Durham University

COMP 3647
Human-AI Interaction Design

Topic 11
Trust and Explainability

Prof. Effie L-C Law

Durham University

Trust/Trustworthiness: Human-Human Relationships

Definition of Trust:

the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party.

Definition of Trustworthiness:

The extent to which an actor has the ability to execute relevant tasks, demonstrates integrity, and is benevolent towards fellow team members.

(Mayer et al., 1995; *Annual Management Review*).

Trust in Automation

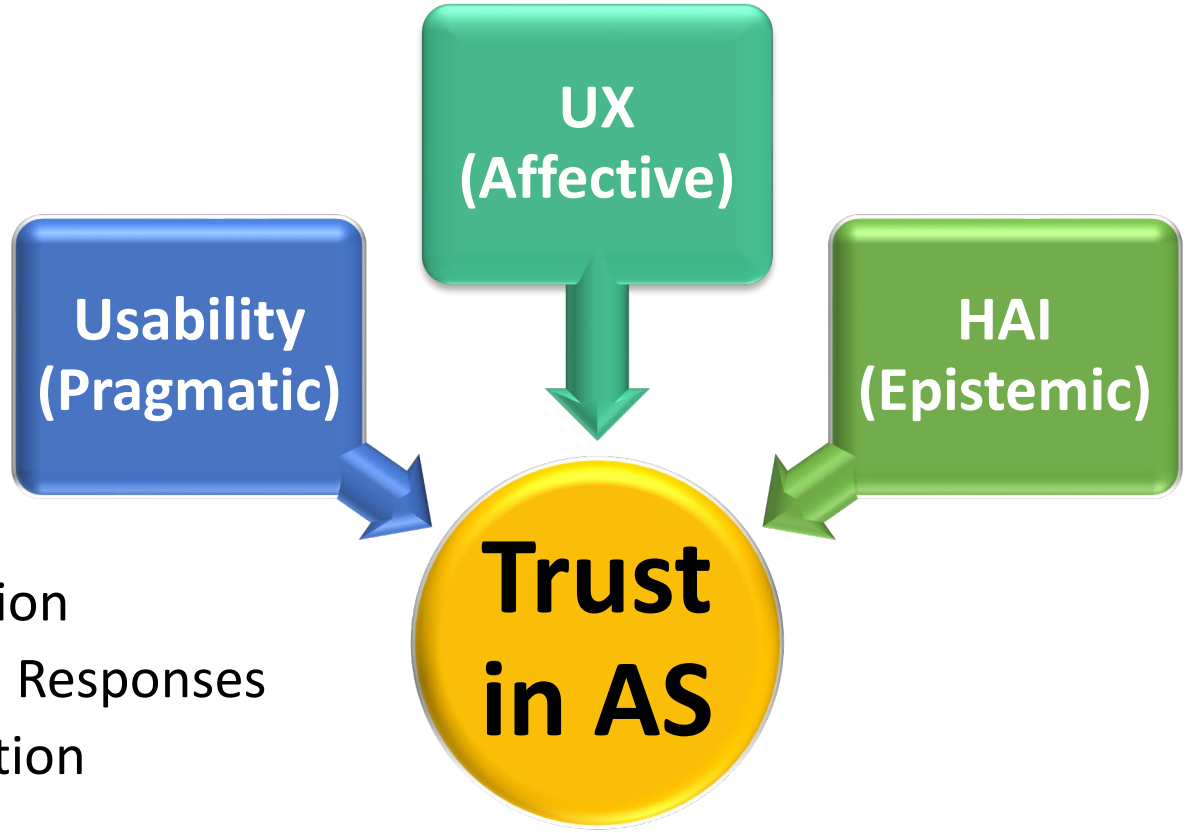
The attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability.

(Lee & See, 2004, Human Factors)

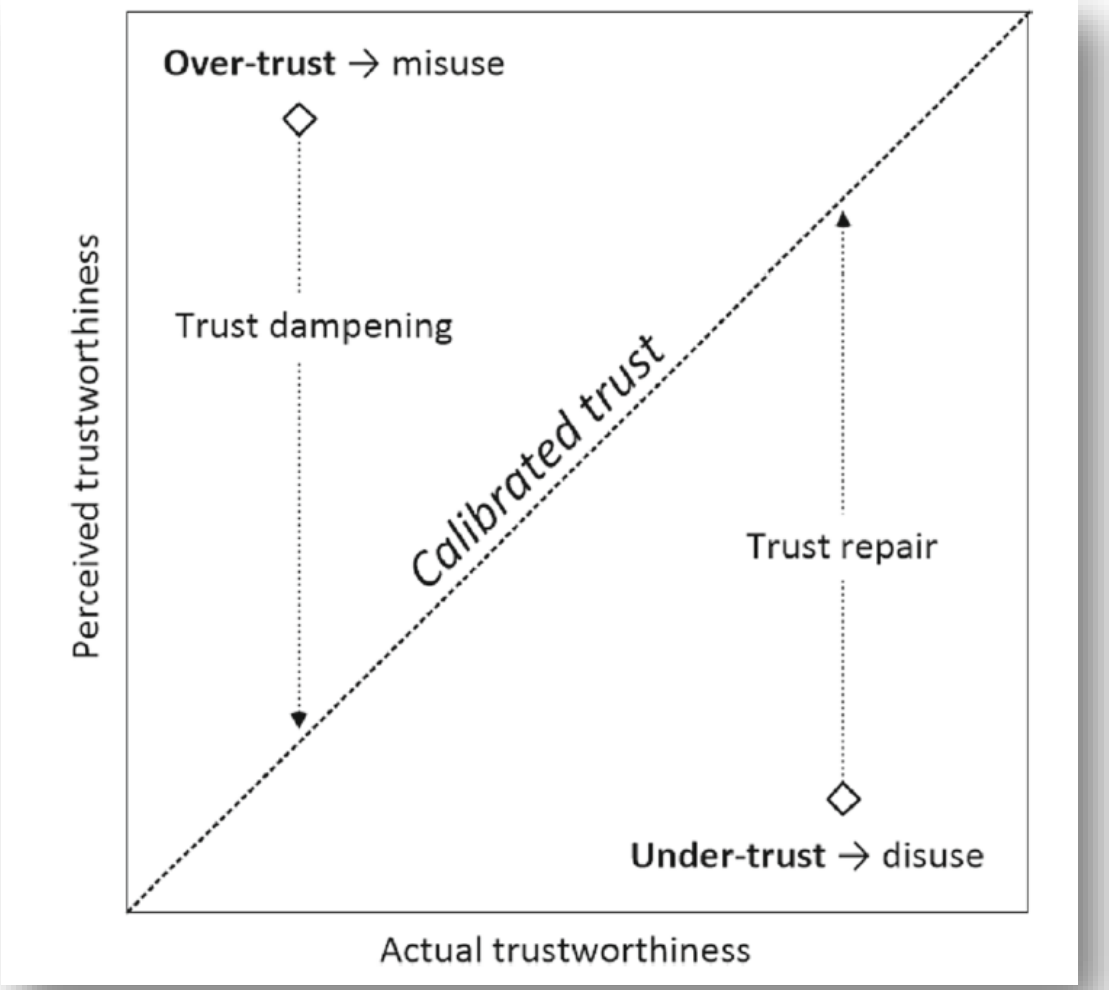
- Trust is NOT: a trait, static
- Trust is: an attitude, dynamic

Three major types of quality:

- **Usability**: Effectiveness, Efficiency, Satisfaction
- **User Experience**: Perceptual and Emotional Responses
- **Human-AI**: Knowledge, Reasoning, Explanation



Calibration of Trust



De Visser et al. (2020). Int'l J. Social Robotics

Undertrust

- Failure to take full advantage of TAS
- Suboptimal solution/performance
- Lack of communication
- Increased workload on both the human and TAS
- Disuse or micromanagement

Overtrust

- Dangerous with disastrous outcomes

- a lack of guidance and control for systems not fully capable of performing a given task.

Trust in Autonomous Systems

Human Factors and Ergonomics Society (HFES) ('Factors of Trust')

- A valid recommendation from a computer is less appreciated if the operator is capable of performing the task on their own (Yang et al. 2016).
- *Operator fatigue* and the *reliability of the technology* determine actual reliance (Wohleber et al., 2016)
- *Cultural differences* on factors such as individualism and power relations manifest as differing tendencies in the trust of automation (Chien et al., 2018)

Process Models of Trust in Autonomous Systems

Human Factors and Ergonomics Society

- **Mathematical instantiations** designed to predict or estimate values or levels of trust for automation-dependent judgments.
- **Conceptual models** depict a process by which trust results from *the causal influence* of mediating variables, and in turn leads to action.

(1) Information ---> (2) Operator's belief ---> (3) Operator's trust --->

(4) Operator's intention ---> (5) Operator's action ---> (6) Operator's action --->

(7) Automation's action ---> (8) Display of resultants ---> Back to (1)

Measurability of Trust

- **Self-reported**
  - Trust Measurement Scale for XAI (Hoffman et al. 2018)
    - Predictability, Reliance, Efficiency, Believability
- **Sensor-based psycho-physiological** (Akash et al. 2018)
  - Trust-Sensor Model: Galvanic skin response (GSR), Electroencephalography (EEG)
- **Temporality of Trust** (Yang et al. 2017)
  - $Trust_{end}$ = Trust measured at the END of a session (***retrospective***)
  - $Trust_{AUTC}$ = Trust measured repeatedly over time DURING a session (*momentary*)

$$Trust_{end} = Trust_T$$

$$Trust_{AUTC} = \frac{1}{T}\sum_1^T Trust_t, \text{ where T = number of interactions}$$

Challenges and Tasks for Human Trust in AS

- **Ontology and Model Development**
  - Providing the vocabulary and semantics of multimodal human-robot team communications
  - A computational model of trust development and repair, prediction of trust violation and impact
- **Trust measurement**
  - Sensor fusion: behavioural, self-reported, observational, psychophysiological
- **Verification and Validation**
  - Empirical studies with end-users
  - Software modules for artificial agents to reason about trust

Transparency

"the quality of an interface pertaining to its ability to afford an operator's comprehension about an intelligent agent's intent, performance, future plans and reasoning process."

(Mercado et al. 2016, Human Factors)

- Convey a history of performance
- Provide performance feedback
- Convey uncertainty directly
- Provide verification methods
- Enhance mode awareness
- Show critical states
- Explain why things fail
- Mimic the social behaviour of a user
- Equip machine with social behaviours

TAS Conveying Uncertainty?

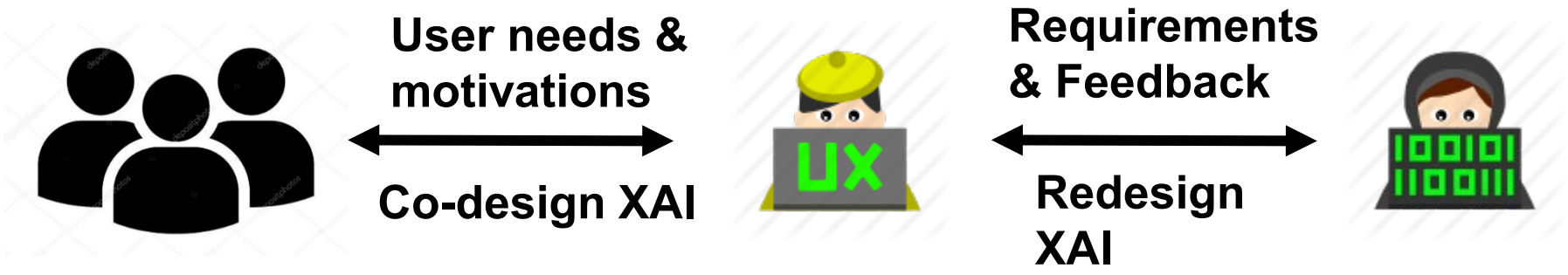
- Conveying automation uncertainty could improve driver-automation interaction; Situation awareness (SA) and trust were higher when uncertainty information was displayed (Beller et al., 2013):

- Users could take over control quicker, but their participants trusted the automation less when shown uncertainty information (Helldin et al. 2013):

Design Principles for HiL

- **Value human values.** Incorporate human preference, taste, and judgement.
- **Continuum rather than all-or-none.** A Big Red Button leaves little room to have fine-grained control. Break up the task to incorporate human interaction.
- **Reduce dependence.** enabling to learn how to use rather than always providing ready-made solutions
- **Take control.** Interpretability

User-centred Explainable AI (XAI)



Types of Explanation methods

Algorithmic model:

- Complexity
- Specificity
- Maturity

User-centred XAI:

Feature importance, Decision-tree, Rules

- Explain the model
- Explain a prediction
- Inspect counterfactual: *What if?*
- Example-based

XAI Question Bank (Liao et al. 2020)

Input

- **What kind of data does the system learn from?**
- What is the source of the data?
- How were the labels/ground-truth produced?
- * What is the sample size?
- * What data is the system NOT using?
- * What are the limitations/biases of the data?
- * How much data [like this] is the system trained on?

Output

- **What kind of output does the system give?**
- What does the system output mean?
- How can I best utilize the output of the system ?
- * What is the scope of the system's capability? Can it do…?
- * How is the output used for other system component(s) ?

Performance

- **How accurate/precise/reliable are the predictions?**
- How often does the system make mistakes?
- In what situations is the system likely to be correct/incorrect?
- * What are the limitations of the system?
- * What kind of mistakes is the system likely to make?
- * Is the system's performance good enough for…

How (global)

- **How does the system make predictions?**
- What features does the system consider?
  - * Is [feature X] used or not used for the predictions?
- What is the system's overall logic?
  - How does it weigh different features?
  - What rules does it use?
  - How does [feature X] impact its predictions?
  - * What are the top rules/features it uses?
- * What kind of algorithm is used?
  - * How are the parameters set?

Why

- **Why/how is this instance given this prediction?**
- What feature(s) of this instance leads to the system's prediction?
- Why are [instance A and B] given the same prediction?

Why not

- **Why/how is this instance NOT predicted…?**
- Why is this instance predicted P instead of Q?
- Why are [instance A and B] given different predictions?

What If

- **What would the system predict if this instance changes to…?**
- What would the system predict if this feature of the instance changes to…?
- What would the system predict for [a different instance]?

How to be that

- **How should this instance change to get a different prediction?**
- How should this feature change for this instance to get a different prediction?
- What kind of instance gets a different prediction?

How to still be this

- **What is the scope of change permitted to still get the same prediction?**
- What is the [highest/lowest/…] feature(s) one can have to still get the same prediction?
- What is the necessary feature(s) present or absent to guarantee this prediction?
- What kind of instance gets this prediction?

Others

- * How/what/why will the system change/adapt/improve/drift over time? (change)
- * How to improve the system? (change)
- * Why using or not using this feature/rule/data? (follow-up)
- * What does [ML terminology] mean? (terminological)
- * What are the results of other people using the system? (social)

User Motivations for Explanation

- Gain further insights or evidence for the AI system's decision
- Evaluate the capability of the AI system
- Adapt usage or interaction behaviours to better utilize the AI
- Improve AI performance
- Reflect on ethical responsibilities

Implications for XAI design

- Explanations are **selective** and **social** in nature
- Explanation delivery is ***interactive*** and ***conversational***
- Trade-off between XAI and UX – interrupting workflow

TAS– Trust Relationships (Hoffman et al. 2018)

- **Absolute Trusting** is when the user takes the computer's assertions (data, claims) as valid and true in all circumstances.
- **Contingent Trusting** is when the user can take some of the computer's presentations or assertions as valid and true under certain circumstances.
- **Progressive Trusting** is when the user takes more of the machine's presentations or assertions as valid and true over time or across experiences.
- **Digressive Trusting** is when the user takes fewer of the machine's presentations or assertions as valid and true over time and across experiences.

TAS– Mistrust Relationships (Hoffman et al. 2018)

- **Mistrusting** is the belief that the computer might do things that are not in the user's interest.
- **Distrusting** is the belief that the computer may or may not do things that are in the user's interest.
- **Anti-trusting** is the belief that the computer will do things that are not in the user's interest.
- **Counter-trusting** is the belief that the computer must not be relied upon because the machine is presenting information that suggests it should be trusted.

TAS– Measuring Trust

A trust scale can have two dimensions:

Trust:

- *Do you trust the machine's outputs?*
- *How much do you trust?*
- *Under which condition do you trust it?*

Reliance

- *Would you follow the machine's advice?*
- *Does it act as part of a team?*
- *Is it friendly?*
- *Are you prepared to rely on it?*

Durham University

Measuring Trust - Madsen-Gregor Scale 1st Subscale

Perceived Reliability

- The system always provides the advice I require to make my decision.
- The system performs reliably.
- The system responds the same way under the same conditions at different times.
- I can rely on the system to function properly.
- The system analyses problems consistently.

Measuring Trust - Madsen-Gregor Scale 2nd Subscale

Perceived Technical Competence

- The system uses appropriate methods to reach decisions.
- The system has sound knowledge about this type of problem built into it.
- The advice the system produces is as good as that which a highly competent person could produce.
- The system correctly uses the information I enter.
- The system makes use of all the knowledge and information available to
- it to produce its solution to the problem

Measuring Trust - Madsen-Gregor Scale 3rd Subscale

Perceived Understandability

- I know what will happen the next time I use the system because I understand how it behaves.
- I understand how the system will assist me with decisions I have to make.
- Although I may not know exactly how the system works, I know how to use it to make decisions about the problem.
- It is easy to follow what the system does.
- I recognize what I should do to get the advice I need from the system the next time I use it.

Measuring Trust - Madsen-Gregor Scale 4th Subscale

Faith

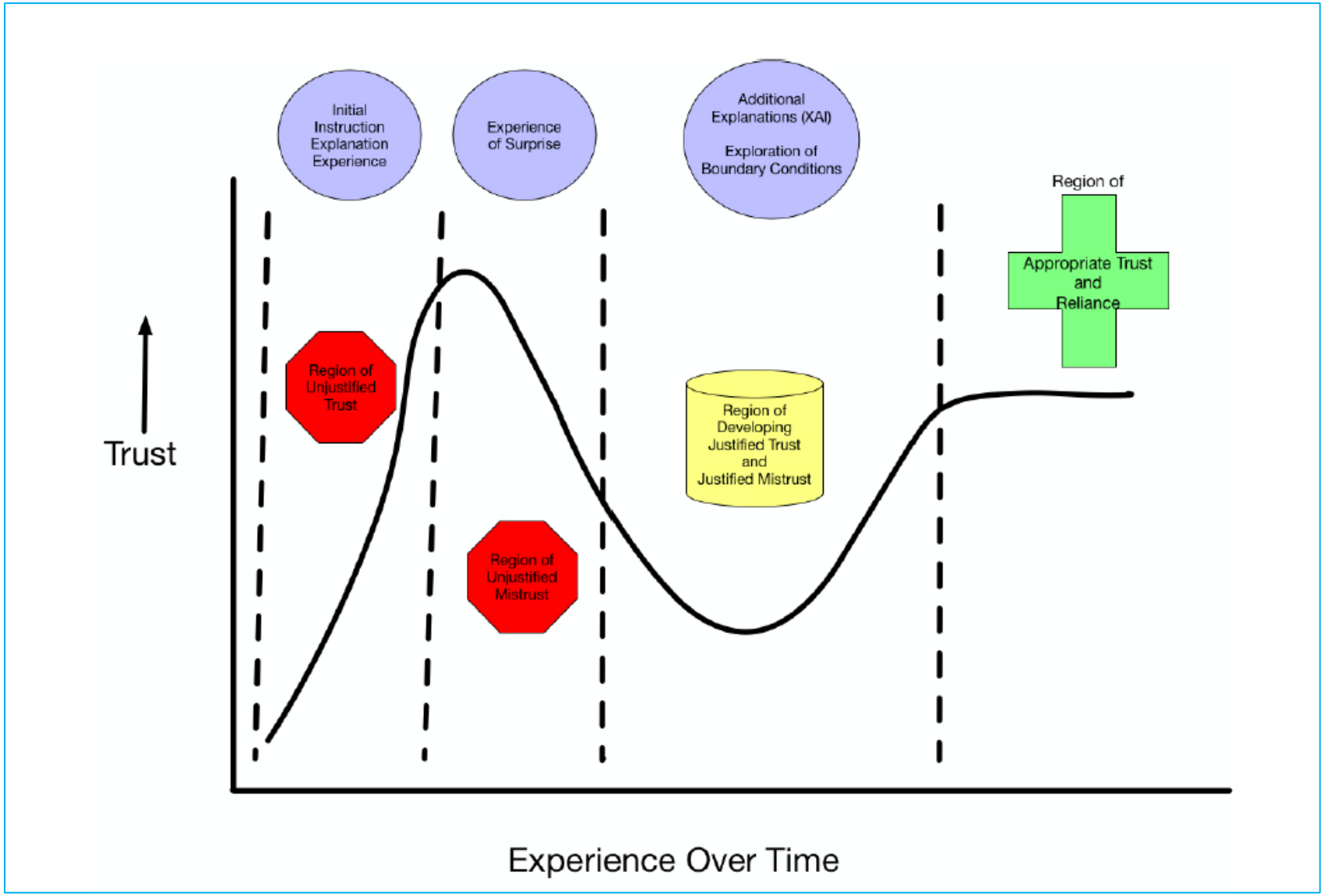
- I believe advice from the system even when I don't know for certain that it is correct.
- When I am uncertain about a decision I believe the system rather than myself.
- If I am not sure about a decision, I have faith that the system will provide the best solution.
- When the system gives unusual advice I am confident that the advice is correct.
- Even if I have no reason to expect the system will be able to solve a difficult problem, I still feel certain that it will.

Measuring Trust - Madsen-Gregor Scale 5th Subscale

Personal Attachment

- I would feel a sense of loss if the system was unavailable and I could no longer use it.
- I feel a sense of attachment to using the system.
- I find the system suitable to my style of decision making.
- I like using the system for decision making.
- I have a personal preference for making decisions with the system.

TAS – Trust in XAI Context



Durham University

Trust and Explainability

As 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.

Explainability and **trust** are inherently linked!

- How and when to explain what AI does
- What data it uses to make decisions
- The confidence level of model output

Key considerations for explaining AI systems

- Identify what goes into user trust
- Help users calibrate their trust
- Calibrate trust throughout the product experience
- Optimise for understanding
- Manage influence on user decisions

Identify what goes into user trust

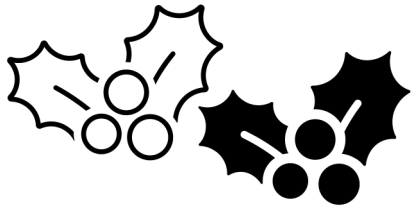
Trust	the willingness to take a risk based on the expectation of a benefit.

Contributing factors:

- **Ability**: a product's competence to get the job done.
  Strive for a product that provides meaningful value that is easy to recognize.
- **Reliability**: how consistently a product delivers on its abilities.
  Only launch if we can meet the bar set and described transparently to the user.
- **Benevolence**: the belief that the trusted party wants to do good for the user
  Be honest and up-front about this.

Example

An app that can identify plants…



A user will **calibrate** their trust in the app based on

- understanding of how well the app can recognise 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
- 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.

Help users calibrate their trust

Users shouldn't implicitly trust AI system in all circumstances, but rather **calibrate** their trust correctly.

- "algorithm aversion", where people are suspicious of AI systems
- over-trusting an AI system to do something that it cannot

Ideally, users have the appropriate level of trust given what the system can & cannot do.

Indicating a prediction could be wrong may cause the user to trust that particular prediction less, but, in the long term, users may come to use or rely on our product more, because they're less likely to over-trust our system and be disappointed.

Help users calibrate their trust

1. Articulate data sources

Data sources have to be part of explanations.

there may be legal, fairness, and ethical considerations for collecting and communicating about data sources used in AI.

Users surprised by their own info when seeing it in a new context.

Seeing their data used in a way that appears as if it isn't private or seeing data they didn't know the system had access to -> **erode trust**. To avoid this:

- Explain to users where their data is coming from and how it is being used.

Tell users what data the model is using can help them know when they have a critical piece of info that the AI does not.

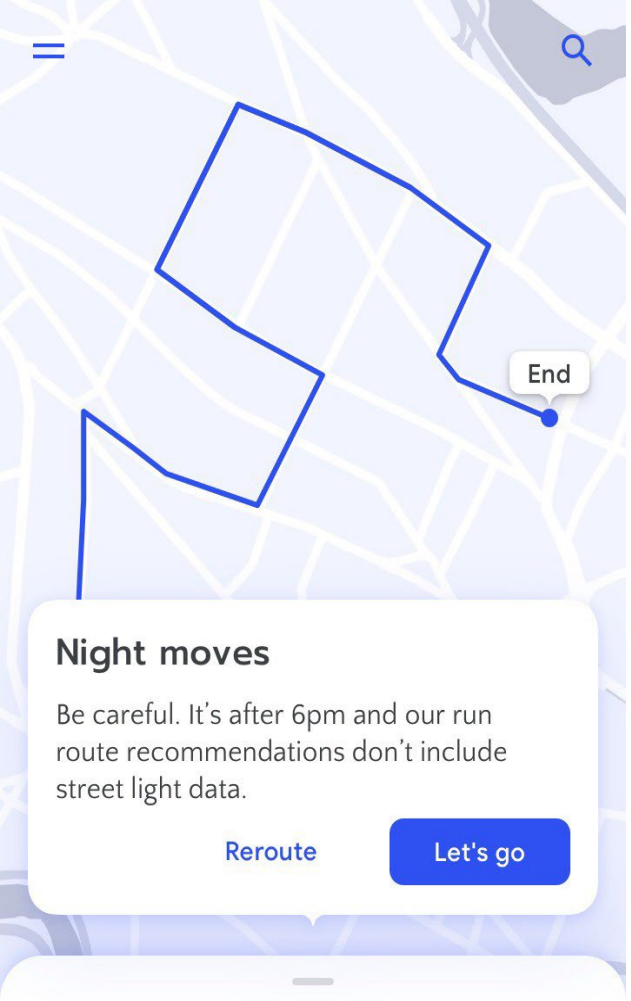
- Can help the user avoid over-trusting the system in certain situations.

Help users calibrate their trust

1. Articulate data sources

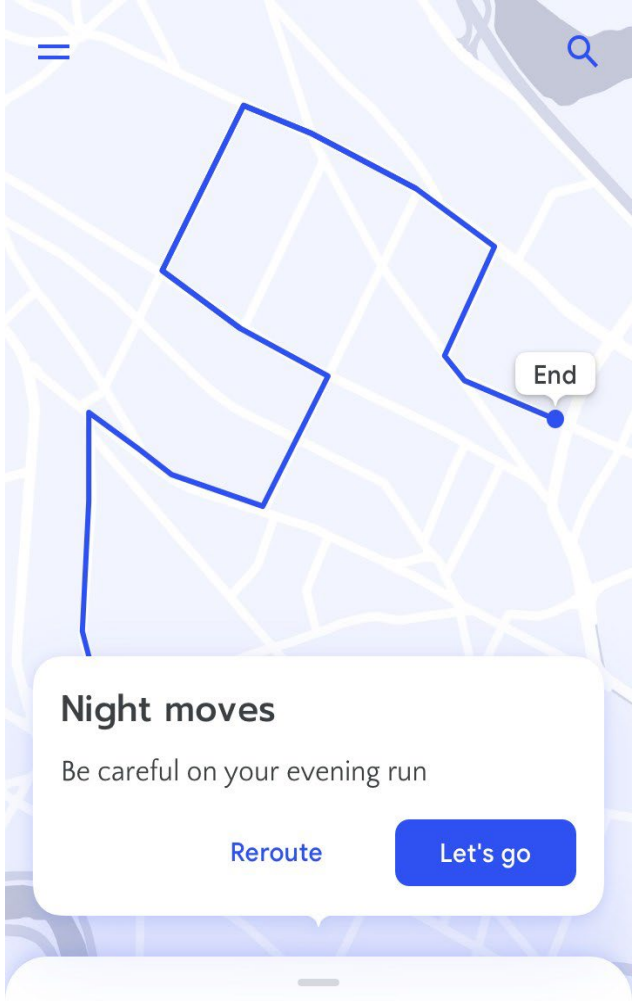
AI system should explain the following aspects about data use:

- **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 personalised 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.



Aim for

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



Avoid

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

Help users calibrate their trust

2. 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**.

The perfect time to show explanations is in response to a user's action.

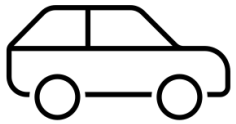
- Not respond or respond in an unexpected way -> an explanation can go a long way in building or recovering a user's trust.
- When the system is working well, responding to users' actions is a great time to tell the user what they can do to help the system continue to be reliable.

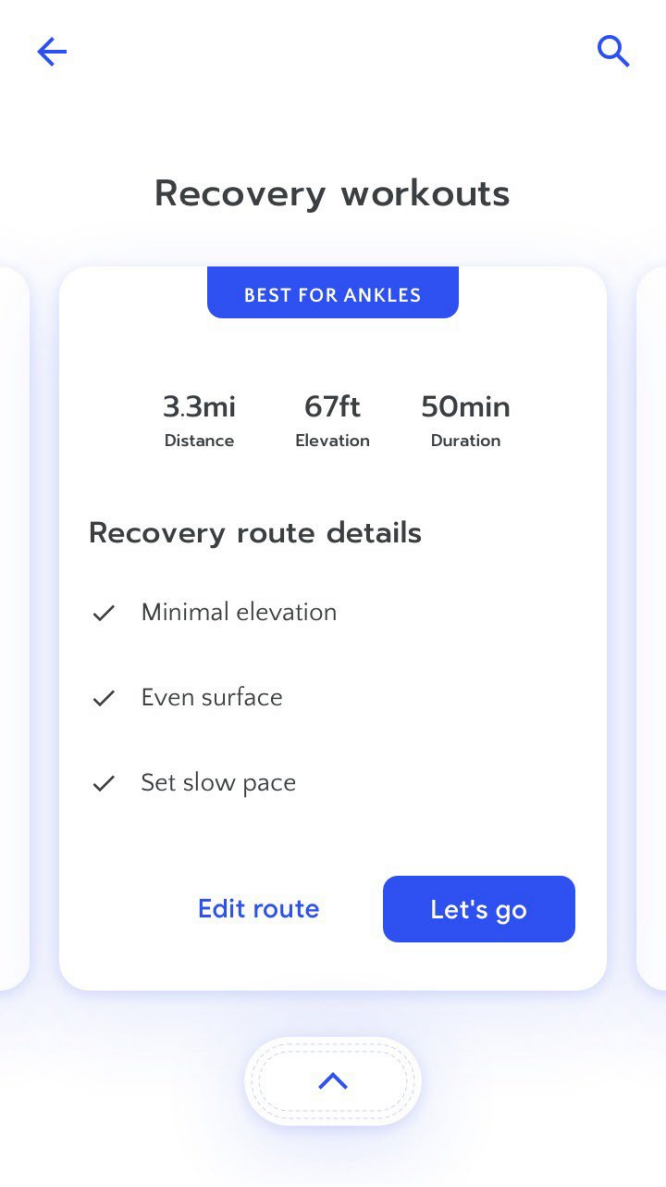
Help users calibrate their trust

3. Account for situational stakes

Use explanations to encourage users to trust an output more or less depending on the situation and potential consequences.

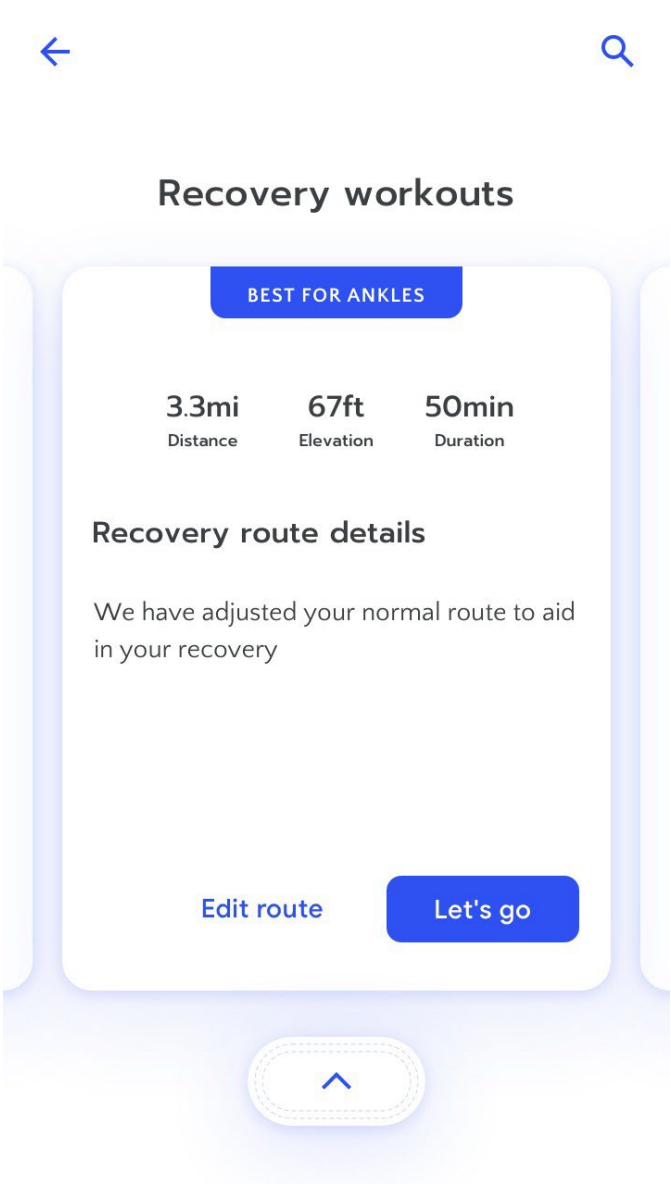
Consider risks of a user trusting a false positive, false negative, or a prediction that’s off by a certain percent.





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.



Avoid

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

Help users calibrate their trust

Example: Running app

As a team, brainstorm what kinds of interactions, results, and corresponding explanations would decrease, maintain, or inflate trust in the AI system. These should fall somewhere along a trust spectrum of “No trust” to “Too much trust”.

- 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 app’s suggestion for a recovery run, but it’s too difficult to complete. ✗

Calibrate trust throughout the product experience

To help users set 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 we 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.

Calibrate trust throughout the product experience

Guidance for building trust at specific phases of the product experience

1. Establish trust from the beginning
2. Grow trust early on
3. Maintain trust
4. Regain or prevent lost trust

Calibrate trust throughout the product experience

1. Establish trust from the beginning

To user a new product? Users may have certain concerns…

what the system can and can't do, how it works, how they should interact with it. can be trusted, etc.

Best practices to start establishing trust with users before they use the product:

- Communicate the product's capabilities and limitations clearly to set expectations, and do this early.
- Highlight what's familiar.
- Contextualise recommendations with third-party sources.

Calibrate trust throughout the product experience

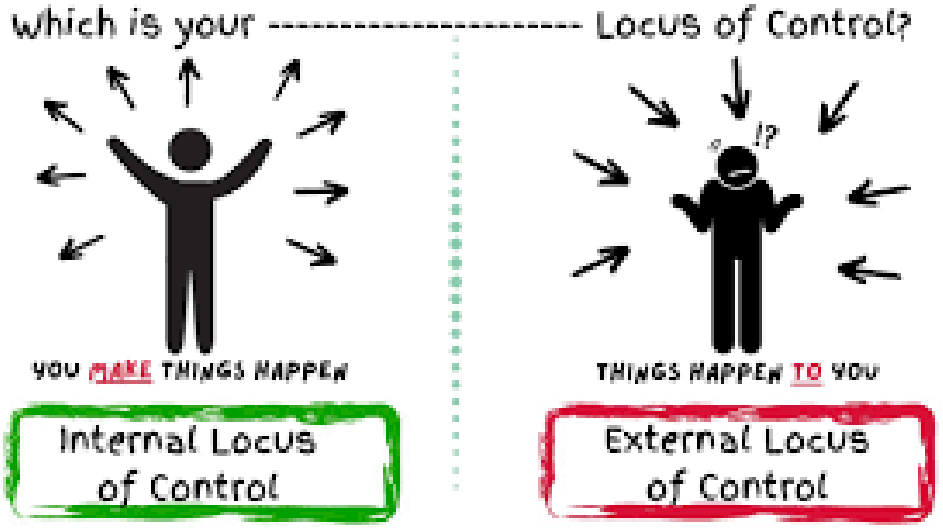
2. Grow trust early on

When onboarding to product, users will likely have a new set of concerns…

- want to understand which settings they can edit, especially those controlling privacy and security.
- want to understand 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.
- Make it easy to try the product first.
- Engage users and give them some control as they get started.



Calibrate trust throughout the product experience

2. 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, we'll want to make sure to address some common concerns that may come up as the user-product relationship matures.

To maintain trust with users as they continue using the product:

- Progressively increase automation under user guidance.
- Continue to communicate clearly about permissions and settings.

Calibrate trust throughout the product experience

3. Regain or prevent lost trust

During the course of the product experience, users may run into various errors. The nature of the error and a product's ability to recover from it will impact users' trust.

To maintain trust with users as they run into errors while using product:

- Communicate with appropriate responses.
- Give users a way forward, according to the severity of possible outcomes.
  - Address the error in the moment
  - Prevent the error from recurring

Optimise for understanding

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

In many cases the best approach is not to attempt to **explain everything** – just the aspects that impact user trust and **decision-making**.

Techniques to consider:

1. Explain what's important
2. Describe the system or explain the output
3. Data sources
4. Model confidence displays
5. Example-based explanations
6. Explanation via interaction

Optimise for understanding

1. 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.
- Intentionally leave out parts of the system's function that are unknown, highly complex, or simply not useful.
- Progressive disclosure can be used together with partial explanations to give curious users more details.

Optimise for understanding

2. Describe the system or explain the output

General system explanations

- Describe how the whole system behaves, regardless of the specific input.
- Can explain the types of data used, what the system is optimising for, and how the system was trained.

Specific output explanations

- Should explain the rationale behind a specific output for a specific user.
- Output explanations are useful because they connect explanations directly to actions and can help resolve confusion in the context of user tasks.

Optimise for understanding

3. 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 and then describe the influential feature(s) 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.

Optimise for understanding

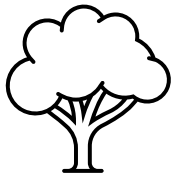
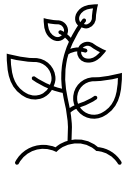
3. Data sources

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 colour, leaf shape, and other factors to identify plants”.

Optimise for understanding

4. Model confidence displays

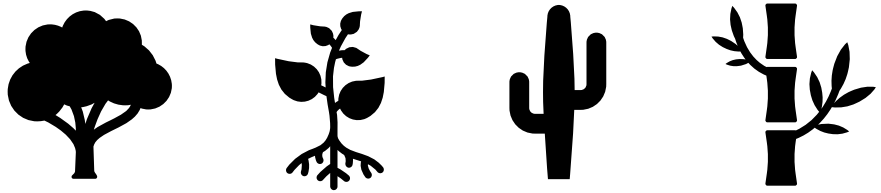
Rather than stating why or how the AI came to a certain decision, model confidence displays explain how certain the AI is in its prediction, and the alternatives it considered.

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 categorises images with 80% confidence on average



Durham University

Optimise for understanding

4. Model confidence displays

Confidence displays help users gauge how much trust to put in the AI output

Confidence can be displayed in many different ways, and statistical information like confidence scores can be challenging for users to understand.

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.

Optimise for understanding

5. Example-based explanations

When it's tricky to explain the reasons behind the AI's predictions, give users examples from the model's training set that are relevant to 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 analyse the examples and decide how much to trust the classification.

Specific output

Displays most-similar images of poison oak and most-similar images of other leaves.

General system

Shows sets of image examples it tends to make errors on / perform well on.

Optimise for understanding

6. Explanation via interaction

Help users build mental models by letting them experiment with the AI on-the-fly.

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.

Specific output

- Does the system give too much weight to the leaf colour of a bush, which led to a misclassification? 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

- Can't be used for the entire app generally. It requires a specific output to play with.

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.

Displaying model confidence to help users calibrate their trust and make better decisions, but it's not always actionable.

When and how to show the confidence levels behind a model's predictions?

1. Determine if we should show confidence
2. Decide how best to show model confidence

Manage influence on user decisions

1. Determine if you should show confidence

Not easy to make model confidence intuitive…

- Even if we're sure that our user has enough knowledge to properly interpret confidence displays, consider how it will affect usability and comprehension of the system.
- There is always a risk that confidence displays is distracting / misinterpreted.

Test if showing model confidence is beneficial for users and the product or feature.

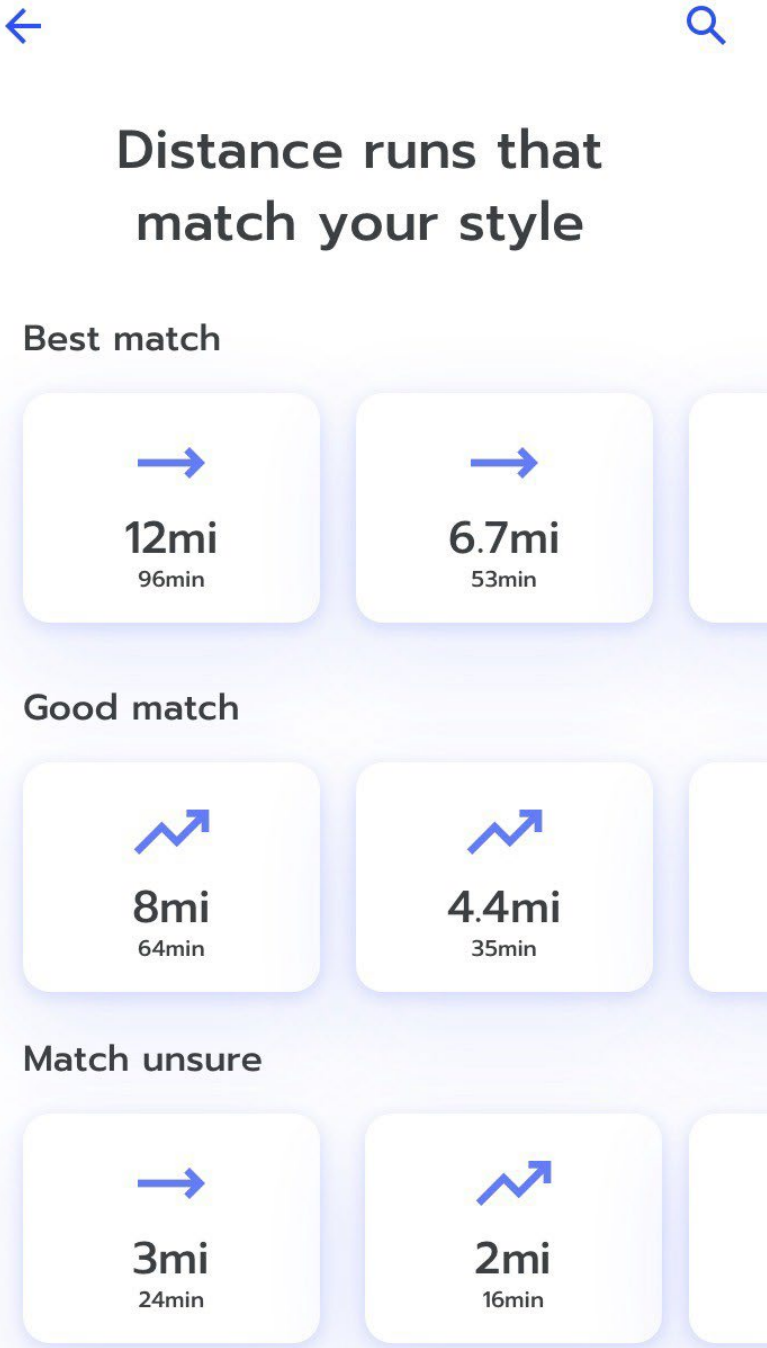
Choose NOT to indicate model confidence if:

- The confidence level isn't impactful.
- Showing confidence could create mistrust.

Manage influence on user decisions

2. Decide how best to show model confidence

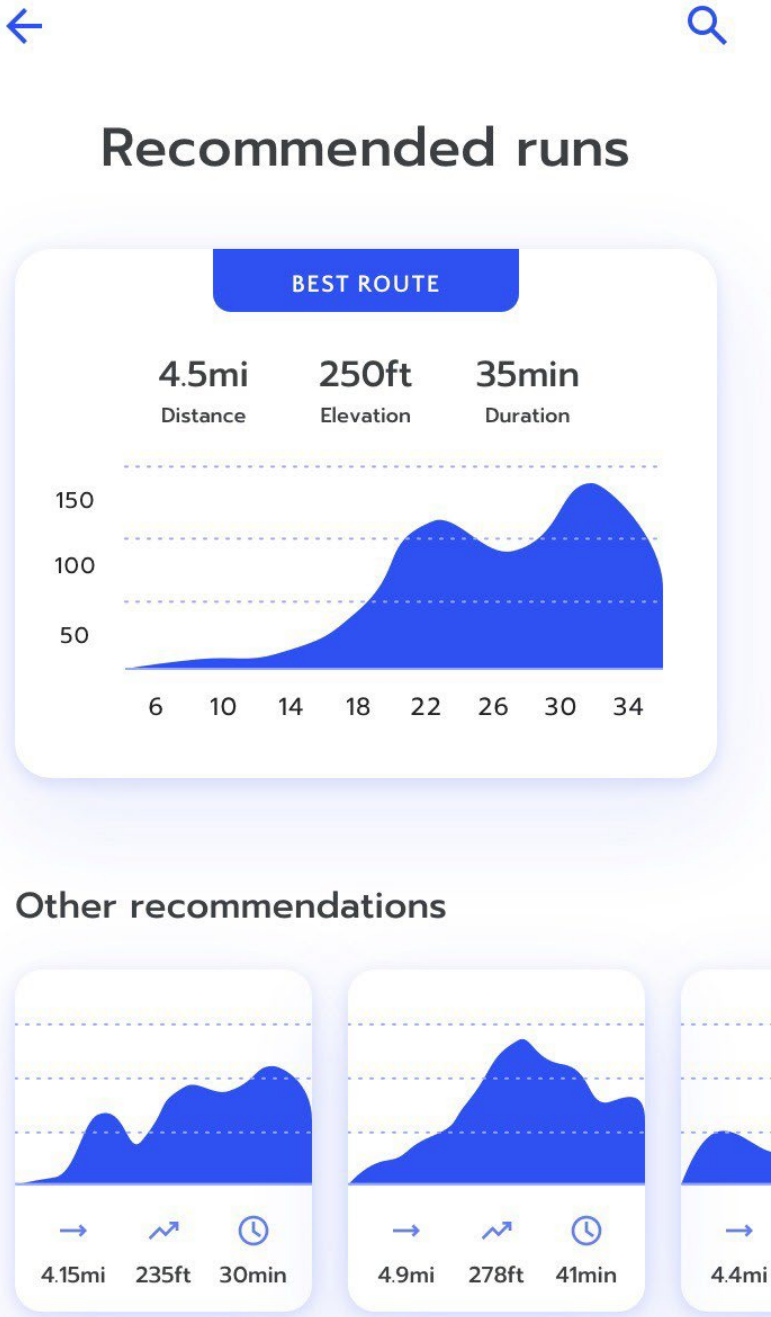
- **Categorical**. These visualisations categorise confidence values into buckets, such as <u>High / Medium / Low</u> and show the category rather than the numerical value.
  - Determine cut-off points for the categories. Think about their meanings & how many there should be.
  - Clearly indicate what action a user should take under each category of confidence.



Manage influence on user decisions

2. Decide how best to show model confidence

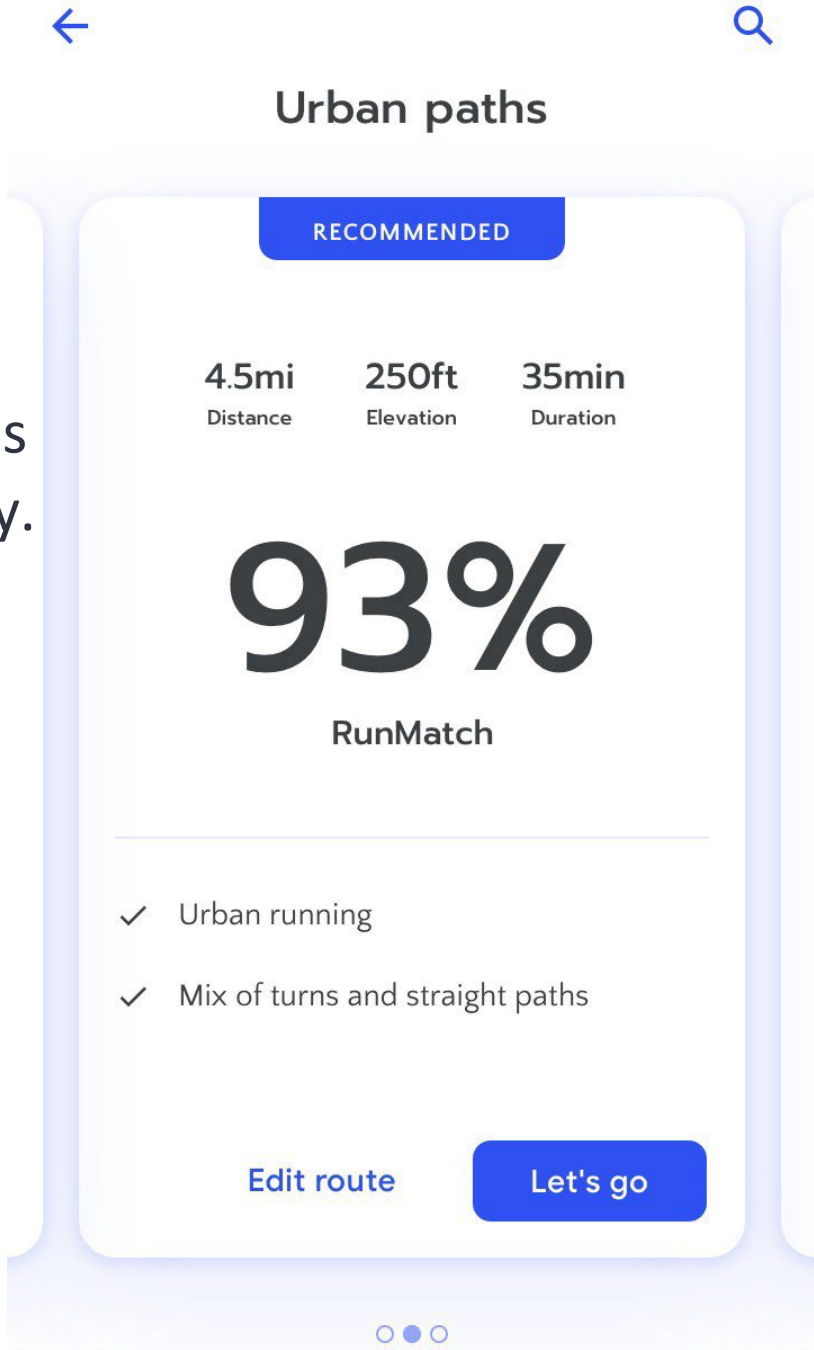
- **N-best alternatives**. E.g., “This photo might be of London, Shanghai, or São Paulo.”
  - Useful in low-confidence situations. Showing multiple options prompts the user to rely on their own judgement. Also helps people build a mental model of how the system relates different options.
  - Determine how many alternatives to show requires user testing and iteration.



Manage influence on user decisions

2. Decide how best to show model confidence

- **Numeric**. E.g., a simple percentage. Risky as it presumes users have a good baseline understanding of probability.
    - Give enough context for users to understand what the percentage means.
    - Might confuse users for outputs they consider to be a sure thing.



Reference

https://pair.withgoogle.com/chapter/explainability-trust/

https://journals.sagepub.com/doi/10.1177/0018720816681350

https://journals.sagepub.com/doi/10.1177/0018720816681350

https://hbr.org/2018/07/we-need-transparency-in-algorithms-but-too-much-can-backfire

https://medium.com/zetta-venture-partners/gdpr-panic-may-spur-data-and-ai-innovation-3fd8b0df16fd

https://www.thedrum.com/news/2019/07/31/how-brands-can-regain-consumer-trust

https://www.seas.harvard.edu/news/2016/05/automaton-we-trust

Questions?