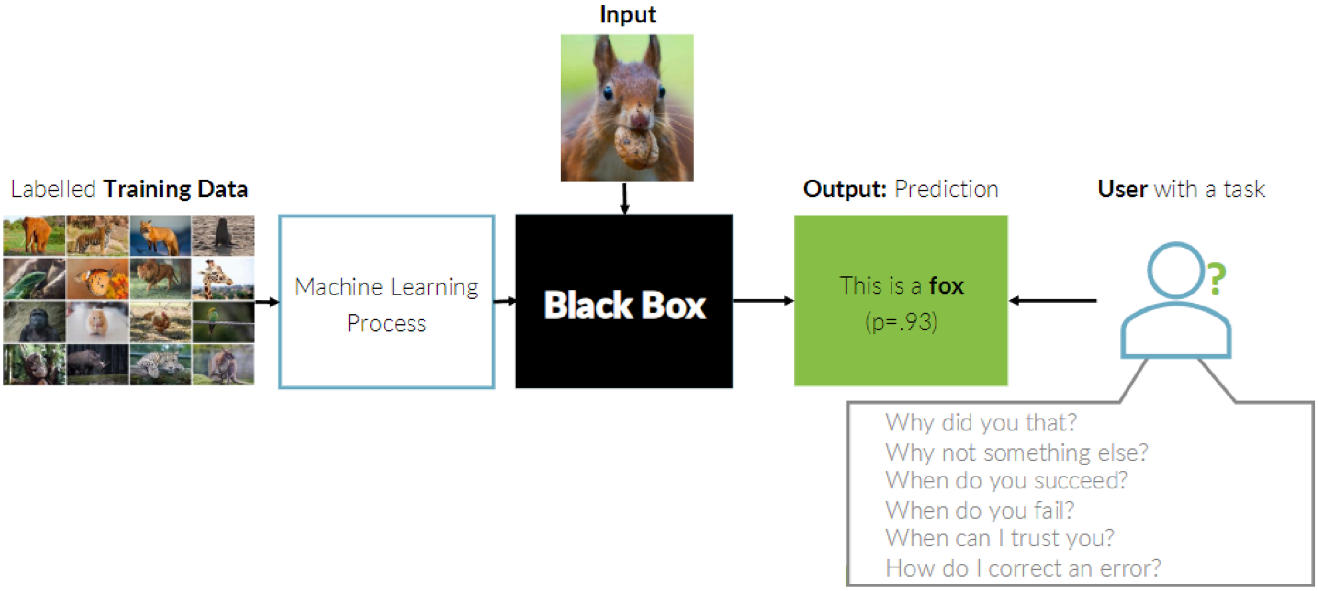
Lecture 10 - Explainable AI

Instead of accepting opaque “black box” outputs, Explainable AI (XAI) strives to provide insights into how decisions are made. This not only increases user trust but also supports error diagnosis, bias detection, and regulatory compliance.

> **Yudkowsky 2008**
>
> "By far the greatest danger of Artificial Intelligence is that people conclude too early that they understand it."

The Black Box Problem and the Clever Hans Phenomenon

Modern machine learning models often operate as black boxes—their internal decision processes remain hidden from end users. This opacity can lead to misinterpretations, as exemplified by the “Clever Hans” problem, where a model might appear to understand a task while actually exploiting spurious patterns.



> 🔎 **Black Box Problem**
> Refers to the challenge of interpreting complex algorithms whose internal logic is not directly accessible or understandable, making it difficult to predict how and why specific decisions are made.

There has always been proprietary, non-interpretable knowledge. What is different now?

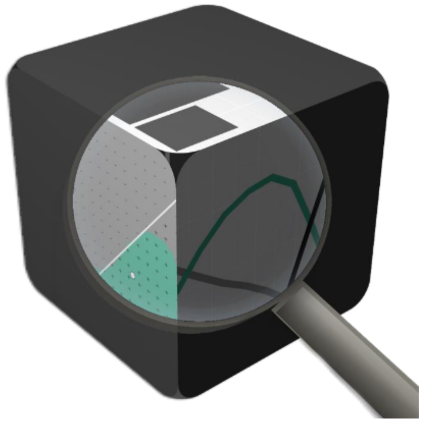
- **Scale and Impact:** Modern ML models operate at a vast scale and are embedded in critical systems (e.g., healthcare, finance, autonomous vehicles), making their opaque decision processes have widespread societal consequences.
- **Need for Accountability:** Unlike traditional proprietary knowledge that was confined to specific domains, today’s data-driven algorithms must be interpretable to ensure fairness, prevent bias, and allow for regulatory oversight.

We do not need to understand how a motor works to drive a car – why do we need to understand ML models now?

- **Dynamic and Uncertain Behavior:** While the mechanics of a motor are stable and predictable, ML models are probabilistic and can evolve with new data, leading to unexpected behavior that impacts decision-making.
- **High-Stakes Decisions:** ML models often underpin systems that directly affect human lives and societal functions, so understanding their inner workings is crucial for ensuring transparency, trust, and proper accountability.

What is Explanability?

"Explainability," "Interpretability," and "Transparency" are often used interchangeably.

	**Possible Definitions:** **- "… the ability to explain or to present in understandable terms to a human" [Doshi-Velez & Kim 2017]** **- "… is the degree to which a human can understand the cause of a decision" [Miller 2017]** **- "… is the degree to which a human can consistently predict the model's result" [Kim et al. 2016]**

Societal Challenges and Ethical Implications

In sensitive applications like judicial systems, finance, or recruitment, lack of transparency can result in biased outcomes and erode public trust.
GDPR's right to explanation, mandate that automated decisions include human-understandable justifications.

Balancing Innovation and Accountability

AI systems must not only be high-performing but also transparent and fair, ensuring that users are not subjected to unexplained or biased decisions.

What Constitutes a Good Explanation?

Good explanations in AI should:

- **Clarify the decision process:** Helping users understand why a particular output was generated.
- **Build trust:** Enabling users to calibrate their confidence in the system.

- **Support user intervention:** Allowing users to correct or challenge decisions.
- **Be concise yet comprehensive:** Striking the right balance between detail and understandability.

Researchers have offered various definitions:

- “The ability to explain or to present in understandable terms to a human.” ([Doshi-Velez & Kim 2017])
- “The degree to which a human can understand the cause of a decision.” ([Miller 2017])

Interpretability in Machine Learning

Interpretability refers to how well a human can comprehend the reasoning behind a model's output. Two key dimensions are:

Local vs. Global Interpretability

Scope	Focus	Key Question
Local	Individual predictions	"Why did this specific decision occur?"
Global	Overall model behavior	"How does the model generally operate?"

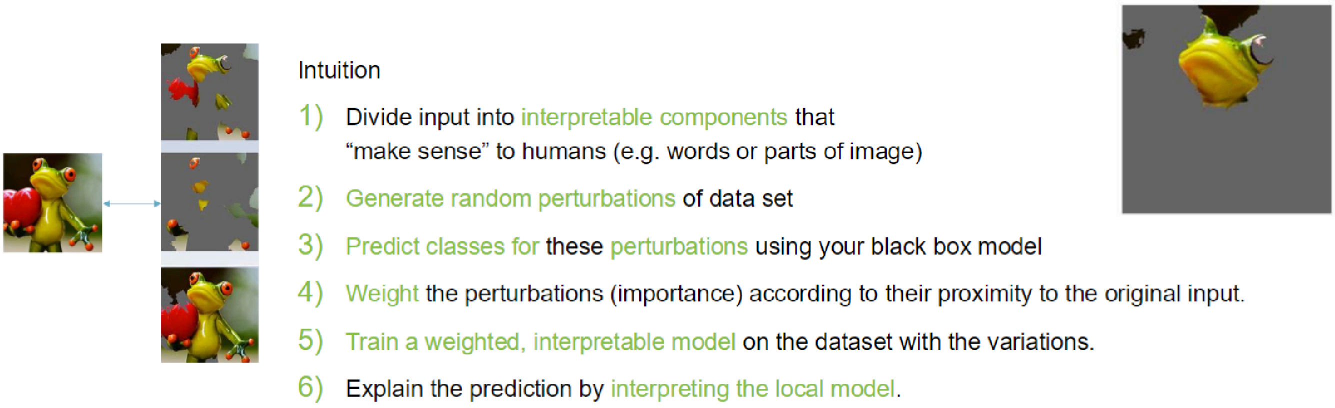
Local explanations help users understand individual outcomes, while global explanations reveal the overarching mechanics of a model.

Intrinsic vs. Post-hoc Interpretability

Interpretability Type	Description	Example Methods
Intrinsic	Models designed to be interpretable by nature	Decision Trees, Linear Models
Post-hoc	Techniques applied after model training to extract explanations	LIME, SHAP

Intrinsic methods build interpretability into the model structure, whereas post-hoc methods generate explanations for otherwise opaque models.

Explanation Methods: Local Interpretable Model-Agnostic Explanations (LIME)



Intuition

1) Divide input into interpretable components that "make sense" to humans (e.g. words or parts of image)

2) Generate random perturbations of data set

3) Predict classes for these perturbations using your black box model

4) Weight the perturbations (importance) according to their proximity to the original input.

5) Train a weighted, interpretable model on the dataset with the variations.

6) Explain the prediction by interpreting the local model.

LIME is a popular post-hoc method that provides local interpretability.

1. **Decomposition:** Dividing an input into interpretable components (e.g., words or image segments).
2. **Perturbation:** Generating variations of the input data.
3. **Prediction:** Using the original black box model to predict outcomes for these perturbations.
4. **Weighting:** Assigning importance based on the similarity to the original input.
5. **Local Modeling:** Training a simple, interpretable model on the weighted perturbed data.
6. **Explanation:** Presenting the factors that most influenced the model's prediction.

Applications of Explainability

Explainability in AI serves multiple functions:

- **Model Validation:** Detecting and eliminating bias in training data.
- **Model Debugging:** Identifying reasons behind misclassifications or errors, including adversarial attacks.
- **Knowledge Discovery:** Uncovering new insights from data by revealing hidden patterns and correlations.

Each of these applications contributes to a more reliable and accountable AI system, benefiting both developers and end users.

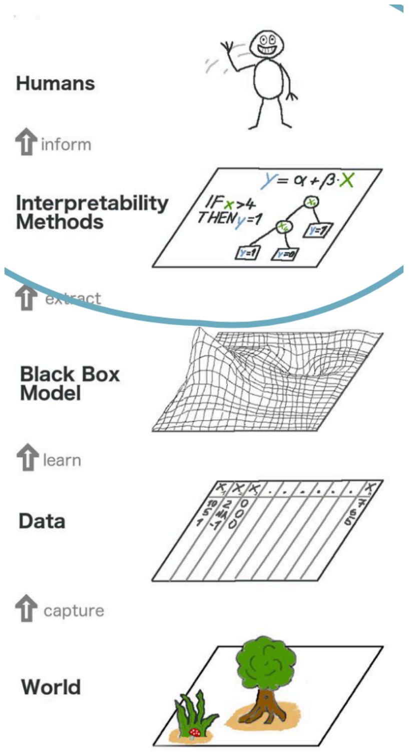
Human-Centered AI and HCI Challenges

Human-Centered Artificial Intelligence (HCAI) prioritizes enhancing human performance and ensuring that systems are safe, reliable, and trustworthy. Key challenges include:

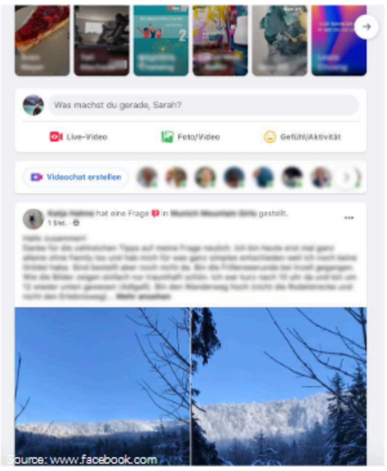
- **Understanding:** Helping users build accurate mental models of how AI systems operate.

- **Trust:** Enabling users to gauge when to rely on the system and when to question its outputs.

- **Control:** Allowing users to provide feedback and correct system errors.

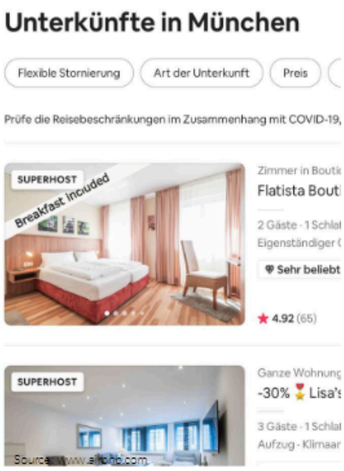


Source: [Molnar 2019]



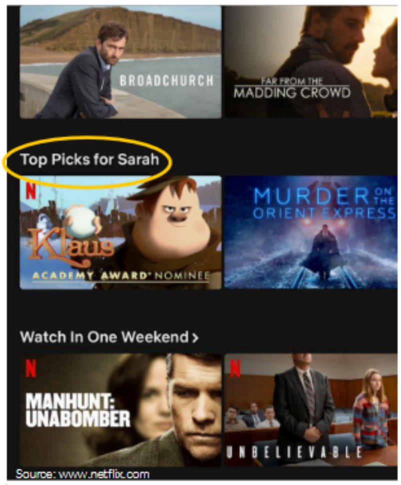
Source: www.facebook.com

Lack of Algorithmic Awareness



Source: www.airbnb.com

Algorithmic Anxiety



Source: www.netflix.com

Intransparent Recommendations

Enhancing User Trust

Clear, transparent explanations empower users to engage confidently with AI systems and mitigate algorithmic anxiety.

Support Strategies and User Control

Real-world systems, such as recommendation engines or navigation apps, often suffer from issues like:

- **Lack of Feedback:** Users have few opportunities to inform the system of errors or misinterpretations.
- **Limited User Control:** Systems may override user preferences without clear justification.

- **Intransparent Recommendations:** Explanations that fail to reveal the reasoning behind suggestions.

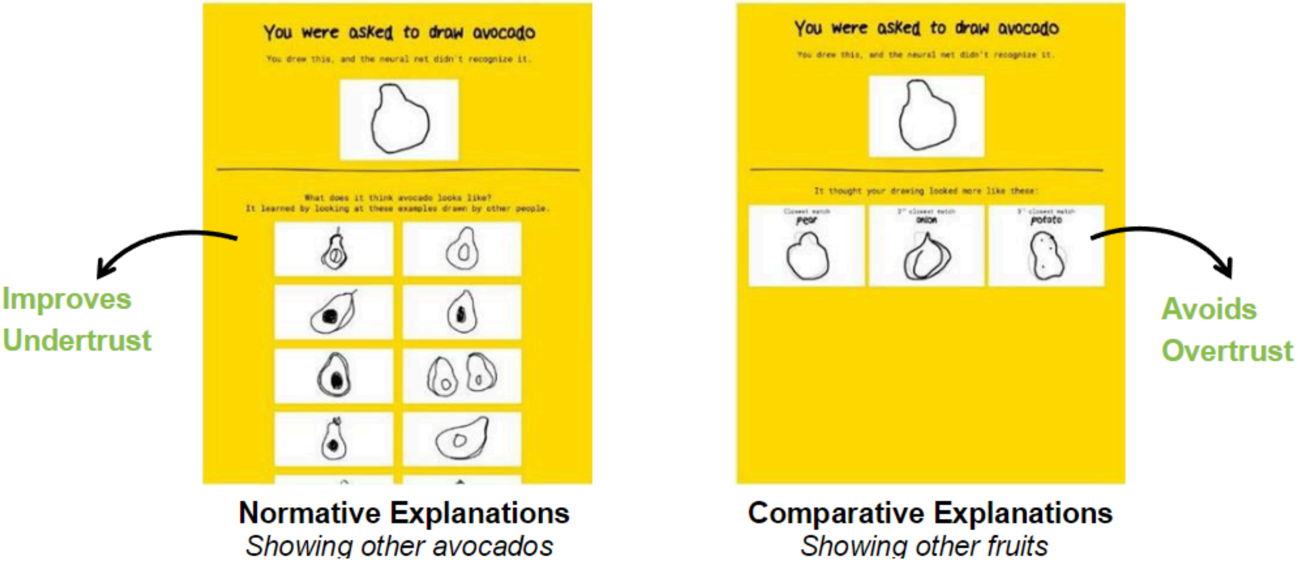
Improving these areas requires designing interfaces that not only provide explanations but also invite interactive user feedback and correction.

Discussion and Future Directions

As AI systems continue to permeate various aspects of daily life, the need for robust, user-friendly explanations becomes increasingly critical.

- **Interactive Explanations:** Allowing users to ask follow-up questions and explore alternative scenarios.
- **Placebo vs. Actual Explanations:** Differentiating between superficial explanations and those that truly enhance understanding.
- **Integration of Social Science Insights:** Leveraging research on human cognition to design explanations that are contrastive, selective, credible, and conversational.

Contrastive Insights



You were asked to draw avocado

You drew this, and the neural net didn't recognize it.

What does it think avocado looks like?
It learned by looking at these examples drawn by other people.

You were asked to draw avocado

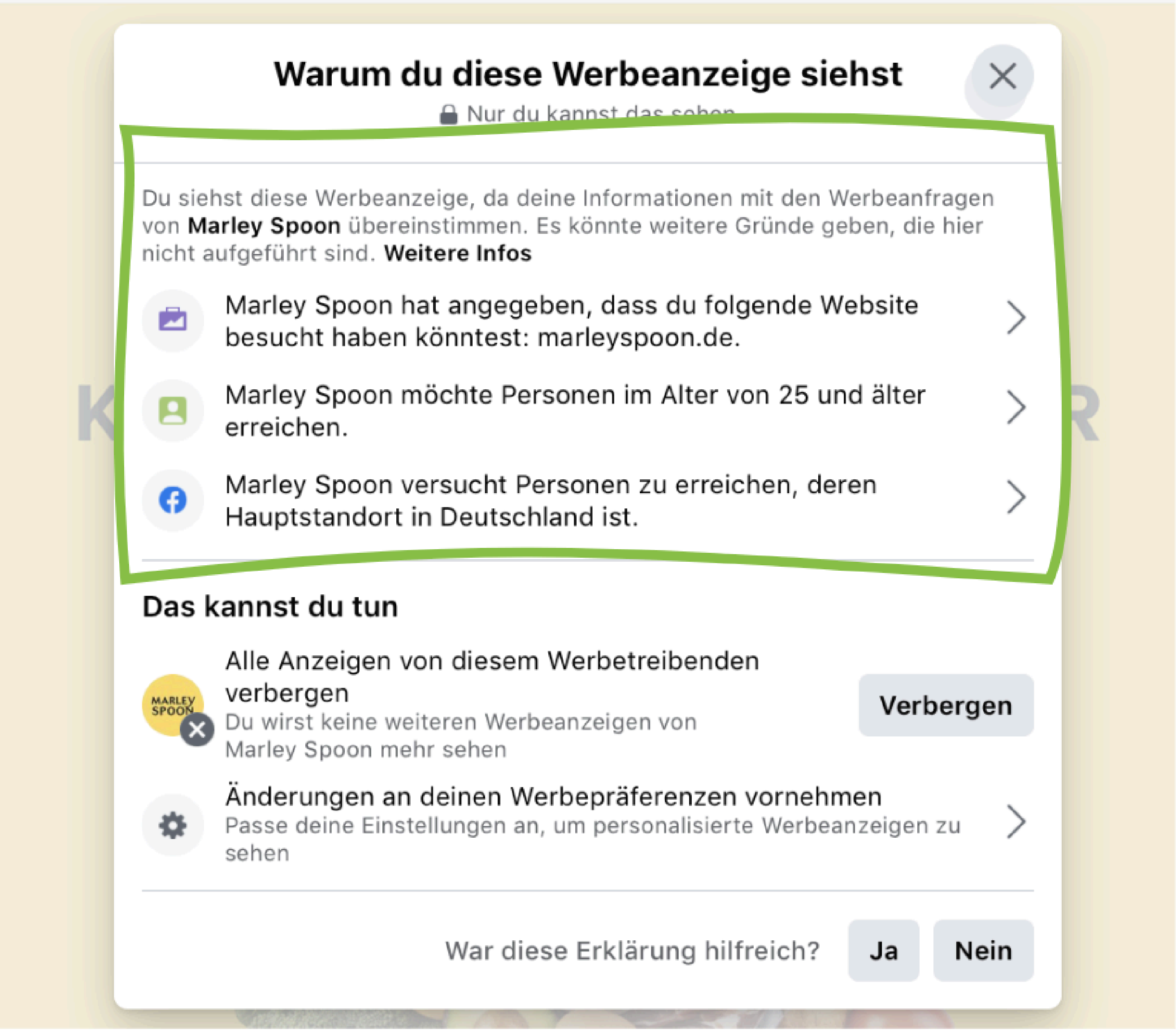
You drew this, and the neural net didn't recognize it.

It thought your drawing looked more like these:

Normative Explanations
Showing other avocados

Comparative Explanations
Showing other fruits

Selective Insights



Warum du diese Werbeanzeige siehst

Nur du kannst das sehen

Du siehst diese Werbeanzeige, da deine Informationen mit den Werbeanfragen von **Marley Spoon** übereinstimmen. Es könnte weitere Gründe geben, die hier nicht aufgeführt sind. **Weitere Infos**

Marley Spoon hat angegeben, dass du folgende Website besucht haben könntest: marleyspoon.de.

Marley Spoon möchte Personen im Alter von 25 und älter erreichen.

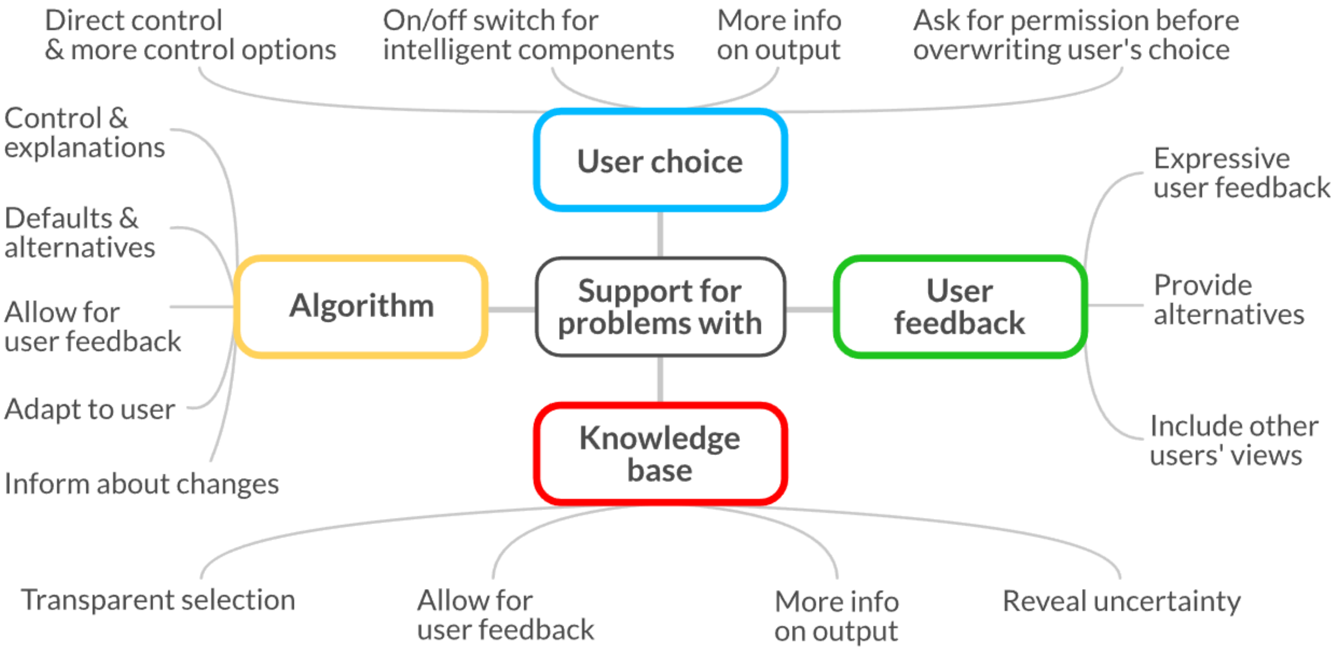
Marley Spoon versucht Personen zu erreichen, deren Hauptstandort in Deutschland ist.

Das kannst du tun

Alle Anzeigen von diesem Werbetreibenden verbergen
Du wirst keine weiteren Werbeanzeigen von Marley Spoon mehr sehen

Änderungen an deinen Werbepräferenzen vornehmen
Passe deine Einstellungen an, um personalisierte Werbeanzeigen zu sehen

War diese Erklärung hilfreich?



These directions promise to not only improve system transparency but also to foster greater user engagement and trust.