sure that they can provide feedback to guide this process. The key is to take small steps, and make sure that they land well and add value. Automate more when trust is high, or risk of error is low.
- **Continue to communicate clearly about permissions and settings**. Ask users for permissions early. Think about when the user might want to review preferences they've set in the past, and consider reminding them of these settings when they shift into different contexts, and may have different needs. They may also forget what they're sharing and why, so explain the reasons and benefits.

For example, if a Run app user started by setting themselves a daily run time goal of 20 minutes on running routes marked "easy", and they've been reaching this goal every day for several months, you may ask them if they'd like to upgrade the difficulty level of the routes that the app recommends to them.

## Regain or prevent lost trust

During the course of the product experience, users may run into various errors, which you can learn more about in the Errors + Graceful Failure chapter. Because AI is based on probability and uncertainty, it will sometimes get things wrong. At such critical points, users will often be concerned about their ability to signal to the system that an error has occurred, and their ability to continue safely once the error has occurred. The nature of the error and your product's ability to recover from it will impact users' trust. To maintain trust with users as they run into errors while using your product, you can:

- **Communicate with appropriate responses**. Let the users know ahead of time that there is a recovery plan in case something goes wrong, and in the moment, when something doesn't go as expected. Advance knowledge of how the system may behave in such cases can help users make more informed decisions that they feel more comfortable about. For example, if a user purchases an eligible itinerary through an online flight booking service, and the price for the same itinerary on the same booking partner later drops below their purchase price, then the booking service may notify the user that they are eligible for a payout equal to the difference between the lowest price and what they paid.

- **Give users a way forward, according to the severity of possible outcomes**. Your approach should provide a way to deal with the existing error, and learn to prevent it happening again in the product.
    - **Address the error in the moment**: provide a remittance plan that lets users know how the problem will be addressed. For example, if the Run app recommended a running route in a location where the user was vacationing the previous week, but is no longer there, you may notify the user that running recommendations in that city have been removed from their current recommendation list.
    - **Prevent the error from recurring**: give users the opportunity to teach the system the prediction that they were expecting, or in the case of high-risk outcomes, completely shift away from automation to manual control. For additional guidance on balancing control and automation, see the Feedback + Control chapter.

# ③ Optimize for understanding

As described above, explanations are crucial for building calibrated trust. However, offering an explanation of an AI system can be a challenge in and of itself. Because AI is inherently probabilistic, extremely complicated, and making decisions based on multiple signals, it can limit the types of possible explanations.

Often, the rationale behind a particular AI prediction is unknown or too complex to be summarized into a simple sentence that users with limited technical knowledge can readily understand. In many cases the best approach is not to attempt to explain everything – just the aspects that impact user trust and decision-making. Even this can be hard to do, but there are lots of techniques to consider.

## Explain what’s important

Partial explanations clarify a key element of how the system works or expose some of the data sources used for certain predictions. Partial explanations intentionally leave out parts of the system’s function that are unknown, highly complex, or simply not useful. Note that progressive disclosures can also be used together with partial explanations to give curious users more detail.

You can see some example partial explanations below for an AI-powered plant classification app.

## Describe the system or explain the output

General system explanations talk about how the whole system behaves, regardless of the specific input. They can explain the types of data used, what the system is optimizing for, and how the system was trained.

Specific output explanations should explain the rationale behind a specific output for a specific user, for example, why it predicted a specific plant picture to be poison oak. Output explanations are useful because they connect explanations directly to actions and can help resolve confusion in the context of user tasks.

## Data sources

Simple models such as regressions can often surface which data sources had the greatest influence on the system output. Identifying influential data sources for complex models is still a growing area of active research, but can sometimes be done. In cases where it can, the influential feature(s) can then be described for the user in a simple sentence or illustration. Another way of explaining data sources is counterfactuals, which tell the user why the AI did not make a certain decision or prediction.

### Specific output

“This plant is most likely poison oak because it has XYZ features”.

“This tree field guide was created for you because you submit lots of pictures of maple and oak trees in North America”.

“This leaf is not a maple because it doesn’t have 5 points”.

### General system

“This app uses color, leaf shape, and other factors to identify plants”.

# Model confidence displays

Rather than stating why or how the AI came to a certain decision, model confidence displays explain how certain the AI is in its prediction, and the alternatives it considered. As most models can output n-best classifications and confidence scores, model confidence displays are often a readily-available explanation.

## Specific output

**N-best most-likely classifications**

Most likely plant:

- Poison oak
- Maple leaf
- Blackberry leaf

**Numeric confidence level**

Prediction: Poison oak (80%)

## General system

**Numeric confidence level**

This app categorizes images with 80% confidence on average.

Confidence displays help users gauge how much trust to put in the AI output. However, confidence can be displayed in many different ways, and statistical information like confidence scores can be challenging for users to understand. Because different user groups may be more or less familiar with what confidence and probability mean, it's best to test different types of displays early in the product development process. There's more guidance about confidence displays and their role in user experiences in Section 3 of this chapter.

## Example-based explanations

Example-based explanations are useful in cases where it's tricky to explain the reasons behind the AI's predictions. This approach gives users examples from the model's training set that are relevant to the decision being made. Examples can help users understand surprising AI results, or intuit why the AI might have behaved the way it did. These explanations rely on human intelligence to analyze the examples and decide how much to trust the classification.

### Specific output

To help the user decide whether to trust a "poison oak" classification, the system displays most-similar images of poison oak as well as most-similar images of other leaves.

### General system

The AI shows sets of image examples it tends to make errors on, and examples of images it tends to perform well on.

## Explanation via interaction

Another way to explain the AI and help users build mental models is by letting users experiment with the AI on-the-fly, as a way of asking “what if?”. People will often test why an algorithm behaves the way it does and find the system’s limits, for example by asking an AI voice assistant impossible questions. Be intentional about letting users engage with the AI on their own terms to both increase usability and build trust.

### Specific output

A user suspects the system gave too much weight to the leaf color of a bush, which led to a mis-classification.
To test this, the user changes the lighting to yield a more uniform brightness to the bush’s leaves to see whether that changes the classification.

### General system

This type of explanation can’t be used for the entire app generally. It requires a specific output to play with.

It’s important to note that developing any explanation is challenging, and will likely require multiple rounds of user testing. There’s more information on introducing AI systems to users in the chapter on Mental Models.

## Note special cases of absent or comprehensive explanation

In select cases, there’s no benefit to including any kind of explanation in the user interface. If the way an AI works fits a common mental model and matches user expectations for function and reliability, then there may not be anything to explain in the interaction. For example, if a cell phone camera automatically adjusts to lighting, it would be distracting to describe when and how that happens as you’re using it. It’s also wise to avoid explanations that would reveal proprietary techniques or private data. However, before abandoning explanations for these reasons, consider using partial explanations and weigh the impact on user trust.

In other situations, it makes sense, or is required by law, to give a complete explanation — one so detailed that a third party could replicate the results. For example, in software used by the government to sentence criminals, it would be reasonable to expect complete disclosure of every detail of the system. Nothing less than total accountability would be sufficient for a fair, contestable decision.

### Key concept

Think about how an explanation for each critical interaction could decrease, maintain, or increase trust. Then, decide which situations need explanations, and what kind. The best explanation is likely a partial one.

There are lots of options for providing a partial explanation, which intentionally leave out parts of the system's function that are unknown, too complex to explain, or simply not useful. Partial explanations can be:

- **General system**. Explaining how the AI system works in general terms
- **Specific output**. Explaining why the AI provided a particular output at a particular time

Apply the concepts from this section in Exercise 3 in the worksheet.

# ④ Manage influence on user decisions

One of the most exciting opportunities for AI is being able to help people make better decisions more often. The best AI-human partnerships enable better decisions than either party could make on their own. For example, a commuter can augment their local knowledge with traffic predictions to take the best route home. A doctor could use a medical diagnosis model to supplement their historical knowledge of their patient. For this kind of collaboration to be effective, people need to know if and when to trust a system's predictions.

As described in section 2 above, model confidence indicates how certain the system is in the accuracy of its results. Displaying model confidence can sometimes help users calibrate their trust and make better decisions, but it's not always actionable. In this section, we'll discuss when and how to show the confidence levels behind a model's predictions.

## Determine if you should show confidence

It's not easy to make model confidence intuitive. There's still active research around the best ways to display confidence and explain what it means so that people can actually use it in their decision making. Even if you're sure that your user has enough knowledge to properly interpret your confidence displays, consider how it will improve usability and comprehension of the system – if at all. There's always a risk that confidence displays will be distracting, or worse, misinterpreted.

Be sure to set aside lots of time to test if showing model confidence is beneficial for your users and your product or feature. You might choose not to indicate model confidence if:

- **The confidence level isn't impactful**. If it doesn't make an impact on user decision making, consider not showing it. Counterintuitively, showing more granular confidence can be confusing if the impact isn't clear — what should I do when the system is 85.8% certain vs. 87% certain?
- **Showing confidence could create mistrust**. If the confidence level could be misleading for less-savvy users, reconsider how it's displayed, or whether to display it at all. A misleadingly high confidence, for example, may cause users to blindly accept a result.

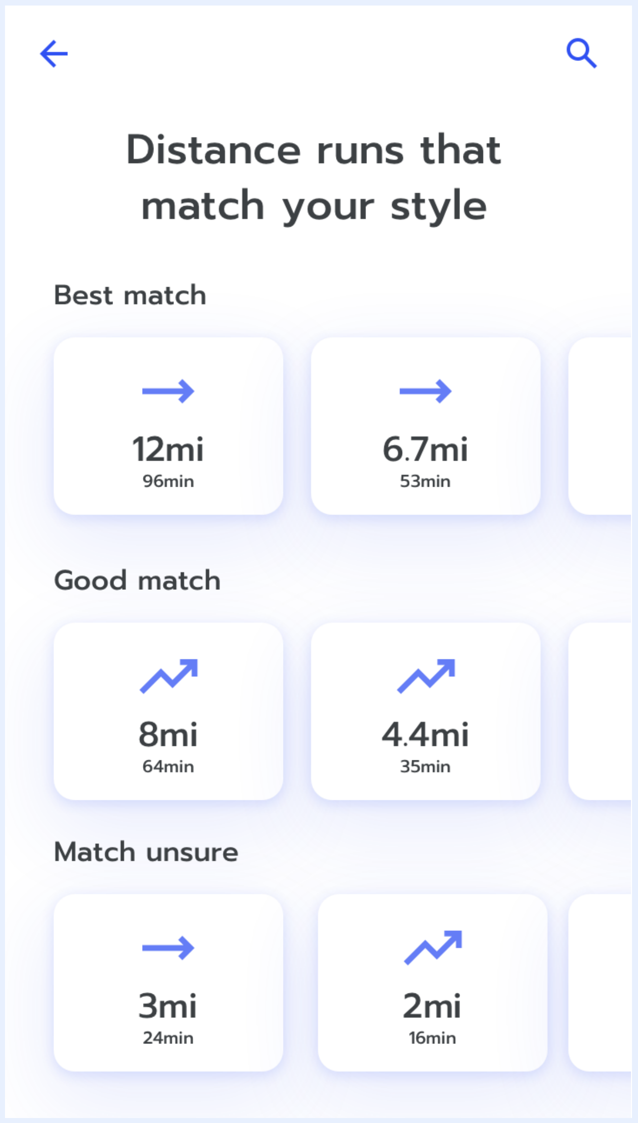
## Decide how best to show model confidence

If your research confirms that displaying model confidence improves decision making, the next step is choosing an appropriate visualization. To come up with the best way to display model confidence, think about what user action this information should inform. Types of visualizations include:

## Categorical

These visualizations categorize confidence values into buckets, such as High / Medium / Low and show the category rather than the numerical value. Considerations:

- Your team will determine cutoff points for the categories, so it's important to think carefully about their meaning and about how many there should be.
- Clearly indicate what action a user should take under each category of confidence.

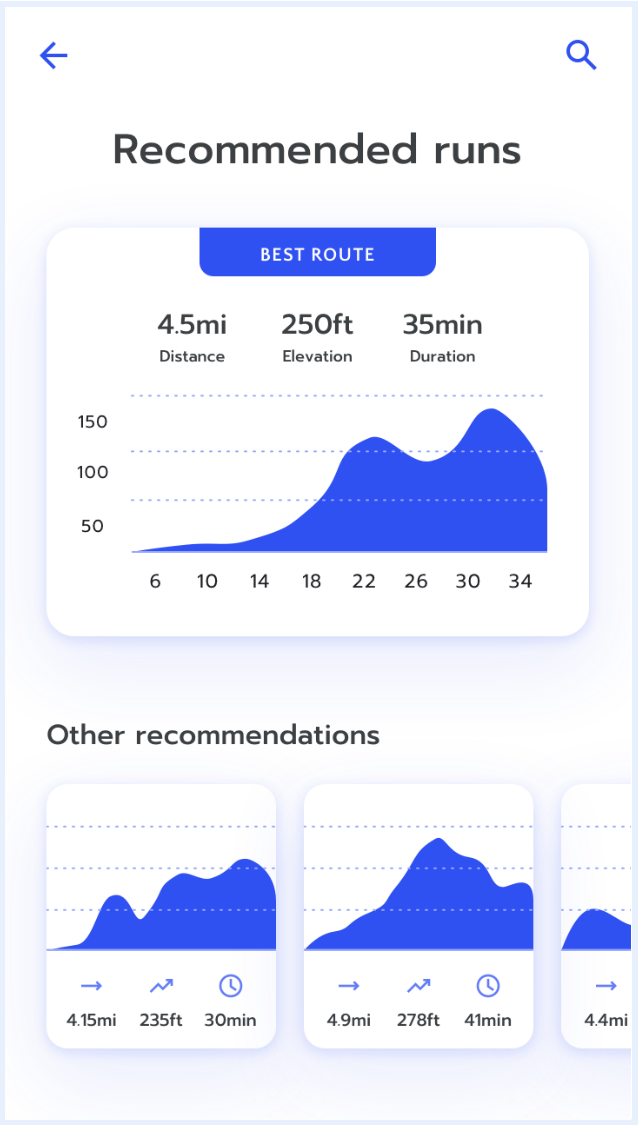


Categorical model confidence visualization. Learn more

## N-best alternatives

Rather than providing an explicit indicator of confidence, the system can display the N-best alternative results. For example, “This photo might be of New York, Tokyo, or Los Angeles.” Considerations:

- This approach can be especially useful in low-confidence situations. Showing multiple options prompts the user to rely on their own judgement. It also helps people build a mental model of how the system relates different options.
- Determining how many alternatives you show will require user testing and iteration.

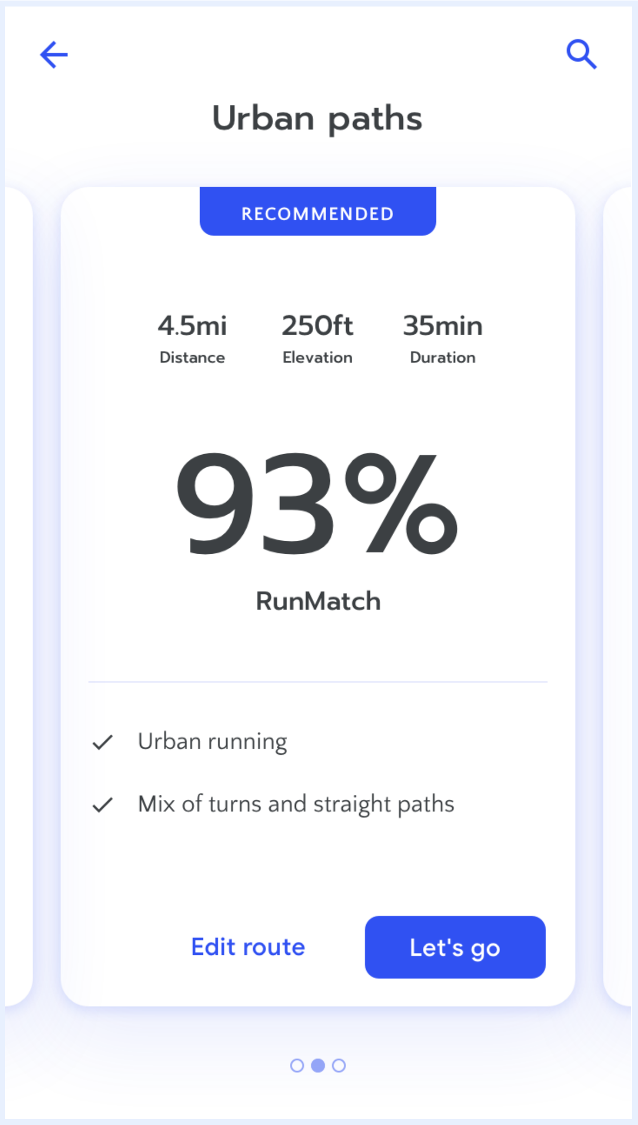


N-best model confidence visualization. Learn more

## Numeric

A common form of this is a simple percentage. Numeric confidence indicators are risky because they presume your users have a good baseline understanding of probability. Additional considerations:

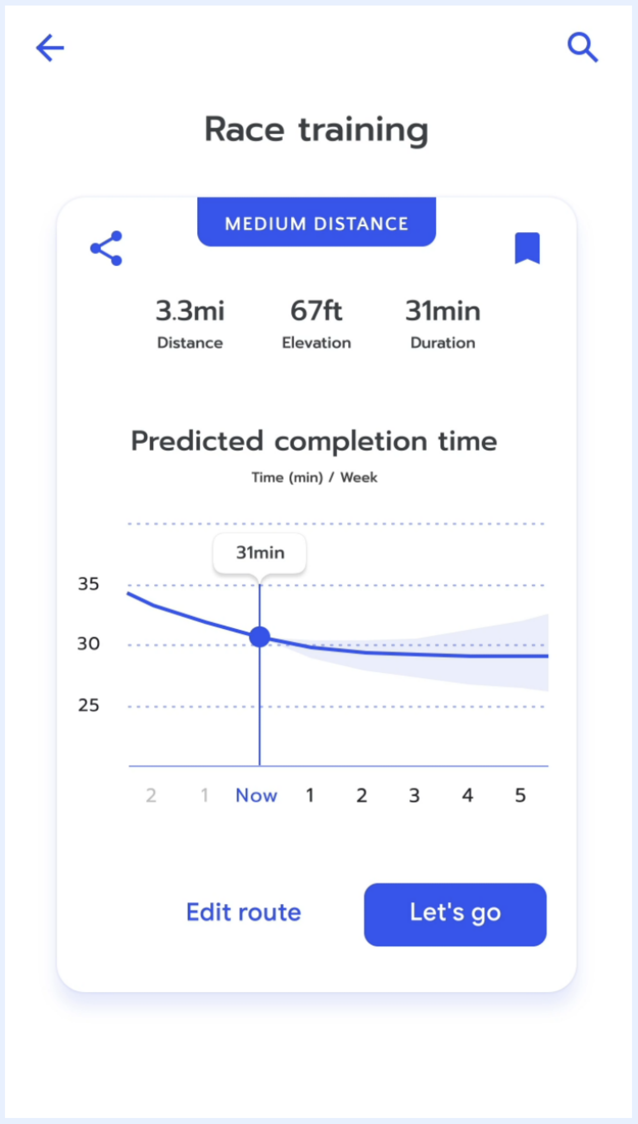
- Make sure to give enough context for users to understand what the percentage means. Novice users may not know whether a value like 80% is low or high for a certain context, or what that means for them.
- Because most AI models will never make a prediction with 100% confidence, showing numeric model confidence might confuse users for outputs they consider to be a sure thing. For example, if a user has listened to a song multiple times, the system might still show it as a 97% match rather than a 100% match.



Numeric model confidence visualization. Learn more

## Data visualizations

These are graphic-based indications of certainty – for example, a financial forecast could include error bars or shaded areas indicating the range of alternative outcomes based on the system's confidence level. Keep in mind, however, that some common data visualizations are best understood by expert users in specific domains.



Data visualization of model confidence. Learn more

## Key concept

To assess whether or not showing model confidence increases trust and makes it easier for people to make decisions, you can conduct user research with people who reflect the diversity of your audience. Here are some examples of the types of questions you could ask:

- “On this scale, show me how trusting you are of this recommendation.”
- “What questions do you have about how the app came to this recommendation?”
- “What, if anything, would increase your trust in this recommendation?”
- “How satisfied or dissatisfied are you with the explanation written here?”

Once you’re sure that displaying model confidence is needed for your AI product or feature, test and iterate to determine what is the right way to show it.
Apply the concepts from this section in Exercise 4 in the worksheet.

# Summary

If and how you offer explanations of the inner-workings of your AI system can profoundly influence the user’s experience with your system and its usefulness in their decision-making. The three main considerations unique to AI covered in this chapter were:

① **Help users calibrate their trust. The** goal of the system should be for the user to trust it in some situations, but to double-check it when needed. Factors influencing calibrated trust are:

- **Articulate data sources**: Telling the user what data are being used in the AI’s prediction can help your product avoid contextual surprises and privacy suspicion and help the user know when to apply their own judgment.

- **Tie explanations to user actions**: Showing clear cause-effect relationships between user actions and system outputs with explanations can help users develop the right level of trust over time.

- **Account for situational stakes**: Providing detailed explanations, prompting the user to check the output in low-confidence/high-stakes situations, and revealing the rationale behind high-confidence predictions can bolster user trust.

② **Calibrate user trust throughout the product experience.** The process to build the right level of trust with users is slow and deliberate, and happens before and throughout the user's interaction with the product. There are a variety of approaches that you can use within and around the product (education, onboarding) to calibrate trust at each stage.

③ **Optimize for understanding.** In some cases, there may be no way to offer an explicit, comprehensive explanation. The calculations behind an output may be inscrutable, even to the developers of those systems. In other cases, it may be possible to surface the reasoning behind a prediction, but it may not be easy to explain to users in terms they will understand. In these cases, use partial explanations.

④ **Manage influence on user decisions.** When a user needs to make a decision based on model output, when and how you display model confidence can play a role in what action they take. There are multiple ways to communicate model confidence, each with its own tradeoffs and considerations.

Want to drive discussions, speed iteration, and avoid pitfalls? Use the worksheet