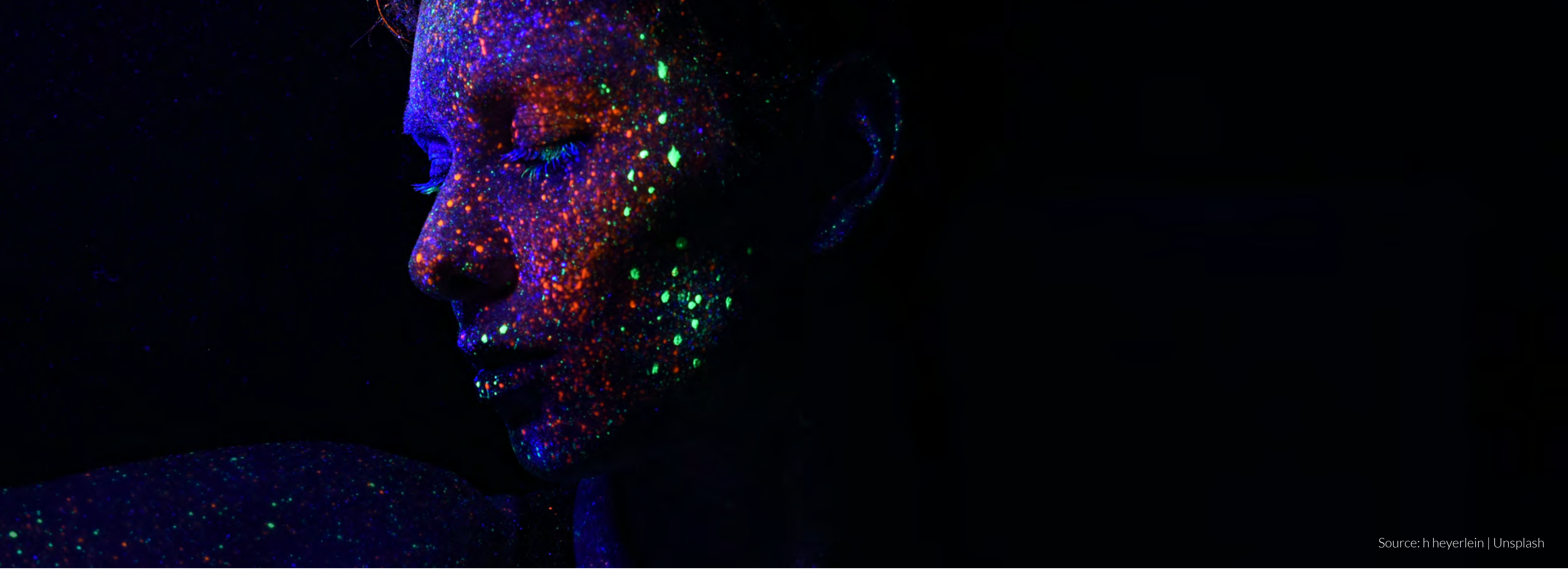


Introduction to Intelligent User Interfaces | Sarah Theres Völkel | January 21st, 2020

Explainable AI

About Me

Sarah Theres Völkel

- PhD student at Media Informatics Group
- Contact: sarah.voelkel@ifi.lmu.de

- Research Interests:
  - Personalisation of Voice User Interfaces
  - Personality-tailored Personalisation
  - Transparency of intelligent systems

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

[Yudkowsky 2008]

Overview

1 **Transparency for Intelligent Systems**

- The Black Box Problem
- Resulting Challenges for Society
- Explainable AI
- What Makes a Good Explanation
- User Problems and Support

2 **Transparency for Personality-Targeting**

- Personality and Personality-Targeting
- Requirements for Explanations for Personality-Targeting
- How to Trick AI

Online

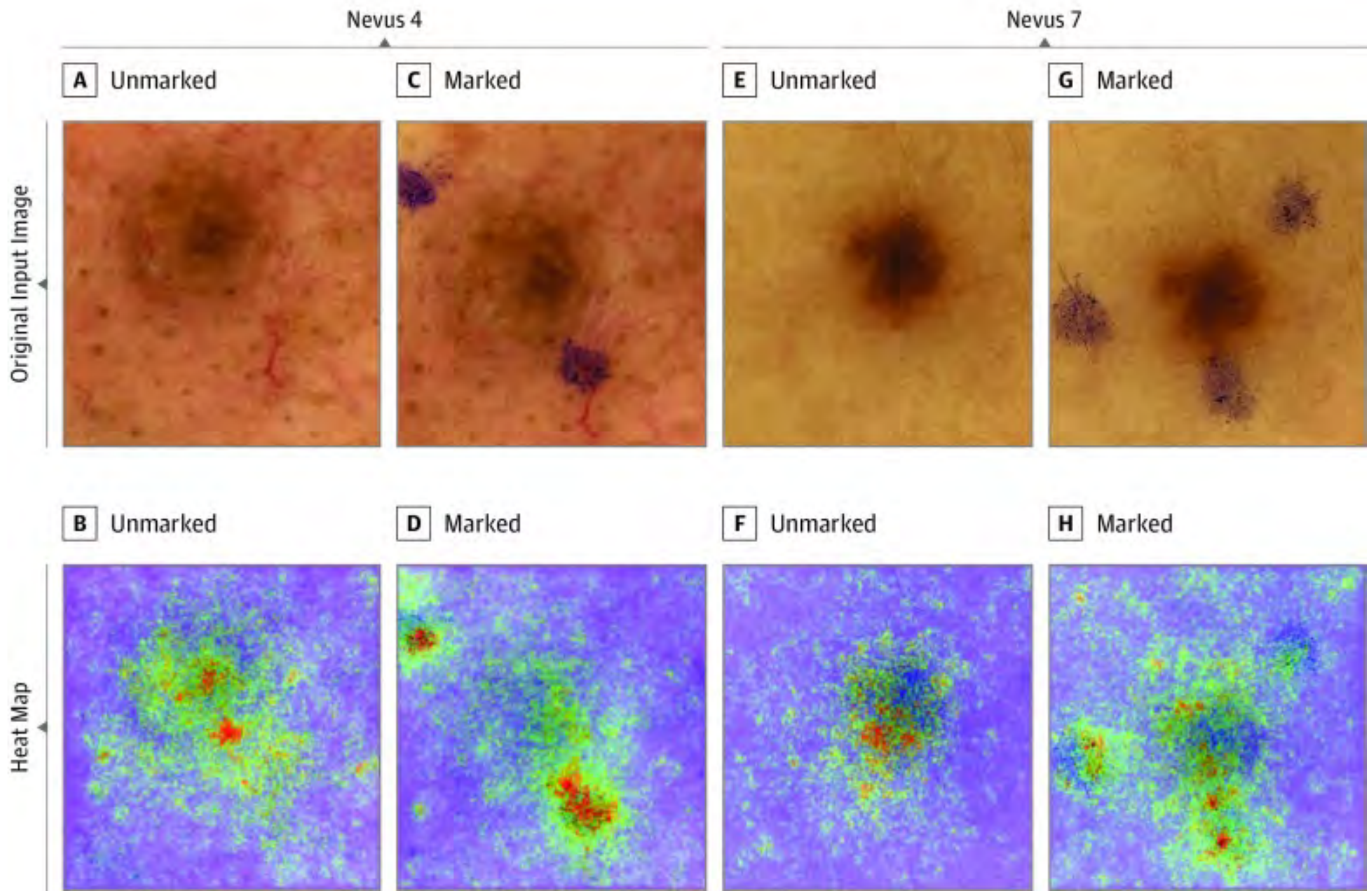
Transparency for Intelligent Systems

Overview

The “Clever Hans” Problem

Source: Unknown Author, Public domain, via Wikimedia Commons

The "Clever Hans" Problem



Source: Winkler et al. 2019 American Medical Association

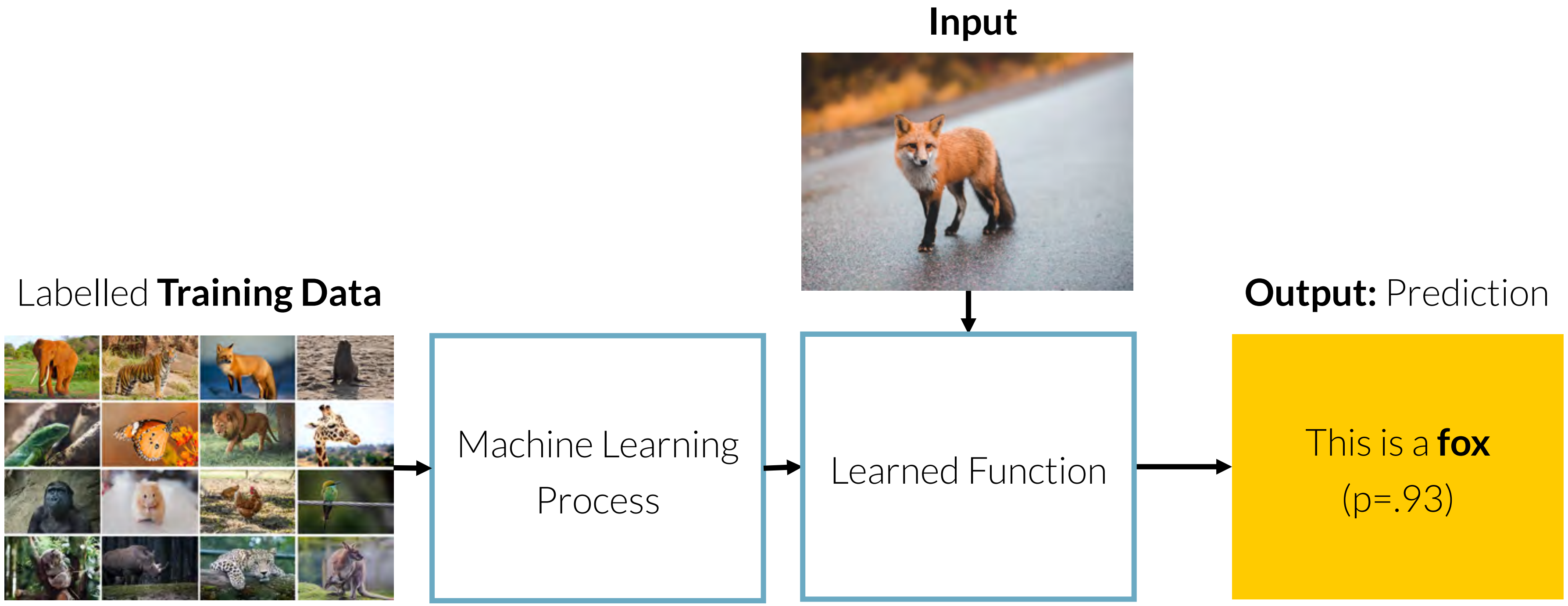
The Black Box Problem of Machine Learning

Source: Courtesy of Quay Au

"[…] stems from the ***mismatch*** *between mathematical optimization in high-dimensionality* ***characteristic of machine learning*** *and the* ***demands of human-scale reasoning*** *and styles of semantic interpretation."*

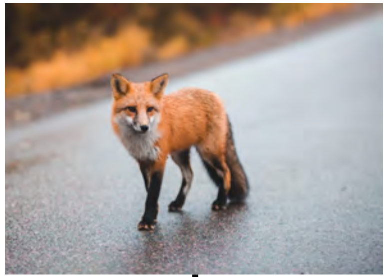
[Burrell 2016]

The Black Box Problem of Machine Learning



The Black Box Problem of Machine Learning

Input



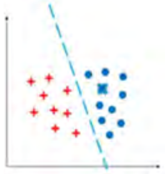
Labelled **Training Data**

Machine Learning Process

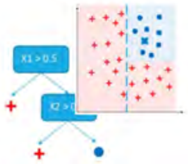
Black Box

Output: Prediction

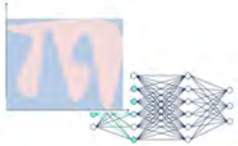
This is a **fox**
(p=.93)



Lineare Modelle

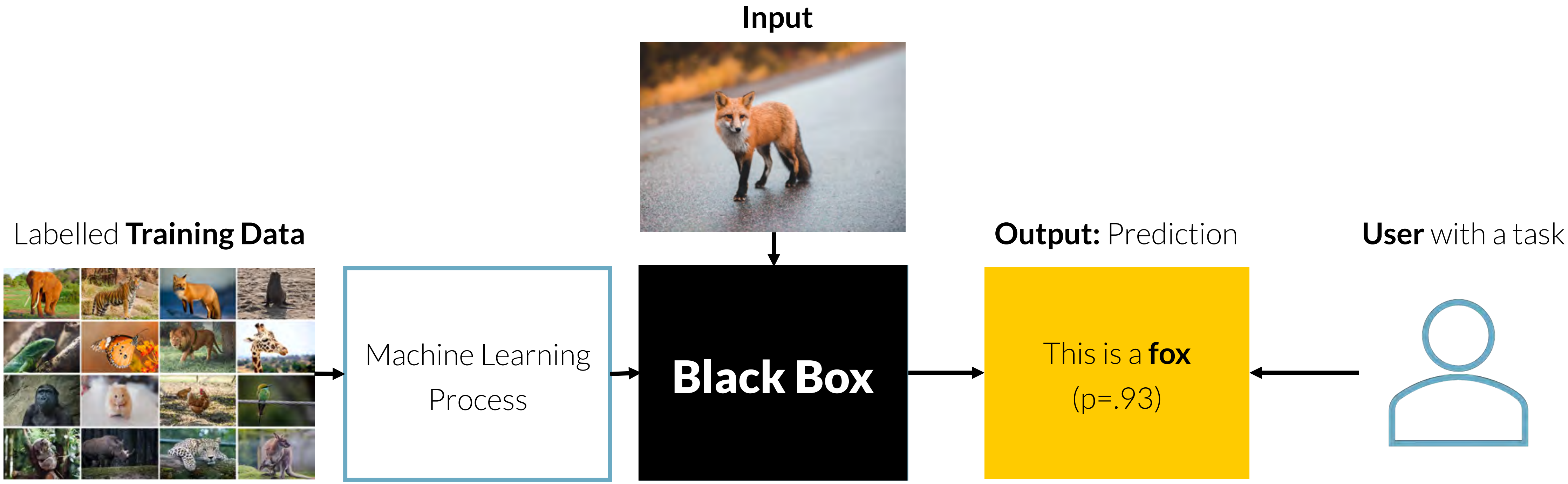


Decision Trees / Ensembles

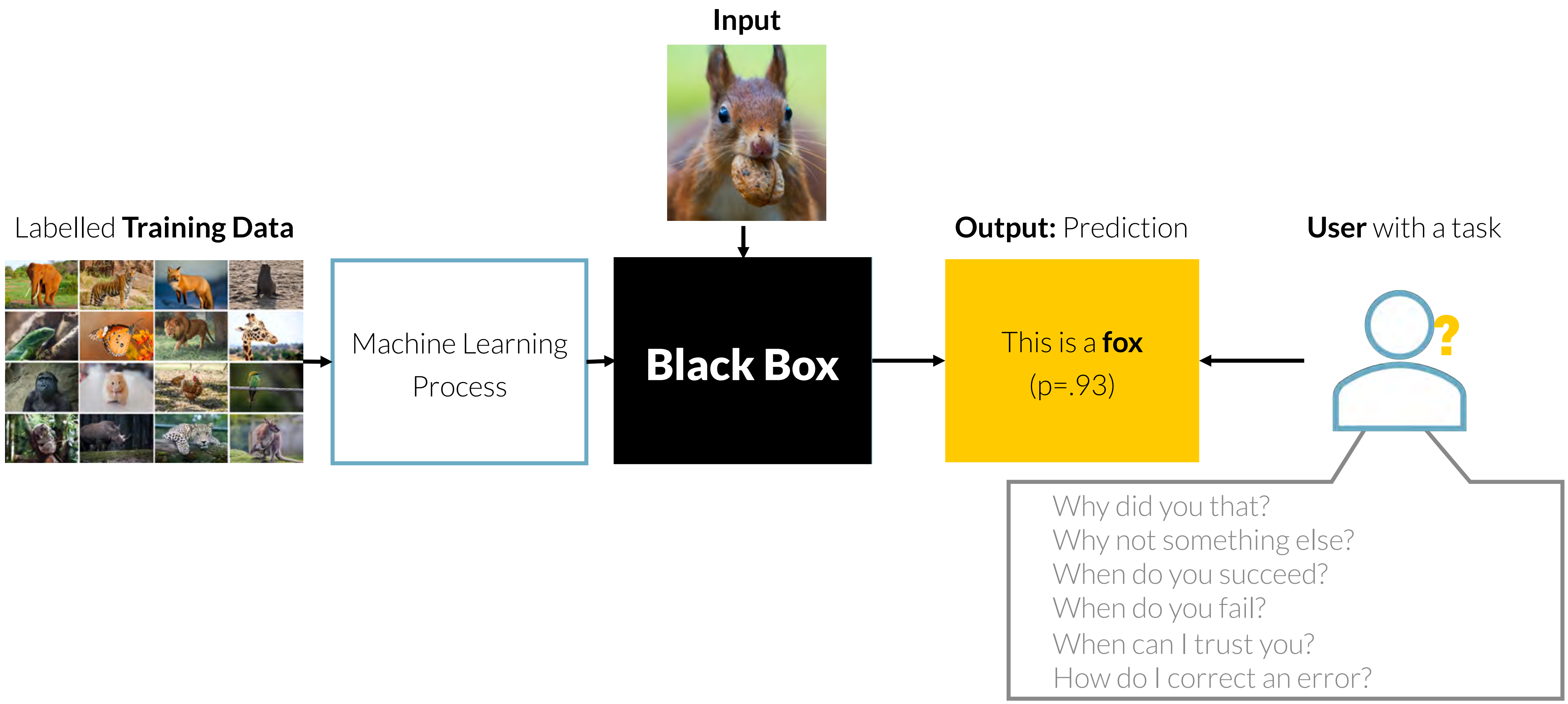


Nicht-Lineare Modelle (DNN, SVM)

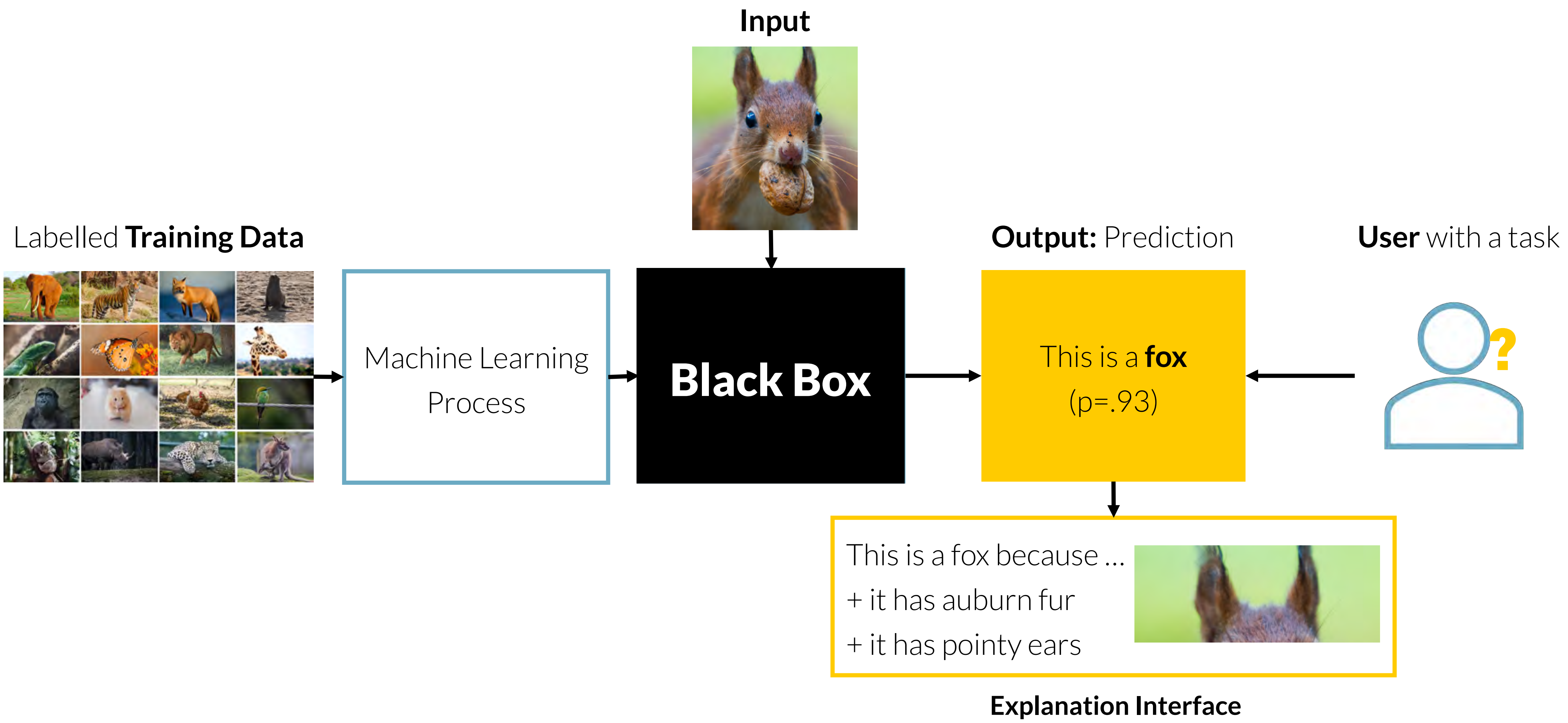
The Black Box Problem of Machine Learning



The Black Box Problem of Machine Learning



The Black Box Problem of Machine Learning



Discussion

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

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

Discuss for 5min in breakout rooms

Overview

1 **Transparency for Intelligent Systems**

- The Black Box Problem
- Resulting Challenges for Society
- Explainable AI
- What Makes a Good Explanation
- User Problems and Support

2 **Transparency for Personality-Targeting**

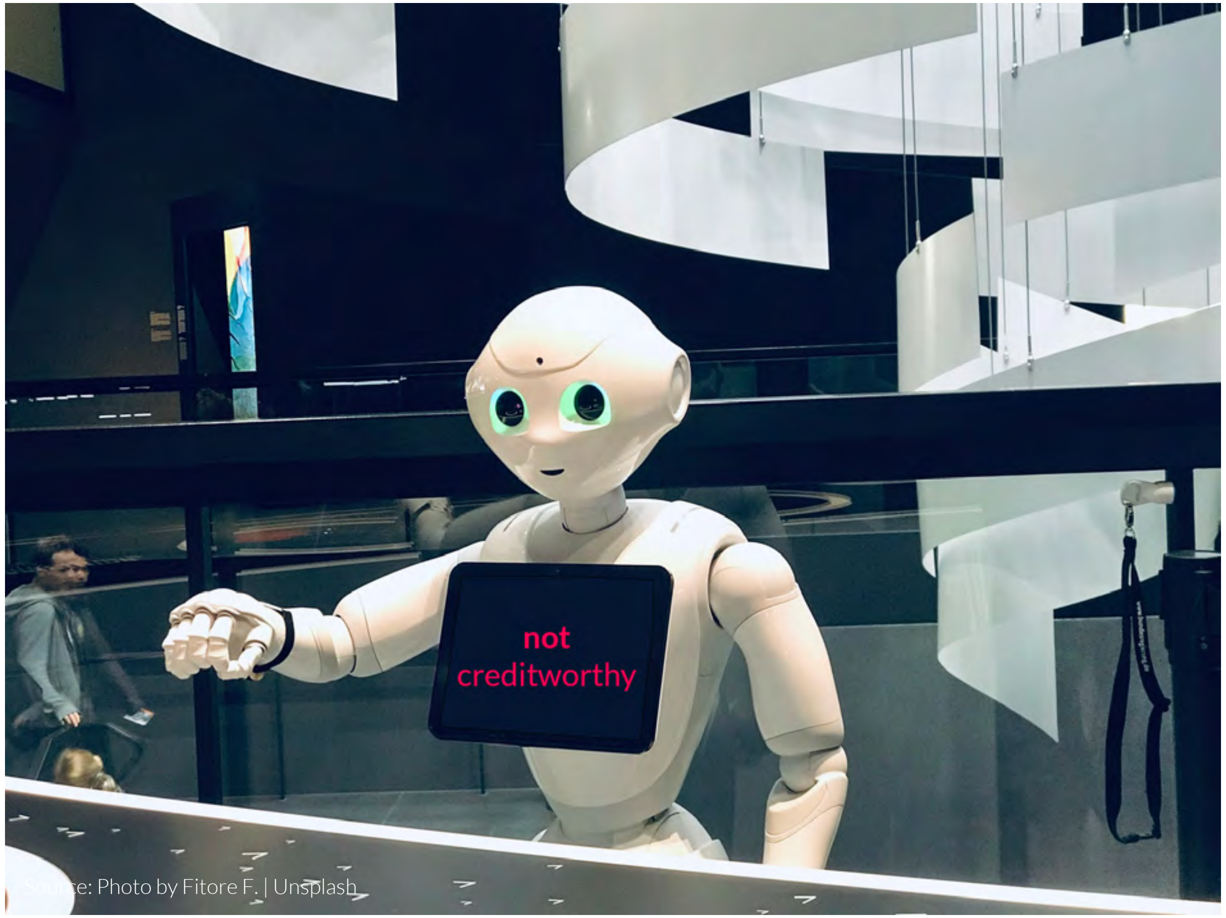
- Personality and Personality-Targeting
- Requirements for Explanations for Personality-Targeting
- How to Trick AI

AI in the Courtroom


Source: Andrey Popov in Kugler et al. 2018

Bias in training data set

AI in Financing



Lack of transparency

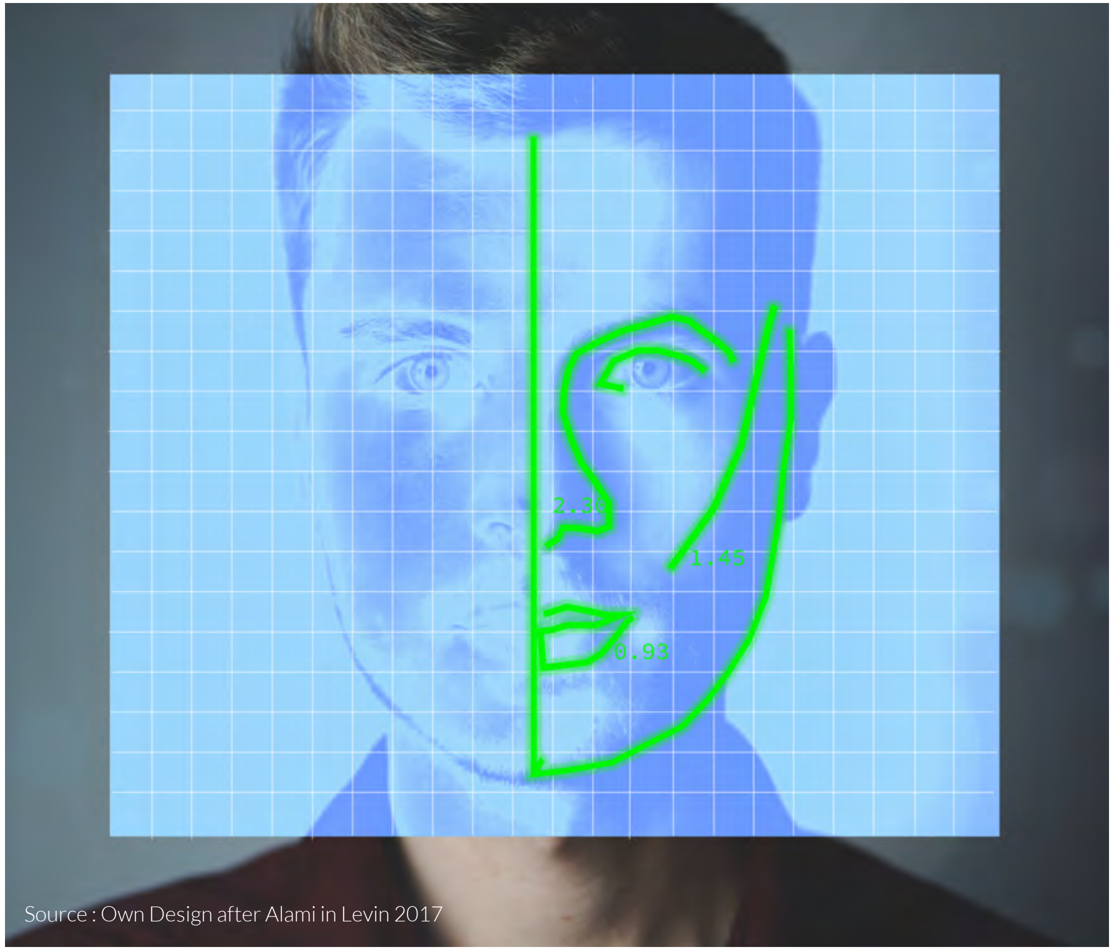
[Kayser-Bril 2020b, Phaure & Robin 2020]

AI in Recruiting

Discrimination due to bias in training data set

[Kayser-Bril 2020b, Phaure & Robin 2020]

Does AI Have a “Gaydar”?

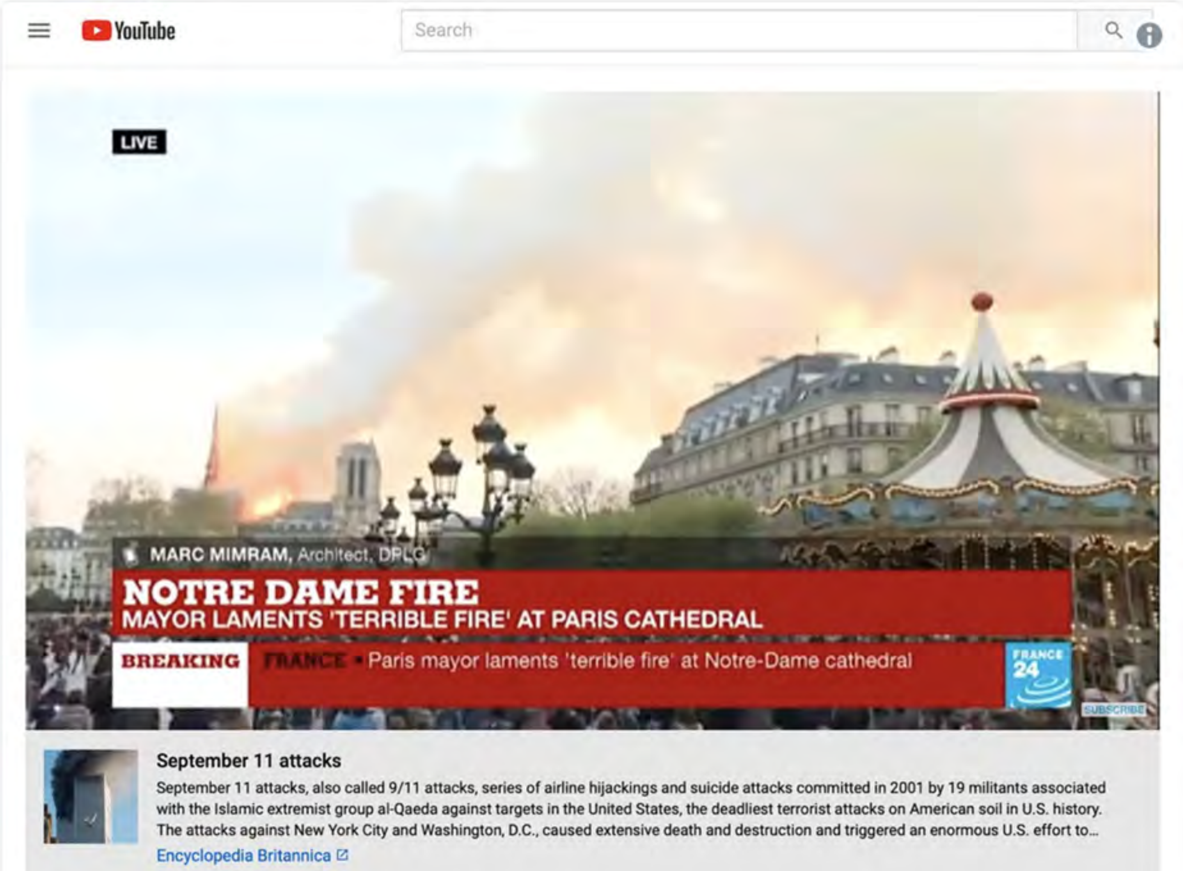


Source : Own Design after Alami in Levin 2017

Lack of interpretability

[Agüera y Arcas et al. 2018, Yilun & Kosinski 2018]

AI Acting Information Control?



Source : Benton 2019

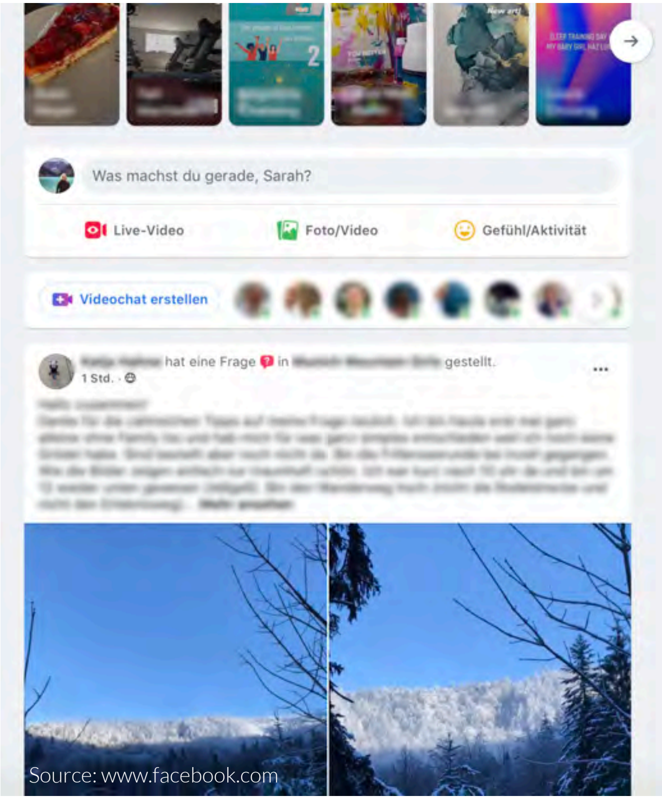
Lack of feedback and correction

AI as Translator?

Source: www.translate.google.com

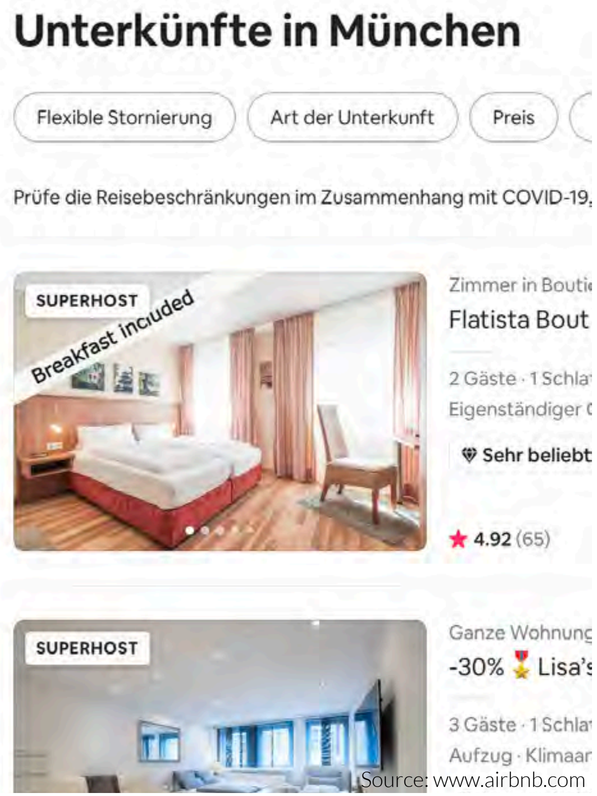
Lack of transparency about algorithm limitations

Everyday Challenges with Intelligent Systems



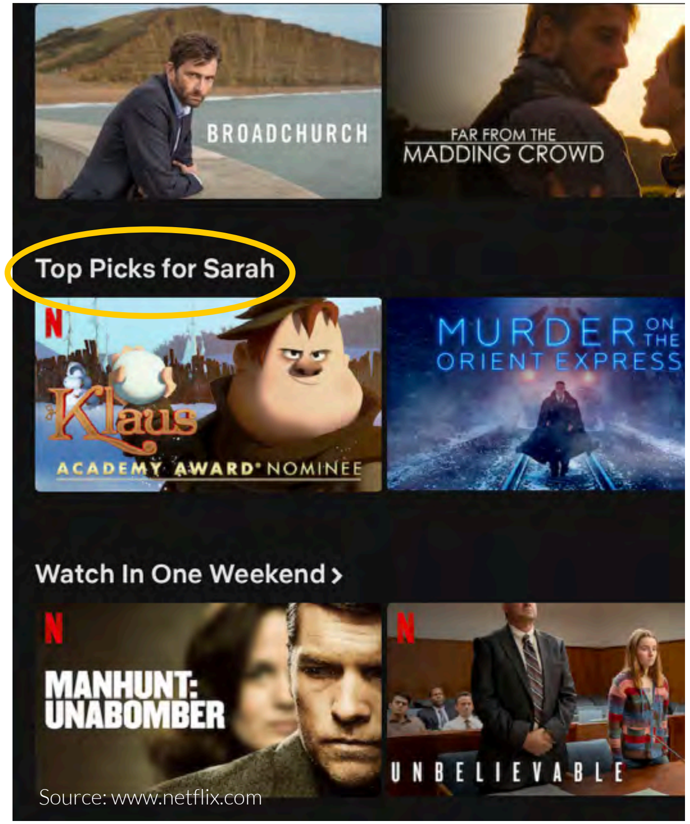
Source: www.facebook.com

Lack of Algorithmic Awareness



Source: www.airbnb.com

Algorithmic Anxiety



Source: www.netflix.com

Intransparent Recommendations

Right to Explanation in the GDPR

Article 22

The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention.

[…]

In any case, such processing should be subject to suitable safeguards, which should include **specific information to the data subject** and the right to **obtain human intervention**, to **express his or her point of view**, to **obtain an explanation** of the decision reached after such assessment and to **challenge the decision**.

Overview

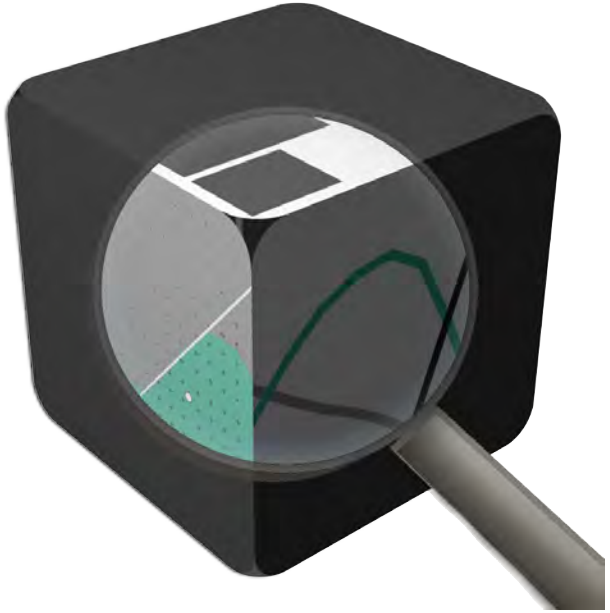
Human-Centred Artificial Intelligence

Source: Andy Kelly | Unsplash

“Human-Centred Artificial Intelligence (HCAI) focuses on **amplifying, augmenting, and enhancing human performance** in ways that make systems **reliable, safe, and trustworthy**. These systems also **support** human self-efficacy, encourage creativity, clarify responsibility, and facilitate social participation.”

[Shneiderman 2020]

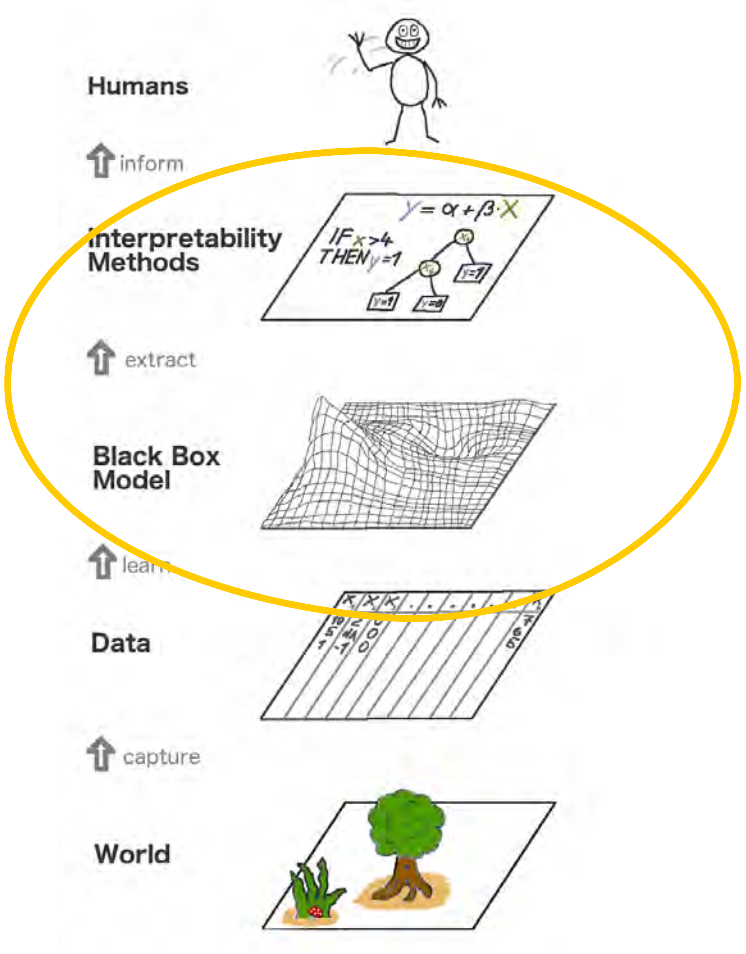
What is Explainability?



Source: Courtesy of Quay Au

- “… 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]
- “**Explainability**”, “**Interpretability**”, and “**Transparency**” are often used interchangeably

Applications of Explainability



Source: [Molnar 2019]

1. **Model Validation:** Eliminate bias in the training data
2. **Model Debugging:** Debug models and analyse wrong predictions
3. **Knowledge Discovery:** Gain new insights through the analysis

[Du et al. 2020] |

Model Validation

Classified as Dog

Classified as Wolf

Model Validation

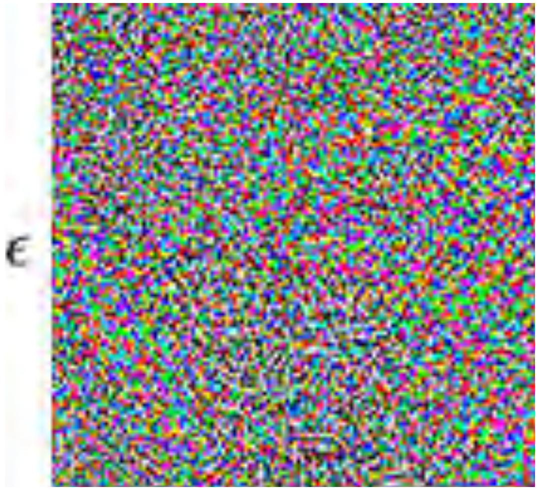
Classified as Wolf

LIME-Explanation (idealised)

Model Debugging

Adversarial Attacks

+



=

“panda”
57.7% confidence

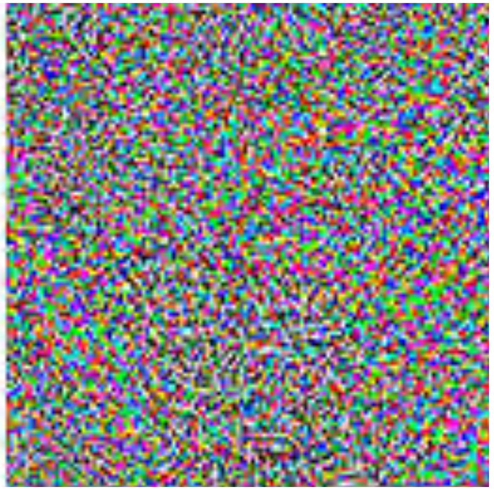
“gibbon”
99.3% confidence

Image Source: Own design after Goodfellow et al. 2014
Photo: Mélody P. | Unsplash

Model Debugging

Adversarial Attacks in Traffic

$+ \; \epsilon$

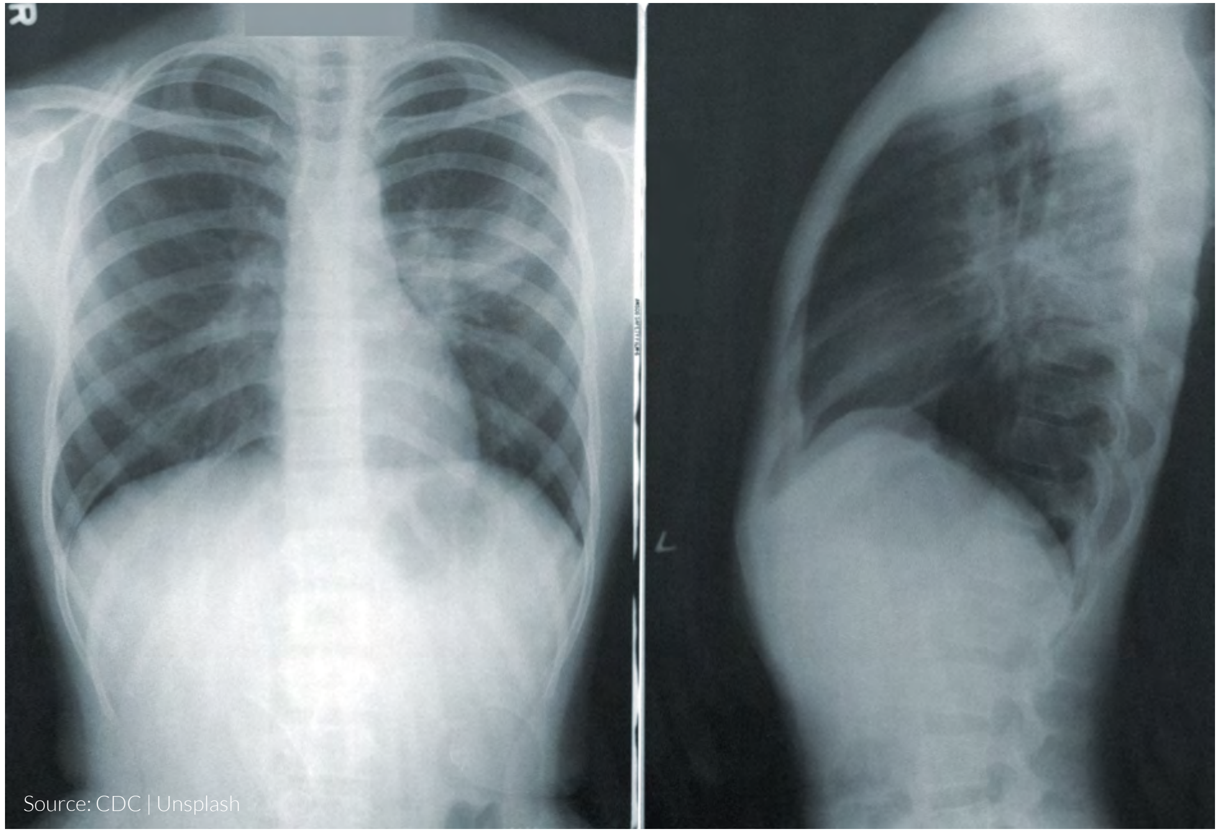


$=$

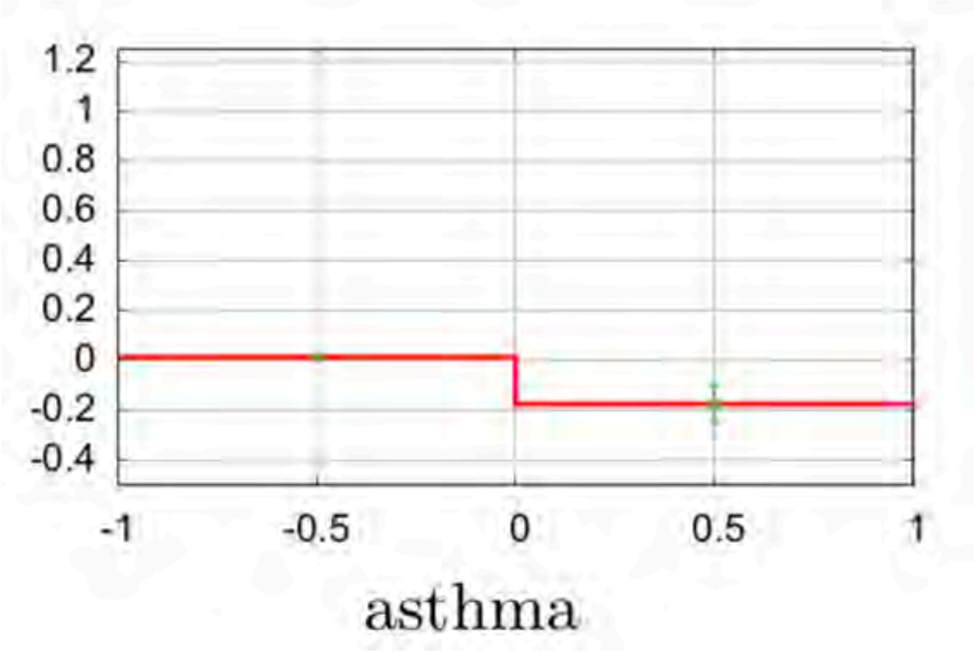
“stop sign”
76.0% confidence

“no stop sign”
97.3% confidence

Knowledge Discovery



Source: CDC | Unsplash

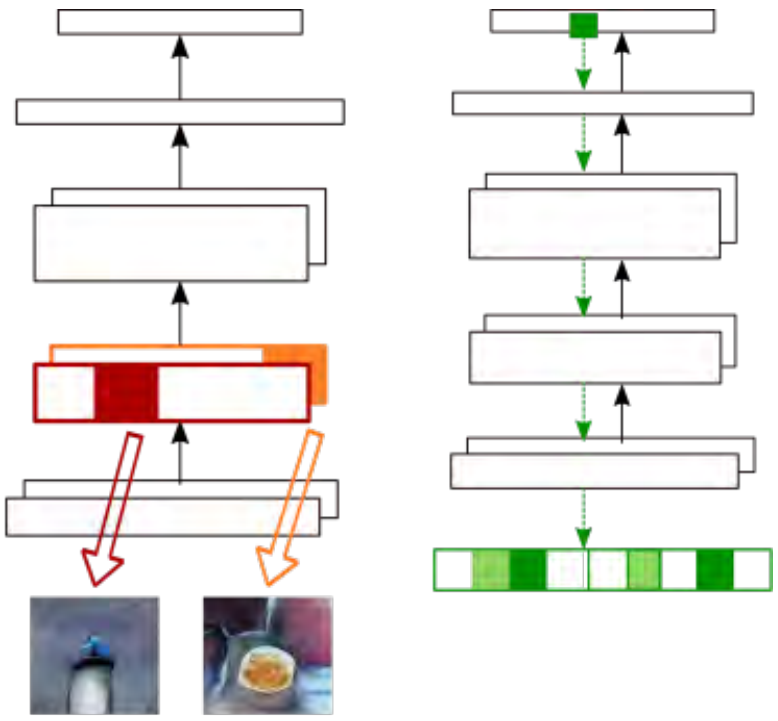


Source: [Caruana et al. 2015]

Local vs Global Interpretability

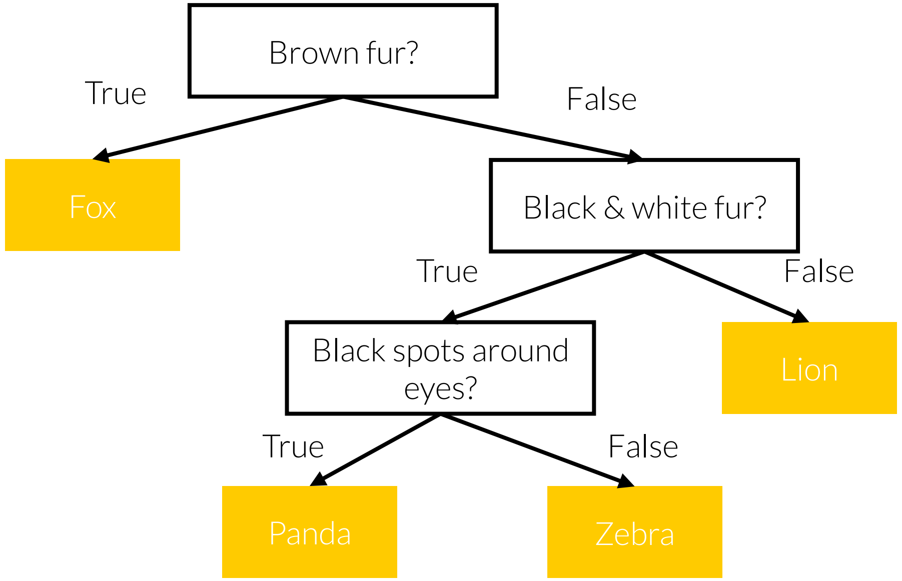
- **Local Interpretability:** Explain **individual predictions** (causal relations between input and corresponding output) → why a certain prediction?

- **Global Interpretability:** Explain **structures and parameters** for a global understanding (inner workings & mechanisms) → how are predictions made?

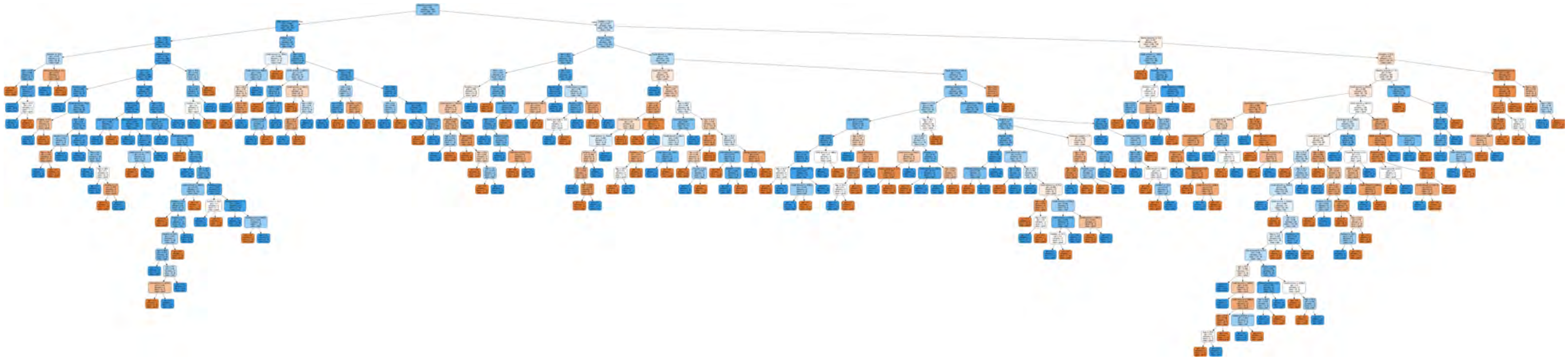


Intrinsic vs Post-hoc Interpretability

Intrinsic Interpretability:
self-explanatory models which integrate interpretability directly in the structure

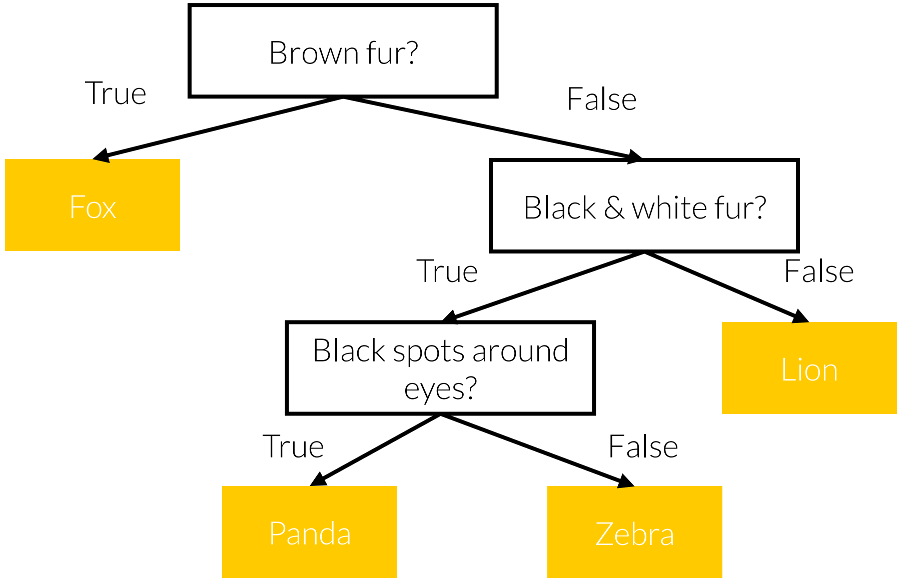


Intrinsic Interpretability



Intrinsic vs Post-hoc Interpretability

Intrinsic Interpretability:
self-explanatory models which integrate interpretability directly in the structure



Post-hoc Interpretability:
a second model is needed that creates explanations for the existing model

Source: [Ribeiro et al. 2016]

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)

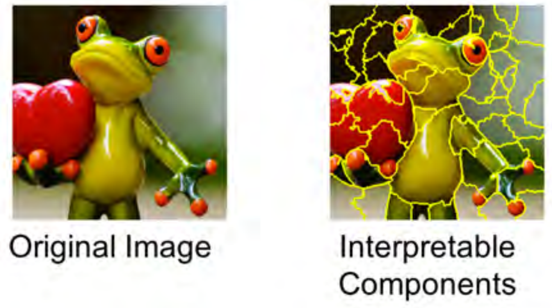


Image & Content: [Ribeiro et al. 2016]

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

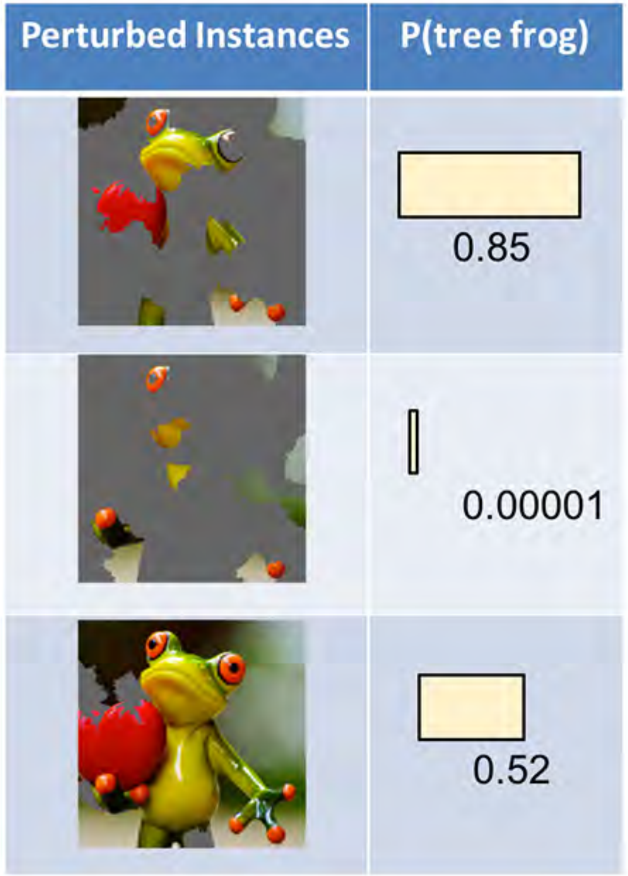
Local Interpretable Model-Agnostic Explanations (LIME)

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

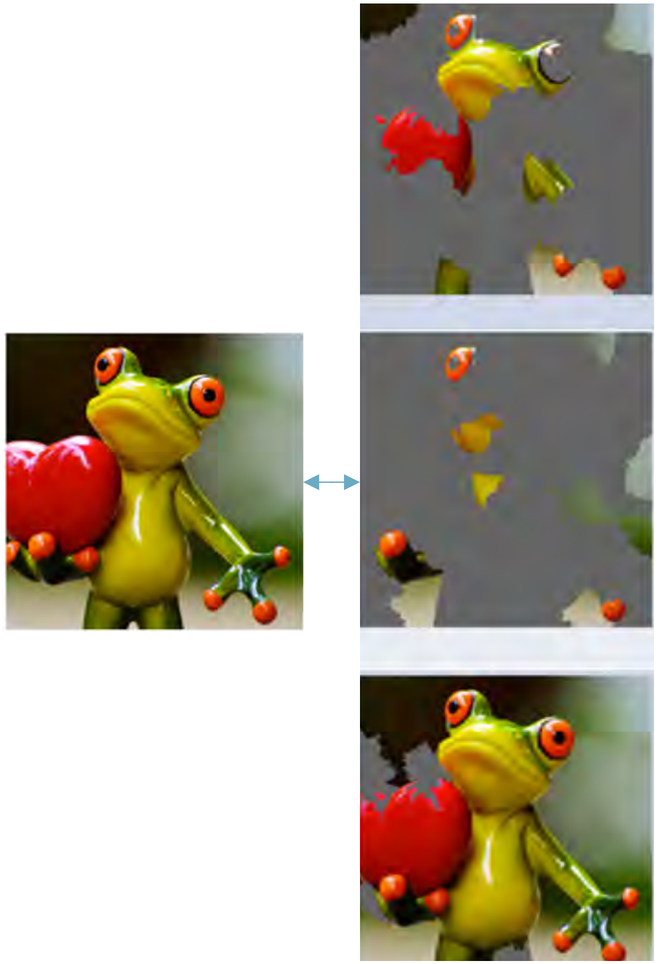
3) **Predict classes for** these **perturbations** using your black box model



Local Interpretable Model-Agnostic Explanations (LIME)

Intuition

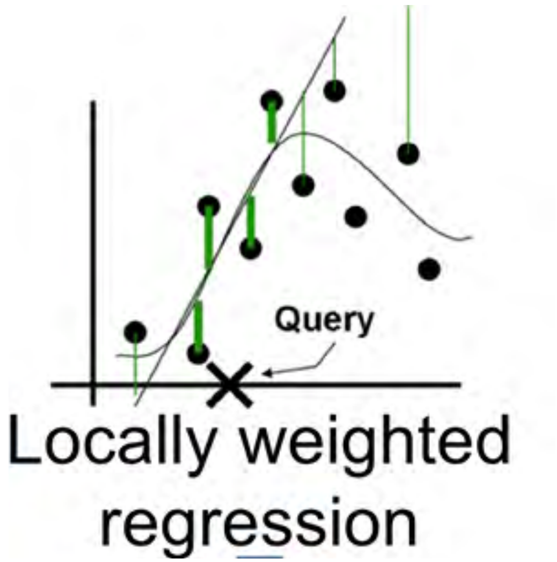
1) Divide input into **interpretable components** that "make sense" to humans (e.g. words or parts of image)
2) **Generate random perturbations** of data set
3) **Predict classes for** these **perturbations** using your black box model
4) **Weight** the perturbations (importance) according to their proximity to the original input.



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.



Local Interpretable Model-Agnostic Explanations (LIME)

Intuition

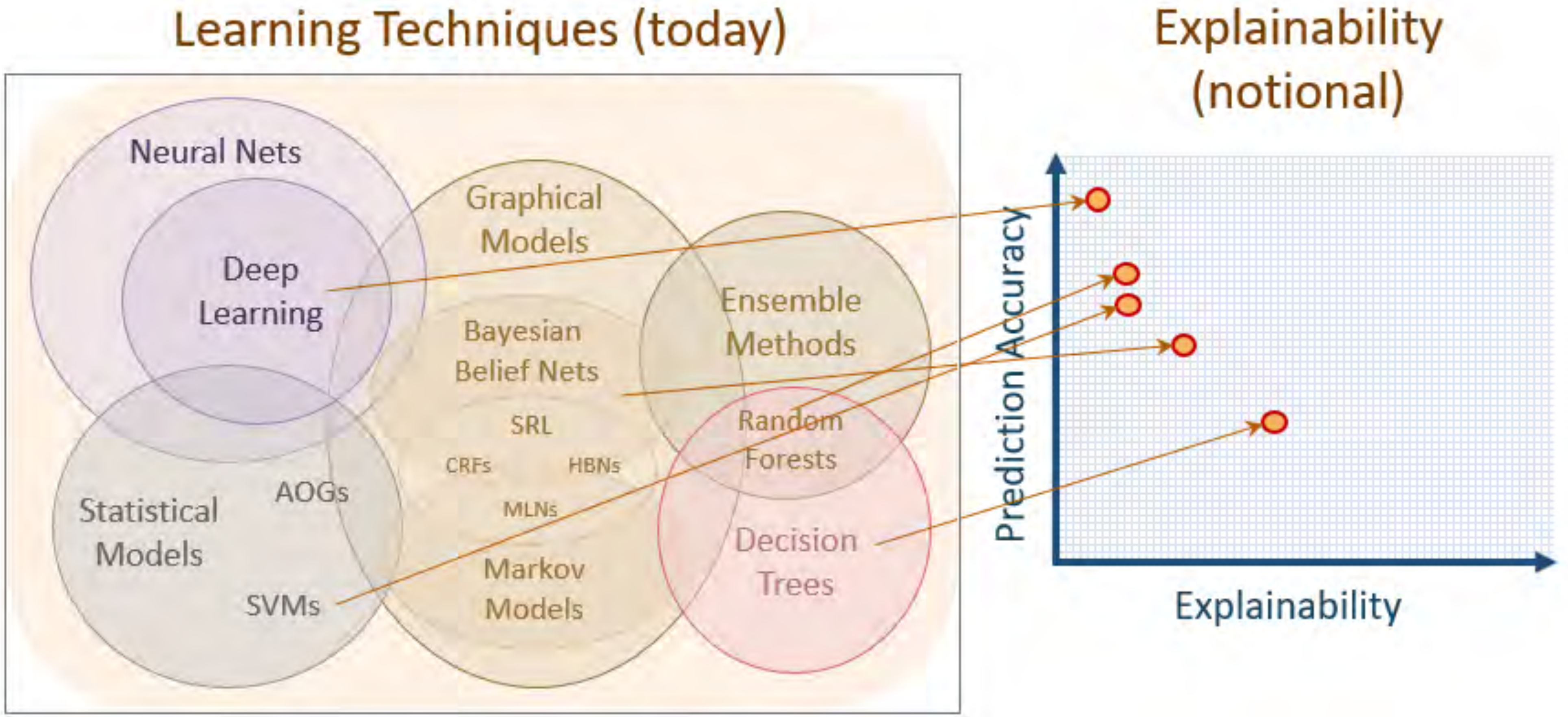
1) Divide input into **interpretable components** that "make sense" to humans (e.g. words or parts of image)
2) **Generate random perturbations** of data set
3) **Predict classes for** these **perturbations** using your black box model
4) **Weight** the perturbations (importance) according to their proximity 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**.

Local Interpretable Model-Agnostic Explanations (LIME)

Practical Example:

https://colab.research.google.com/github/arteagac/arteagac.github.io/blob/master/blog/lime_image.ipynb

Trade-Off Interpretability & Accuracy



Overview

1

Transparency for Intelligent Systems

- The Black Box Problem
- Resulting Challenges for Society
- Explainable AI
- What Makes a Good Explanation
- User Problems and Support

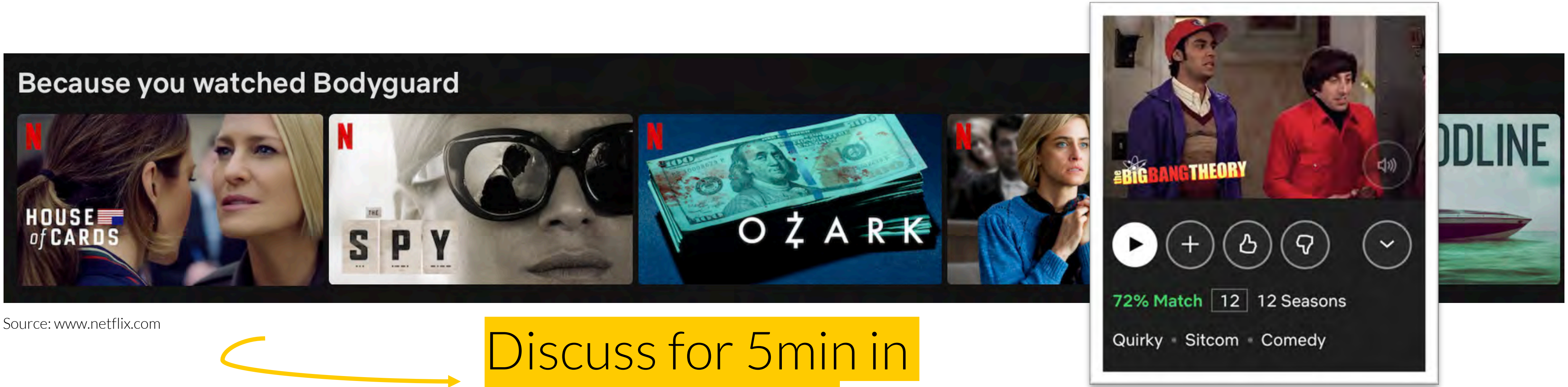
2

Transparency for Personality-Targeting

- Personality and Personality-Targeting
- Requirements for Explanations for Personality-Targeting
- How to Trick AI

Discussion

1) How does Netflix explain why a movie / TV show is recommended to the user?
2) Do you think this explanation helps users?

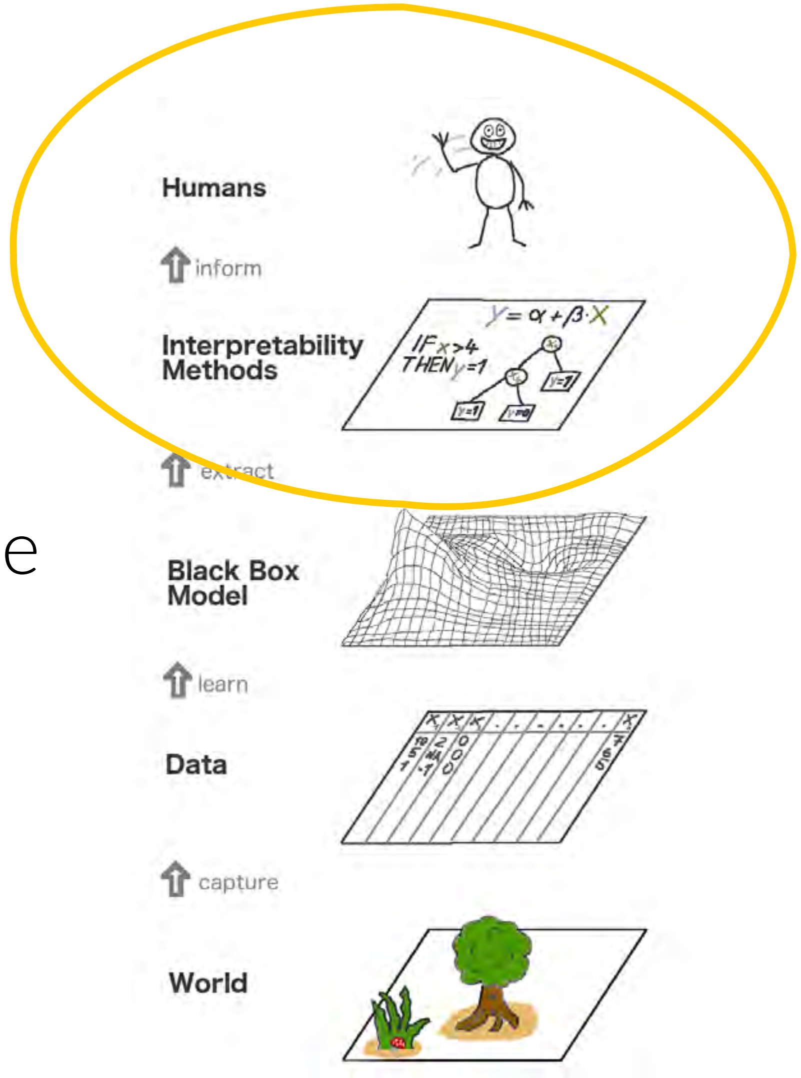


Source: www.netflix.com

Discuss for 5min in breakout rooms

Challenges for HCI Research

- **Understand:** Enable users to develop an appropriate mental model
- **Trust:** Enable users to calibrate their trust in the model
- **Correct:** Enable users to correct the model



Source: [Molnar 2019]

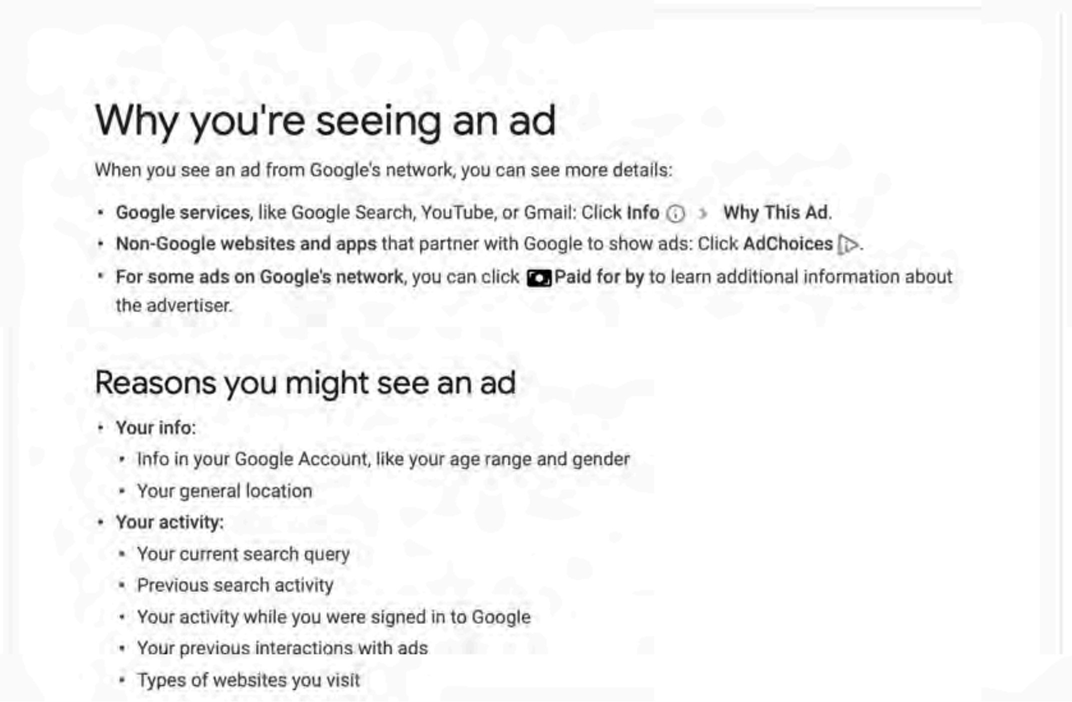
Local vs Global Explanation

Local Explanation

Source: mail.google.com

Local vs Global Explanation

Global Explanation

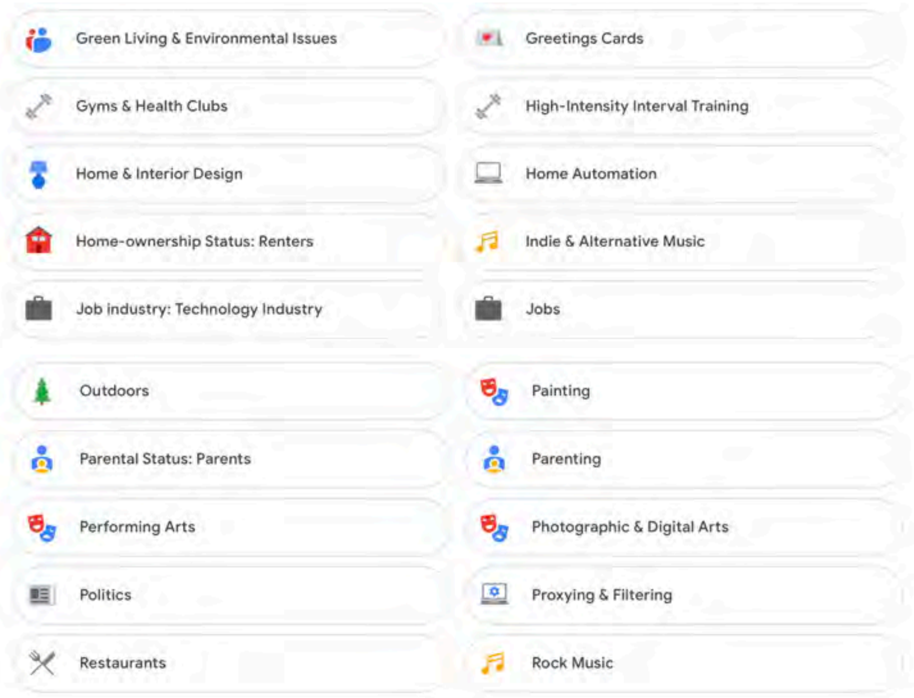

Why you're seeing an ad

When you see an ad from Google's network, you can see more details:

- **Google services**, like Google Search, YouTube, or Gmail: Click **Info** > **Why This Ad**.
- **Non-Google websites and apps** that partner with Google to show ads: Click **AdChoices**.
- **For some ads on Google's network**, you can click **Paid for by** to learn additional information about the advertiser.

Reasons you might see an ad

- **Your info:**
  - Info in your Google Account, like your age range and gender
  - Your general location
- **Your activity:**
  - Your current search query
  - Previous search activity
  - Your activity while you were signed in to Google
  - Your previous interactions with ads
  - Types of websites you visit



Source: https://adssettings.google.com

What to Explain

Explanation Types

- What?
- Why?
- Why not?
- How to?
- Inputs?
- Outputs?
- What if?
- Certainty?

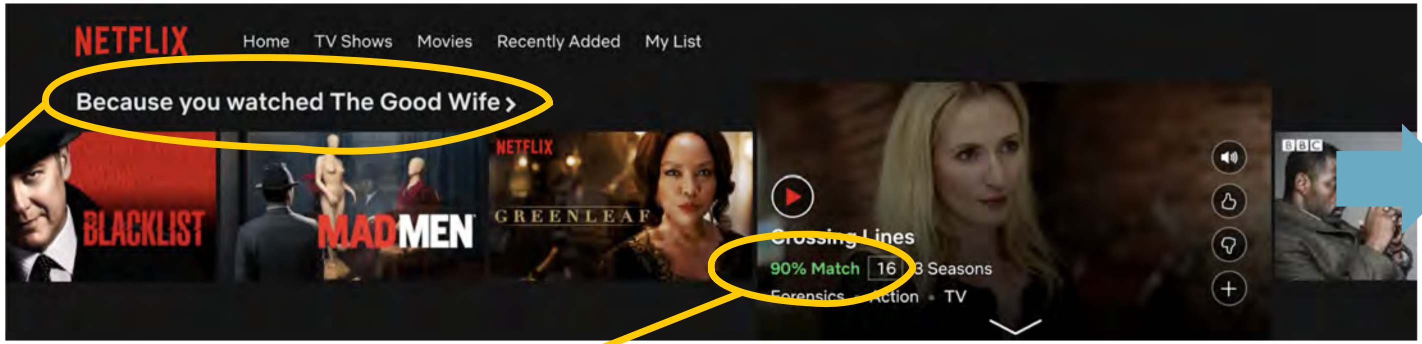
[Lim & Dey 2009, 2010, 2011]

Goals of Explanations

- Transparency
- Scrutability
- Trust
- Effectiveness
- Persuasiveness
- Efficiency
- Satisfaction

[Tintarev & Masthoff 2012]

Explanations in Today's Systems



Source: www.netflix.com

"Why"
Explanation

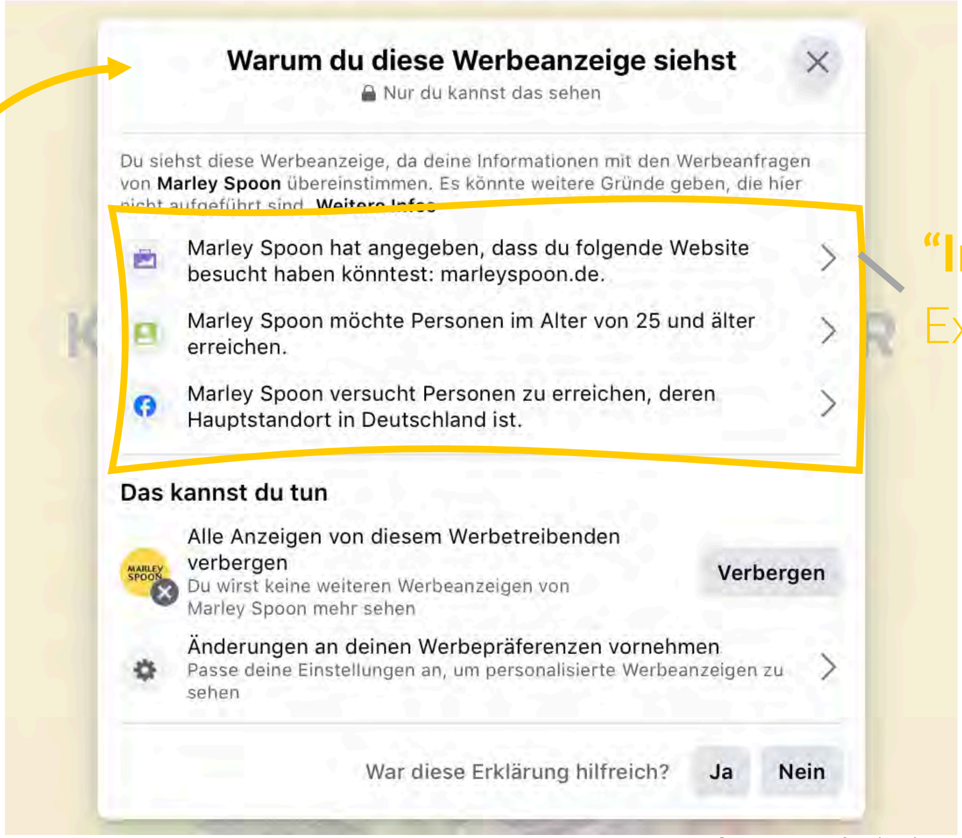
"Certainty"
Explanation

Source: www.amazon.de

Transparency
Trust
Effectiveness
Persuasiveness
Satisfaction

Explanations in Today's Systems

"Why"
Explanation



"Inputs"
Explanation

Source: www.facebook.com

Transparency
Scrutability

Persuasiveness
Satisfaction

Which Questions Do Users Have?

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]?

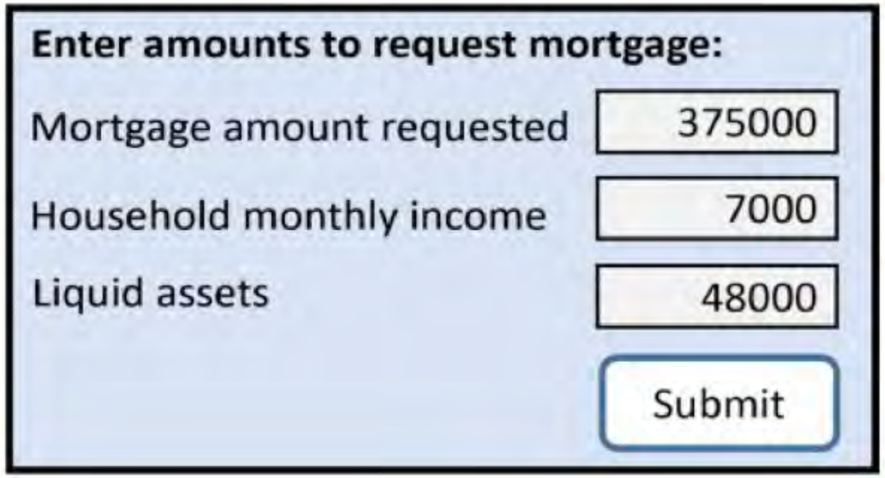
How to be that
- **How should this instance change to get a different prediction?**
- How should this feature change for this instance to get a different prediction?
- What kind of instance gets a different prediction?

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)

Insights from Social Sciences

- **Explanations are contrastive:** Why X instead of Y?

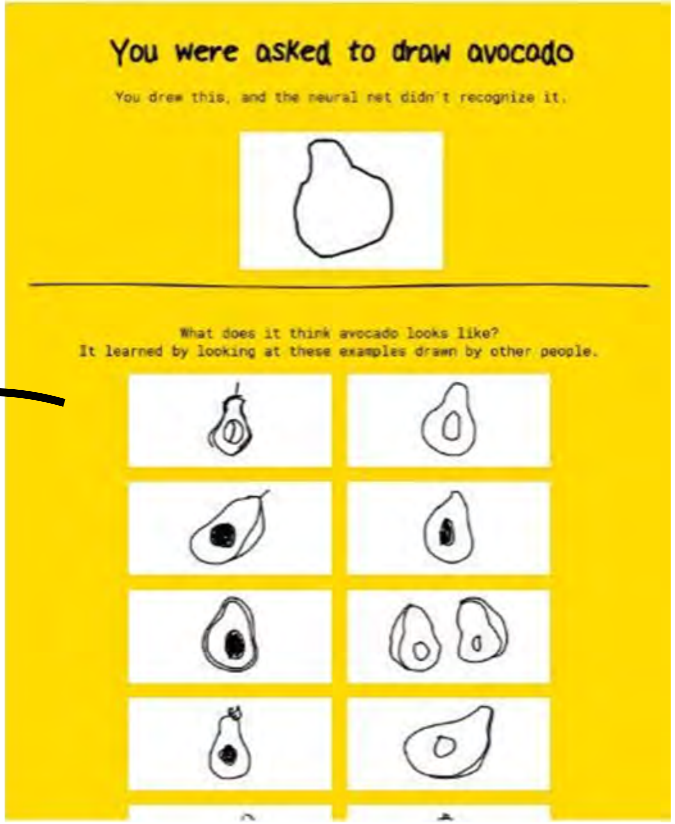


Source: [Shneiderman 2020]

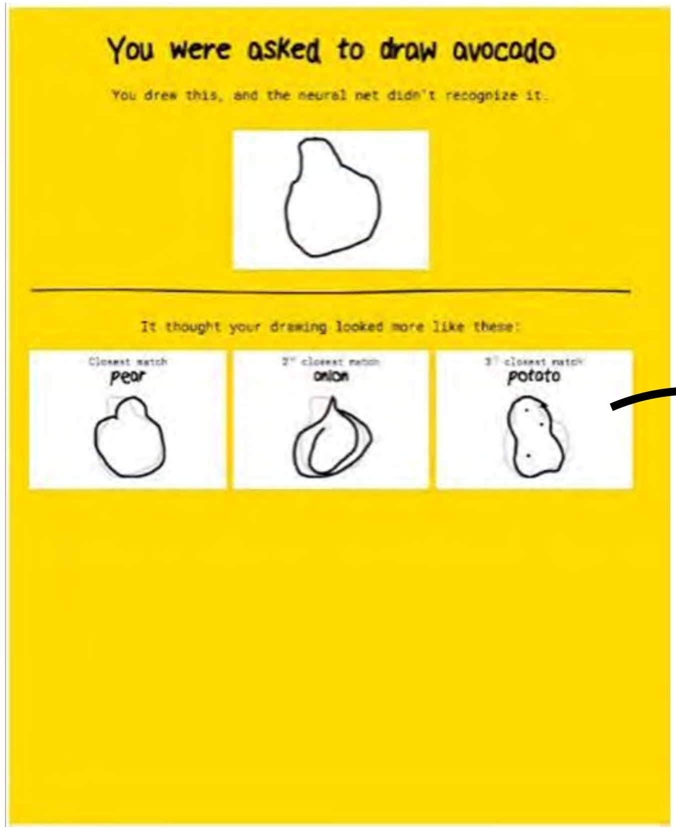
Your Mortgage was rejected since your monthly income is smaller than your neighbour's.

Contrastive example-based Explanations

Improves Undertrust



Normative Explanations



Comparative Explanations

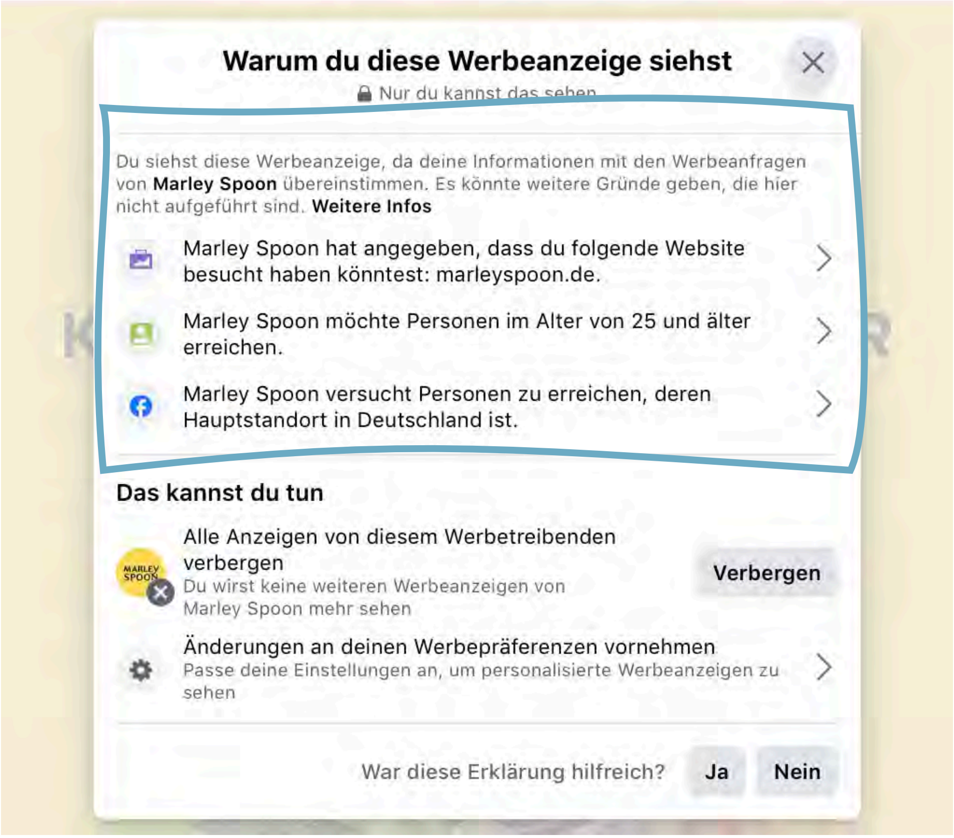
Avoids Overtrust

Insights from Social Sciences

- **Explanations are contrastive:** Why C instead of Y?
- **Explanations are selective:** Show the most important information that contributed to a decision (at the cost of completeness)

[Du 2020, Miller 2017; Molnar 2019]

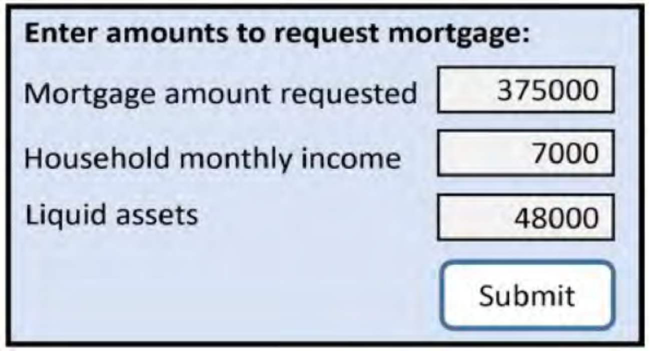
Explanations Are Selective



Source: www.facebook.com

Insights from Social Sciences

- **Explanations are contrastive:** Why C instead of Y?
- **Explanations are selective:** Show the most important information that contributed to a decision (at the cost of completeness)
- **Explanations are credible:** Be consistent with users' prior knowledge



Source: [Shneiderman 2020]

Your mortgage was rejected because you have an A-level degree.

Insights from Social Sciences

- **Explanations are contrastive:** Why C instead of Y?
- **Explanations are selective:** Show the most important information that contributed to a decision (at the cost of completeness)
- **Explanations are credible:** Be consistent with users' prior knowledge
- **Explanations are conversational:** Who reads an explanation? Allow users to raise queries

[Du 2020, Miller 2017; Molnar 2019]

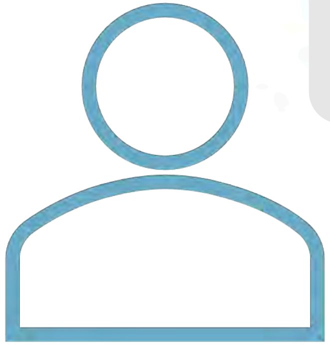
Explanations Are Conversational

Source: www.amazon.de

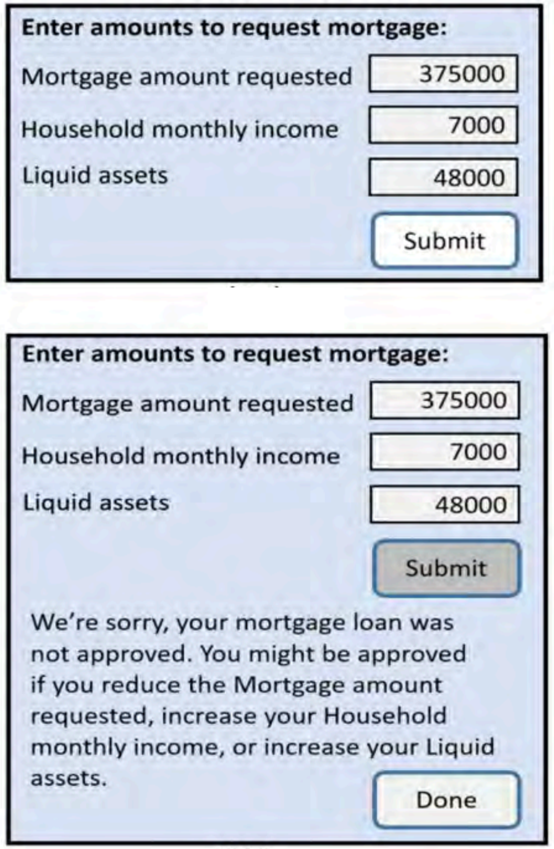
What happens if...

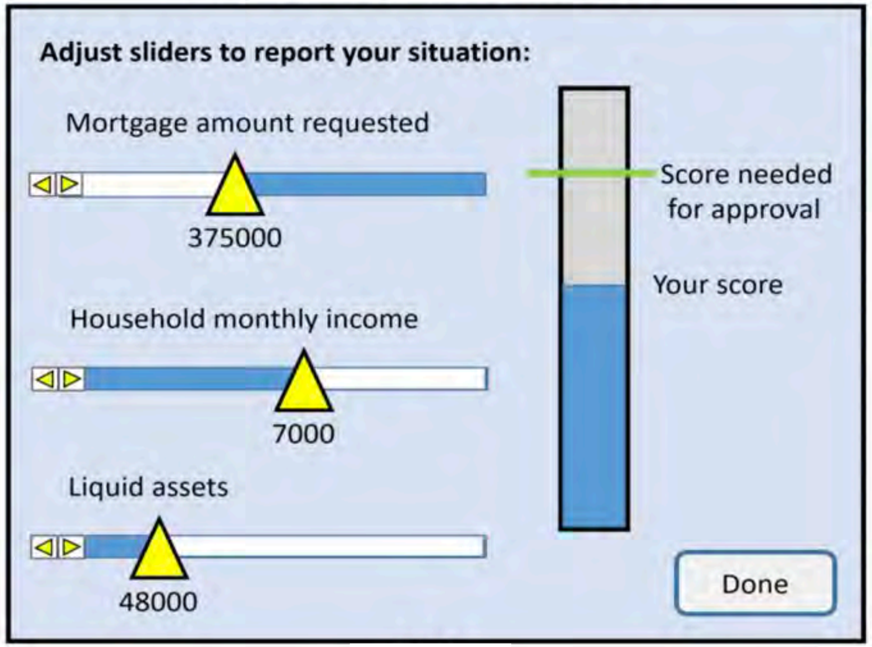
Why not also book X?

Why this book first?

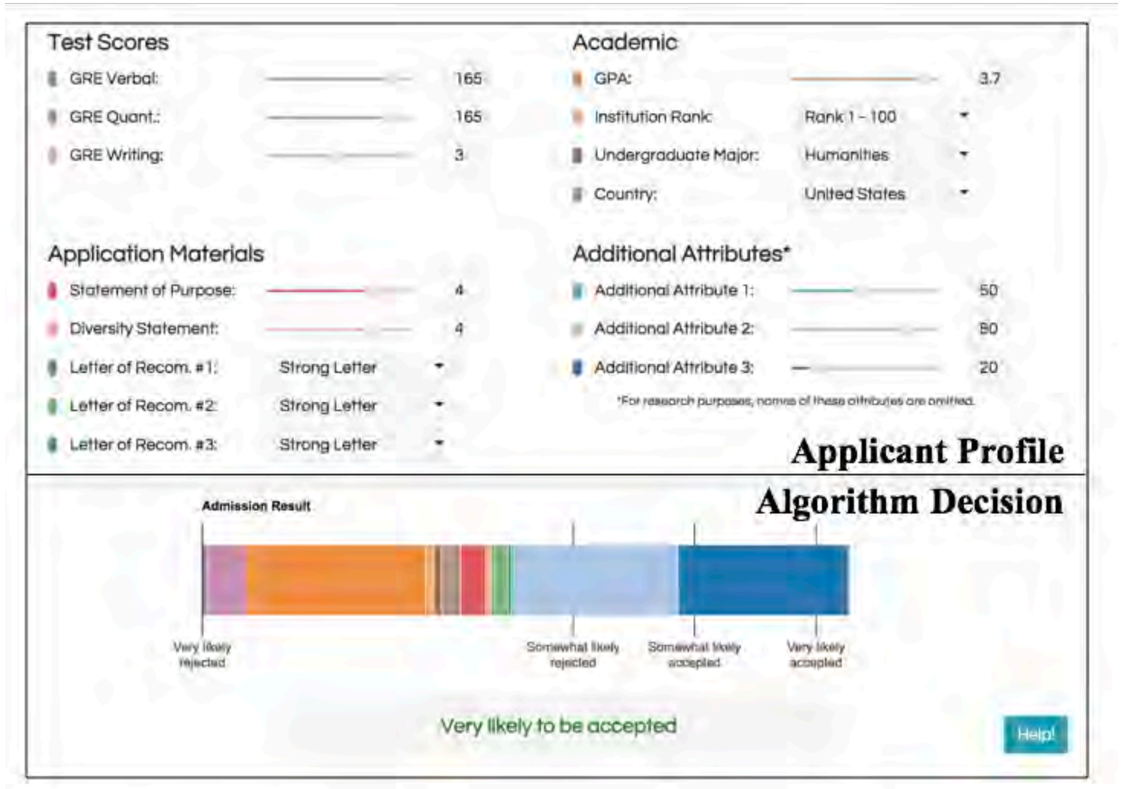


Post-hoc vs Interactive Explanations



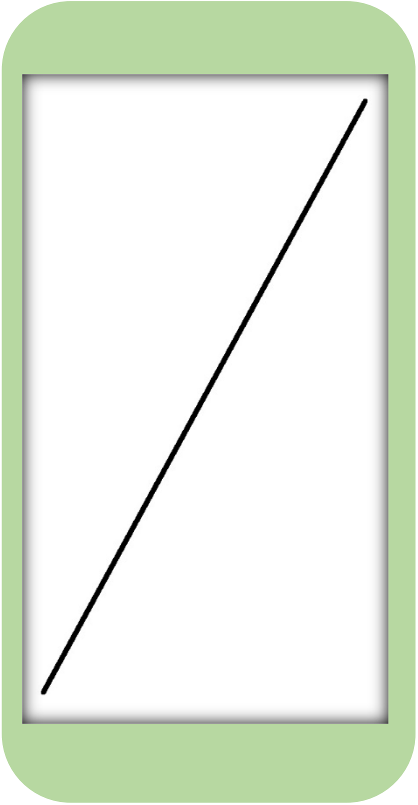


Interactive Explanations

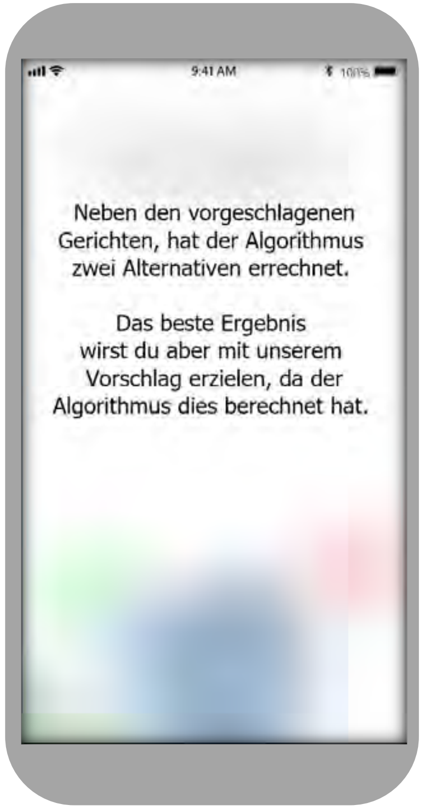


Placebo Explanations

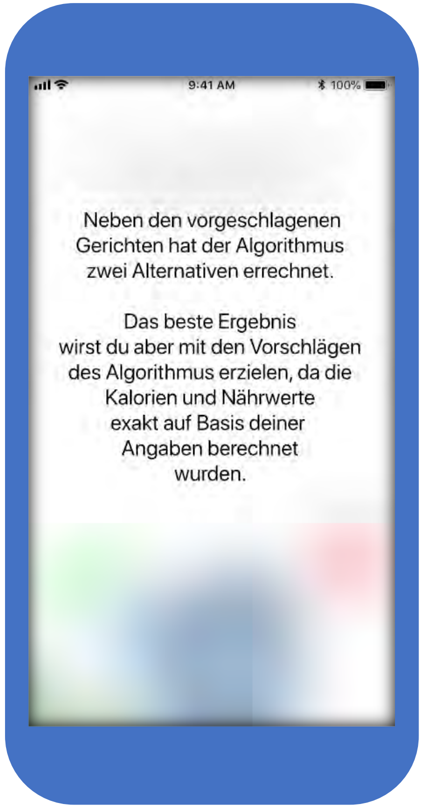
No Explanation

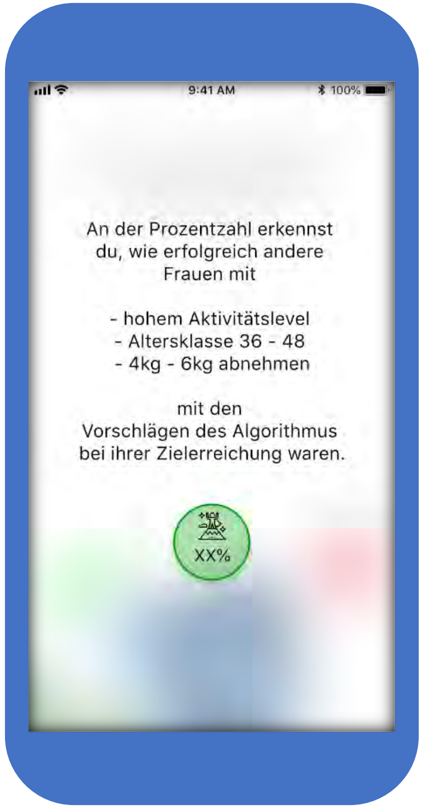


Placebo Explanation



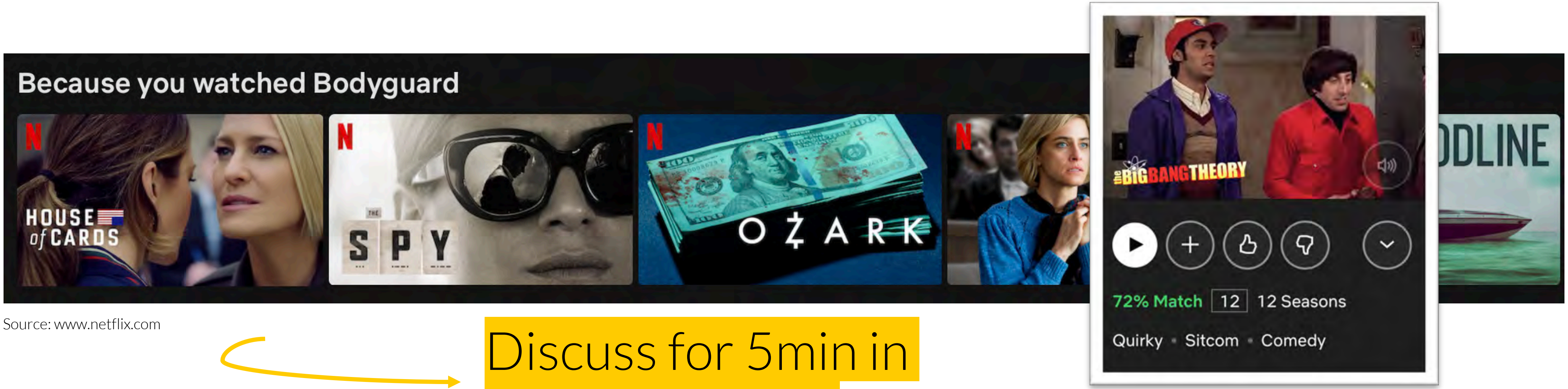
Actual Explanation





Discussion

How would you improve Netflix' explanation of why a particular movie was recommended?



Source: www.netflix.com

Discuss for 5min in breakout rooms

Overview

Which Problems Do Users Face?

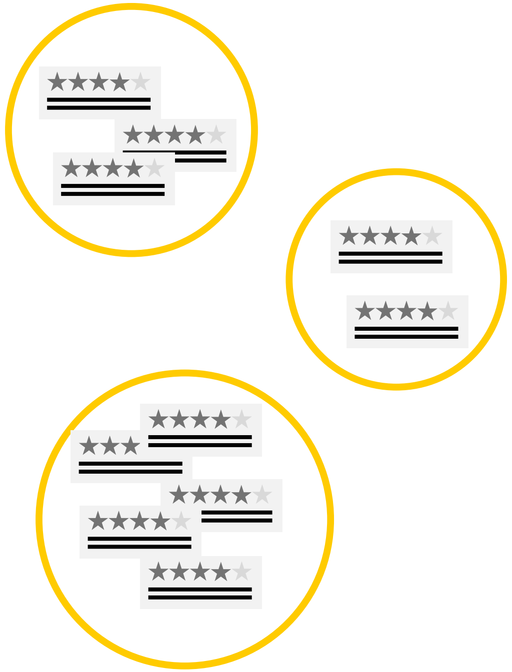
[Eiband et al. 2019a]

Research Design

1 **Reviews**

Source: play.google.com

2 **Topic Modeling**



3 **Problems**

Problem 1

Problem 2

Problem 3

...

Problem n

Research Design

1 **Reviews**

Source: play.google.com

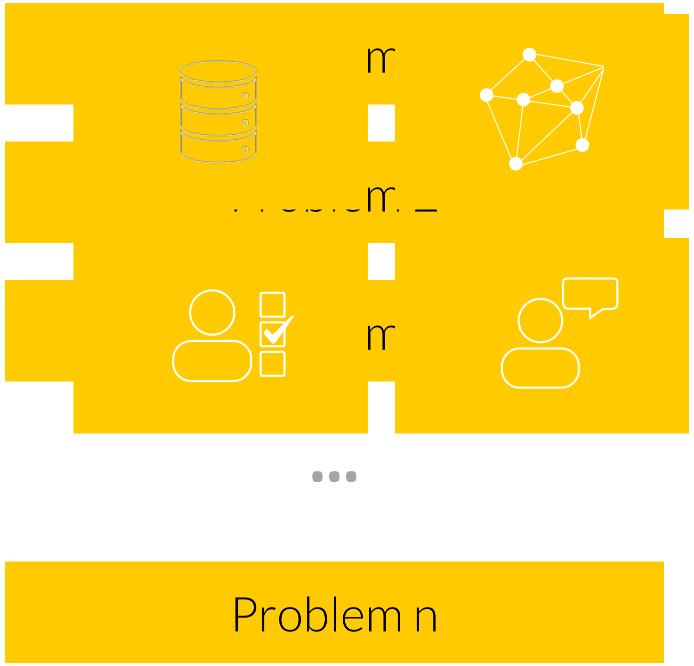
2 **Topic Modeling** / **Online Survey**

1 How often did you encounter similar situation?

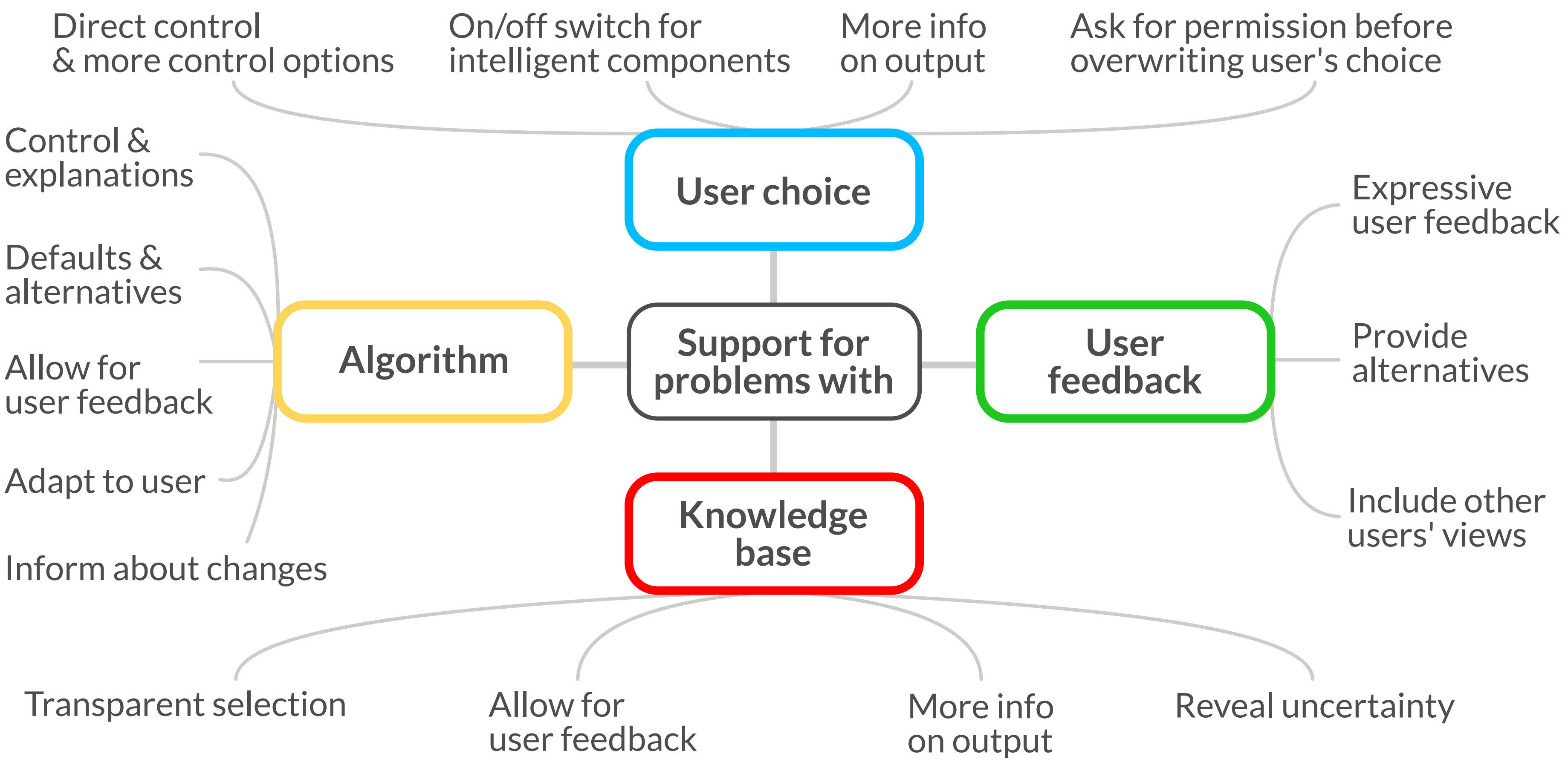
2 How did you cope with this situation?

3 How could the app support you in this situation?

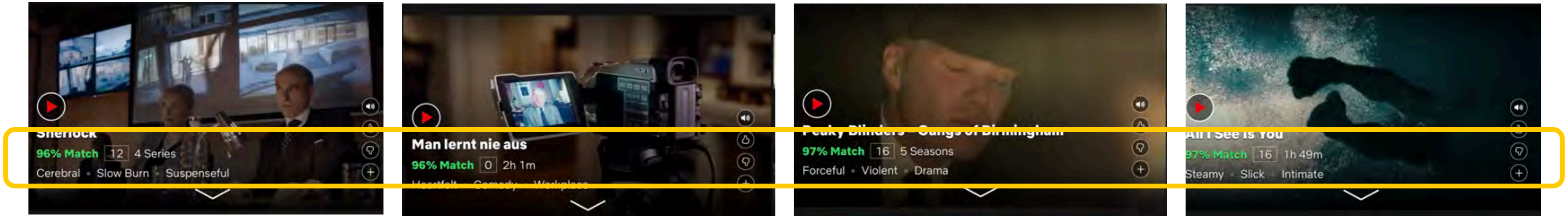
3 **Problems** / **System Support**



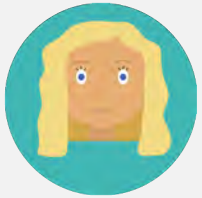
Support Strategies



Lack of Feedback Opportunities



Source: www.netflix.com



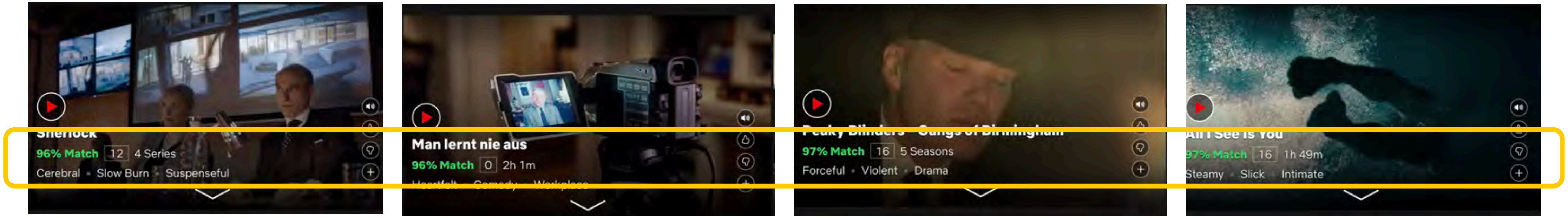
Eden Thomson

August 3, 2018

The **rating system is still horrible**, every movie I look at says 98% match like how am I supposed to **know if I should actually watch the movie if every movie is a match**. Bring back the **star system**. [...]

Bildquelle: eigene Anfertigung

Lack of Feedback Opportunities

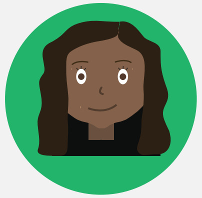


Source: www.netflix.com

“Suggest movies which only match my movies by 50% but have been received good ratings (by other users).“

“The system should show me more TV shows that all people like [….], not only those that I will probably like.”

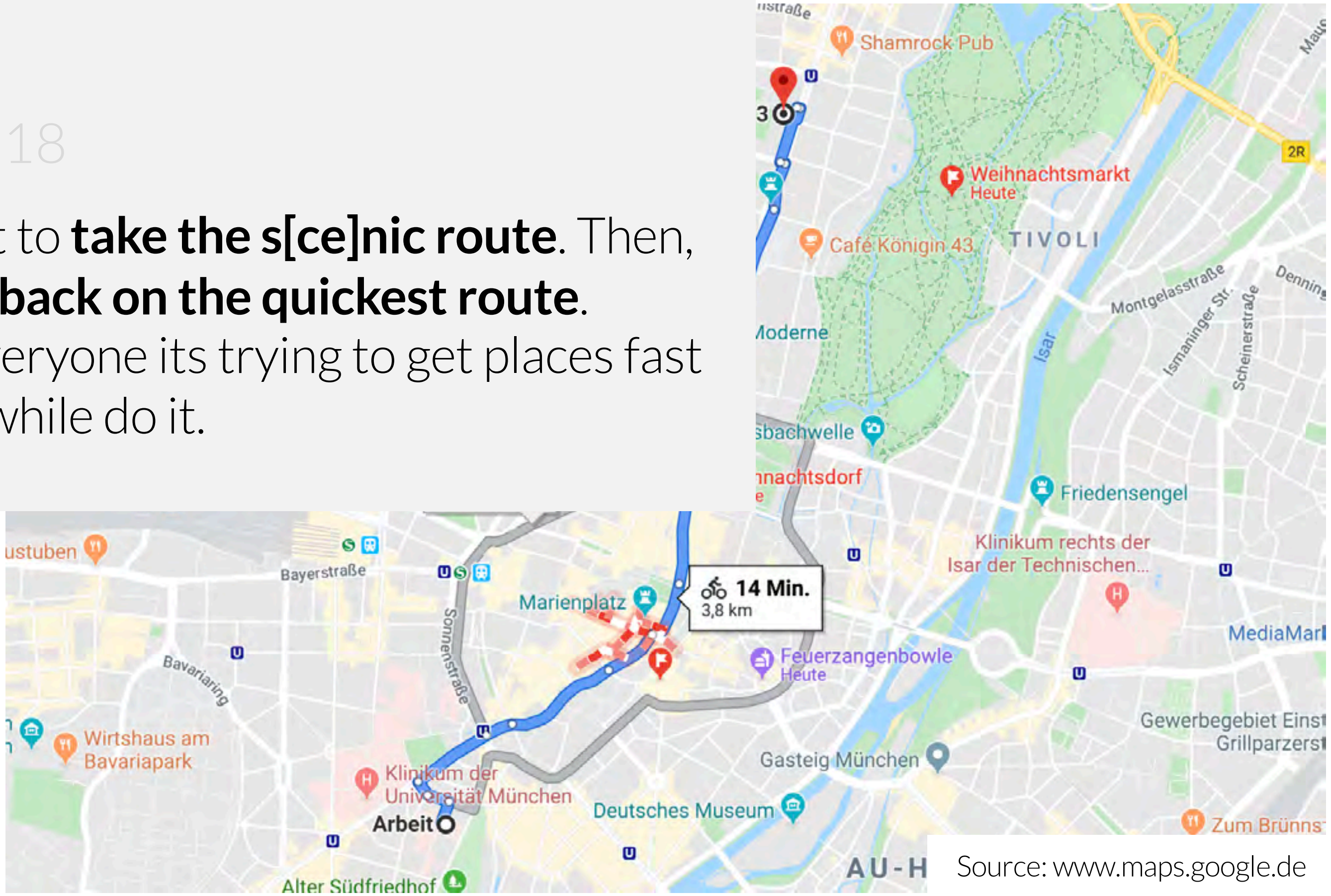
Lack of User Control



Charlotte Brooks

July 20, 2018

I choose [a route] because I want to **take the s[ce]nic route**. Then, **without telling me** just puts me **back on the quickest route**. Which **drives me insane** - not everyone its trying to get places fast some of us like to see the world while do it.

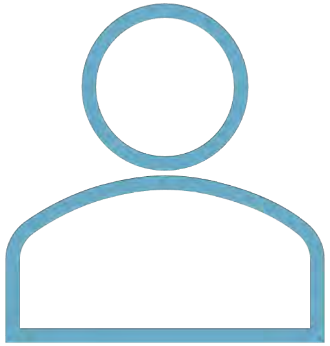


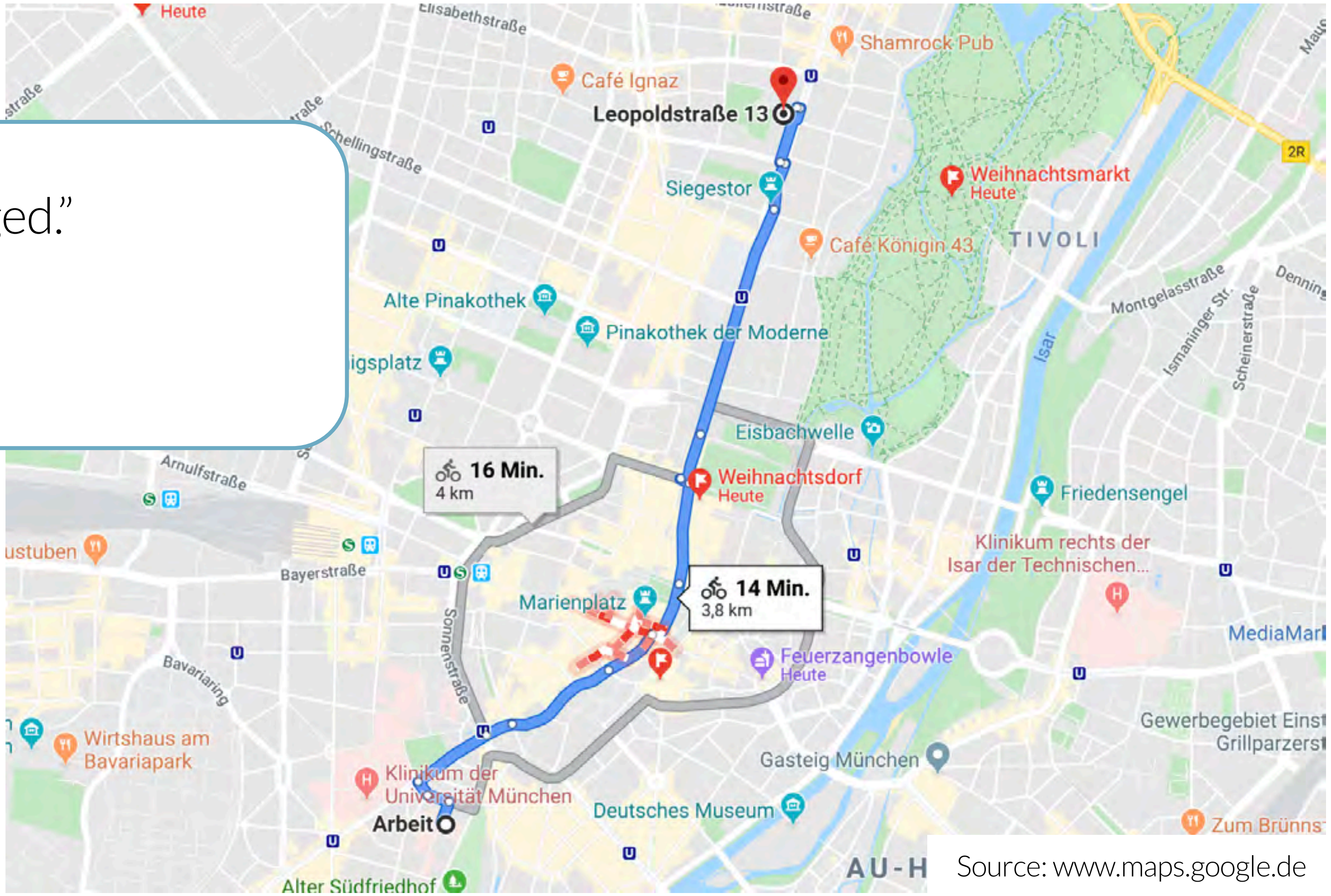
Source: www.maps.google.de

Lack of User Control

"**Ask for permission** bevor the route is changed."

"At least **offer the option** "Don't change"."



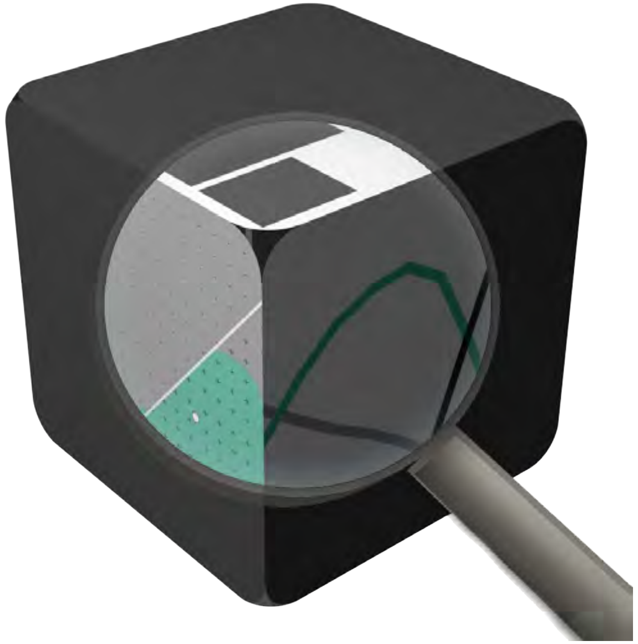


Source: www.maps.google.de

Take Aways

- Machine learning models are **black boxes** which are opaque to developers and end users
- As a consequence, there are **several challenges** for individual users as well as society when employing machine learning
- Machine learning models have to be **explainable** – either by choosing **intrinsic** or **post-hoc models**
- **Explanations** have to be designed carefully to be **easily understandable**

Beyond Explainability



Source: Courtesy of Quay Au

Trustworthy Certification:
External Reviews

Safety Culture:
Organisational Design

Reliable Systems:
Software Engineering

Overview

1 **Transparency for Intelligent Systems**

- The Black Box Problem
- Resulting Challenges for Society
- Explainable AI
- What Makes a Good Explanation
- User Problems and Support

2 **Transparency for Personality-Targeting**

- Personality and Personality-Targeting
- Requirements for Explanations for Personality-Targeting
- How to Trick AI

Online

Thank you to Malin Eiband and Michael Chromik

– who contributed to earlier versions of this slide deck

References

- Balise Agüera y Arcas, Alexander Todorov, and Margaret Mitchell. 2018. Do algorithms reveal sexual orientation or just expose our stereotypes? https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477
- Danny Azucar, Davide Marengo, and Michele Settanni. 2018. Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. Personality and individual differences 124: 150-159. DOI: https://doi.org/10.1016/j.paid.2017.12.018
- Joshua Benton. 2019. As Notre Dame burned, an algorithmic error at YouTube put information about 9/11 under news videos. https://www.niemanlab.org/2019/04/as-notre-dame-burned-an-algorithmic-error-at-youtube-put-information-about-9-11-under-news-videos/
- Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society 3, no. 1. DOI: https://doi.org/10.1177/2053951715622512
- Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19). Association for Computing Machinery, New York, NY, USA, 258–262. DOI:https://doi.org/10.1145/3301275.3302289
- Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15). Association for Computing Machinery, New York, NY, USA, 1721–1730. DOI:https://doi.org/10.1145/2783258.2788613
- Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery, New York, NY, USA, Paper 559, 1–12. DOI:https://doi.org/10.1145/3290605.3300789
- Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. https://arxiv.org/abs/1702.08608
- Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for Interpretable Machine Learning. Commun. ACM 63, 1 (December 2019), 68–77. DOI: https://doi.org/10.1145/3359786
- Malin Eiband, Sarah Theres Völkel, Daniel Buschek, Sophia Cook, and Heinrich Hussmann. 2019. When People and Algorithms Meet: User-reported Problems in Intelligent Everyday Applications. In Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19). ACM, New York, NY, USA, 96–106. DOI: http://dx.doi.org/10.1145/3301275.3302262

References

- Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The Impact of Placebic Explanations on Trust in Intelligent Systems. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19). Association for Computing Machinery, New York, NY, USA, Paper LBW0243, 1–6. DOI: https://doi.org/10.1145/3290607.3312787
- Motahhare Eslami, Sneha R. Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. 2018. Communicating Algorithmic Process in Online Behavioral Advertising. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). ACM, New York, NY, USA, Article Paper 432, 13 pages. DOI: http://dx.doi.org/10.1145/3173574.3174006
- Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust physical-world attacks on deep learning visual classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1625-1634. http://openaccess.thecvf.com/content_cvpr_2018/papers/Eykholt_Robust_Physical-World_Attacks_CVPR_2018_paper.pdf
- Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572. https://arxiv.org/pdf/1412.6572.pdf
- Guardian. 2018. https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine
- David Gunning. 2017. Explainable Artificial Intelligence (XAI). Defense Advanced Research Projects Agency (DARPA), nd Web, 2. https://www.darpa.mil/attachments/XAIProgramUpdate.pdf
- Karen Hao. 2019. AI is sending people to jail—and getting it wrong. https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/
- Simon Hurtz. 2019. Youtube verwechselt Feuer in Paris mit 9/11. Süddeutsche Zeitung. https://www.sueddeutsche.de/digital/paris-notre-dame-youtube-algorithmus-filter-1.4411910
- Shagun Jhaver, Yoni Karpfen, and Judd Antin. 2018. Algorithmic Anxiety and Coping Strategies of Airbnb Hosts. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). ACM, New York, NY, USA, Article 421, 12 pages. DOI: https://doi.org/10.1145/3173574.3173995

- Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust physical-world attacks on deep learning visual classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1625-1634. http://openaccess.thecvf.com/content_cvpr_2018/papers/Eykholt_Robust_Physical-World_Attacks_CVPR_2018_paper.pdf
- Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572. https://arxiv.org/pdf/1412.6572.pdf

References

- Guardian. 2018. https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine
- David Gunning. 2017. Explainable Artificial Intelligence (XAI). Defense Advanced Research Projects Agency (DARPA), nd Web, 2. https://www.darpa.mil/attachments/XAIProgramUpdate.pdf
- Karen Hao. 2019. AI is sending people to jail—and getting it wrong. https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/
- Simon Hurtz. 2019. Youtube verwechselt Feuer in Paris mit 9/11. Süddeutsche Zeitung. https://www.sueddeutsche.de/digital/paris-notre-dame-youtube-algorithmus-filter-1.4411910
- Shagun Jhaver, Yoni Karpfen, and Judd Antin. 2018. Algorithmic Anxiety and Coping Strategies of Airbnb Hosts. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). ACM, New York, NY, USA, Article 421, 12 pages. DOI: https://doi.org/10.1145/3173574.3173995
- Kayser-Brill, Nicolas. 2020a. Female historians and male nurses do not exist, Google Translate tells its European users. Algorithm Watch. https://algorithmwatch.org/en/story/google-translate-gender-bias/
- Kayser-Brill, Nicolas. 2020b. Dutch city uses algorithm to assess home value, but has no idea how it works. Algorithm Watch. https://algorithmwatch.org/en/story/woz-castricum-gdpr-art-22/
- Been Kim, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! criticism for interpretability." Advances in neural information processing systems 29 (2016): 2280-2288.
- Logan Kugler. 2018. AI judges and juries. Commun. ACM 61, 12 (December 2018), 19–21. DOI:https://doi.org/10.1145/3283222
- Sam Levin. 2017. New AI can guess whether you're gay or straight from a photograph. Guardian. https://www.theguardian.com/technology/2017/sep/07/new-artificial-intelligence-can-tell-whether-youre-gay-or-straight-from-a-photograph
- Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 1–15. DOI:https://doi.org/10.1145/3313831.3376590
- Brian Y. Lim and Anind K. Dey. 2009. Assessing demand for intelligibility in context-aware applications. In Proceedings of the 11th international conference on Ubiquitous computing (UbiComp '09). ACM, New York, NY, USA, 195-204. DOI: https://doi.org/10.1145/1620545.1620576

References

- Brian Y. Lim and Anind K. Dey. 2010. Toolkit to support intelligibility in context-aware applications. In Proceedings of the 12th ACM international conference on Ubiquitous computing (UbiComp '10). ACM, New York, NY, USA, 13-22. DOI: https://doi.org/10.1145/1864349.1864353
- Brian Y. Lim and Anind K. Dey. 2011. Design of an Intelligible Mobile Context- aware Application. In Proceedings of the 2011 International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '11). ACM, New York, NY, USA, 157–166. https://doi.org/10.1145/2037373.2037399
- Sandra C. Matz, Michal Kosinski, Gideon Nave, and David J. Stillwell. 2017. Psychological targeting as an effective approach to digital mass persuasion. Proceedings of the national academy of sciences 114, no. 48: 12714-12719. https://doi.org/10.1073/pnas.1710966114
- Robert R. McCrae and Paul T. Costa, Jr. 2008. The five-factor theory of personality. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), Handbook of personality: Theory and research (p. 159–181). The Guilford Press.
- Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267: 1-38. DOI: https://doi.org/10.1016/j.artint.2018.07.007
- Christoph Molnar. 2019. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/
- Hervé Phaure and Erwan Robin. 2020. Explain Artificial Intelligence for Credit Risk Management. Deloitte. https://www2.deloitte.com/content/dam/Deloitte/fr/Documents/risk/Publications/deloitte_artificial-intelligence-credit-risk.pdf
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135–1144. DOI:https://doi.org/10.1145/2939672.2939778
- Ben Shneiderman. 2020. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. ACM Trans. Interact. Intell. Syst.10, 4, Article 26 (December 2020), 31 pages. DOI:https://doi.org/10.1145/3419764
- Clemens Stachl, Quay Au, Ramona Schoedel, Samuel D. Gosling, Gabriella M. Harari, Daniel Buschek, Sarah Theres Völkel et al. 2020. Predicting personality from patterns of behavior collected with smartphones. Proceedings of the National Academy of Sciences 117, no. 30: 17680-17687. DOI: https://doi.org/10.1073/pnas.1920484117

References

- Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. User Modeling and User-Adapted Interaction 22.4-5: 399-439. DOI: https://doi.org/10.1007/s11257-011-9117-5
- Sarah Theres Völkel, Renate Haeuslschmid, Anna Werner, Heinrich Hussmann, and Andreas Butz. 2020. How to Trick AI: Users' Strategies for Protecting Themselves from Automatic Personality Assessment. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 1–15. DOI:https://doi.org/10.1145/3313831.3376877
- Yilun Wang and Michal Kosinski. 2018. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. Journal of personality and social psychology 114.2: 246. DOI: https://doi.org/10.1037/pspa0000098
- Julia K. Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas et al. 2019. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA dermatology 155, no. 10: 1135-1141. DOI:10.1001/jamadermatol.2019.1735
- Yudkowsky, Eliezer. 2008. Artificial Intelligence as a Positive and Negative Factor in Global Risk. In Global Catastrophic Risks, edited by Nick Bostrom and Milan M. Ćirković, 308–345. New York: Oxford University Press. https://intelligence.org/files/AIPosNegFactor.pdf