«Human-Centered Data Science»

Post-hoc Interpretability: Evaluating Explanation UIs

Prof. Dr. Claudia Müller-Birn

Human-Centered Computing, Institute of Computer Science

Freie Universität Berlin

June 30, 2022

Lecture Overview

Recap

Evaluating Explanation UIs (Evaluating XAI Interfaces, XAI Target Users, Evaluation Measures)

Integrating Evaluation in AI System Design (Nested Model, Iterative Process, Types of Evaluations)

Heuristic Evaluation (Process, Heuristics)

Break

Usability Testing (Process, Measuring Satisfaction)

Applying Evaluation Methods in System Design (Design Complexity Map)

Recap

Intrinsic vs. Post-hoc Interpretability Techniques

Use models that are intrinsically interpretable and known to be easy for humans to understand.

This is a fair decision.

$$P(d = 1 \mid L = l, G = m) = P(d = 1 \mid L = l, G = f)$$

Train a black box model and apply post-hoc interpretability techniques to provide explanations.

Intrinsic Interpretability Techniques

Post-hoc Interpretability Techniques

Questions adapted from D. Gunning, Explainable artificial intelligence (xAI), Tech. rep., Defense Advanced Research Projects Agency (DARPA) (2017).
https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

A Participatory Process for Interpretability Techniques

WHAT to explain?

Expert Mental Model

- What happens to the best of our knowledge?
- What can be explained?
- What does an expert mental model of the system look like?

User Mental Model

- How do users currently make sense of the system?
- What is the user mental model of the system based on its current UI?
- How does it differ from the expert mental model?

Target Mental Model

- Which key components of the algorithm do users want to be made transparent in the UI?
- To what extent are users actually interested in the rationale behind the algorithms?

HOW to explain?

User Mental Model

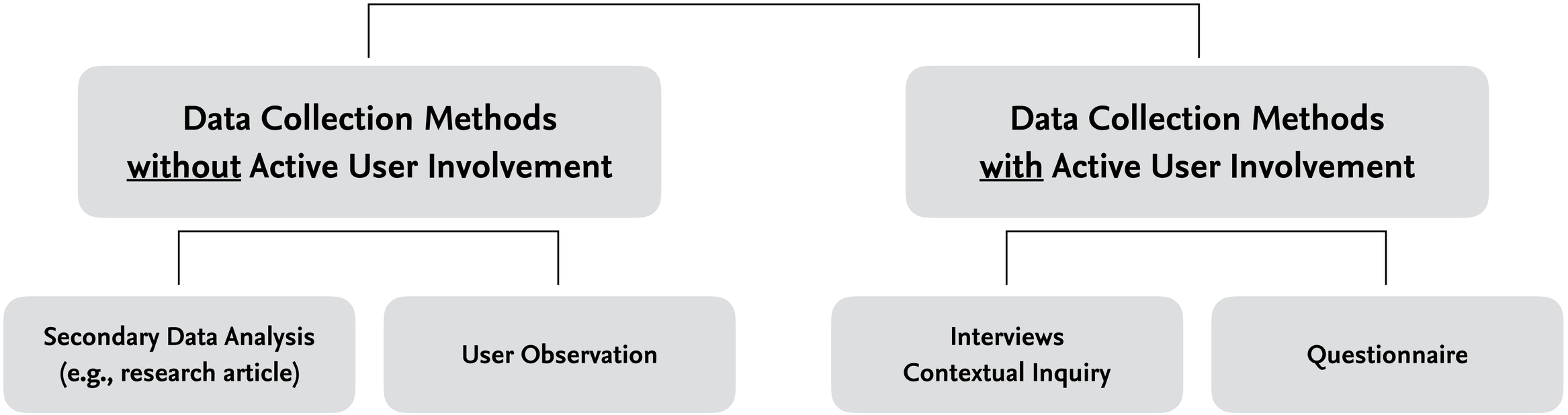
- How can the target mental model be reached through UI design?
- How and where can transparency be integrated into the UI of the system?

Target Mental Model

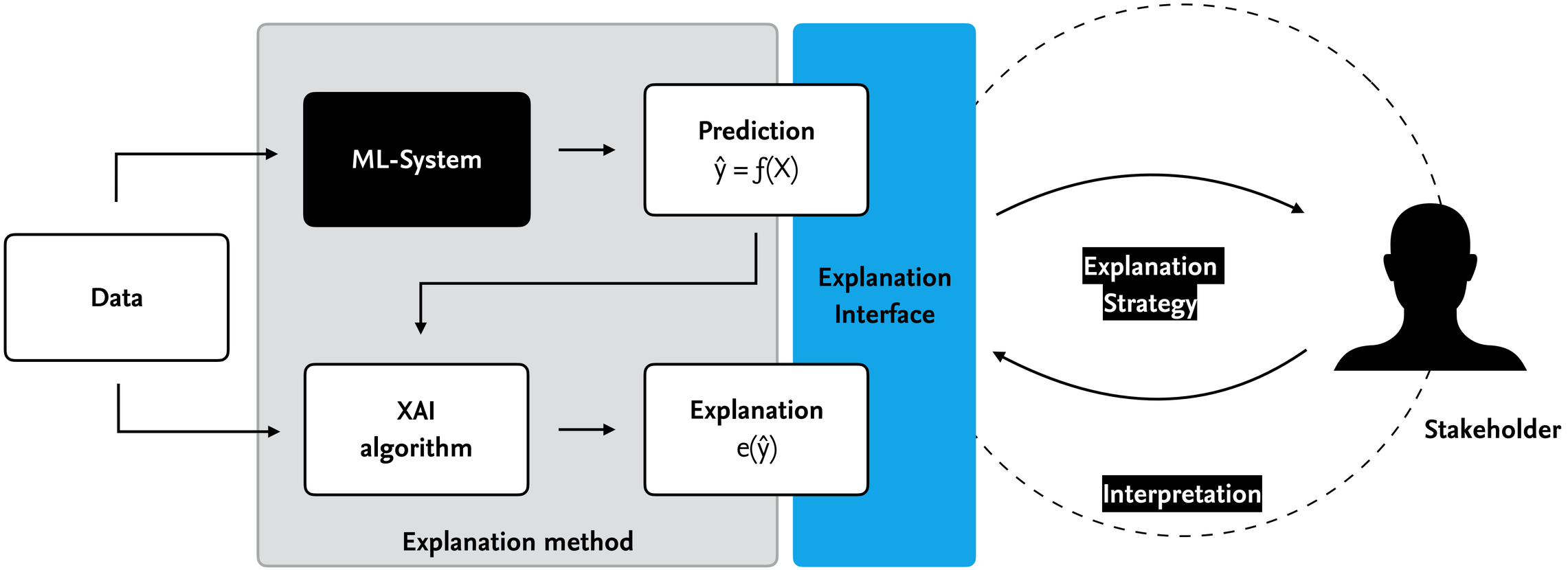
- How has the users mental model developed?
- Has the target mental model been reached?

Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In Conf. on Intelligent User Interfaces, 211–223.

Eliciting Explanation Needs



Ensuring Interpretability by Explanation User Interfaces



Jesse Josua Benjamin, Christoph Kinkeldey, Claudia Müller-Birn, Tim Korjakow, and Eva-Maria Herbst. 2022. Explanation Strategies as an Empirical-Analytical Lens for Socio-Technical Contextualization of Machine Learning Interpretability. Proc. ACM Hum.-Comput. Interact. 6, GROUP.

Principles of Good Design

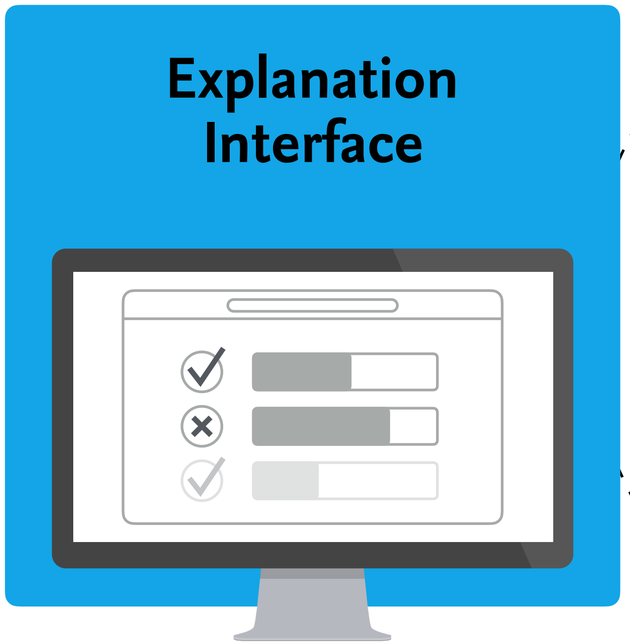
Aim of a good design is to minimize the gulfs of execution and evaluation.

In order to do this the design should

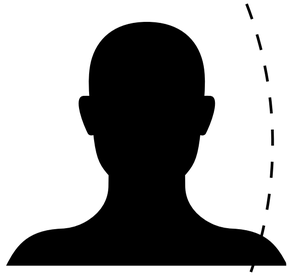
- Help the user build the correct conceptual model of the system
- Make the right parts visible
- Provide memory aids to the user
- Provide good feedback
- Accommodate errors

Norman, D. (2013). The design of everyday things: Revised and expanded edition. Basic books.

Possible Outcome of Using Explanations



Interpretation



User follows correct suggestions
(appropriate reliance)

User ignores correct suggestions
(self-reliance)

User follows incorrect suggestions
(over-reliance)

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. 2015 International Conference on Healthcare Informatics, 160–169. ICHI.2015.26

Improving XAI Outcome by Using Interventions

Automatic Thinking

» Intuitive understanding

» Decisions are made instinctively, emotionally and unconsciously

» Finds application in repeated and practiced actions

Reflective Thinking

» Logical Reasoning

» Decisions are based on a rational process that is consciously carried out

» Finds application in deliberate, slow and strenuous actions

Kahneman, Daniel. Thinking, Fast and Slow. London: Penguin Books, 2012.

Reduce Overreliance on AI in AI-assisted Decision-making

Human's **cognitive motivation influences the effectiveness** of XAI solutions.

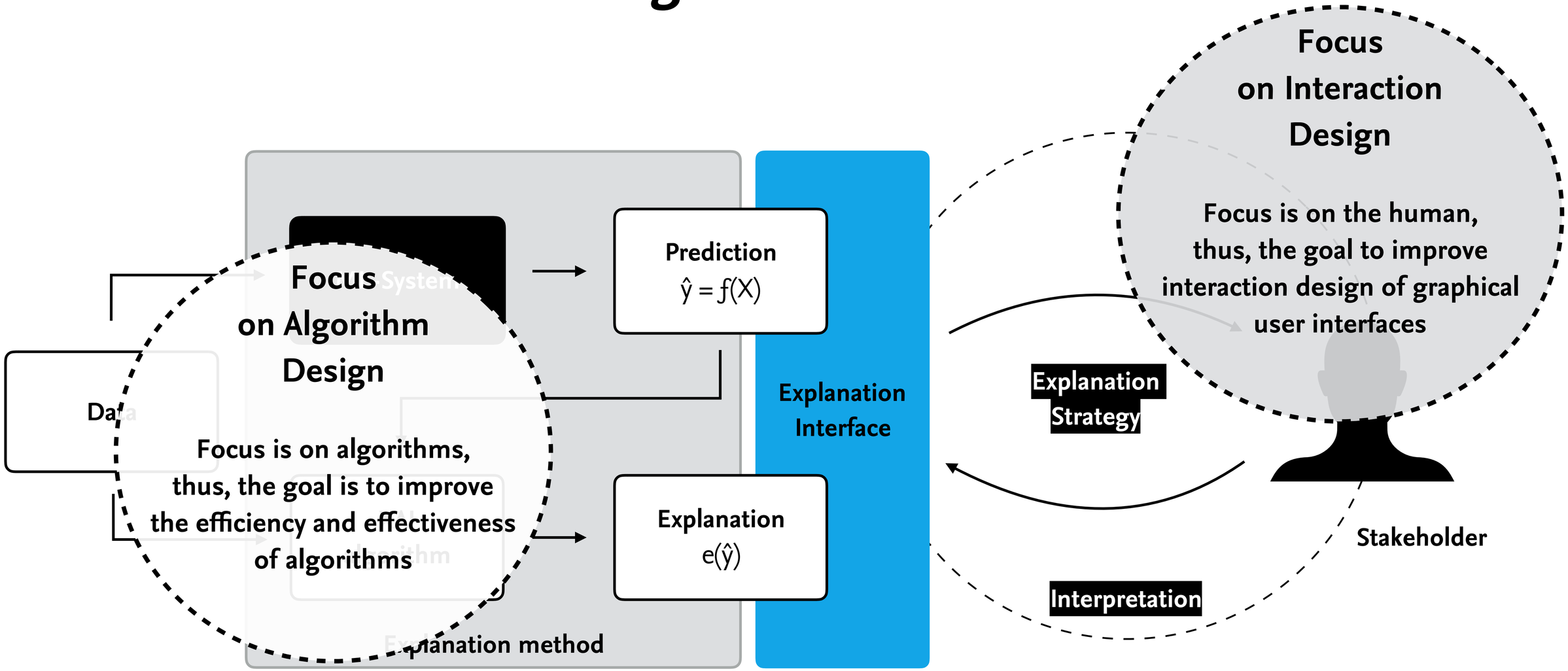
Future interventions might be tailored to **account for the differences in intrinsic cognitive motivation:** stricter interventions might benefit more and still be accepted by people with lower intrinsic cognitive motivation.

Developing adaptive strategies for providing different interventions based on models that predict the performance of human+AI teams on particular task instances.

Buçinca, Zana, Maja Barbara Malaya, und Krzysztof Z. Gajos. „To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making". Proceedings of the ACM on Human-Computer Interaction_ 5, Nr. CSCW (2021): 188:1-188:21.

Evaluating Explanation UIs

Possible Foci in Evaluating XAI Interfaces



Possible Foci in Evaluating XAI Interfaces (*cont.*)

AI Perspective

» Evaluation focuses on algorithms

» Offline experiments that measures, for example, precision and recall

HCI Perspective

» Evaluation focuses on interaction design

» Studies for measuring, for example, usability, or experiments evaluating, for example, the mental load of users

The centrality of context.
Putting human, technologies and researchers in their place.
Explicit focus on values in design.

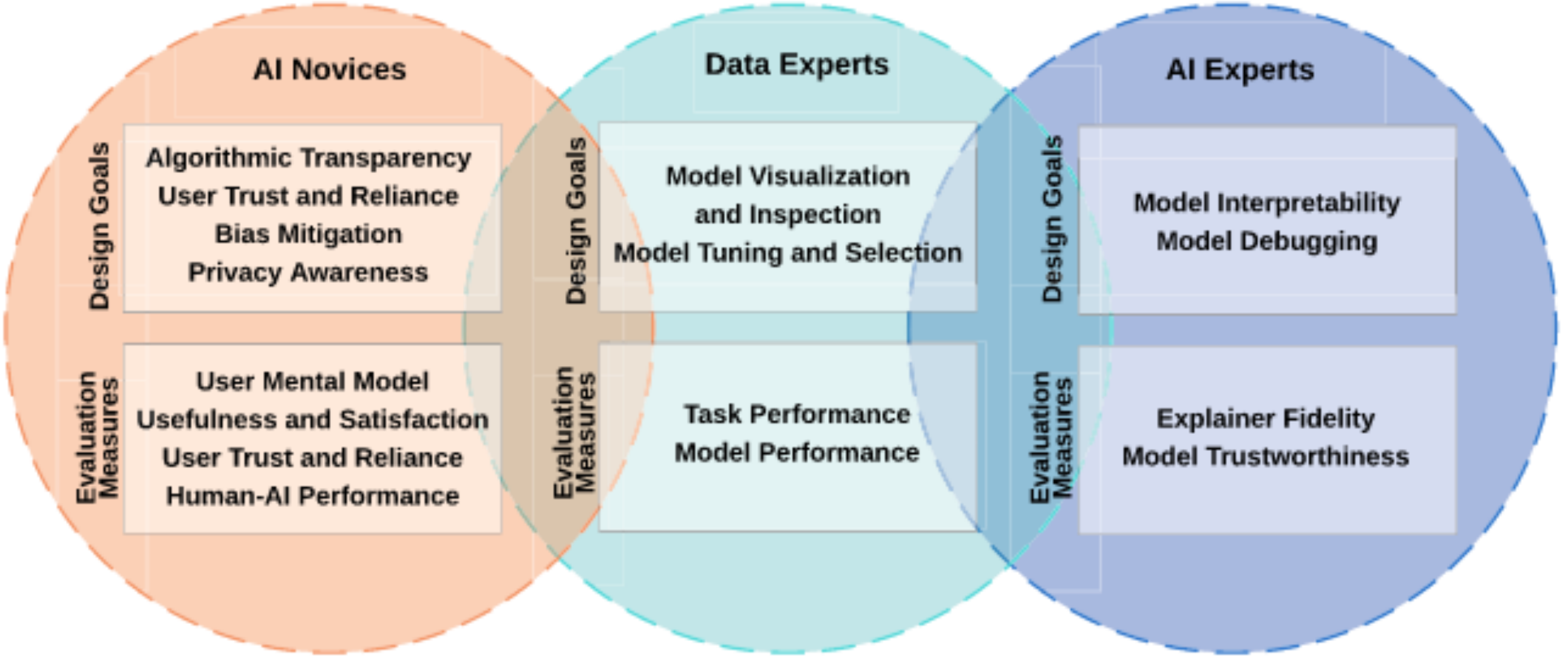
Frauenberger, C., & Purgathofer, P. (2019). Ways of thinking in informatics. Commun. ACM, 62(7), 58–64.
Jameson, A., & Riedl, J. (2011). Introduction to the Transactions on Interactive Intelligent Systems. TiiS, 1(1), 1–6.
Harrison, S., Sengers, P., & Tatar, D. G. (2011). Making epistemological trouble - Third-paradigm HCI as successor science. Interacting with Computers, 23(5), 385–392.

Defining Evaluation

> “Evaluation is integral to the design process. Evaluators collect information about users’ or potential users’ experiences when interacting with a prototype, an app, [...] an application, or a design artifact such as a screen sketch. [...]
>
> (Preece et al. 2002. p. 433)

Rogers, Y., Sharp, H., & Preece, J. (2002). Interaction Design: Beyond Human-Computer Interaction, Jon Wiley & Sons.

XAI Target Groups



Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Transactions on Interactive Intelligent Systems, 11(3–4), 1–45. https://doi.org/10.1145/3387166

XAI Evaluation Measures

Explanation Usefulness and Satisfaction, such as

» User Satisfaction (Interview and Self-report, likert-scale questionnaire, expert case study)

» Explanation Usefulness (task duration, cognitive load, engagement with explanations)

Mental Model, such as

» User Understanding of Model (interview, self-explanations)

» Model Output Prediction (user prediction of model output)

» Model Failure Prediction (user prediction of model failure)

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Transactions on Interactive Intelligent Systems, 11(3–4), 1–45. https://doi.org/10.1145/3387166

XAI Evaluation Measures (*cont.*)

User Trust and Reliance, such as

» Subjective Measures (self-explanation and interview, likert-scale questionnaire)

» Objective Measures (user perceived system competence, user compliance with system, user perceived understandability)

Human-AI Task Performance, such as

» User Performance (task performance, task throughput, model failure prediction)

» Model Performance (model accuracy, model tuning and selection)

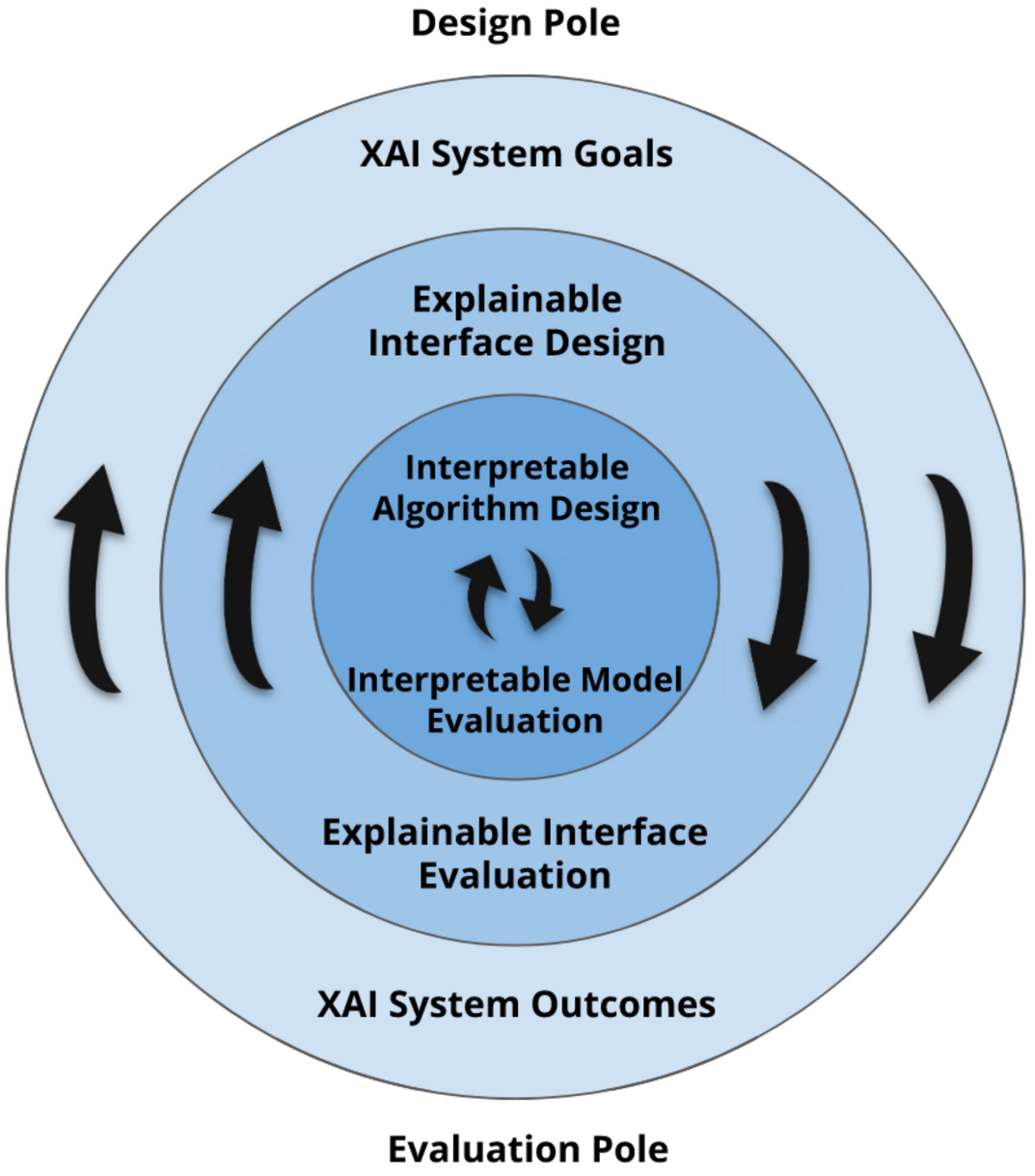
Computational Measures, such as

» Explainer Fidelity (simulated experiments, sanity checks, comparative evaluation)

» Model Trustworthiness (debugging model and training, human-grounded evaluation)

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Transactions on Interactive Intelligent Systems, 11(3–4), 1–45. https://doi.org/10.1145/3387166

Integrating Evaluation in AI System Design



Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Transactions on Interactive Intelligent Systems, 11(3–4), 1–45.

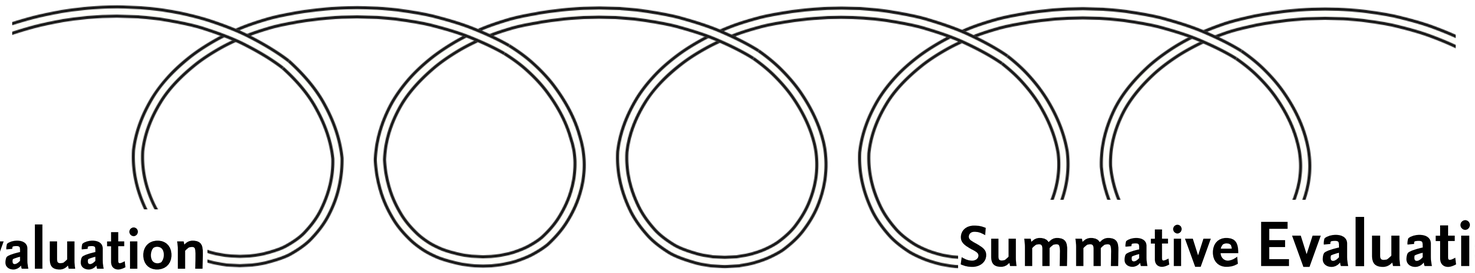
Iterative Process of Evaluation

1
Prototype
conceptual model

2
Detailed Model

3
Integrated Product
online help, documentation, etc.

4
“Out-of-Box” Experience
getting started



Formative Evaluation

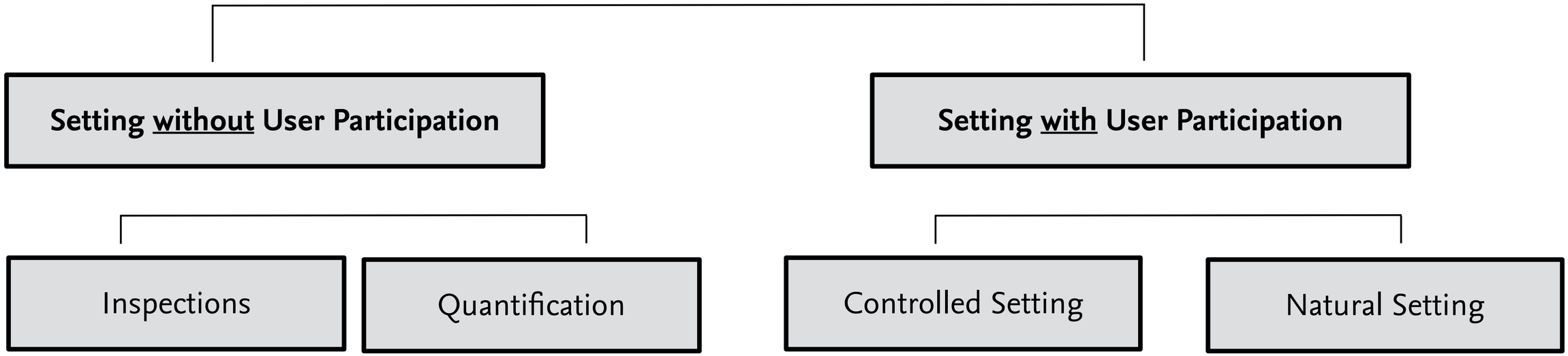
- Evaluate new system as it is being designed and developed
- Assumption is that the system is incomplete and that feedback can be quickly folded back into design process

Summative Evaluation

- Evaluation performed on a completed system
- Provides information regarding how well system meets expectations and goals previously decided upon
- Used to assess competing, completed products

Barnum, C. M. (2011). Usability Testing Essentials. Elsevier. https://www.oreilly.com/library/view/usability-testing-essentials/9780123750921/

Types of Evaluation Methods



» Heuristic evaluation

What is Heuristic Evaluation?

Heuristic evaluation is a systematic inspection of a user interface design for usability.

The goal is to find the usability problems in a user interface design as part of an iterative design process.

It involves a small set of evaluators examine the interface and judge its compliance with recognized usability principles.

Selected Sets of Heuristics Principles

» Amershi, S. et al. (2019) Guidelines for Human-AI Interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–13.

» Langevin, R., Lordon, R. J., Avrahami, T., Cowan, B. R., Hirsch, T., & Hsieh, G. (2021). Heuristic Evaluation of Conversational Agents. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–15.

Nielsen, J., and Molich, R.: Heuristic evaluation of user interfaces, Proc. ACM CHI'90 Conf. (Seattle, WA, 1-5 April), 249-256. (1990).
Nielsen, Jakob. "How to conduct a heuristic evaluation." Nielsen Norman Group 1 (1995): 1-8.

18 Guidelines for Human-AI Interaction

1. **Make clear what the system can do.**
   Help the user understand what the AI system is capable of doing.
2. **Make clear how well the system can do what it can do.**
   Help the user understand how often the AI system may make mistakes.
3. **Time services based on context.**
   Time when to act or interrupt based on the user's current task and environment.
4. **Show contextually relevant information.**
   Display information relevant to the user's current task and environment.
5. **Match relevant social norms.**
   Ensure the experience is delivered in a way that users would expect, given their social and cultural context.
6. **Mitigate social biases.**
   Ensure the AI system's language and behaviors do not reinforce undesirable and unfair stereotypes and biases.
7. **Support efficient invocation.**
   Make it easy to invoke or request the AI system's services when needed.
8. **Support efficient dismissal.**
   Make it easy to dismiss or ignore undesired AI system services.
9. **Support efficient correction.**
   Make it easy to edit, refine, or recover when the AI system is wrong.
10. **Scope services when in doubt.**
    Engage in disambiguation or gracefully degrade the AI system's services when uncertain about a user's goals.
11. **Make clear why the system did what it did.**
    Enable the user to access an explanation of why the AI system behaved as it did.
12. **Remember recent interactions.**
    Maintain short term memory and allow the user to make efficient references to that memory.
13. **Learn from user behavior.**
    Personalize the user's experience by learning from their actions over time.
14. **Update and adapt cautiously.**
    Limit disruptive changes when updating and adapting the AI system's behaviors.
15. **Encourage granular feedback.**
    Enable the user to provide feedback indicating their preferences during regular interaction with the AI system.
16. **Convey the consequences of user actions.**
    Immediately update or convey how user actions will impact future behaviors of the AI system.
17. **Provide global controls.**
    Allow the user to globally customize what the AI system monitors and how it behaves.
18. **Notify users about changes.**
    Inform the user when the AI system adds or updates its capabilities.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3290605.3300233

18 Guidelines for Human-AI Interaction

6 **Mitigate social biases.**
Ensure the AI system's language and behaviors do not reinforce undesirable and unfair stereotypes and biases.

Example applications:
[Autocomplete, Product #2] "The autocomplete feature clearly suggests both genders [him, her] without any bias while suggesting the text to complete."

Example violations:
[Navigation, Product #1] "Regards the "Walking" transport there's no way to set an average walking speed. [The product] assumes users to be healthy."

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3290605.3300233

18 Guidelines for Human-AI Interaction

9 Support efficient correction.
Make it easy to edit, refine, or wrong.

Example applications:
[Navigation, Product #1] “If [the product] is wrong about where I parked my car, it provides an easy way to edit the location by dragging on the map.”

Example violations:
[Web Search, Product #1] “Searches can be easily corrected with a new query (which are sometimes suggested by [the product] itself). However, editing a seemingly AI system override to interpret ‘Sea of’ to ‘SEA to’ is not possible.”

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3290605.3300233

18 Guidelines for Human-AI Interaction

15 Encourage granular feedback.
Enable the user to provide feedback indicating their preference during interaction with the AI system.

Example applications:
[Email, Product #1] “The user can directly mark something as important, when the AI hadn’t marked it as that previously.”

Example violations:
[Voice Assistants, Product #2] “Once [the assistant] performed the task I had asked of it, there was no additional ability to customize the experience or give feedback on my satisfaction; even when I chose to remove the reminder right after I verbally requested it.”

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3290605.3300233

18 Guidelines for Human-AI Interaction

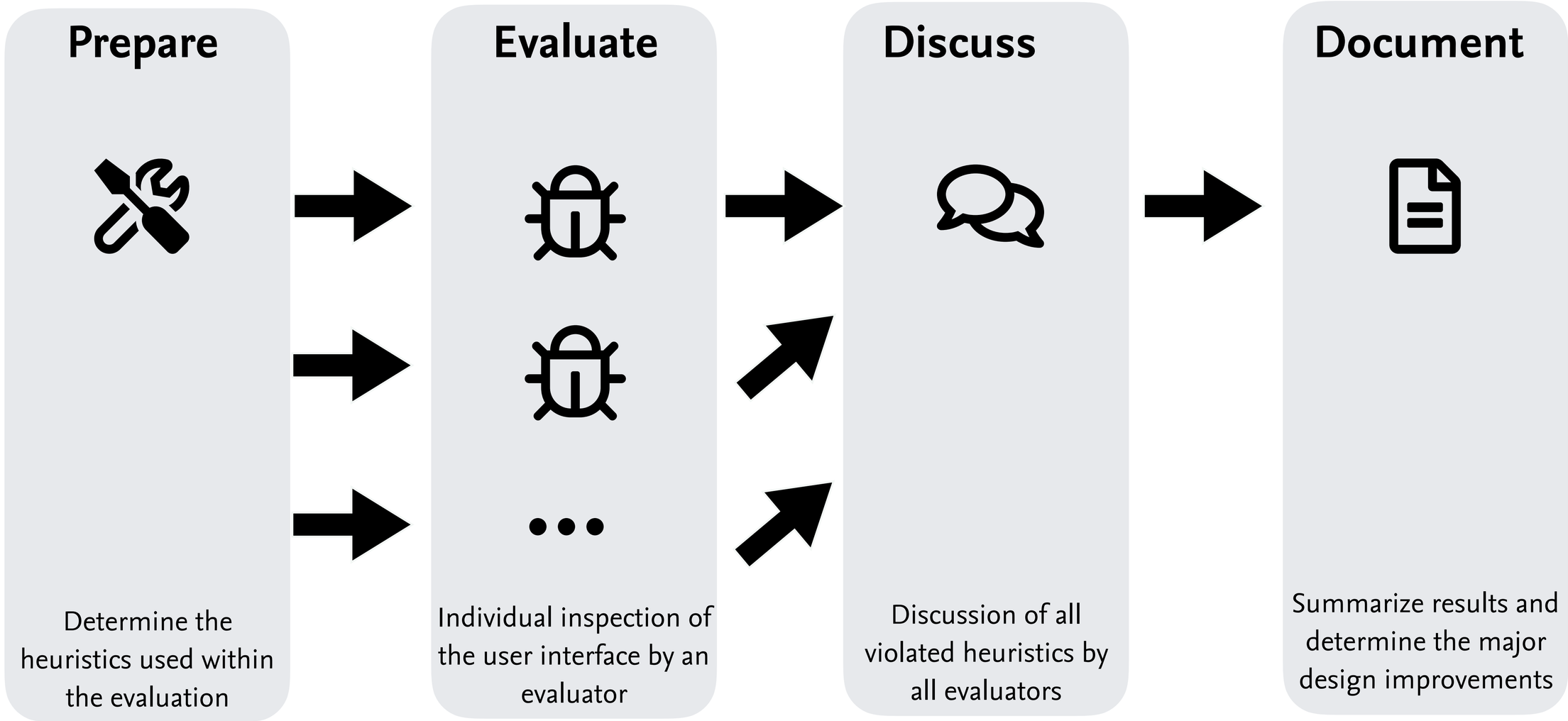
Example applications:
[Web Search, Product #1] "With different filters, the search results are auto updated."

16 **Convey the consequences of user actions.**
Immediately update or convey how user actions will impact future behaviors of the AI system.

Example violations:
[Email, Product #1] "It's clear there is some consequence for the user actions, but it's not well-defined. So while it exists, the lack of clarity on the exact outcomes for behavior are unclear."

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3290605.3300233

Process of Heuristic Evaluation



Hints for better Heuristics Evaluation

Use multiple evaluators

» Different evaluators find different problems

» The more the better, but diminishing returns. (Nielsen recommends 3-5)

Alternate heuristic evaluation with user testing

» Each method finds different problems

» Heuristic evaluation is cheaper

It’s OK for observer to help evaluator

» As long as the problem has already been noted

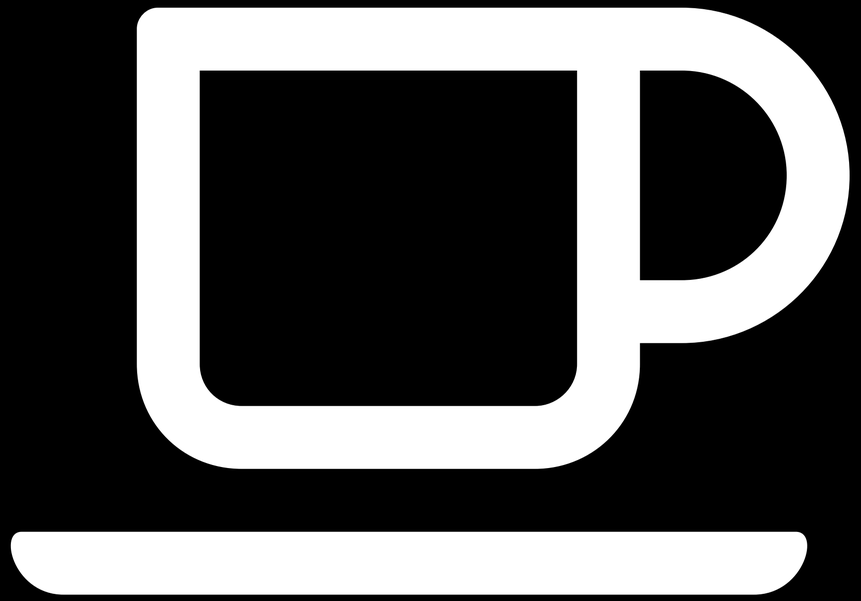
» This wouldn’t be OK in a user test

Nielsen, J. 1992. Finding usability problems through heuristic evaluation. Proceedings ACM CHI'92 Conference (Monterey, CA, May 3-7), 373-380.

Pro & Cons of Heuristic Evaluation

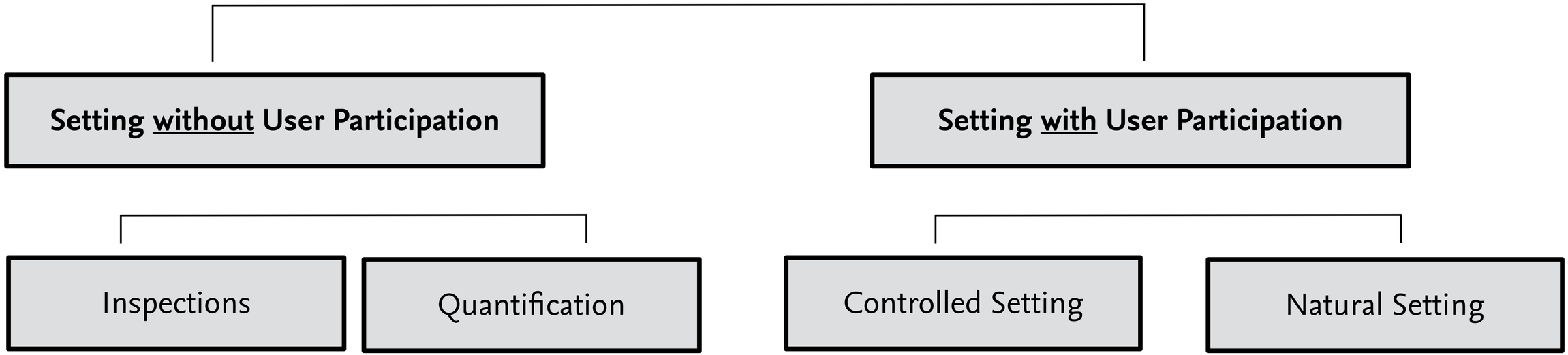
- No user involvement
- Identifies major problems well
- Easy to learn to do
- Can train developers to do it
- Cheap
- Gives pointers for improvements – but is not explicit
- Can be applied to interfaces in varying states of readiness, including paper prototypes

- Relies on interpretation of heuristics
- Only as good as the evaluator applying them
- Broad-based guidelines – but sometimes it is necessary to break the rules
- Can't cover every type of interface that might get built – different application areas might have very different concerns
- Doesn't involve users!



5 minutes break

Types of Evaluation Methods



» Heuristic evaluation

» Usability Testing

Usability and Usability Testing

> The usability of software is the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

> Usability testing [is] the activity that focuses on observing users working with a product, performing tasks that are real and meaningful to them.

DIS, I. (2010). 9241-210: 2010. Ergonomics of human system interaction-Part 210: Human-centred design for interactive systems (formerly known as 13407). International Standardization Organization (ISO). Switzerland.
Barnum, C. M. (2011). Usability Testing Essentials. Elsevier. https://www.oreilly.com/library/view/usability-testing-essentials/9780123750921/

Process of Usability Testing

1. Define the purpose, and goals of the test
2. Define the needed participant characteristics
3. Select an appropriate method (Test Design)
4. Define a task list
5. Setting up the test environment, equipment and logistics
6. Describe the data to be collected and the evaluation measures
7. Prepare a consent form and check ethical issues
8. Report the results

Measuring Satisfaction

User satisfaction is measured through standardized satisfaction questionnaires which can be administered after each task and/or after the usability test session.

Types:

- *Task Level Satisfaction*
  - After users attempt a task (whether completed or not), they should immediately be given a questionnaire so as to measure how difficult that task was.
- *Test Level Satisfaction*
  - Is measured by giving a formalized questionnaire to each test participant at the end of the test session.

http://usabilitygeek.com/usability-metrics-a-guide-to-quantify-system-usability/

Task Level Satisfaction

These post-task questionnaires often take the form of Likert scale ratings and their goal is to provide insight into task difficulty as seen from the participants' perspective.

The most popular post-task questionnaires are:

» **ASQ**: After Scenario Questionnaire (3 questions)

» **NASA-TLX**: NASA's task load index is a measure of mental effort (5 questions)

» **SMEQ**: Subjective Mental Effort Questionnaire (1 question)

» **UME**: Usability Magnitude Estimation (1 question)

» **SEQ**: Single Ease Question (1 question)

http://usabilitygeek.com/usability-metrics-a-guide-to-quantify-system-usability/

Test Level Satisfaction

It serves to measure their impression of the overall ease of use of the system being tested.

Following questionnaires can be used:

- **SUS**: System Usability Scale (10 questions)
- **SUPR-Q**: Standardized User Experience Percentile Rank Questionnaire (8 questions)
- **QUIS**: Questionnaire For User Interaction Satisfaction (24 questions)
- **SUMI**: Software Usability Measurement Inventory (50 questions)
- **UMEX:** Usability Metric for User Experience (4 items)
- **UEQ**: User Experience Questionnaire (26 items)
- **AttrakDiff2**: (28 items)

http://usabilitygeek.com/usability-metrics-a-guide-to-quantify-system-usability/

Objective of Test Level Questionnaires

SUMI

SUS

QUIS

SUPR-Q

UMUX

UEQ

AttrakDiff2

Traditional

Web-focussed

Includes Usefulness

Assessing System Usability

Includes Pragmatic and Hedonic Usability

Image adapted from https://measuringu.com/three-branches-ux/

Which Questionnaire Should I Use?

Each study is different and you will select a different combination of methods and metrics.

The choice depends on the:

» Budget allocated for measuring user satisfaction
» Importance that the user's perceived satisfaction has on the overall project

For example, if users' satisfaction is very important and there is a large budget than use SUMI but if the budget is small the SUS might be a better choice.

In the German HCI community, the UEQ and AttrakDiff2 is a widely accepted instrument.

Testing does not Replace User Research

User Research: Data acquisition and analysis to develop a software.

Usability Testing: Evaluation of a software under certain criteria.

Challenging aspects of Usability Testing (UT) as an analysis-tool:

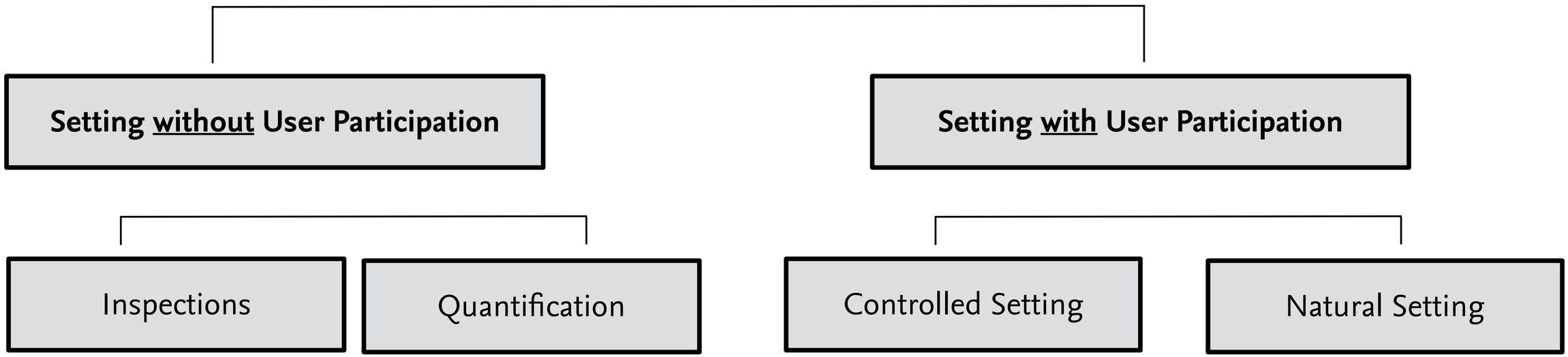
» UT is a reactive analysis.

» UT helps to identify problems with the design, but does neither show the exact problem nor a way to resolve them.

» One can only observe the user's reactions; the causes or reasons for them remain unclear.

» The results of a UT are limited to the current task and can not generalized to the whole application.

Types of Evaluations



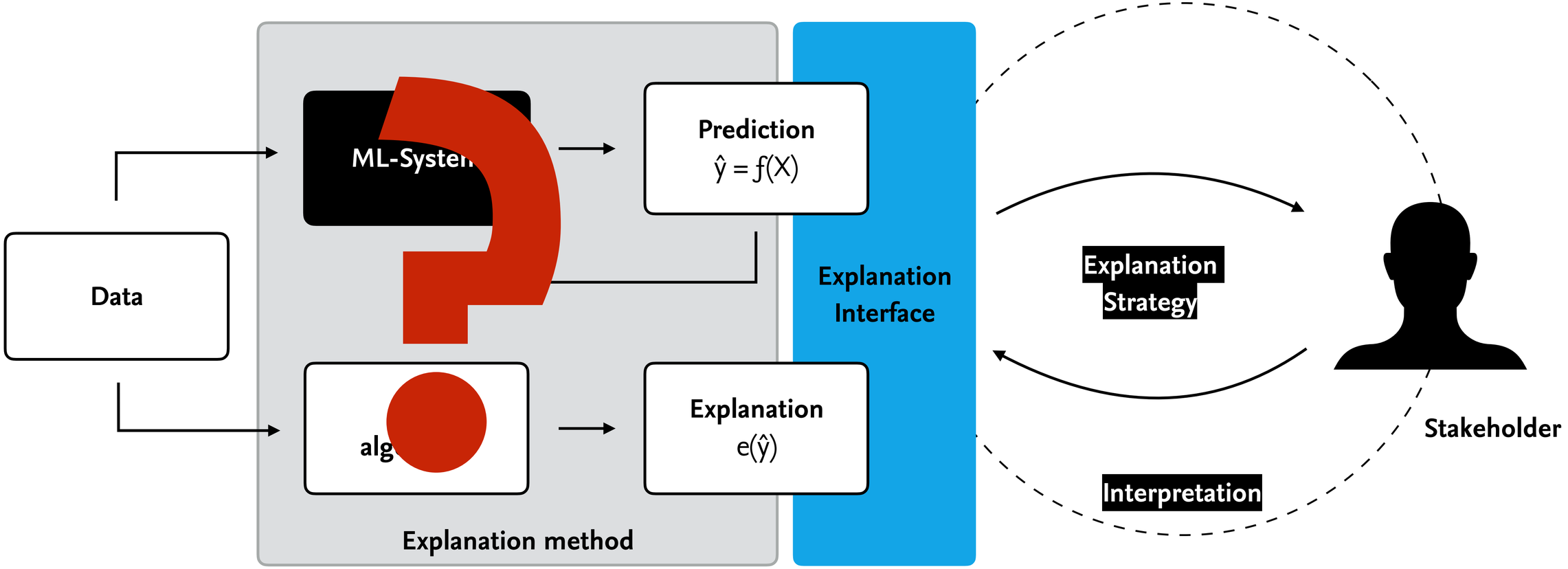
» Heuristic evaluation

» Usability Testing

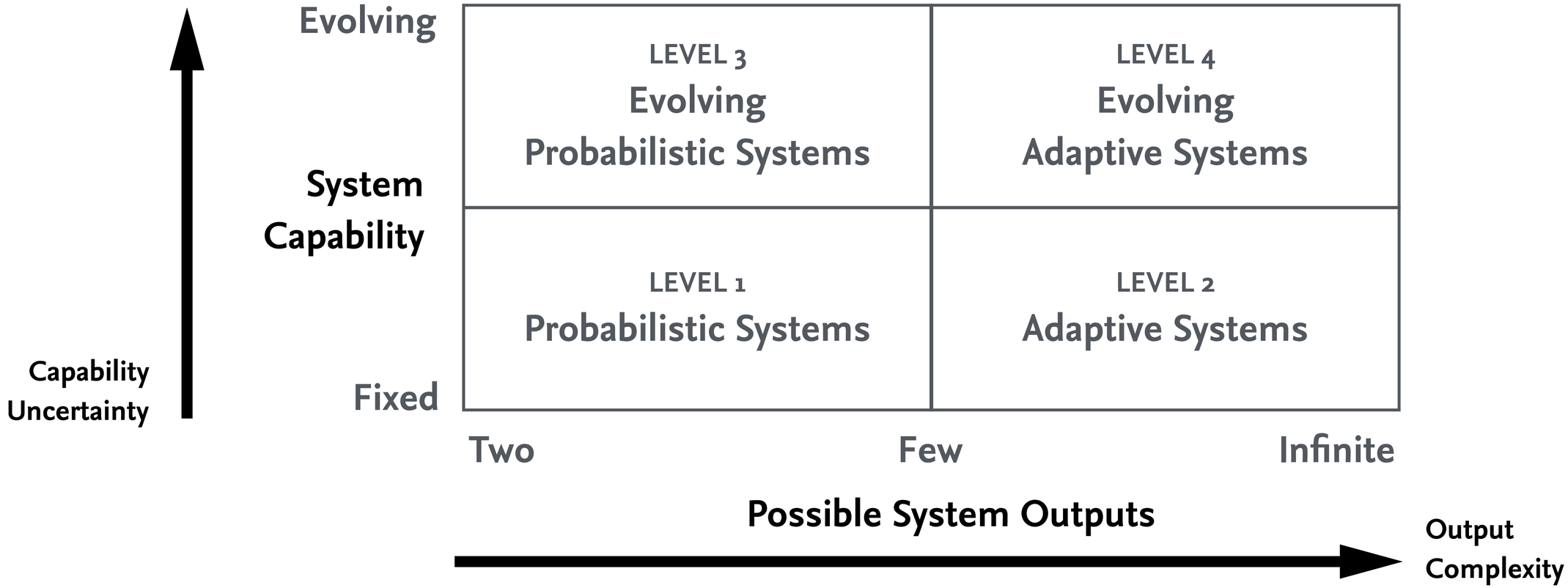
» Remote Testing
» Field Testing

Applying Evaluation Methods in System Design

System Design and Explanation User Interface

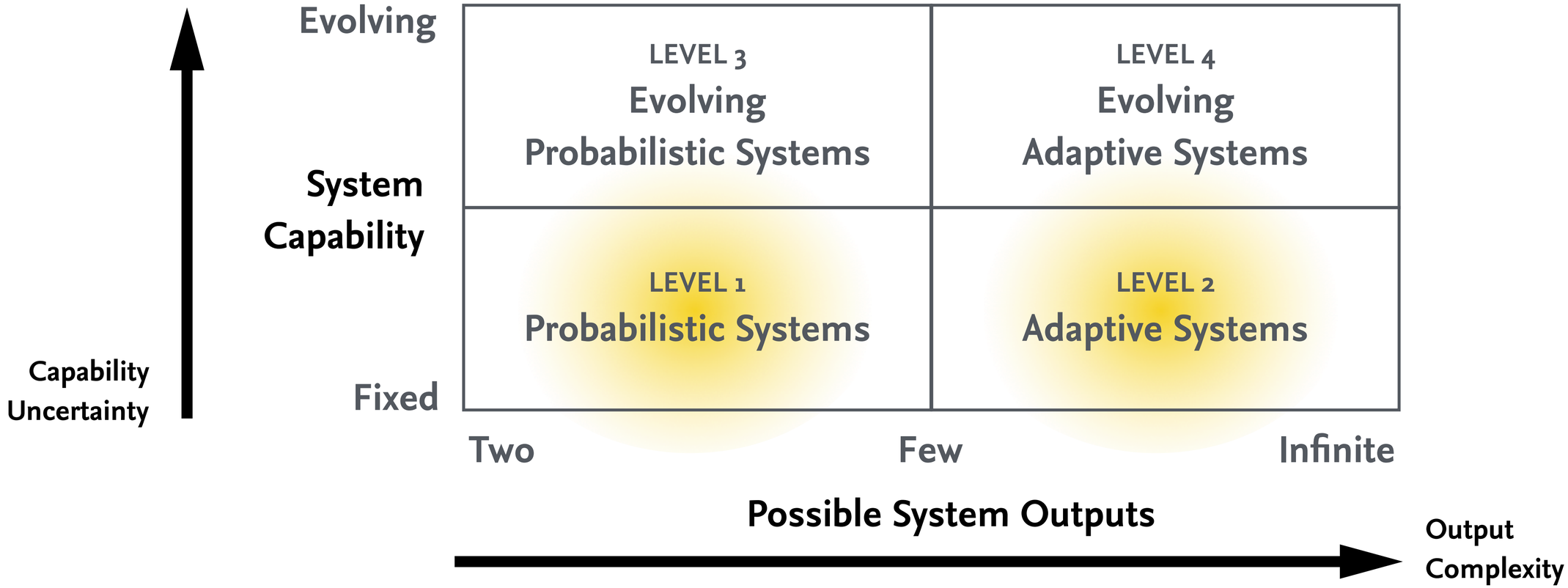


Design Complexity Map



Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design (Vol. 1, pp. 1–13). Presented at the CHI '20: CHI Conference on Human Factors in Computing Systems, New York, NY, USA: ACM. http://doi.org/10.1145/3313831.3376301

Design Complexity Map



Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design (Vol. 1, pp. 1–13). Presented at the CHI '20: CHI Conference on Human Factors in Computing Systems, New York, NY, USA: ACM. http://doi.org/10.1145/3313831.3376301

What is a Scenario?

"Scenarios are stories. They are stories about people and their activities."

It is a narrative that describes how a specific person behaves as a sequence of events by specifying the what and where. Scenarios can:

» evoke reflections about design issues,

» fix an interpretation and a solution and can be easily revised,

» support participation among stakeholders in the design process, and

» motivate the choice of use-cases.

The level of detail present in a scenario varies depending on where in the development process they are being used.

Carrol, John M. "Five reasons for scenario-based design." Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. IEEE, 1999.

Explainability Scenarios

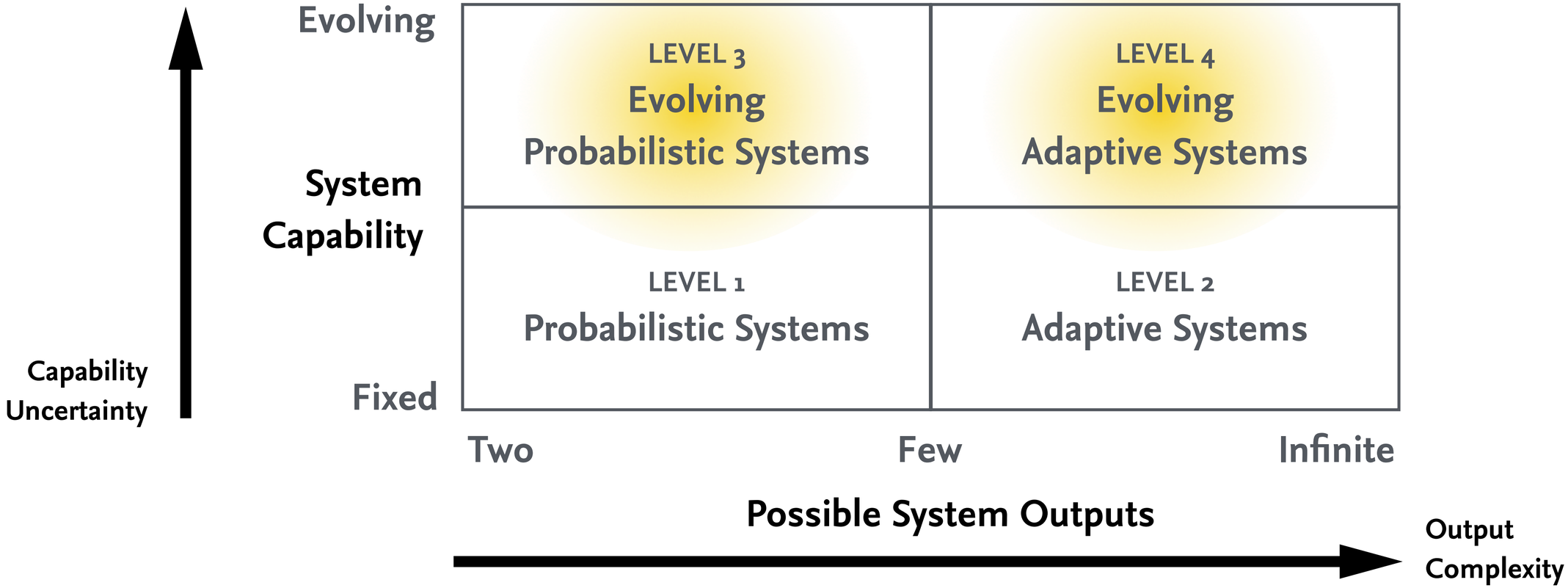
Wulf suggests the concept of “explainability scenarios” as resource for designing for interpretability. Explainability scenarios focus on what people need to understand to act on system outputs.

Instead of asking *what might an AI system be capable of explaining*, a scenario perspective asks: *what types of explanation might users need in the course of using AI systems?*

Explanation scenarios re-orient the design focus on possible practices (and problems) of use, which become resources to then brainstorm possible technological development.

Wolf, Christine T. "Explainability scenarios: towards scenario-based XAI design." Proceedings of the 24th International Conference on Intelligent User Interfaces. 2019. http://doi.org/10.1145/3301275.3302317

Design Complexity Map



Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design (Vol. 1, pp. 1–13). Presented at the CHI '20: CHI Conference on Human Factors in Computing Systems, New York, NY, USA: ACM. http://doi.org/10.1145/3313831.3376301

Check your Insights

Why is it necessary to evaluate explanation user interfaces? What are the different perspectives of such evaluation?

How does the selected user group impacts the evaluation measures?

When is the process of heuristic evaluation adequate in the process?

What measures can you generally considers when evaluating XAI user interfaces?

How do you measure for usability? Which dimensions need to be considered?

Which questionnaire should you use?

Why does it make sense to differentiate system capability and system output?

«Human-Centered Data Science»

Next week: Privacy - Protecting Individuals' Sensitive Information in Data

Dr. Daniel Franzen, Lars Sipos

Human-Centered Computing, Institute of Computer Science

Freie Universität Berlin

July 7, 2022