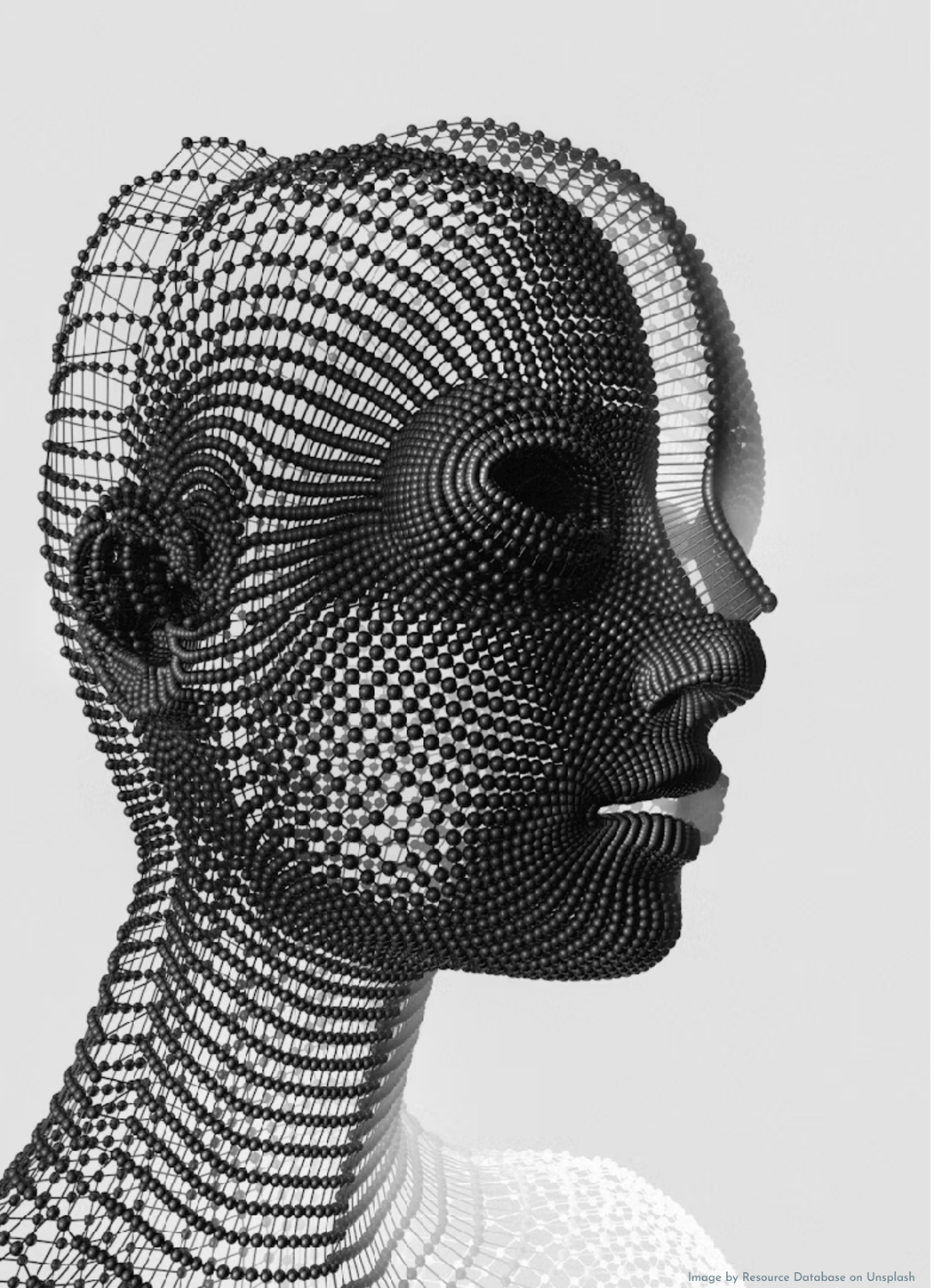


Image by Resource Database on Unsplash

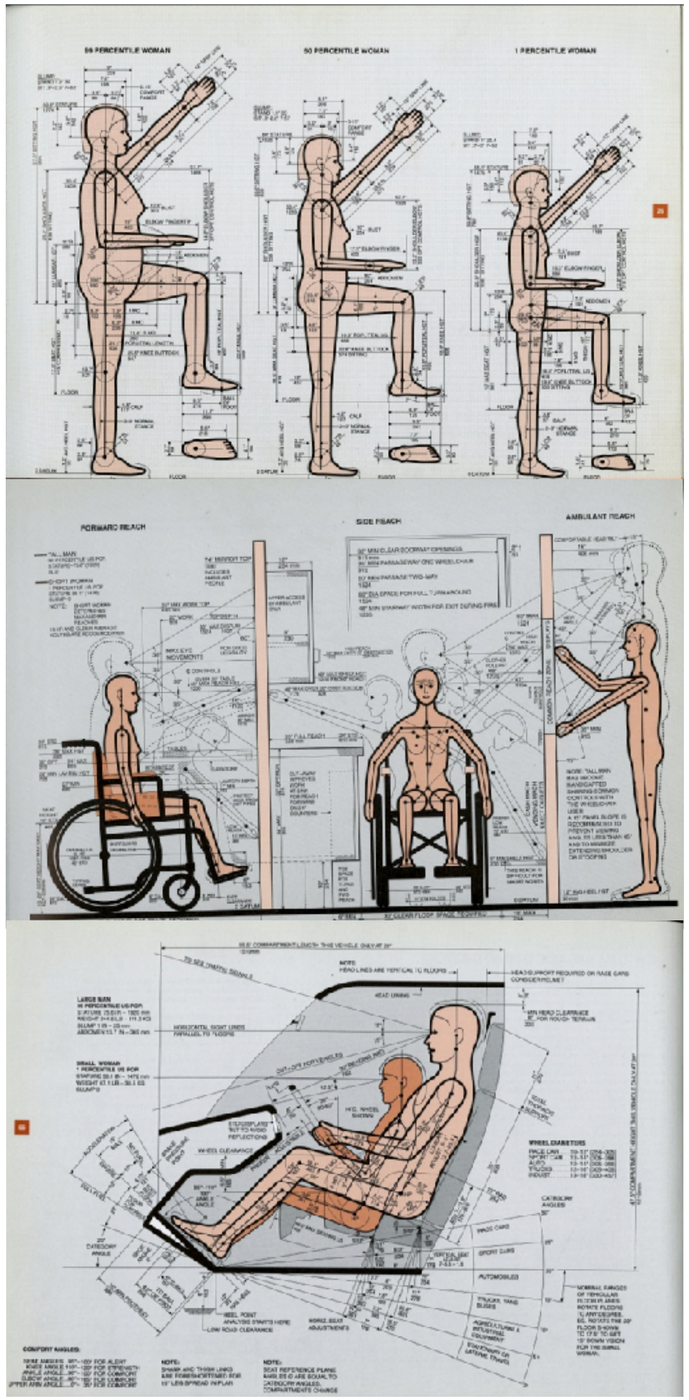
di lab

User-Centered Design in Artificial Intelligence & Machine Learning

Why People Matter in Technology Design & How to Involve Them

dr. Katja Rogers
Assistant professor, Digital Interactions Lab, UvA

Why does it matter how technology is designed?



frustration

lost time, lack of motivation, lost income

behaviour & actions unwanted by ...

... the designer

... the user

exclusion

The Measure Of Man And Woman—Human Factors In Design by Alvin R. Tilley and Henry Dreyfuss Associates

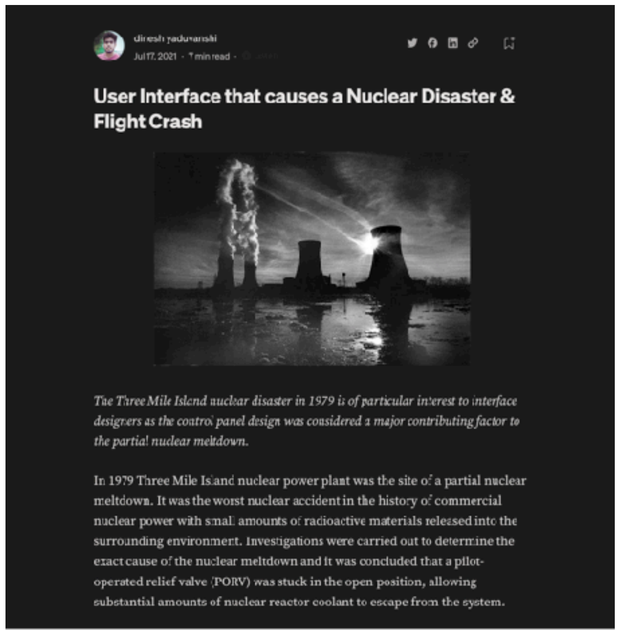
Consider the design of critical systems:

… for charting patient status and medical interventions

(e.g., Therac-25 radiation therapy machine)

… for piloting/steering of transportation

(e.g., in interfaces on planes or ferries)



https://hackaday.com/2015/10/26/killed-by-a-machine-the-therac-25/
https://dineshyaduvanshi.medium.com/user-interface-that-causes-a-nuclear-disaster-flight-crash-2585669d84f7

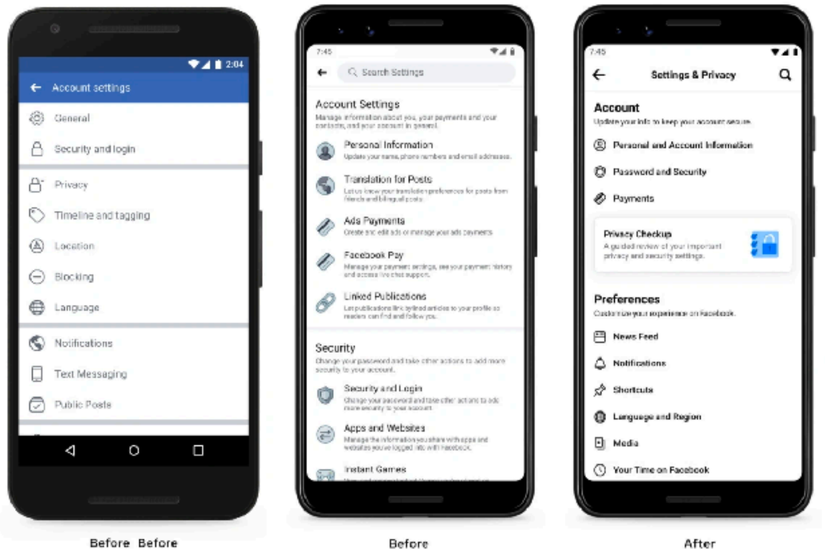
But also in non-critical systems:

… privacy settings in Facebook

(e.g., personal characteristics like sexual orientation or workplaces, group memberships, …)

… persuading users to perform actions not in their interest, or actively against their own interests

(e.g., through click advertising or choosing certain data handling settings - see Dark Patterns)



Before Before Before After

Devin Coldewey, 2021. Oh, Facebook changed its privacy settings again. TechCrunch. https://techcrunch.com/2021/08/04/oh-facebook-changed-its-privacy-settings-again/

https://twitter.com/OyooQuartey/status/602062838090375168/photo/1

di lab

Polite technologies:

respect the users’ time and decisions,

ask for consent,

anticipate input & actions

“Taking responsibility is not a nice-to-have design skill.

[…]

You are responsible for the work you put into the world,

and [… its] effects […] on the world.”

Mike Monteiro

How Designers Destroyed the World - Mike Monteiro, at USI
https://www.youtube.com/watch?v=qIcM21l61TE

white, male, affluent default

e.g., “nude” bandaids etc. are for white skin tones

e.g., train/bus routes in some places are designed to omit certain areas > decreases access and impacts health

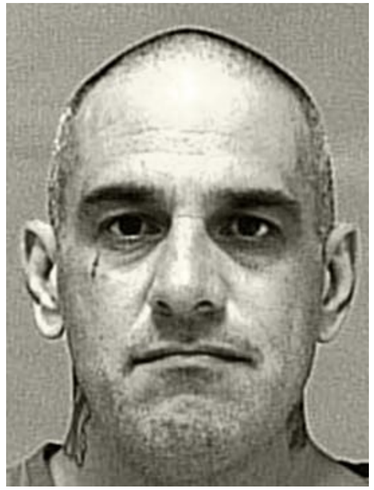
e.g., some automatic soap dispensers are unable to detect dark skin

e.g., facial recognition software historically has struggled with mistaking Black people for each other

e.g., voice detection software historically has struggled with female voices (higher pitch) and accents

e.g., educational curriculums are designed to teach a specific lens

COMPAS - Propublica report



LOW RISK	3	HIGH RISK	8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of $80.

Compare their crime with a similar one: The previous summer, 41-year-old Vernon Prater was picked up for shoplifting $86.35 worth of tools from a nearby Home Depot store.

Prater was the more seasoned criminal. He had already been convicted of armed robbery and attempted armed robbery, for which he served five years in prison, in addition to another armed robbery charge. Borden had a record, too, but it was for misdemeanors committed when she was a juvenile.

Yet something odd happened when Borden and Prater were booked into jail: A computer program spat out a score predicting the likelihood of each committing a future crime. Borden — who is black — was rated a high risk. Prater — who is white — was rated a low risk.

Two years later, we know the computer algorithm got it exactly backward. Borden has not been charged with any new crimes. Prater is serving an eight-year prison term for subsequently breaking into a warehouse and stealing thousands of dollars' worth of electronics.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

COMPAS - Propublica report

Black defendants were often predicted to be at a higher risk of recidivism than they actually were. Our analysis found that black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45 percent vs. 23 percent).

White defendants were often predicted to be less risky than they were. Our analysis found that white defendants who re-offended within the next two years were mistakenly labeled low risk almost twice as often as black re-offenders (48 percent vs. 28 percent).

The analysis also showed that even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 45 percent more likely to be assigned higher risk scores than white defendants.

https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

difficult to determine the extent to which this racial bias results from bias in the dataset used for analysis, vs. the COMPAS algorithm itself

the COMPAS algorithm is proprietary and has not been made public > researchers can only study the *results* of the algorithm, not the algorithm itself

if risk factors are made public, the social context of that factor must be considered - e.g., "highest degree attained" as decision variable > what is an individual's opportunity to access education?

Data sources are extremely limited

"Is it ethical for a company to 'commercialise' law?"

https://github.com/violetyao/Re-Exploring-COMPAS-Bias

di lab

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139

Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

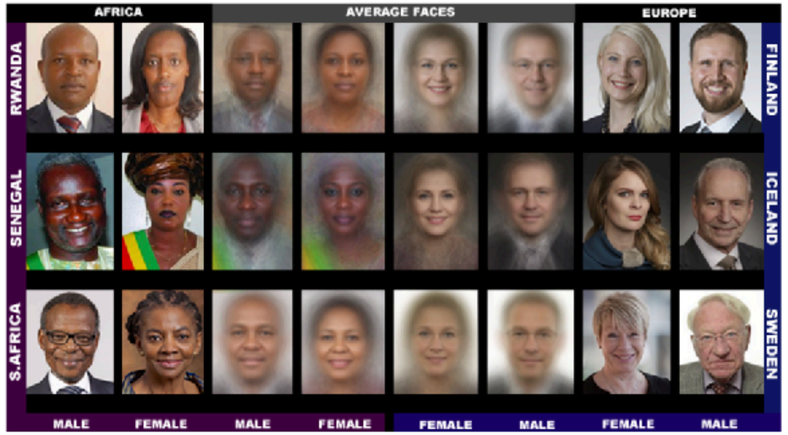


Figure 1: Example images and average faces from the new Pilot Parliaments Benchmark (PPB). As the examples show, the images are constrained with relatively little variation in pose. The subjects are composed of male and female parliamentarians from 6 countries. On average, Senegalese subjects are the darkest skinned while those from Finland and Iceland are the lightest skinned.

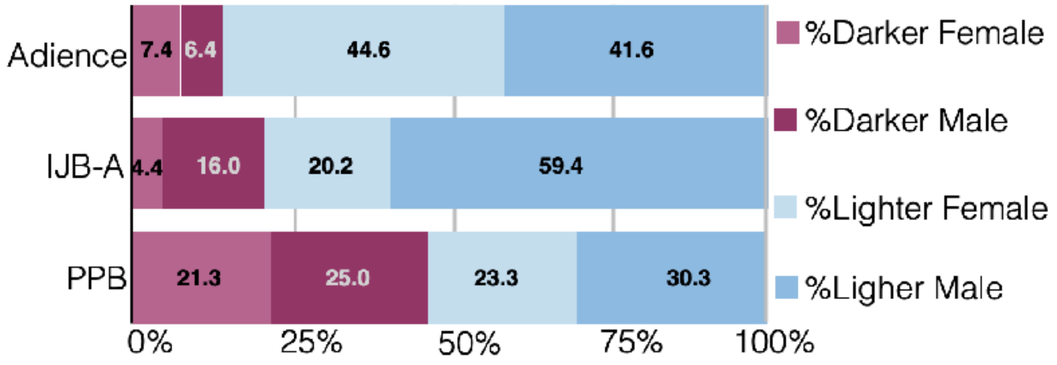


Figure 3: The percentage of darker female, lighter female, darker male, and lighter male subjects in PPB, IJB-A and Adience. Only 4.4% of subjects in Adience are darker-skinned and female in comparison to 21.3% in PPB.

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	TPR(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	PPV (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	2.6	10.7	12.9	0.7	6.0	**20.8**	0.0	1.7
Face++	TPR(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	90.2	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	9.8	0.8
	PPV (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	0.7	21.3	16.5	4.7	0.7	**34.5**	0.8	9.8
IBM	TPR(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	PPV (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	5.6	20.3	22.4	3.2	12.0	**34.7**	0.3	7.1

Table 4: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-TPR), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the PPB dataset. All classifiers have the highest error rates for darker-skinned females (ranging from 20.8% for Microsoft to 34.7% for IBM).

di lab

WORLD NEWS

At least 13 may have killed themselves over UK's Post Office wrongful convictions scandal

A logo of a post office is displayed in London, Wednesday, Jan. 10, 2024. (AP Photo/Kin Cheung, File)

BY SYLVIA HUI
Updated 4:00 PM CET, July 8, 2025

Add AP News on Google | Share

LONDON (AP) — At least 13 people were thought to have taken their own lives as a result of Britain's Post Office scandal, in which almost 1,000 postal employees were wrongly prosecuted or convicted of criminal wrongdoing because of a faulty computer system, a report said Tuesday.

Another 59 people contemplated suicide over the scandal, one of the biggest miscarriages of justice in U.K. history.

From around 1999 to 2015, hundreds of people who worked at Post Office branches were wrongly convicted of theft, fraud and false accounting based on evidence from a defective information technology system. Some went to prison or were forced into bankruptcy. Others lost their homes, suffered health problems or breakdowns in their relationships or became ostracized by their communities.

Retired judge Wyn Williams, who chairs a public inquiry into the scandal, said in a report published Tuesday that 13 people killed themselves as a consequence of a faulty Post Office accounting system "showing an illusory shortfall in branch accounts," according to their families.

https://apnews.com/article/uk-post-office-scandal-suicide-horizon-software-70a6945a3acf945ea9d121425fdd028c

NEWS > TECHNOLOGY

Dutch scandal serves as a warning for Europe over risks of using algorithms

The Dutch tax authority ruined thousands of lives after using an algorithm to spot suspected benefits fraud — and critics say there is little stopping it from happening again.

MARCH 29, 2022 6:14 PM CET
BY MELISSA HEIKKILÄ

Chermaine Leysner's life changed in 2012, when she received a letter from the Dutch tax authority demanding she pay back her child care allowance going back to 2008. Leysner, then a student studying social work, had three children under the age of 6. The tax bill was over €100,000.

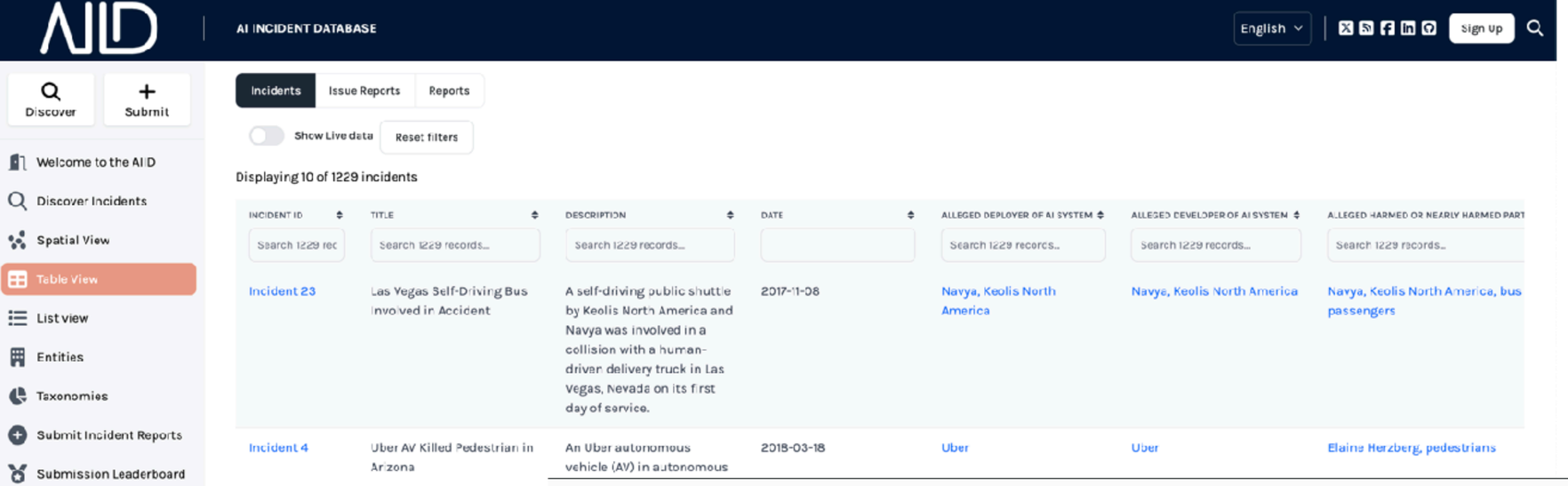
"I thought, 'Don't worry, this is a big mistake.' But it wasn't a mistake. It was the start of something big," she said.

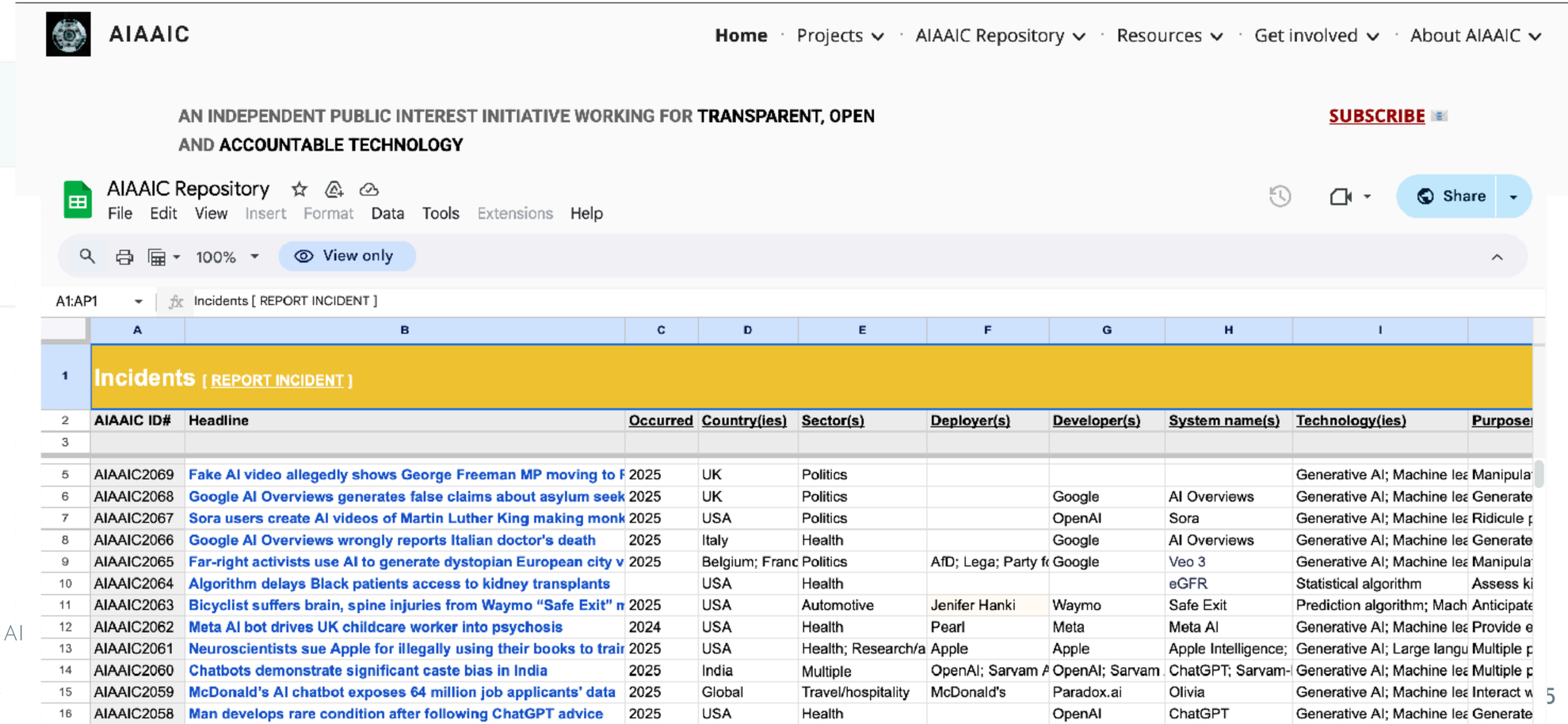
The ordeal took nine years of Leysner's life. The stress caused by the tax bill and her mother's cancer diagnosis drove Leysner into depression and burnout. She ended up separating from her children's father. "I was working like crazy so I could still do something for my children like give them some nice things to eat or buy candy. But I had times that my little boy had to go to school with a hole in his shoe," Leysner said.

Leysner is one of the tens of thousands of victims of what the Dutch have dubbed the "*toeslagenaffaire*," or the child care benefits scandal.

In 2019 it was revealed that the Dutch tax authorities had used a self-learning algorithm to create risk profiles in an effort to spot child care benefits fraud.

Authorities penalized families over a mere suspicion of fraud based on the system's risk indicators. Tens of thousands of families — often with lower incomes or belonging to ethnic minorities — were pushed into poverty because of exorbitant debts to the tax agency. Some victims committed suicide. More than a thousand children were taken into foster care.





https://www.aiaaic.org/home

McGregor 2021. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. https://doi.org/10.1609/aaai.v35i17.17817

people subject to and reliant on AI decisions / predictions:

- often have no say in the design + evaluation of the system: no or limited involvement in goal articulation, and how AI will impact end users & affected communities is often a blind spot
  > co design and/or participatory design; algorithmic impact assessments
- may be part of a group that is not well represented in the data itself, or its curation and labelling: this impacts accuracy, and depending on context can lead to misidentification, access issues and skewed personalization
  > inclusive data annotation; community-based data governance boards
- often are not provided with a way to contest (challenge + correct) AI model outputs
  > designing for failure and including ongoing feedback mechanisms

DOI:10.1145/3458723

Documentation to facilitate communication between dataset creators and consumers.

BY TIMNIT GEBRU, JAMIE MORGENSTERN, BRIANA VECCHIONE, JENNIFER WORTMAN VAUGHAN, HANNA WALLACH, HAL DAUMÉ III, AND KATE CRAWFORD

Datasheets for Datasets

DATA PLAYS A critical role in machine learning. Every machine learning model is trained and evaluated using data, quite often in the form of static datasets. The characteristics of these datasets fundamentally influence a model's behavior: a model is unlikely to perform well in the wild if its deployment context does not match its training or evaluation datasets, or if these datasets reflect unwanted societal biases. Mismatches like this can have especially severe consequences when machine learning models are used in high-stakes domains, such as criminal justice,[1,13,24] hiring,[19] critical infrastructure,[11,21] and finance.[18] Even in other domains, mismatches may lead to loss of revenue or public relations setbacks. Of particular concern are recent examples showing that machine learning models can reproduce or amplify unwanted societal biases reflected in training datasets.[4,5,12] For these and other reasons, the World Economic Forum suggests all entities should document the provenance, creation, and use of machine learning datasets to avoid discriminatory outcomes.[25]

Although data provenance has been studied extensively in the databases community,[3,8] it is rarely discussed in the machine learning community. Documenting the creation and use of datasets has received even less attention. Despite the importance of data to machine learning, there is currently no standardized process for documenting machine learning datasets.

To address this gap, we propose *datasheets for datasets*. In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet describing its operating characteristics, test results, recommended usage, and other information. By analogy, we propose that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on. Datasheets for datasets have the potential to increase transparency and accountability within the machine learning community, mitigate unwanted societal biases in machine learning models, facilitate greater reproducibility of machine learning results, and help researchers and practitioners to select more appropriate datasets for their chosen tasks.

After outlining our objectives, we describe the process by which we developed datasheets for datasets. We then provide a set of questions designed to elicit the information that a datasheet for a dataset might contain, as well as a workflow for dataset creators to use when answering these questions. We conclude with a summary of the impact to date of datasheets for datasets and a discussion of implementation challenges and avenues for future work.

Objectives. Datasheets for datasets are intended to address the needs of two key stakeholder groups: dataset creators and dataset consumers. For dataset creators, the primary objective is to encourage careful reflection on the process of creating, distributing, and maintaining a dataset, including any underlying assumptions, potential risks or harms, and implica-

di lab

- **Datasheets for datasets can increase transparency and accountability within the machine learning community, mitigate unwanted biases in machine learning models, facilitate greater reproducibility of machine learning results, and help researchers and practitioners to choose the right dataset.**
- **Datasheets enable dataset creators to be intentional throughout the dataset creation process.**

1. **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

2. **Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?**

3. **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

4. **Any other comments?**

Composition. Dataset creators should read through these questions prior to any data collection and then provide answers once data collection is complete. Most of the questions here are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for their chosen tasks. Some of the questions are designed to elicit information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

Questions that apply only to datasets that relate to people are grouped together at the end of the section. We recommend taking a broad interpretation of whether a dataset relates to people. For example, any dataset containing text that was written by people relates to people.

5. **What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)?** Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

6. **How many instances are there in total (of each type, if appropriate)?**

7. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).

8. **What data does each instance consist of?** "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.

9. **Is there a label or target associated with each instance?** If so, please provide a description.

10. **Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.

11. **Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.

12. **Are there recommended data splits (for example, training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.

13. **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.

14. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

15. **Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?** If so, please provide a description.

16. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.

If the dataset does not relate to people, you may skip the remaining questions in this section.

17. **Does the dataset identify any subpopulations (for example, by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

18. **Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset?** If so, please describe how.

19. **Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** If so, please provide a description.

20. **Any other comments?**

Collection process. As with the questions in the previous section, dataset creators should read through these questions prior to any data collection to flag potential issues and then provide answers once collection is complete. In addition to the goals outlined earlier, the following questions are designed to elicit information that may help researchers and practitioners to create alternative datasets with similar characteristics. Again, questions that apply only to datasets that relate to people are grouped together at the end of the section.

21. **How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)?** If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

22. **What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** How were these mechanisms or procedures validated?

23. **If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?**

24. **Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?**

25. **Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

26. **Were any ethical review processes conducted (for example, by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

If the dataset does not relate to people, you may skip the remaining questions in this section.

27. **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?**

28. **Were the individuals in question notified about the data collection?** If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

29. **Did the individuals in question consent to the collection and use of their data?** If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

30. **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

31. **Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted?** If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

32. **Any other comments?**

Preprocessing/cleaning/labeling. Dataset creators should read through these questions prior to any preprocessing, cleaning, or labeling and then provide answers once these tasks are complete. The questions in this section are intended to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a "bag-of-words" is not suitable for tasks involving word order.

33. **Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** If so, please provide a description. If not, you may skip the remaining questions in this section.

34. **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)?** If so, please provide a link or other access point to the "raw" data.

35. **Is the software that was used to preprocess/clean/label the data available?** If so, please provide a link or other access point.

36. **Any other comments?**

Uses. The following questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

37. **Has the dataset been used for any tasks already?** If so, please provide a description.

38. **Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.

39. **What (other) tasks could the dataset be used for?**

40. **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

41. **Are there tasks for which the dataset should not be used?** If so, please provide a description.

42. **Any other comments?**

Distribution. Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.

43. **Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.

44. **How will the dataset be distributed (for example, tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?

45. **When will the dataset be distributed?**

46. **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

47. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

48. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

49. **Any other comments?**

Maintenance. As with the previous questions, dataset creators should provide answers to these questions prior to distributing the dataset. The questions in this section are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan to dataset consumers.

50. **Who will be supporting/hosting/maintaining the dataset?**

51. **How can the owner/curator/manager of the dataset be contacted (for example, email address)?**

52. **Is there an erratum?** If so, please provide a link or other access point.

53. **Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?

54. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.

55. **Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

56. **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

57. **Any other comments?**

okay so let's just make AI that is transparent and explainable?

Kocielnik et al. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. CHI 2019.

“Transparency is a complex construct that evades simple definitions. It can refer to explainability, interpretability, openness, accessibility, and visibility, among others”

https://link.springer.com/article/10.1007/S11948-020-00276-4#Sec2

Image by todd kent on Unsplash
https://link.springer.com/article/10.1007/S11948-020-00276-4#Sec2

transparency-as-information:
a means to overcome problematic information asymmetries through clear explanations

transparency-as-inspectability:
a means to allow third parties to examine decision-making

transparency-as-accountability:
a means to ensure that systems and entities that are responsible for them have an interface for questions and requests for justifications, to potentially face consequences

transparency-as-visibility:
a means to elicit trust in the system's ability, integrity, and benevolence

"sunshine is the best disinfectant"

Transparency can be harmful

Transparency can intentionally occlude

Transparency can invoke neoliberal models of agency

Transparency can privilege seeing *over* understanding

"Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability" Ananny & Crawford *New Media & Society* 2018

e.g., privacy, interests of marginalised individuals vs. potentially malevolent authorities

→ *Transparency can be harmful*

Transparency can intentionally occlude

Transparency can invoke neoliberal models of agency

Transparency can privilege seeing *over* understanding

"Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability" Ananny & Crawford *New Media & Society* 2018

e.g., hide important information by purposefully throwing a lot of irrelevant information at the user ("strategic opacity")

→

Transparency can be harmful

Transparency can intentionally occlude

Transparency can invoke neoliberal models of agency

Transparency can privilege seeing *over* understanding

"Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability" Ananny & Crawford *New Media & Society* 2018

Transparency can be harmful

Transparency can intentionally occlude

→ *Transparency can invoke neoliberal models of agency*

Transparency can privilege seeing *over* understanding

e.g., placing the burden of interpreting information and effecting change on the individual

"Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability" Ananny & Crawford *New Media & Society* 2018

Transparency can be harmful

Transparency can intentionally occlude

Transparency can invoke neoliberal models of agency

→ *Transparency can privilege* seeing *over* understanding

e.g., vs. interacting with complex dynamic systems

"Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability" Ananny & Crawford *New Media & Society* 2018

white box model

"[their] internal mechanisms lend themselves to be interpreted in a direct manner"

e.g., decision trees, linear models, ...

generally lower complexity, lower accuracy

black box model

"cannot be understood or interpreted by themselves"

e.g., tree ensembles (increased complexity over decision trees), neural networks, ...

generally higher complexity, often higher accuracy

> post-hoc explanations

Dwivedi et al 2023. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. https://doi.org/10.1145/3561048

XAI Question Bank

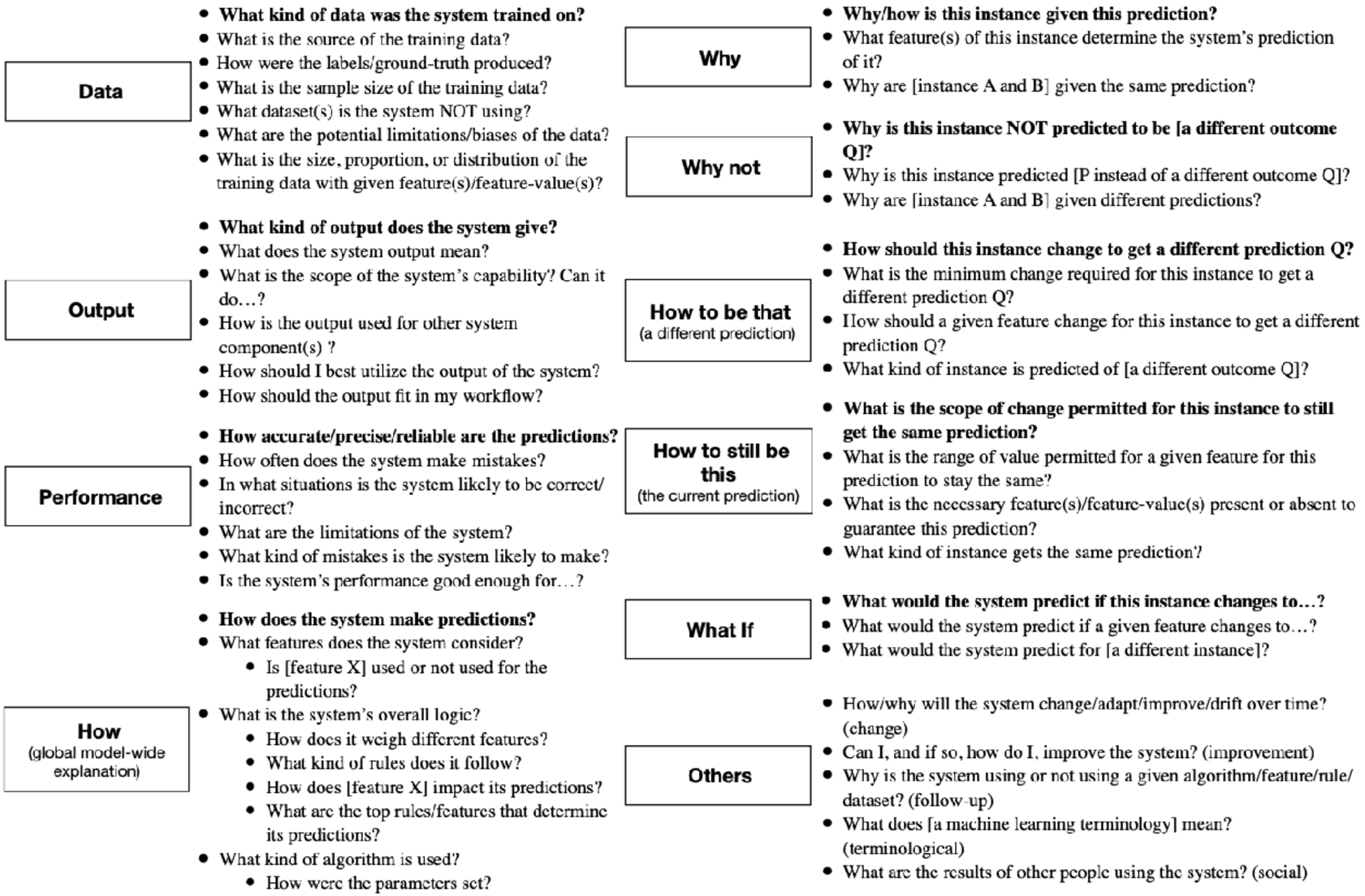
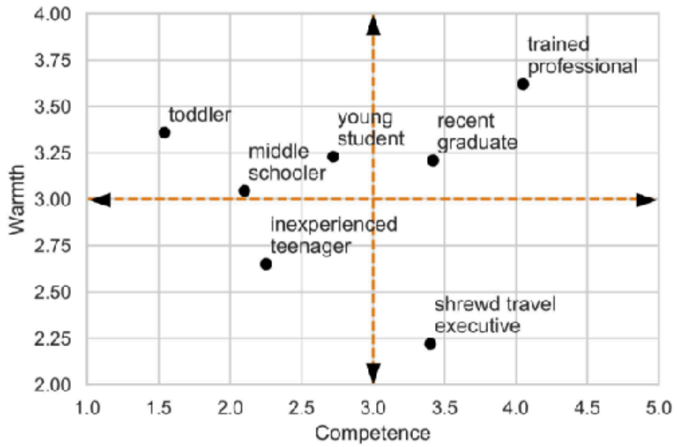
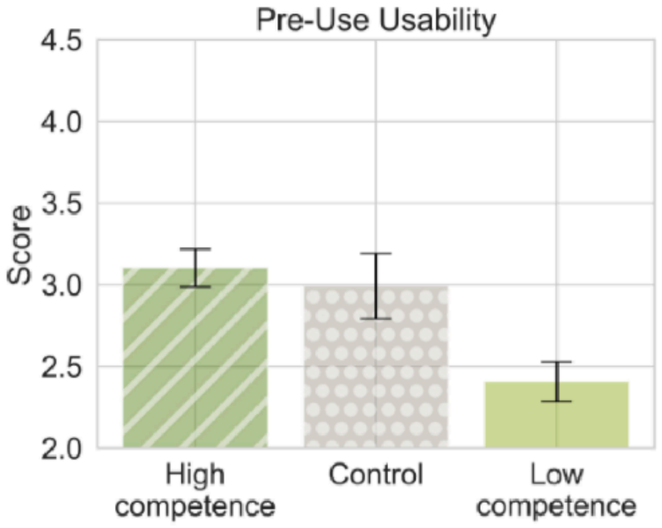


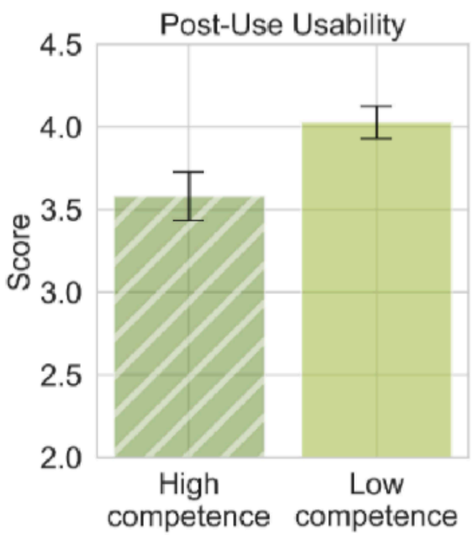
Fig. 1. XAI Question Bank for explaining supervised machine learning systems we developed in [39], with minor updates

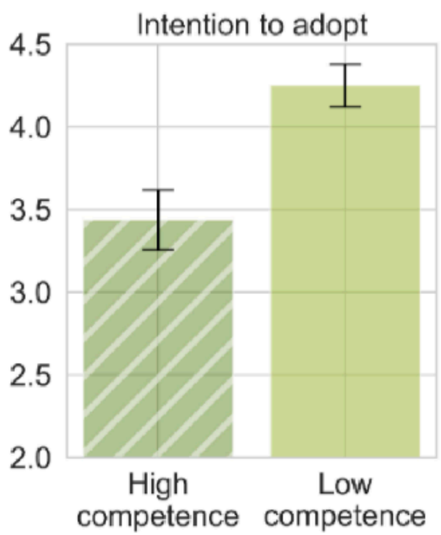
mental models & expectations
are shaped by metaphors

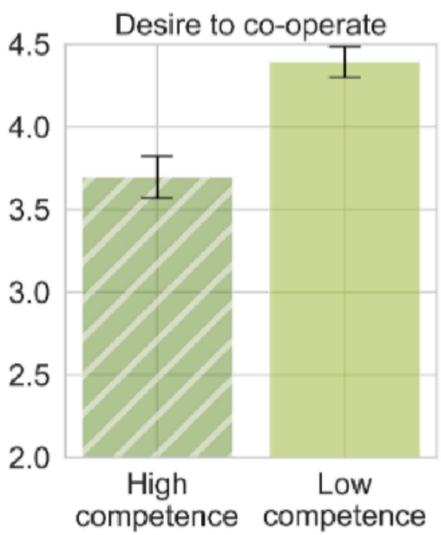
The bot you are about to interact with
is modeled after a "shrewd travel executive".











Khadpe et al. 2020. Conceptual Metaphors Impact Perception of Human-AI Collaboration. CSCW.

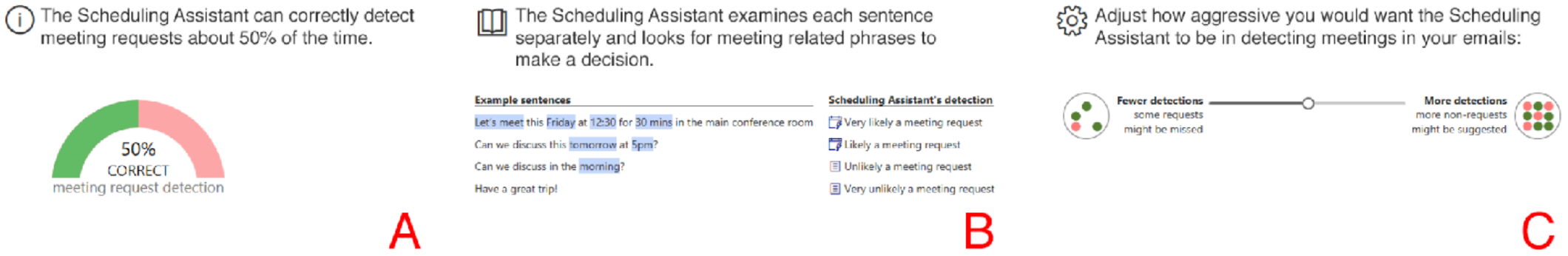


Figure 1: Expectation setting design techniques used prior to interaction with the Scheduling Assistant - an AI system for meeting request detection from free-text of emails. A) Accuracy Indicator - directly communicating to the user the expected accuracy of the AI component, B) Example-based Explanation - helping the user understand the basic principles of how the systems detects meeting requests, C) Control - giving the user control over AI decision making process through detection threshold adjustment.

so what if we adjust the expectations?

incomplete factorial design (N=116):

A
B
A+B
A+C
B+C
A+B+C

Kocielnik et al. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. CHI 2019.

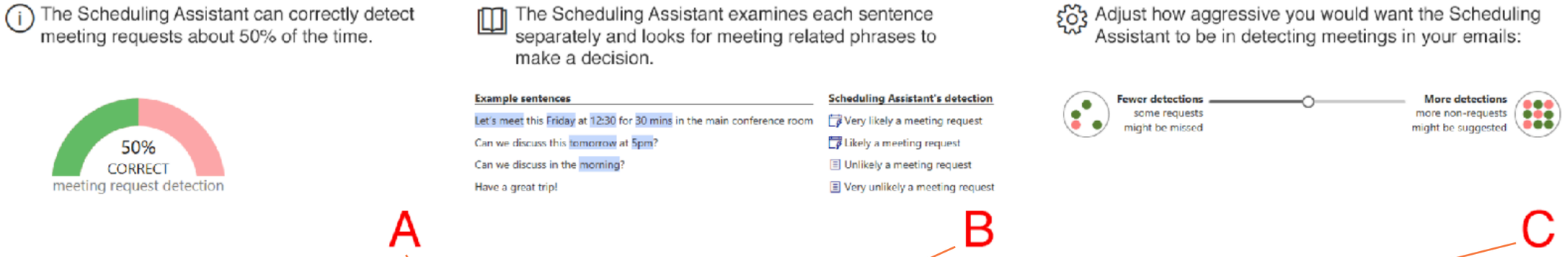


Figure 1: Expectation setting design techniques used prior to interaction with the Scheduling Assistant - an AI system for meeting request detection from free-text of emails. A) Accuracy Indicator - directly communicating to the user the expected accuracy of the AI component, B) Example-based Explanation - helping the user understand the basic principles of how the systems detects meeting requests, C) Control - giving the user control over AI decision making process through detection threshold adjustment.

- **H2.1** Directly communicating AI system accuracy will lead to lower discrepancy between system accuracy and user perception of it.
- **H2.2** Providing explanations will lead to higher perceptions of understanding how the AI system works.
- **H2.3** First-hand experience, through direct impact on the system, will lead to higher perceived level of control over system's behavior.

Kocielnik et al. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. CHI 2019.

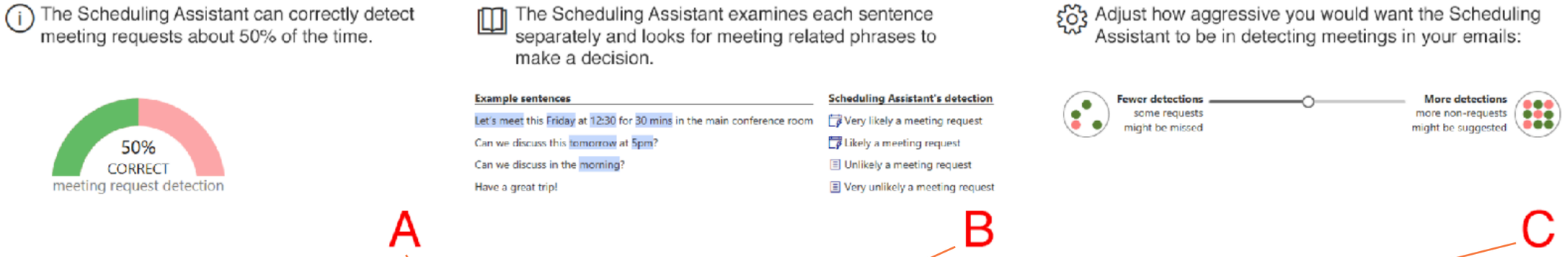


Figure 1: Expectation setting design techniques used prior to interaction with the Scheduling Assistant - an AI system for meeting request detection from free-text of emails. A) Accuracy Indicator - directly communicating to the user the expected accuracy of the AI component, B) Example-based Explanation - helping the user understand the basic principles of how the systems detects meeting requests, C) Control - giving the user control over AI decision making process through detection threshold adjustment.

Perceptions of system accuracy were measured through two questions adopted from the Expectations Confirmation Model [39]: pre-intervention *"How well do you expect the Scheduling Assistant to work"* and post-intervention *"How well do you feel the Scheduling Assistant works"*. Both were answerable on an 11-point scale from *"0% (Never correctly detects meetings)"* to *"100% (Always correctly detects meeting)"* with 10% increments.

- **H2.1** Directly communicating AI system accuracy will lead to lower discrepancy between system accuracy and user perception of it.
- **H2.2** Providing explanations will lead to higher perceptions of understanding how the AI system works.
- **H2.3** First-hand experience, through direct impact on the system, will lead to higher perceived level of control over system's behavior.

Kocielnik et al. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. CHI 2019.

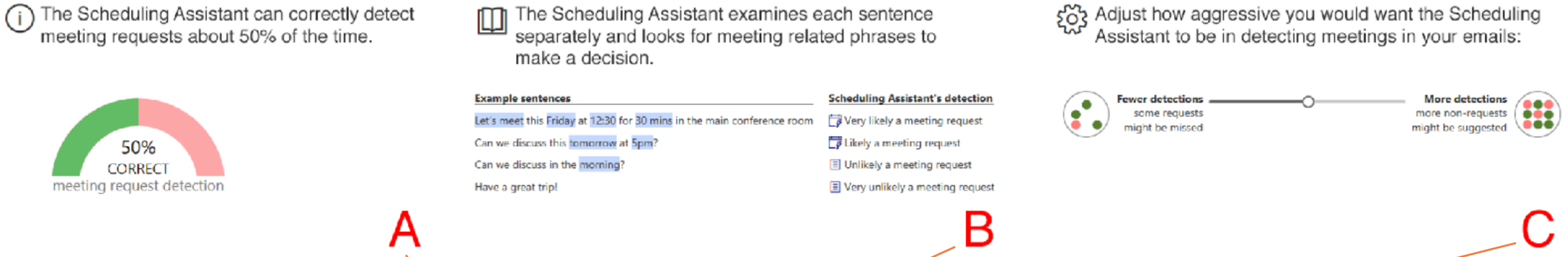


Figure 1: Expectation setting design techniques used prior to interaction with the Scheduling Assistant - an AI system for meeting request detection from free-text of emails. A) Accuracy Indicator - directly communicating to the user the expected accuracy of the AI component, B) Example-based Explanation - helping the user understand the basic principles of how the systems detects meeting requests, C) Control - giving the user control over AI decision making process through detection threshold adjustment.

Understanding of the AI component was measured through two subjective report questions adapted from [42]. One question asked about understanding how the system makes positive detections: *"I feel like I have a good understanding of how the Scheduling Assistant decides whether an email contains a meeting request"*, while the other asked about understanding what kind of mistakes the system can make: *"I feel like I understand what kind of mistakes the Scheduling Assistant is likely to make"*.

- **H2.1** Directly communicating AI system accuracy will lead to lower discrepancy between system accuracy and user perception of it.
- **H2.2** Providing explanations will lead to higher perceptions of understanding how the AI system works.
- **H2.3** First-hand experience, through direct impact on the system, will lead to higher perceived level of control over system's behavior.

Kocielnik et al. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. CHI 2019.

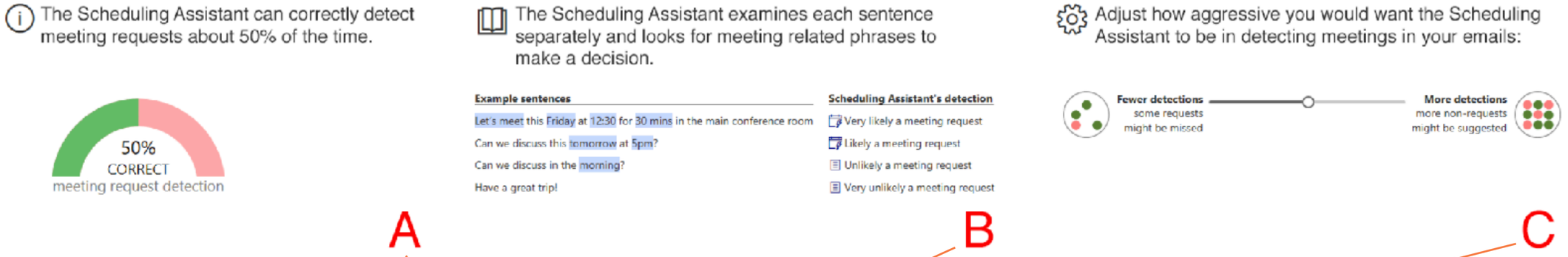


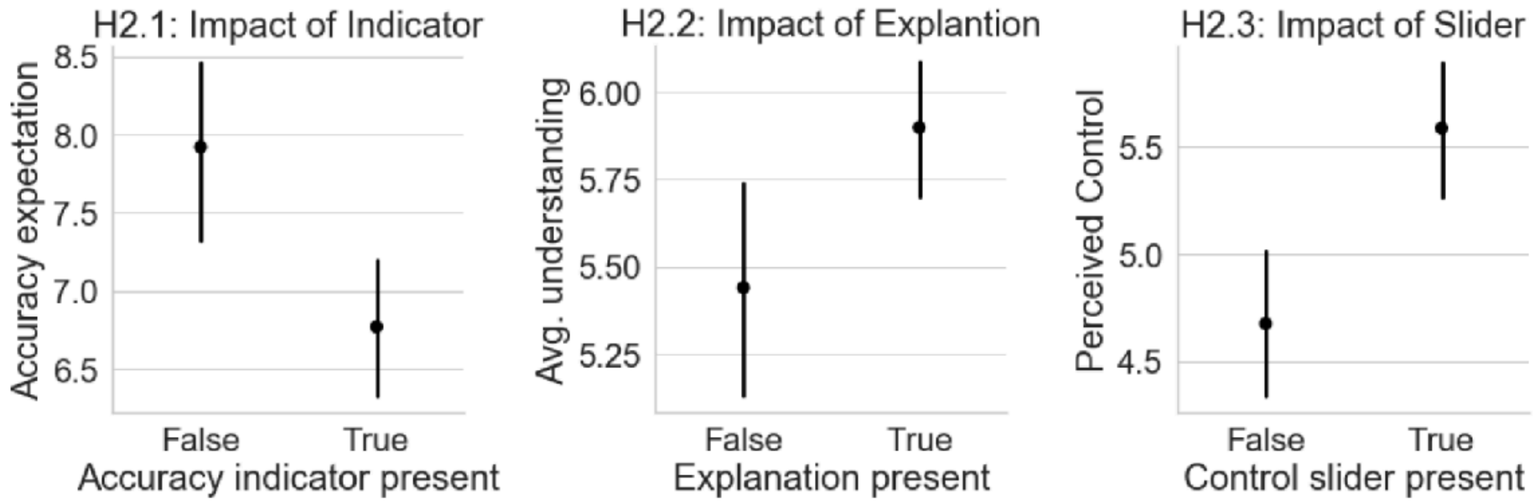
Figure 1: Expectation setting design techniques used prior to interaction with the Scheduling Assistant - an AI system for meeting request detection from free-text of emails. A) Accuracy Indicator - directly communicating to the user the expected accuracy of the AI component, B) Example-based Explanation - helping the user understand the basic principles of how the systems detects meeting requests, C) Control - giving the user control over AI decision making process through detection threshold adjustment.

- **H2.1** Directly communicating AI system accuracy will lead to lower discrepancy between system accuracy and user perception of it.
- **H2.2** Providing explanations will lead to higher perceptions of understanding how the AI system works.
- **H2.3** First-hand experience, through direct impact on the system, will lead to higher perceived level of control over system's behavior.

Subjective perception of control over the system's behavior was measured by a question adapted from [47]: *"I feel like I have some control over the Scheduling Assistant's behavior"*. Answers were given on a 7-point Likert scale from *"Strongly disagree"* to *"Strongly agree"*.

Kocielnik et al. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. CHI 2019.

[50% = groundtruth
i.e., correct accuracy]



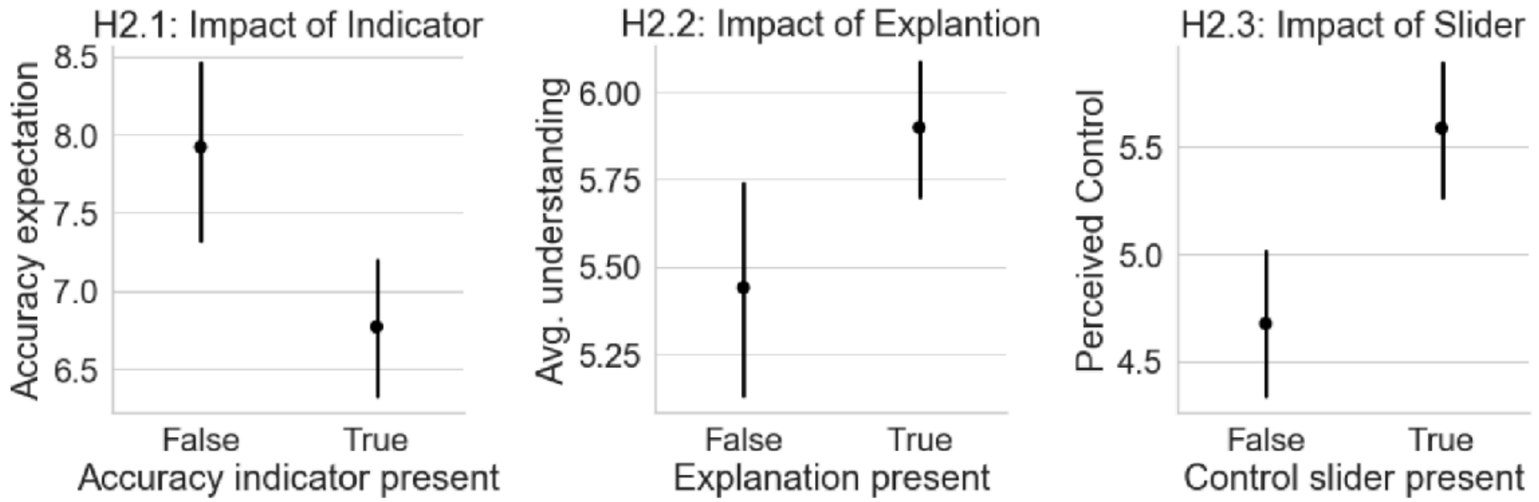
presence of an accuracy indicator brought users' expectations of system accuracy closer in line with actual system accuracy

[though no baseline condition!]

- **H2.1** Directly communicating AI system accuracy will lead to lower discrepancy between system accuracy and user perception of it.
- **H2.2** Providing explanations will lead to higher perceptions of understanding how the AI system works.
- **H2.3** First-hand experience, through direct impact on the system, will lead to higher perceived level of control over system's behavior.

Kocielnik et al. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. CHI 2019.

[50% = groundtruth
i.e., correct accuracy]



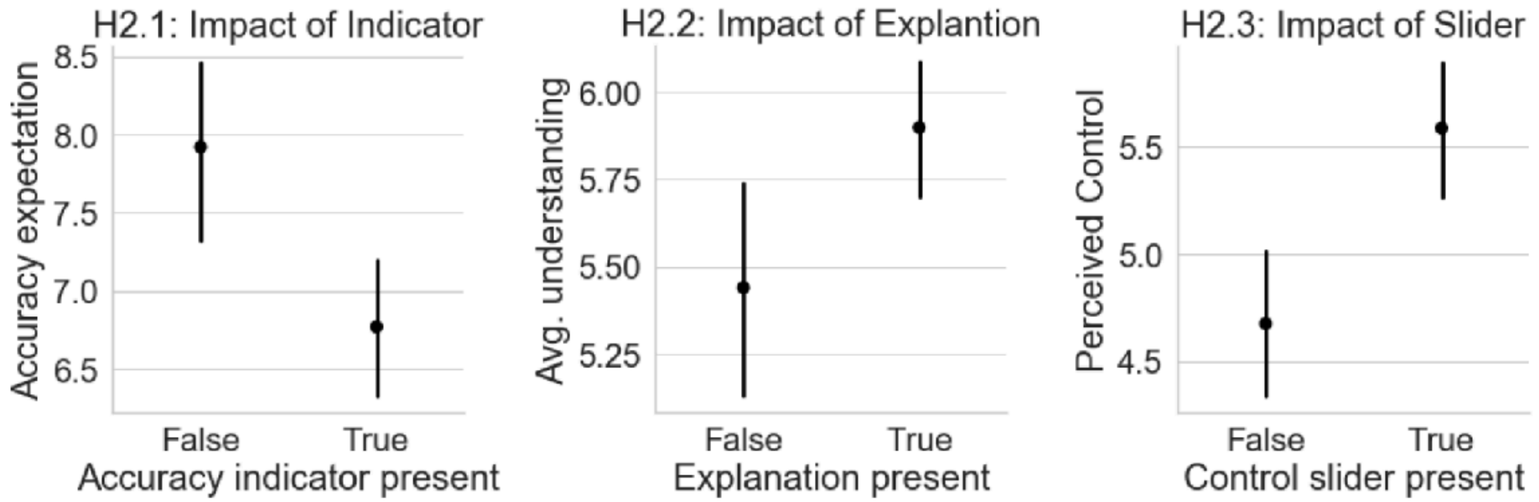
presence of explanations increased users' perceived understanding of the AI system

[though no baseline condition!]

- **H2.1** Directly communicating AI system accuracy will lead to lower discrepancy between system accuracy and user perception of it.
- **H2.2** Providing explanations will lead to higher perceptions of understanding how the AI system works.
- **H2.3** First-hand experience, through direct impact on the system, will lead to higher perceived level of control over system's behavior.

Kocielnik et al. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. CHI 2019.

[50% = groundtruth
i.e., correct accuracy]



presence of a control slider to impact the system's decision-making process increased users' perceived control over the AI system

[though no baseline condition!]

- **H2.1** Directly communicating AI system accuracy will lead to lower discrepancy between system accuracy and user perception of it.
- **H2.2** Providing explanations will lead to higher perceptions of understanding how the AI system works.
- **H2.3** First-hand experience, through direct impact on the system, will lead to higher perceived level of control over system's behavior.

Kocielnik et al. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. CHI 2019.

two systems can have
the same accuracy –

but still vary strongly
in the types of errors
that it makes

Kocielnik et al. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. CHI 2019.

study 2 compared different rates for error types (high recall vs. high precision)

Given an email contains a meeting request, the Scheduling Assistant can correctly determine that it indeed contains a meeting request (*True Positive - TP*) or it can incorrectly determine that it does not contain a meeting request (*False Negative - FN*). Similarly, given an email that does not contain a meeting request, the system can correctly determine that indeed it does not (*True Negative - TN*) or incorrectly determine that the email does contain a meeting request (*False Positive - FP*). A summary of these possible classifications outcomes along with concrete examples can be found in Table 1.

Table 1: A summary of possible correct and erroneous classifications that the Scheduling Assistant's AI component can make

Type	Predicted label	Example
(TP)	< Request	< Let's meet 3:30pm on Friday
(TN)	< No request	< We appreciate all the help
(FP)	< Request	< Yesterday's meeting was good
(FN)	< No request	< How about lunch? Maybe 1:30?

"High Recall" system: minimise FNs

"High Precision" system: minimise FPs

The *High Precision* system minimizes FP types of errors. The *High Recall* system, on the other hand, minimizes FN types of errors. To achieve these versions, we manipulated the classification of 20 email messages obtained from the Enron corpus.

"High Recall" system: minimise FNs
5 TP | 5 TN | 8 FP | 2 FN
accuracy 50%

"High Precision" system: minimise FPs
5 TP | 5 TN | 2 FP | 8 FN
accuracy 50%

H1.1 An AI system focused on High Precision (low False Positives) will result in higher perceptions of accuracy and higher acceptance.

Kocielnik et al. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. CHI 2019.

full factorial design with 16 versions; N=325

8x High Recall:

1. baseline
2. A (accuracy indicator)
3. B (explanations)
4. C (control slider)
5. A+B
6. A+C
7. B+C
8. A+B+C

8x High Precision analogously to the above

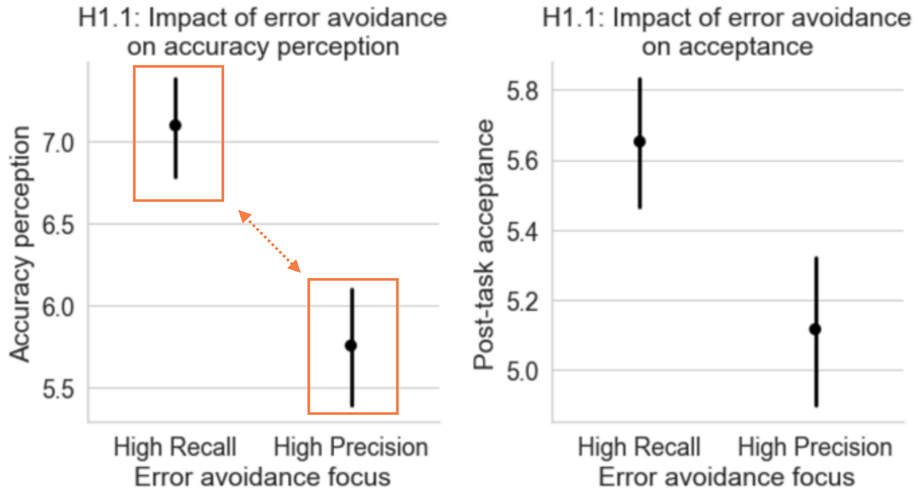


Figure 7: Impact of different focus of AI component on error avoidance on accuracy perceptions. High Recall - low False Negatives rate, High Precision - low False Positives rate

"High Recall" (lower FNs) resulted in significantly higher perceptions of accuracy and significantly higher acceptance compared to "High Precision" (lower FPs)

> H1.1 rejected

Kocielnik et al. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. CHI 2019.

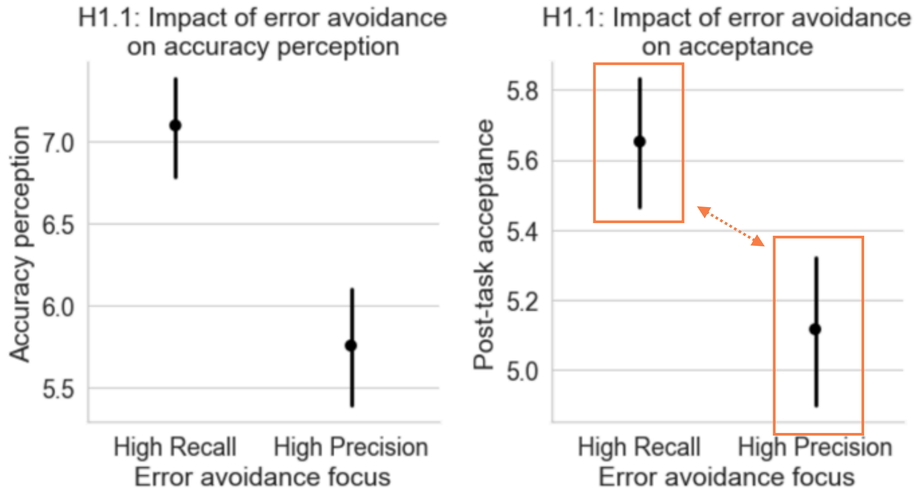


Figure 7: Impact of different focus of AI component on error avoidance on accuracy perceptions. High Recall - low False Negatives rate, High Precision - low False Positives rate

"High Recall" (lower FNs) resulted in significantly higher perceptions of accuracy and significantly higher acceptance compared to "High Precision" (lower FPs)

> H1.1 rejected

Kocielnik et al. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. CHI 2019.

does the type of errors mediate the effect of the expectation adjustment techniques?

"the expectation adjustment techniques have been shown effective in significantly increasing user satisfaction and acceptance, **however only in the High Precision version of the system**.

This is the version of the system that had been perceived as performing worse by our users and our [expectation adjustment] techniques appear to mitigate this effect."

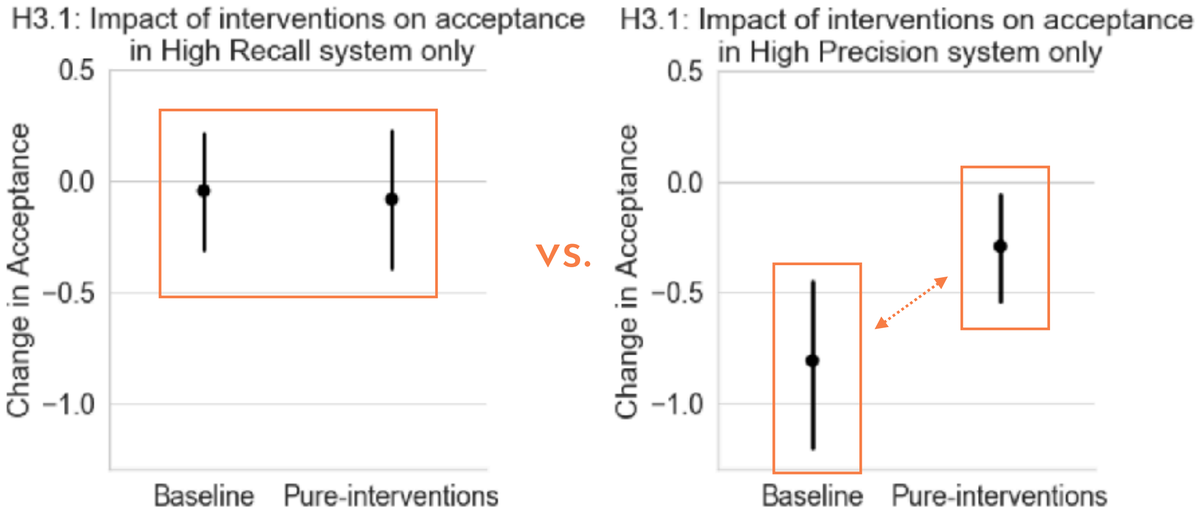


Figure 8: Comparison of impact of Baseline vs Pure-techniques conditions for High Recall (left) and High Precision (right) systems separately

Kocielnik et al. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. CHI 2019.

transparency

visibility into how the system works > inspectability

explainability

communicating the system's workings to users / stakeholders > interpretability

+

expectations

users expect more than just transparency / inspectability and explainability / interpretability: they expect systems to also align with their expectations

e.g., re: causality, morality/fairness, competence, social norms, responsibility, ...

Only 22% of XAI projects evaluate their explanations with humans

"From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI" Nauta et al. *ACM Computing Surveys* 2023

di lab

"studies examining how people actually interact with AI explanations have found popular XAI techniques to be ineffective [...], potentially risky [...], and underused in real-world contexts"

Ehsan et al. 2021. Expanding Explainability: Towards Social Transparency in AI systems. CHI 2021. https://dl.acm.org/doi/pdf/10.1145/3411764.3445188

Photo by Milad Fakurian on Unsplash

“two black boxes” problem

both human cognition and AI are considered as black boxes

AI systems must communicate their internal processes / decisions in human-understandable way
> eXplainable AI (XAI)

resulting communication challenges are a “semantic gap”

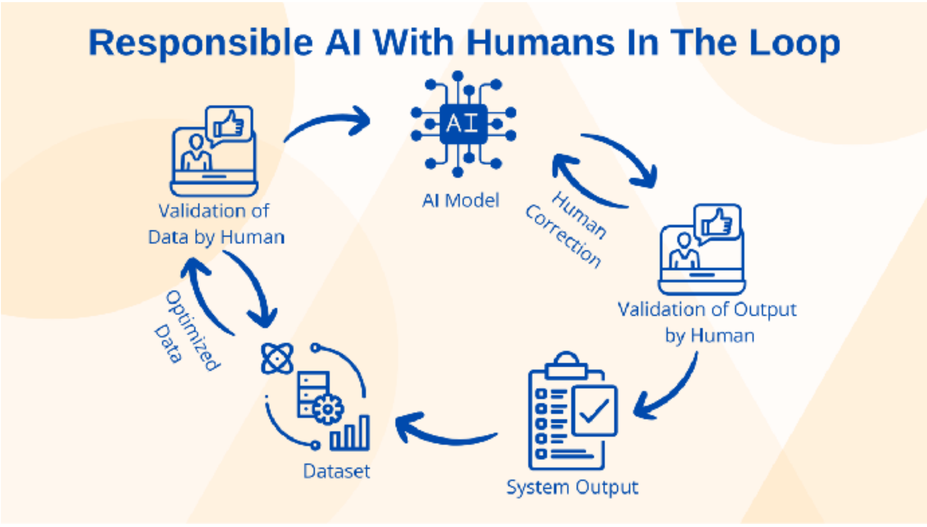
humans goals, decisions, and actions need to be translated to model-/system-understandable parameters, values, and operations
> explainable cognitive intelligence (XCI)

Wenskovitch & North 2020; doi: 10.1109/MC.2020.2996416

human in the loop

"A core concept of [human-centered design] is that of actively involving end-users and appropriate stakeholders in the process.

In the context of AI, this means placing humans in the loop, not only through meaningful human control [...] but also through their active participation in the preparation, learning, and decision-making phases of AI."



Margetis et al. 2021. doi: 10.1002/9781119636113.ch42

human in the loop

increases transparency and human agency

leverages benefits of automation AND human judgement

requires an interface for the human to interact with and shape the AI system

Margetis et al. 2021. doi: 10.1002/9781119636113.ch42

di lab

okay so transparent & explainable is complicated, but still surely we can just identify all the stakeholders and design for*(all the) user(s)?

* user-centred design

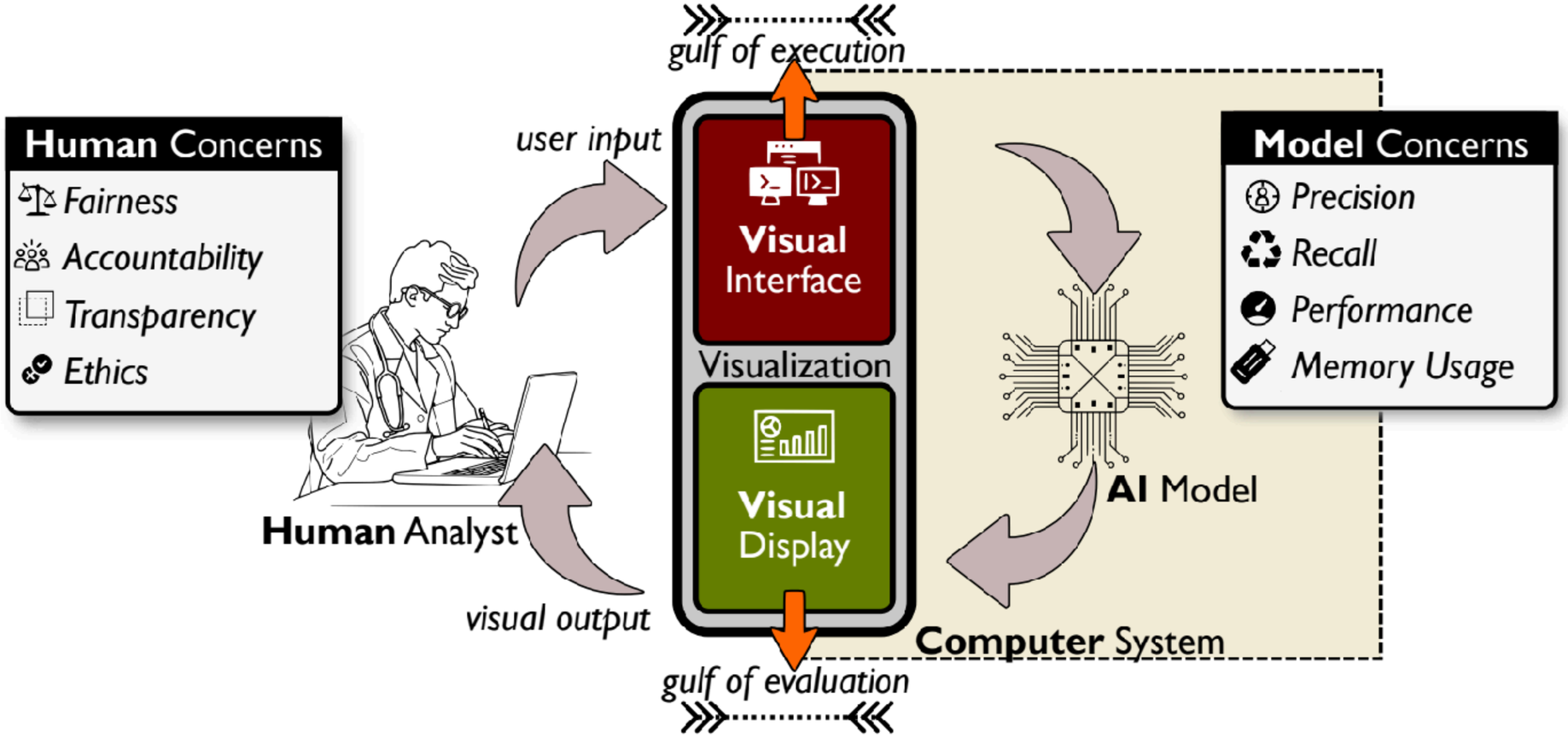
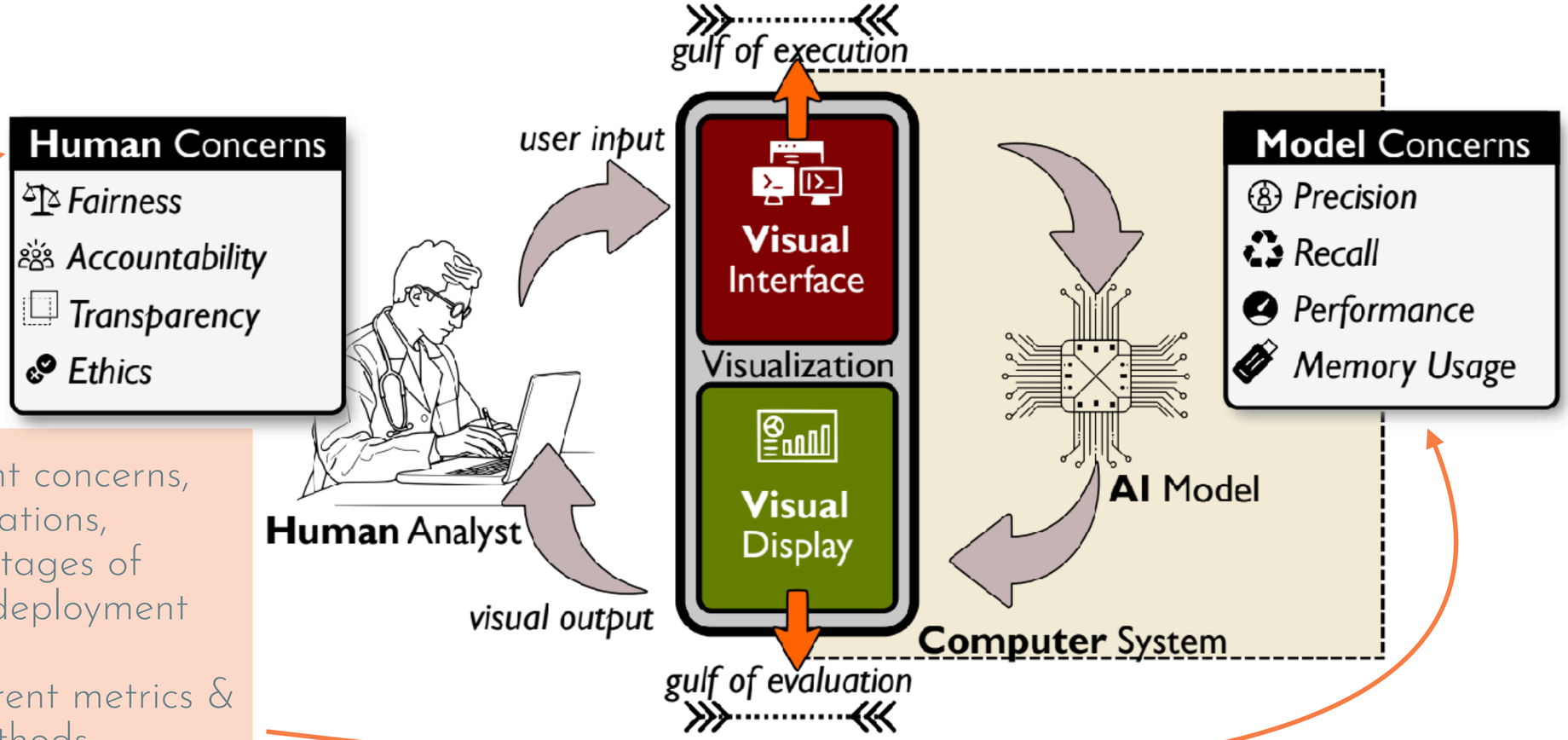


Fig. 1. **Visualization-enabled HCAI tools.** An interactive loop involving a human user and an AI model facilitated by visual interfaces.



starkly different concerns, but also motivations, stakeholders, stages of development/deployment

> starkly different metrics & evaluation methods

Fig. 1. **Visualization-enabled HCAI tools.** An interactive loop involving a human user and an AI model facilitated by visual interfaces.

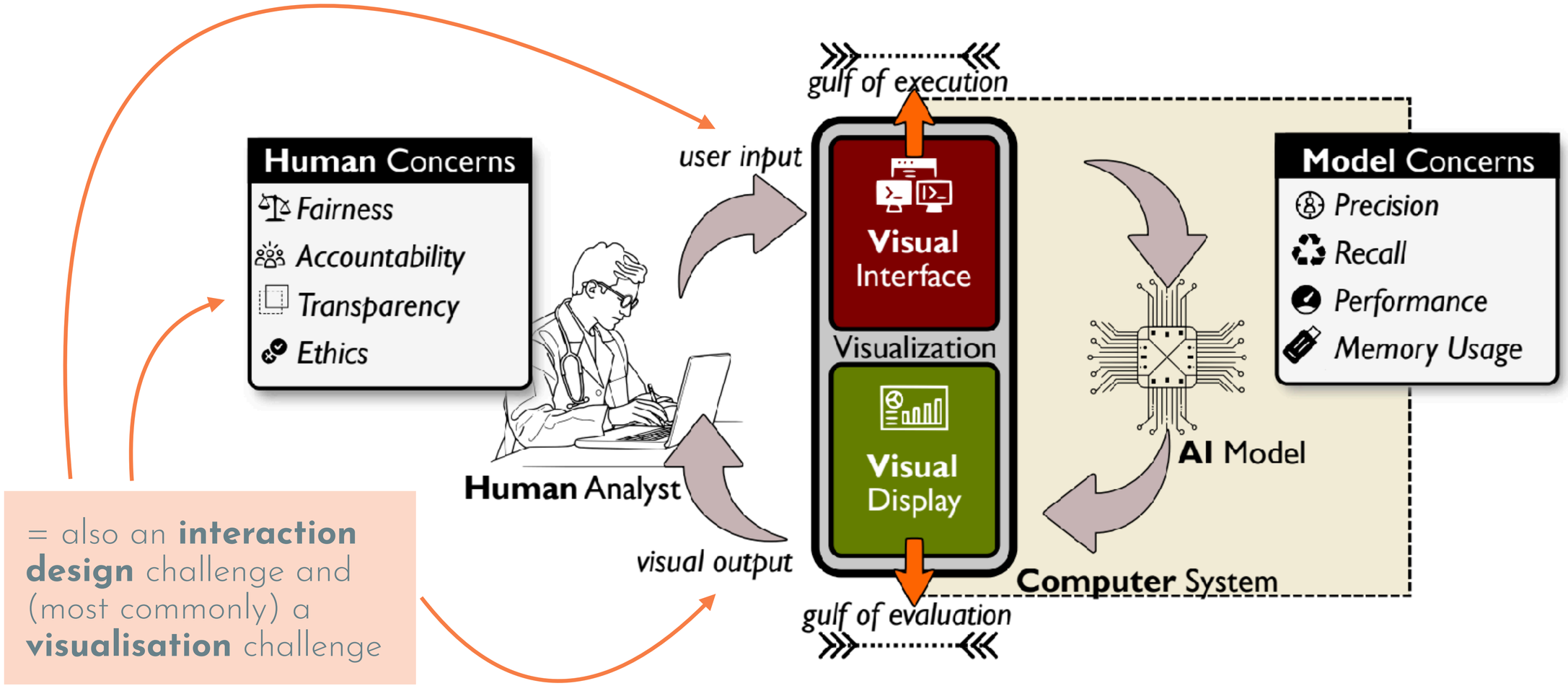


Fig. 1. **Visualization-enabled HCAI tools.** An interactive loop involving a human user and an AI model facilitated by visual interfaces.

Microsoft Research's AI Design Guidelines

INITIALLY

1 INITIALLY

Make clear what the system can do

Help the users understand what the AI system is capable of doing.

2 INITIALLY

Make clear how well the system can do what it can do.

Help the user understand how often the AI system may make mistakes.

DURING INTERACTION

3 DURING INTERACTION

Time services based on context.

Time when to act or interrupt based on the user's current task and environment.

4 DURING INTERACTION

Show contextually relevant information.

Display information relevant to the users' current task and environment.

5 DURING INTERACTION

Match relevant social norms.

Ensure the experience is delivered in a way that users would expect, given their social and cultural context.

6 DURING INTERACTION

Mitigate social biases.

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

WHEN WRONG

7 WHEN WRONG

Support efficient invocation.

Make it easy to invoke or request the AI system's services when needed.

8 WHEN WRONG

Support efficient dismissal.

Make it easy to dismiss or ignore undesired system services.

9 WHEN WRONG

Support efficient correction.

Make it easy to edit, refine, or recover when the AI system is wrong.

10 WHEN WRONG

Scope services when in doubt.

Engage in disambiguation or gracefully degrade the AI system's services when uncertain about a user's goals.

11 WHEN WRONG

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.

OVER TIME

12 OVER TIME

Remember recent interactions.

Maintain short-term memory and allow the user to make efficient references to that memory.

13 OVER TIME

Learn from user behavior.

Personalize the user's experience by learning from their actions over time.

14 OVER TIME

Update and adapt cautiously.

Limit disruptive changes when updating and adapting the AI system's behaviors.

15 OVER TIME

Encourage granular feedback.

Enable the user to provide feedback indicating their preferences during regular interaction with the AI system.

16 OVER TIME

Convey the consequences of user actions.

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

17 OVER TIME

Provide global controls.

Allow the user to globally customize what the AI system monitors and how it behaves.

18 OVER TIME

Notify users about changes.

Inform the user when the AI system adds or updates its capabilities.

Amershi et al. 2019; doi: 10.1145/3290605.3300233 [Microsoft Research]

People + AI Research

Principles

Our product principles provide directional guidance while highlight opportunities and challenges that teams experience when building experiences with AI. Illustrated with hypothetical applications.

Google

USER AUTONOMY

Design for the appropriate level of user autonomy

DATA & MODEL ALIGNMENT

Align AI with real-world behaviors

EVOLVING SAFETY

Treat safety as an evolving endeavor

ADAPT WITH FEEDBACK

Adapt AI with user feedback

HELPFUL AI

Create helpful AI that enhances work and play

People + AI Research

All Patterns

23 PATTERNS

Google

How do I explain my AI system to users?

23 PATTERNS

Set the right expectations

Be transparent with your users about what your AI-powered product can and cannot do.

Explain the benefit, not the technology

Help users understand your product's capabilities rather than what's under the hood.

Determine how to show model confidence, if at all

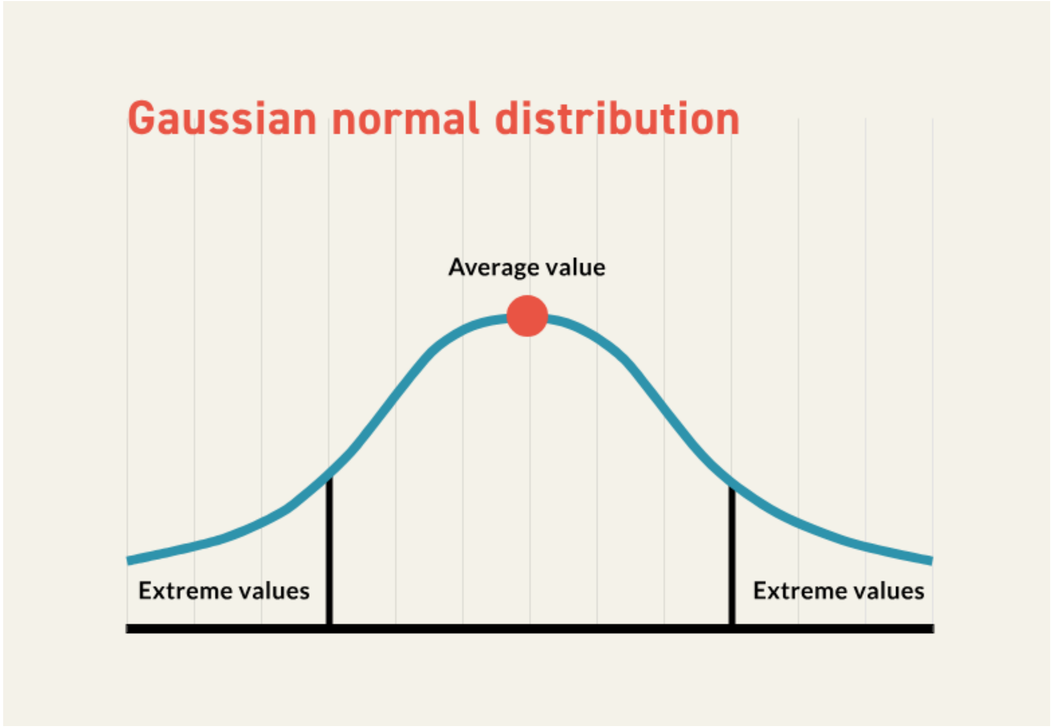
If you decide to show model confidence, make sure it's done in a way that's helpful to your users.

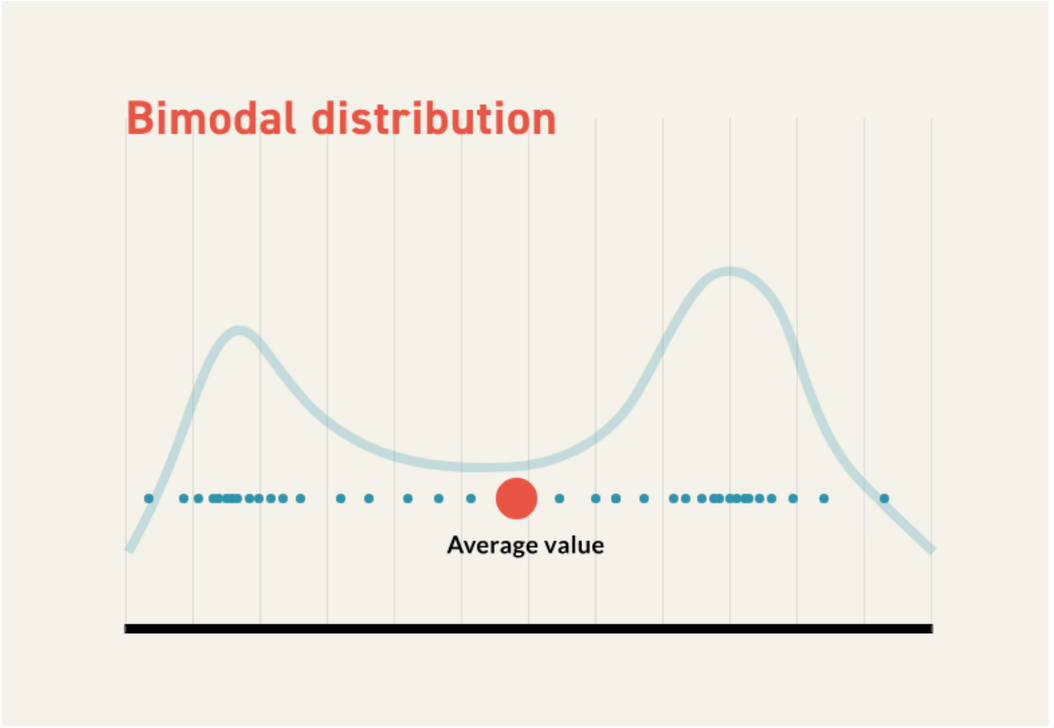
Explain for understanding, not completeness

Focus on giving your users the information they need in the moment, rather than a full run-down of your system.

Go beyond in-the-moment explanations

Help users better understand your product with deeper explanations outside immediate product flows.





The myth of "general / average users"

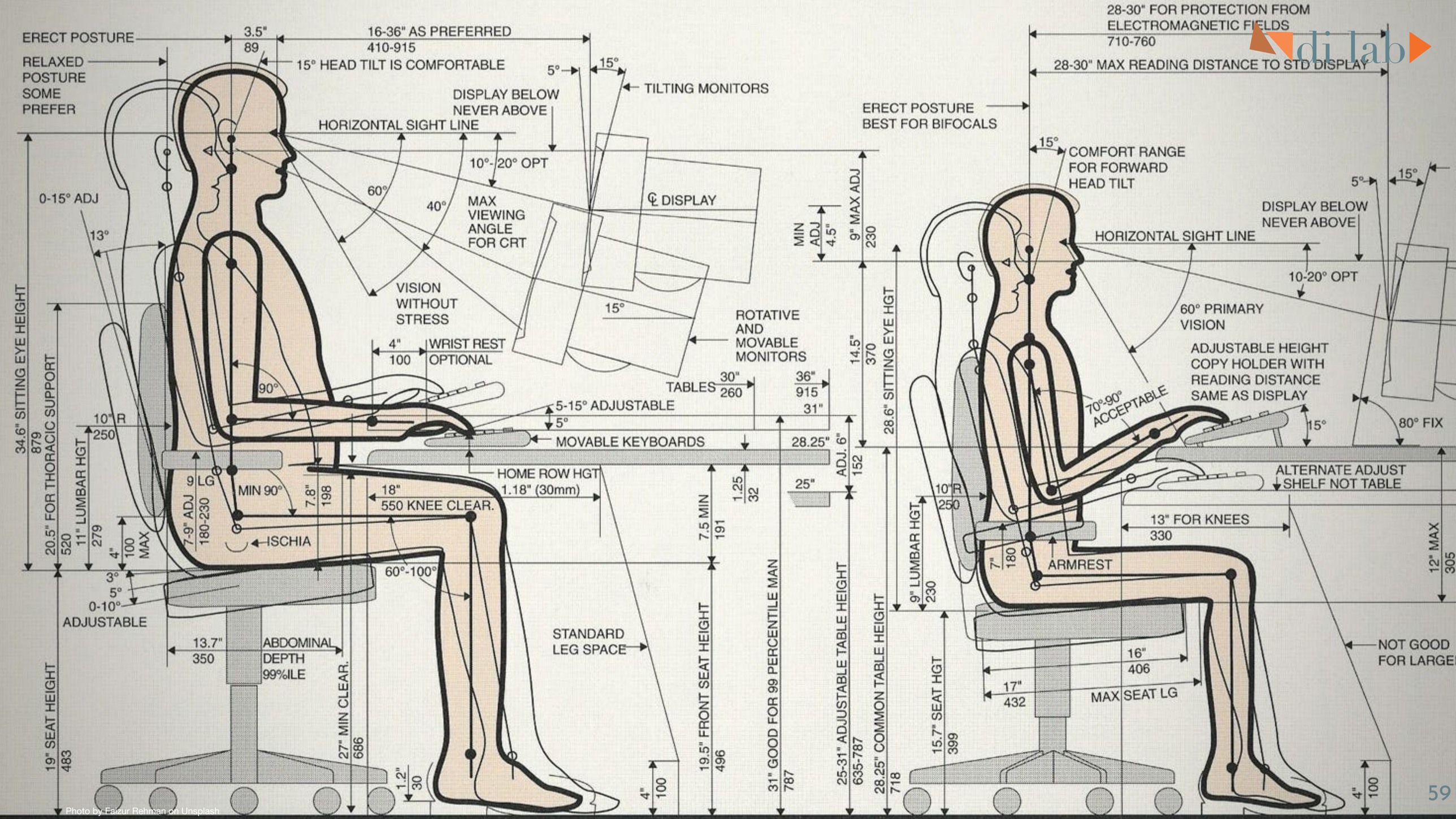


Photo by Faizur Rehman on Unsplash

Photo by Leonel Fernandez on Unsplash

1926:
US Air Force designed cockpits based on an average of pilots' dimensions for an optimal fit

1940s/50s:
pilot error rates are/remain high

1950:
attempts to remeasure - maybe pilots got bigger since 1926?

> measurement of 4k+ pilots across 140 dimensions of size (thumb length, crotch height, distance eye to ear, chest circumference, ...) which they narrowed down to 10 key dimensions most critical to piloting, to recalculate the average

> defined average pilot = within middle 30 percent of the range of values for all 10 dimensions

Measurement	Mean	Std. Dev. (SD)
1. Stature	175.5 cm	6.2 cm
2. Chest Circumference	98.6 cm	6.2 cm
3. Sleeve Length	85.5 cm	3.8 cm
4. Crotch Height	83.4 cm	4.4 cm
5. Vertical Trunk Circ.	164.6 cm	7.3 cm
6. Hip Circumference (Sit.)	105.0 cm	7.2 cm
7. Neck Circumference	38.0 cm	1.9 cm
8. Waist Circumference	81.4 cm	7.7 cm
9. Thigh Circumference	56.9 cm	4.4 cm
10. Crotch Length	71.6 cm	5.1 cm

Photo by Leonel Fernandez on Unsplash

Q. How many pilots in the dataset do you think were average across all 10 dimensions?

di lab

“not a single pilot matched the average measurements across all dimensions”

even when narrowing down further to just three dimensions:
less than 3.5% of pilots

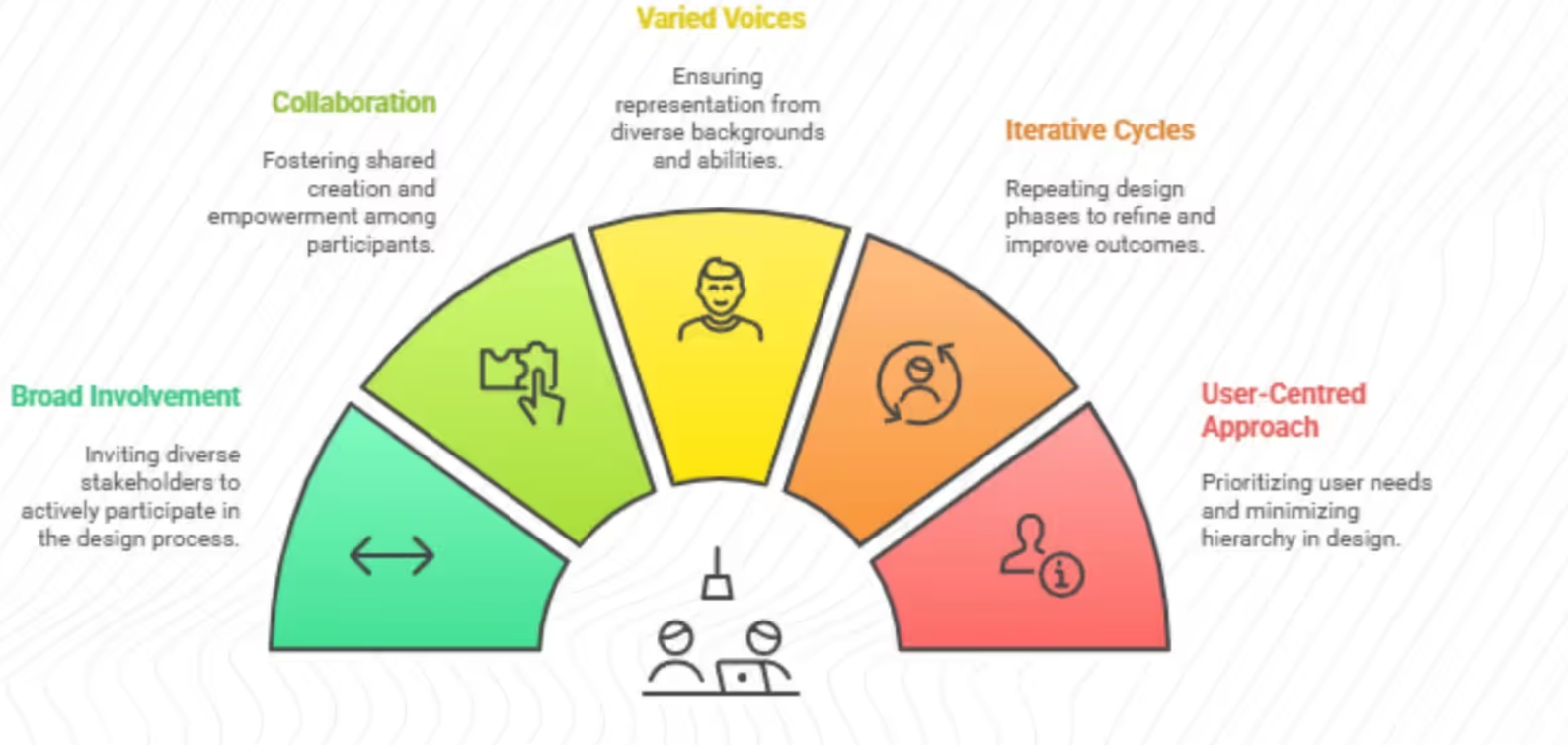
The tendency to think in terms of the "average man" is a pitfall into which many persons blunder when attempting to apply human body size data to design problems. Actually it is virtually impossible to find an "average man" in the Air Force population. This is not because of any unique traits of this group of men, but because of the great variability of bodily dimensions which is characteristic of all men. It is the intent of this Technical Note to point out and explain some of the factors that lead to the difficulties arising from the use of "average" dimensions and to indicate to some extent how they may be avoided.

Daniels 1952. The “Average Man”? Technical Note AD. No. 0/0203, Wright Air Development Center.

di lab

so let's design WITH* the user(s) instead

* participatory design



e.g., stakeholder mapping workshops: who uses output, who is affected, who has power, who bears risk

e.g., value sensitive design methods to elicit stakeholder values from the start

Image from https://www.parallelhq.com/blog/what-participatory-design

The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice

Fernando Delgado*
Cornell University
Ithaca, New York, USA
fad33@cornell.edu

Stephen Yang*†
University of Southern California
Los Angeles, California, USA
stepheny@usc.edu

Michael Madaio*†
Google Research
New York, New York, USA

Qian Yang*
Cornell University
Ithaca, New York, USA
qianyang@cornell.edu

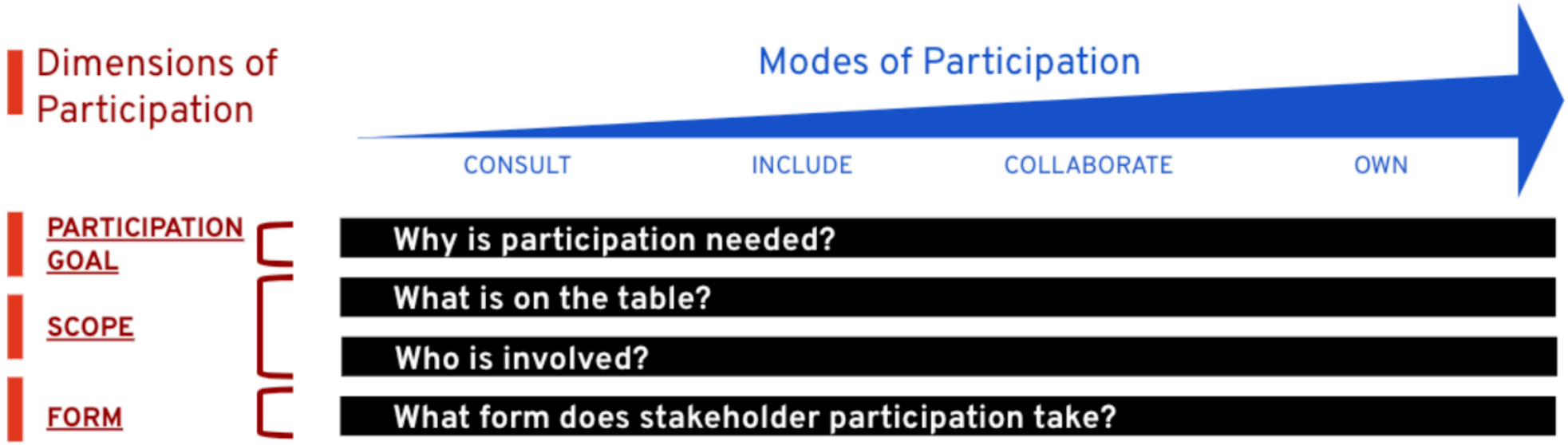


Figure 1: *Parameters of Participation*: a framework derived from a synthesis of prior literature on stakeholder participation

https://dl.acm.org/doi/pdf/10.1145/3617694.3623261

“there have been increasing calls to involve members of communities impacted by AI systems in their design […]. In part, such calls for participation in AI design argue that participation can enable AI systems to better reflect the values, preferences, and needs of users and other impacted stakeholders, or more broadly, that participation will empower stakeholders in shaping the design of AI systems […].

However, despite a growing consensus that stakeholders should participate more in AI design, there is enormous variation in the methods and theories applied to achieve that participation, even with respect to the goals for leveraging participation in the first place.”

https://dl.acm.org/doi/pdf/10.1145/3617694.3623261

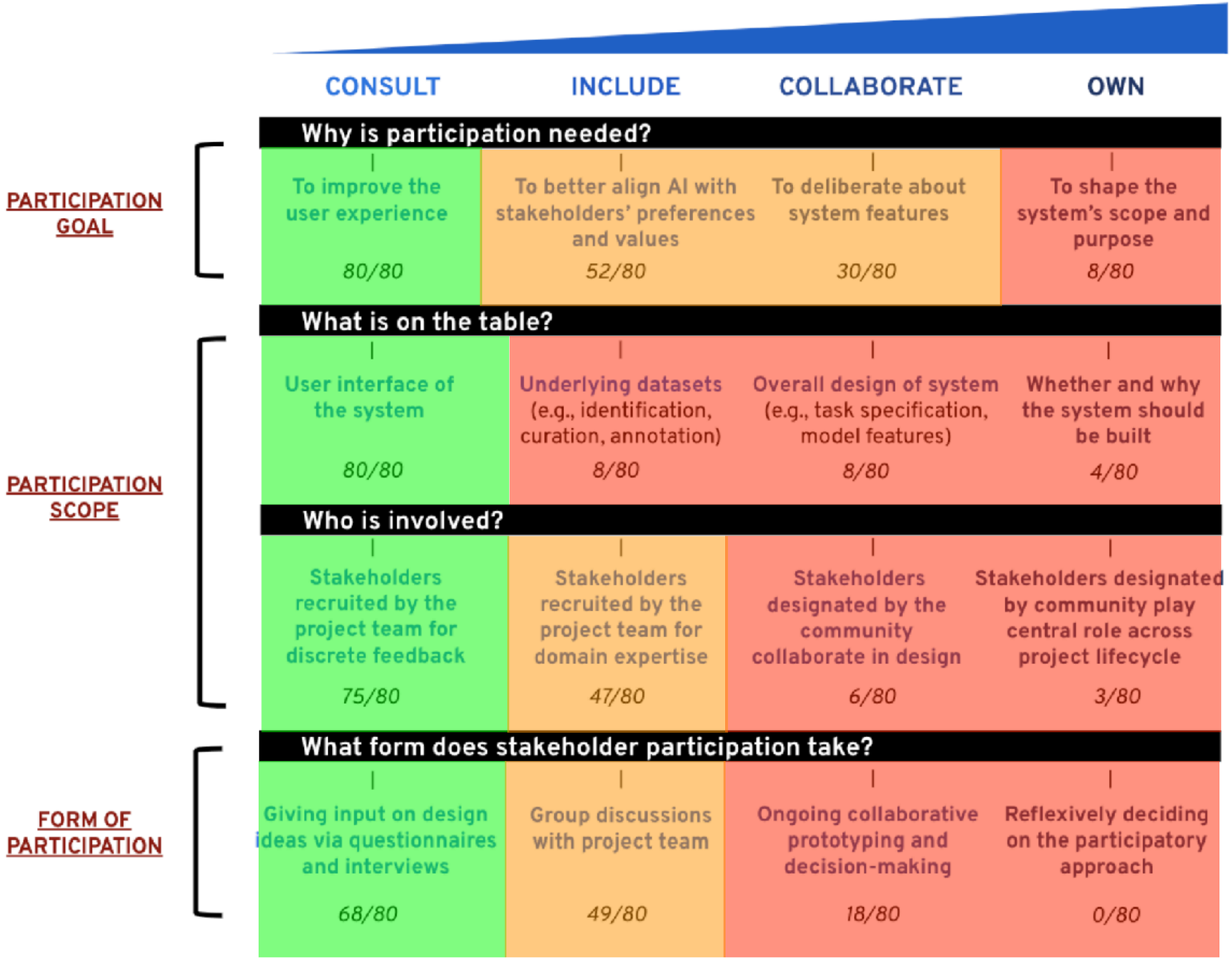


Figure 2: Participatory AI Projects Mapped to Conceptual Framework

https://dl.acm.org/doi/pdf/10.1145/3617694.3623261

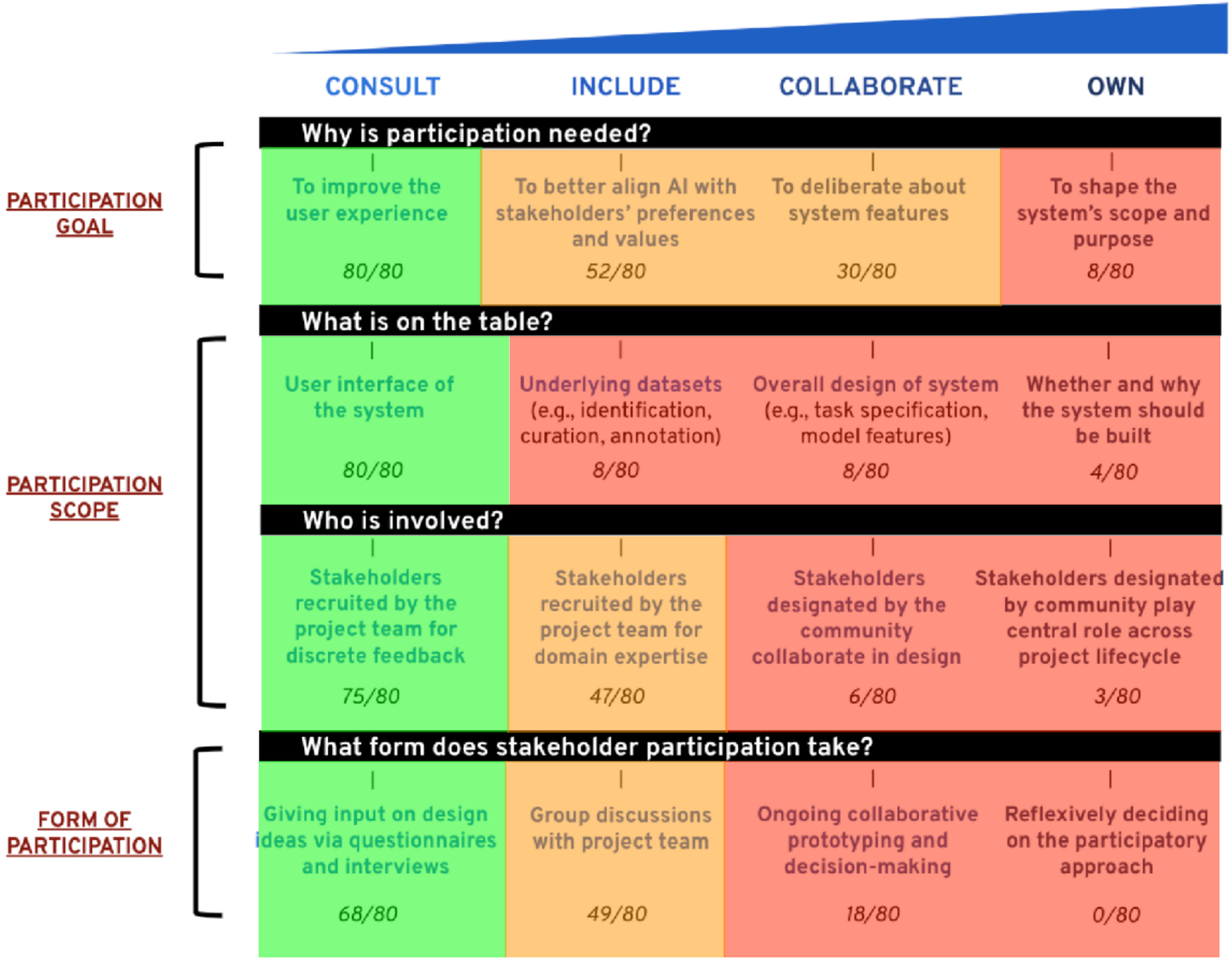


Figure 2: Participatory AI Projects Mapped to Conceptual Framework

"Researchers and practitioners feel they have to substantially scale back their participatory ambitions [...] top-down organizational constraints limited the time and resources available for participatory approaches. "

https://dl.acm.org/doi/pdf/10.1145/3617694.3623261

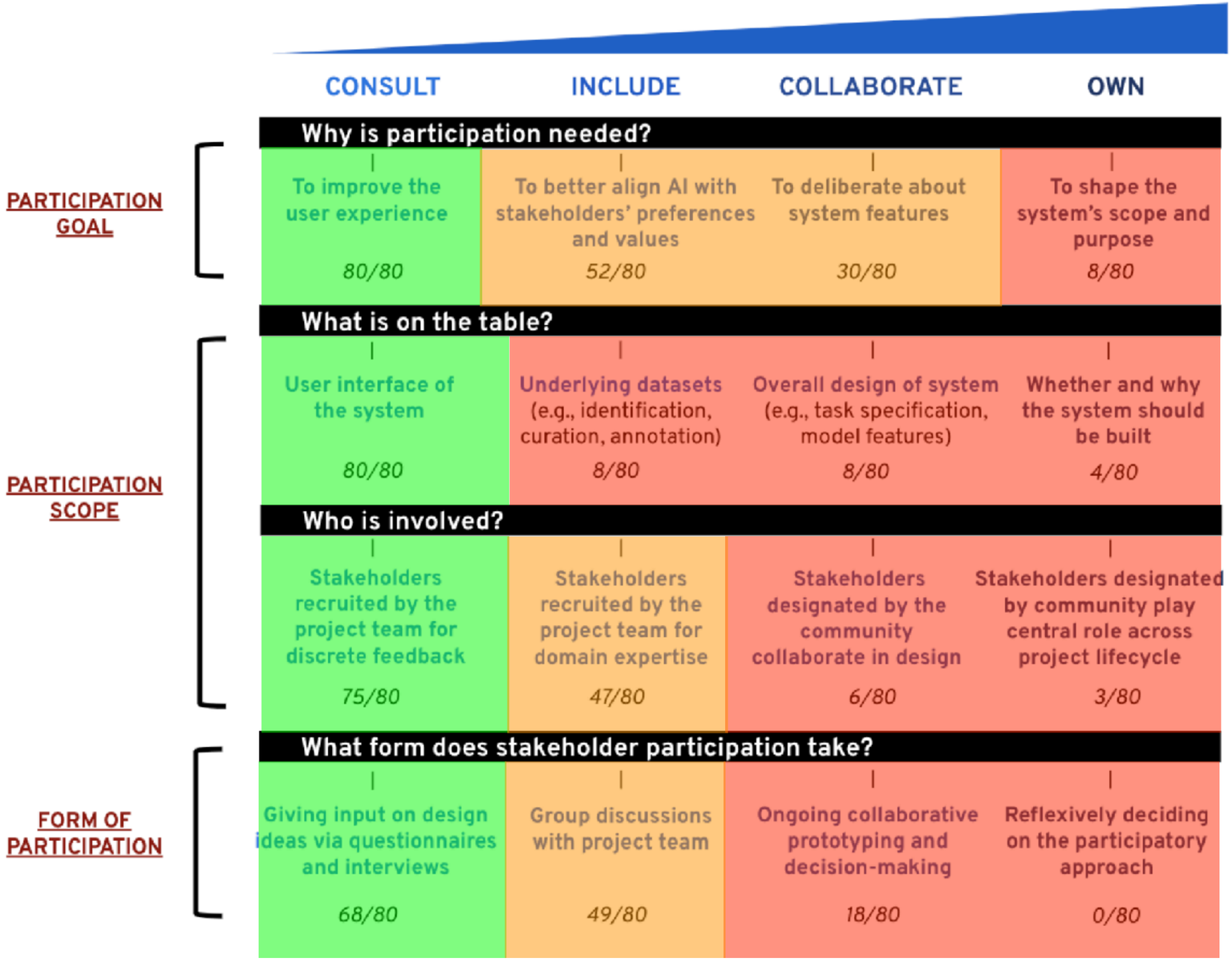


Figure 2: Participatory AI Projects Mapped to Conceptual Framework

"Often, only *how*—not *whether*—AI will be deployed was subject to discussion with stakeholders"

https://dl.acm.org/doi/pdf/10.1145/3617694.3623261

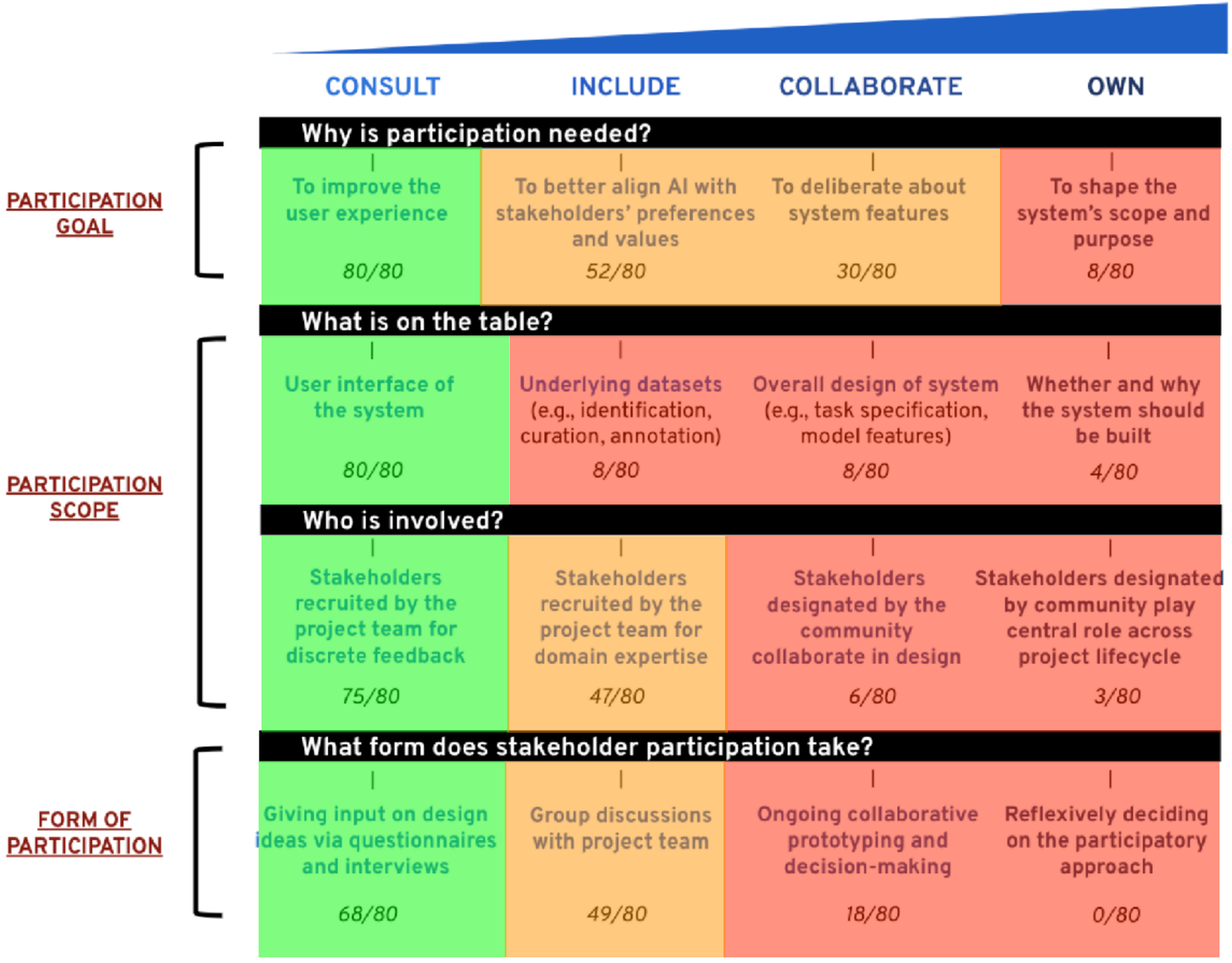


Figure 2: Participatory AI Projects Mapped to Conceptual Framework

"One interviewee went further to describe the notion of full participation as an impossibility, as for them this would require participants "*literally coming to the office with me and making every decision with me and doing all these things; all of a sudden, they don't have a life to live, right?*""

Image by Calum Lewis on Unsplash
https://dl.acm.org/doi/pdf/10.1145/3617694.3623236

"Much like in participatory design more broadly, truly participatory approaches to AI will take more than "*add[ing] diverse users and stir[ring]*" to achieve substantive forms of participation"

concretely, what might substantive participatory AI design look like?

WeBuildAI: Participatory Framework for Algorithmic Governance

MIN KYUNG LEE, University of Texas at Austin & Carnegie Mellon University, USA
DANIEL KUSBIT, Ethics, History & Public Policy, Carnegie Mellon University, USA
ANSON KAHNG, School of Computer Science, Carnegie Mellon University, USA
JI TAE KIM, School of Design, Carnegie Mellon University, USA
XINRAN YUAN, Information Systems, Carnegie Mellon University, USA
ALLISSA CHAN, School of Design, Carnegie Mellon University, USA
DANIEL SEE, Decision Science & Art, Carnegie Mellon University, USA
RITESH NOOTHIGATTU, School of Computer Science, Carnegie Mellon University, USA
SIHEON LEE, Information Systems, Carnegie Mellon University, USA
ALEXANDROS PSOMAS, School of Computer Science, Carnegie Mellon University, USA
ARIEL D. PROCACCIA, School of Computer Science, Carnegie Mellon University, USA

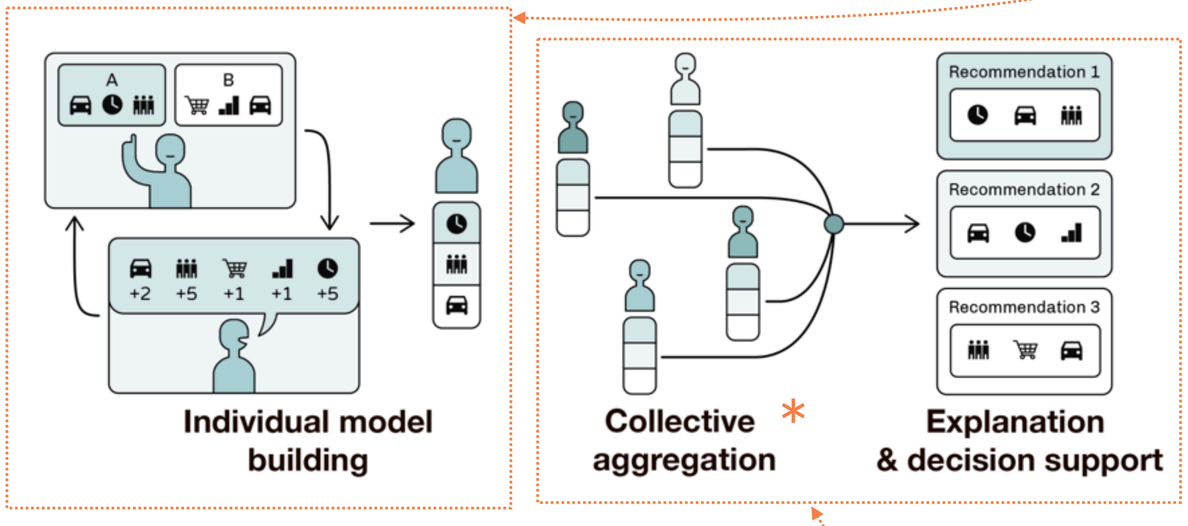


Fig. 1. The WeBuildAI framework allows people to participate in designing algorithmic governance policy. A key aspect of this framework is that individuals create computational models that embody their beliefs on the algorithmic policy in question and vote on the individual's behalf.

"In this framework, stakeholders build an algorithmic model that represents their beliefs about ideal algorithm operation. For each decision task, each individual's model votes on alternatives, and the votes are aggregated to reach a final decision."

* weighted based on stakeholder group (determined by stakeholders together)

https://dl.acm.org/doi/pdf/10.1145/3359283

“As a case study, we designed a matching algorithm that operates 412 Food Rescue’s on-demand transportation service, implementing the framework with their stakeholders: donors, volunteers, recipient organizations, and 412 Food Rescue’s staff.”

Role		Studies Involved
412 Food Rescue.*		
F1		Sessions 1-4
F2		Sessions 1-4
F3		Sessions 1-4, w
Recipient organizations. (Clients served monthly, client neighborhood poverty rate)		
R1	Human services program manager (N=150, 13%)	Sessions 1-4
R2	Shelter & food pantry center director (N=50, 20%)	Sessions 1-4
R3	Food pantry employee (N=200, 53%)	Sessions 1-4
R4	Animal shelter staff	Session 1
R5	Food pantry staff (N=500, 5%)	Sessions 1-4
R6	After-school program employee (N=20, 33%)	Session 1, w
R7	Home-delivered meals delivery manager (N=50, 11%)	Sessions 1-4
R8	Food pantry director (N=200, 14%)	Sessions 1-2
Volunteers.		
V1	White male, 60s	Sessions 1-4, w
V2	White female, 30s	Session 1
V3	White female, 70s	Sessions 1-4, w
V4	White female, 70s (V4a), white male, 70s (V4b) †	Sessions 1-4
V5	White female, 60s	Sessions 1-4
V6	White female, 20s	Sessions 1-4
Donor organizations.		
D1	School A dining service manager	Session 1
D2	School B dining service manager	Sessions 1-4
D3	Produce company marketing coordinator	Session 1
D4	Grocery store manager	Sessions 1-4
D5	Manager at dining and catering service contractor	Session 1
D6	School C dining service employee	Session 1, w

Table 1. Participants. Sessions indicate the study sessions that they participated in: *w* represents a workshop study. *Info excluded for anonymity. † A couple participated together.

https://dl.acm.org/doi/pdf/10.1145/3359283

Factor	Explanation
Travel Time	The expected travel time between a donor and a recipient organization. Indicates time that volunteers would need to spend to complete a rescue. (0-60+ minutes)
Recipient Size	The number of clients that a recipient organization serves every month. (0-1000 people; AVG: 350)
Food Access	USDA-defined food access level in the client neighborhood that a recipient organization serves. Indicates clients' access to fresh and healthy food. (Normal (0), Low (1), Extremely low(2)) [78]
Income Level	The median household income of the client neighborhood that a recipient organization serves (0-100K+, Median=$41,283) [77]. Indicates access to social and institutional resources [69].
Poverty Rate	Percentage of people living under the US Federal poverty threshold in the client neighborhood that a recipient organization serves. (0-60 %; AVG=23% [77])
Last Donation	The number of weeks since the organization last received a donation from 412 Food Rescue. (1 week–12 weeks, never)
Total Donations	The number of donations that an organization has received from 412 Food Rescue in the last three months. (0-12 donations) A unit of donation is a carload of food (60 meals).
Donation Type	Donation types were common or uncommon. Common donations are bread or produce and account for 70% of donations. Uncommon donations include meat, dairy, prepared foods, etc.

Table 2. Factors of matching algorithm decisions. The ranges of the factors are based on their real-world distributions.

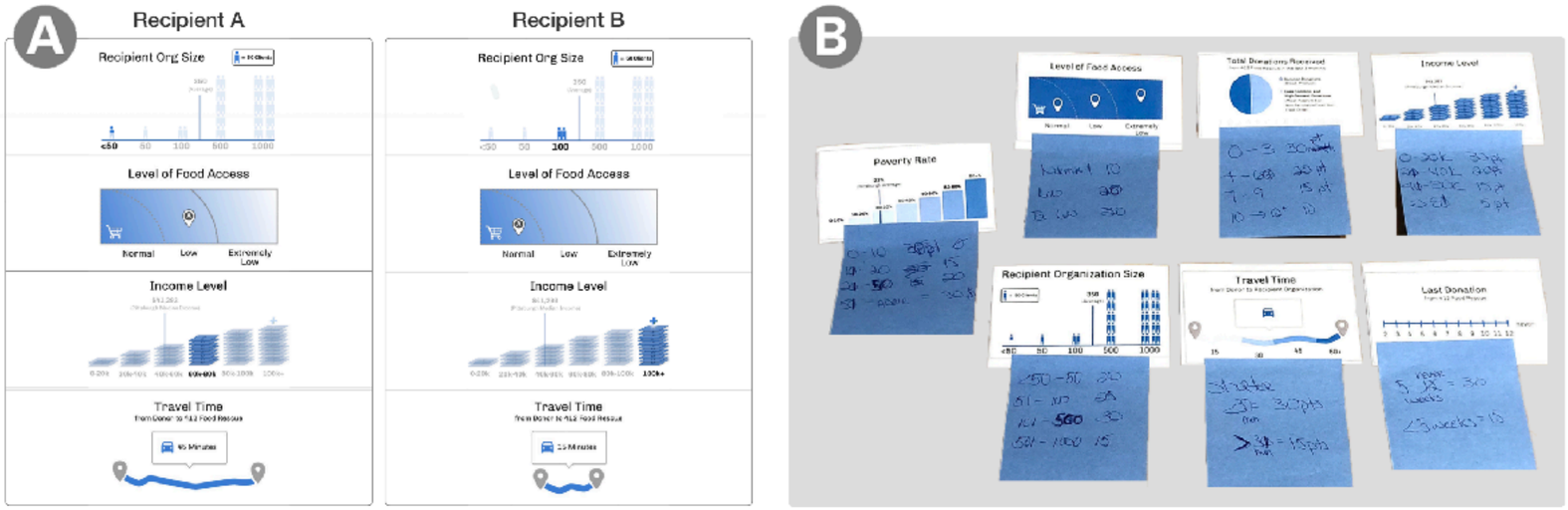


Fig. 2. Two methods of individual model building were used in our study: (a) a machine learning model that participants trained through pairwise comparisons, and (b) an explicit rule model that participants specified by assigning scores to each factor involved in algorithmic decision-making.

	D2	D4	F2	F3	R1	R2	R3	R5	R7	V1	V3	V4	V5	V6
ML	**0.86**	0.78	**0.92**	**0.92**	**0.90**	0.90	**0.78**	**0.94**	0.74	**0.90**	**0.92**	**0.78**	0.56	0.68
ER	0.68	**0.68**	0.68	0.86	0.80	**0.76**	0.70	0.92	**0.74**	0.76	0.82	0.82	**0.80**	**0.88**

Table 3. Accuracy of the Machine Learning (ML) model and the Explicit-Rule (ER) model. Bold denotes the model the participant chose as the one that better represented their belief after seeing both models' explanations (Figure 3) and their predictions on the 50 evaluation pairwise comparisons. F1 chose the machine learning model but did not complete additional survey questions to calculate model agreement, so the result is not included in this table.

https://dl.acm.org/doi/pdf/10.1145/3359283

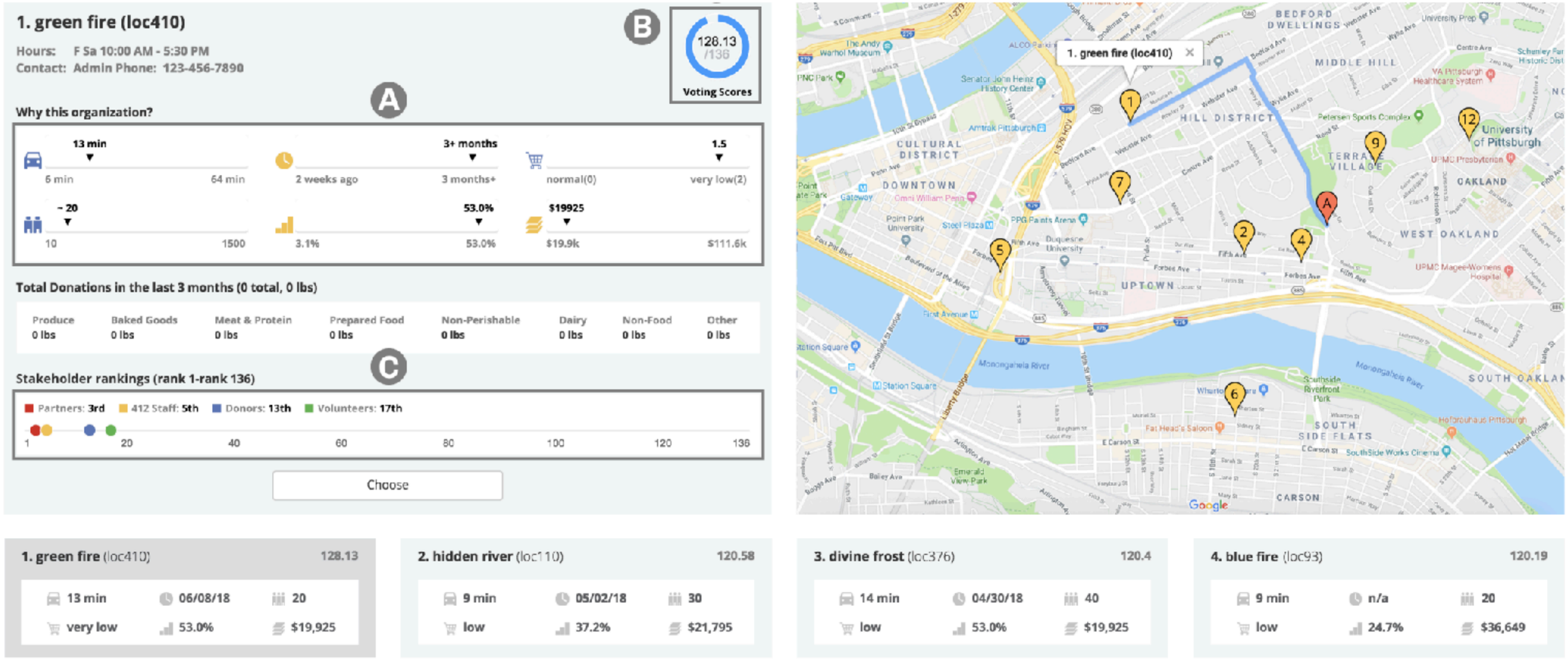
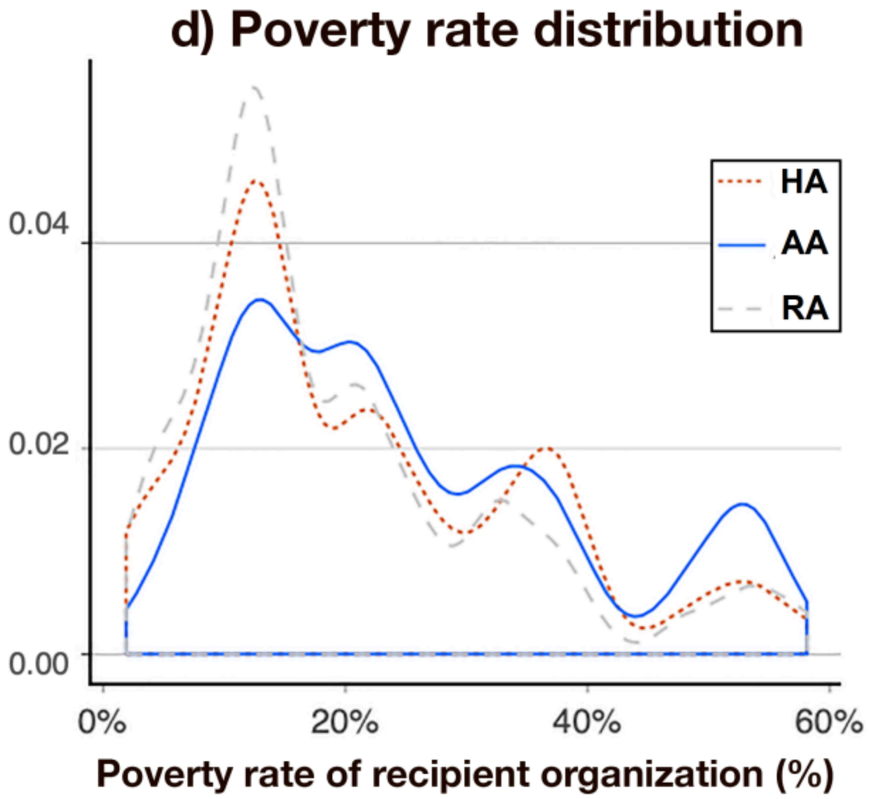


Fig. 4. The decision support tool explains algorithmic recommendations, including the nature of stakeholder participation, stakeholder voting results, and characteristics of each recommendation. The interface highlights the features of the recommended option that led to its selection (marked by A), the Borda scores given to the recommended options in relation to the maximum possible score (marked by B), and how each option was ranked by stakeholder groups (marked by C). All recipient information and locations are fabricated for the purpose of anonymization.

https://dl.acm.org/doi/pdf/10.1145/3359283

“We compared our algorithm (AA) with two benchmarks: human allocations recorded in historical data (HA), and a random algorithm that selected a recipient uniformly at random (RA).”



“We then evaluated the resulting algorithm with historical donation data, which showed that our algorithm leads to a more even donation distribution that prioritizes organizations with lower income, higher poverty rate, and lower food access clients compared to human allocation decisions.

Our findings suggest that the framework improved the perceived fairness of the allocation method. It also increased individuals’ awareness of algorithmic technology as well as the organization’s awareness of the algorithm’s impact and employee decision-making inconsistencies.”

https://dl.acm.org/doi/pdf/10.1145/3359283

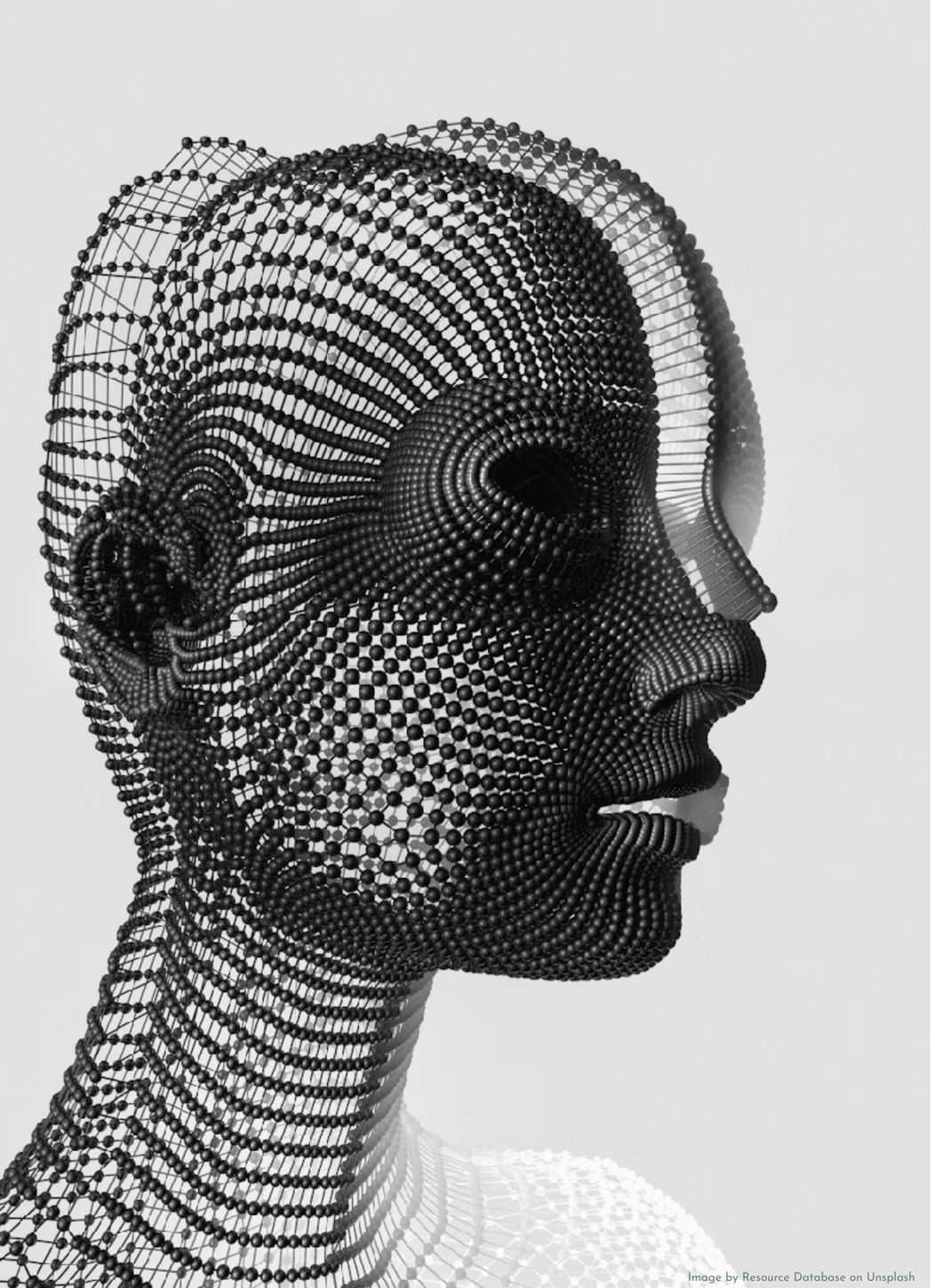


Image by Resource Database on Unsplash

Thanks for listening!

Questions?

dr. Katja Rogers
Assistant professor, Digital Interactions Lab, UvA