Work in progress!

Responsible AI

What is it, why does it matter, and how do we achieve it?

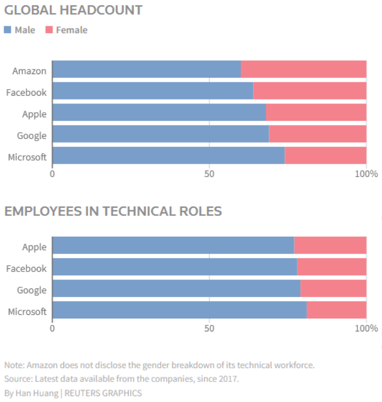
Examples of risks of irresponsible AI

Corona App

An app can tell you if you're at risk of being infected with Corona and whether you should get tested, but such an app might invade your privacy.

Amazon Recruitment

An algorithm determined the fit of an applicant for a job, but used gender as one of its main predictors, due to its use of historical data.



Childcare benefit affair

Missing, faulty, and outdated data resulted in flawed estimations and predictions at the Dutch Tax Services.

SyRi

This social benefit fraud risk detection program was deemed disproportionate by the court of law.

Responsible AI addresses a variety of challenges

Violations of regulations, norms or values (e.g. disproportionate invasion of privacy)

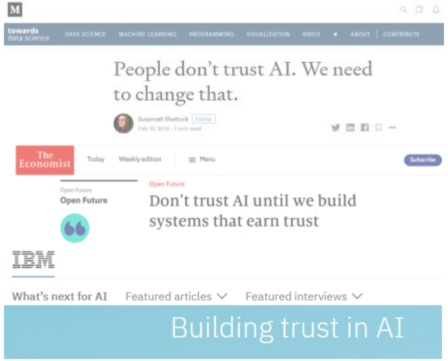
Insufficient understanding of models' short and long term effects

Algorithms and/or models that do not allow for straightforward explanation of the value of predictions

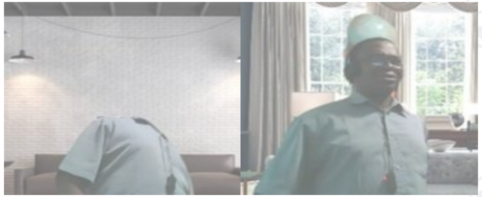
Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By James Vincent | Mar 24, 2016, 6:43am EDT
Via The Guardian | Source TayandYou (Twitter)

Incomprehensible or unacceptable behavior for customers and public



Lack of trust in AI systems



Unfair behavior caused by a variety of biases (e.g. Zoom not recognizing black faces)

Commercial drives to work towards Responsible AI

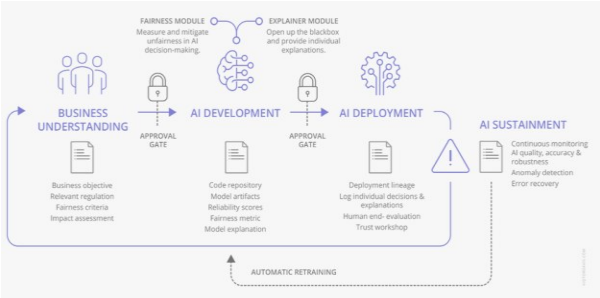
- **Increase the business value of Data Science**
  - Work with the actual predictive features instead of proxies
  - More robust algorithms / solutions (fighting brittleness)
  - Understanding of the models and predictive outcomes, thereby increasing trust, acceptance and use
- **Be prepared for accountability**
  - GDPR
  - Fairness / equal treatment
  - Other legal regulations
- **Safeguard your reputation**
  - Customers / clients
  - Employees
  - Government / society

Responsible AI is trending (and not without reason)

[Price Waterhouse Cooper](#)
Responsible AI Toolkit

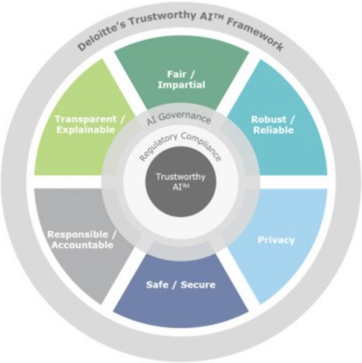
[McKinsey](#)
Articles on Fair and Responsible AI

[ALLAI](#)
Independent organization dedicated to drive and foster Responsible AI

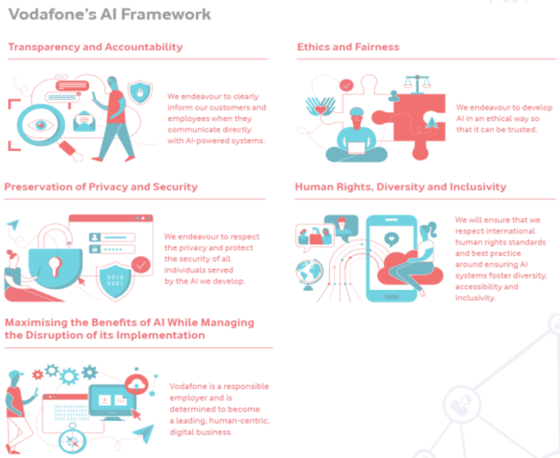


[Vigtor Davis](#)
FACT-AI Framework

[Accenture](#)
Responsible AI Framework



[Deloitte](#)
Trustworthy AI framework



[Vodafone](#)
AI framework

This presentation - an outline

1. Causes of irresponsible AI
   a. Fairness and Bias in Data Science and AI
   b. Misunderstanding and/or misuse of (black box) algorithms
   c. Analysis of stakeholders and potential impact
   d. Limited possibility for humans to provide feedback or input for the system to improve
   e. Unsustainable energy use by (re)training and data use
2. Responsible AI - design method and approach
   a. Three perspectives
   b. Three design methods
3. Xomnia Responsible AI
   a. Way of Working
   b. Roles and Activities related to Responsible AI
   c. Framework for Responsible AI

1. Causes of irresponsible AI

Causes of irresponsible AI

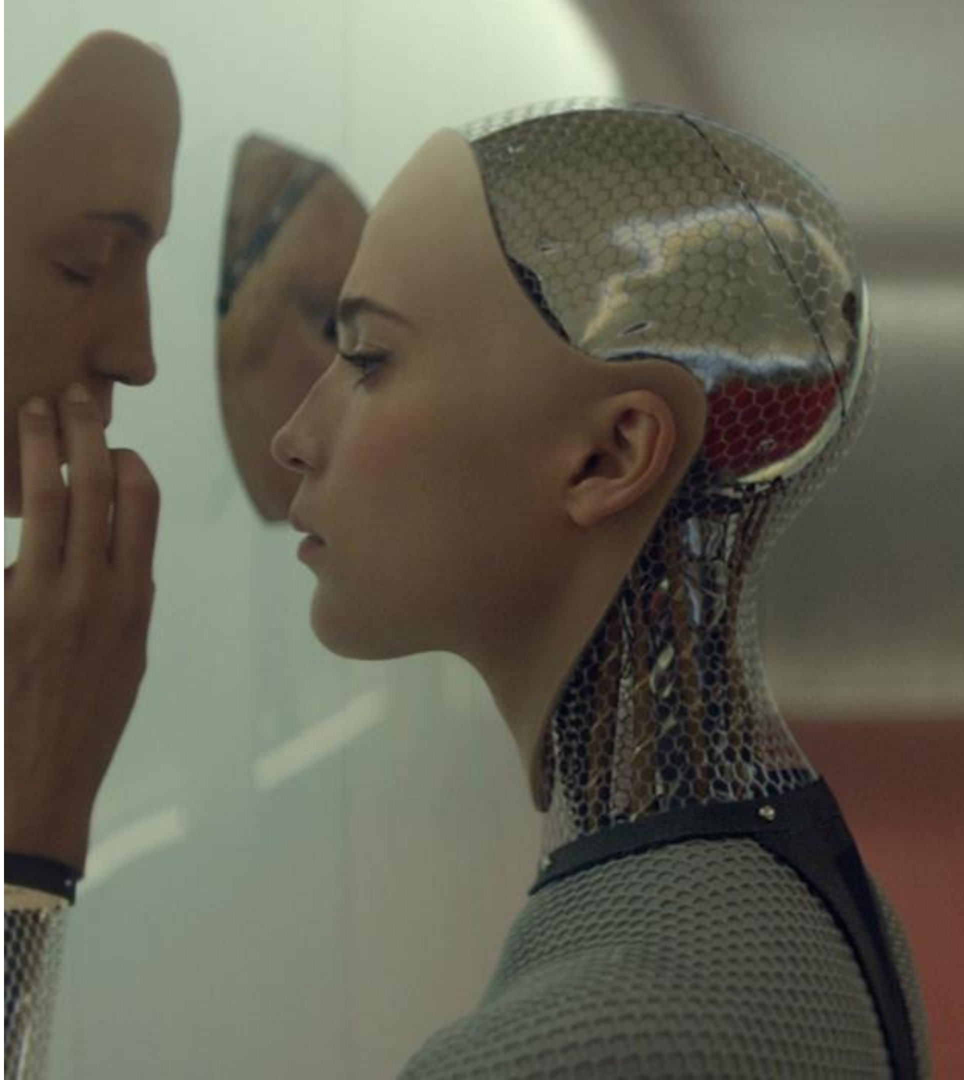
A. Unfairness due to (various kinds of) bias

B. Misunderstanding and/or misuse of (black box) algorithms

C. Focus on the merits / benefits / impact for only a subset of the system's stakeholders

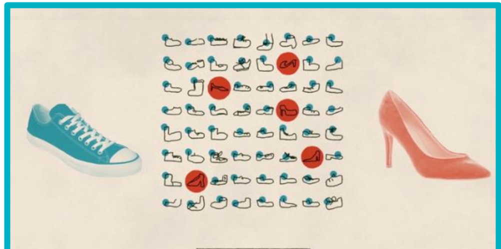
D. Limited possibility for humans to provide feedback or input for the system to improve

E. Unsustainable energy use by (re)training and data use



A. Fairness and Bias in Data Science and AI

A. Fairness and bias in data



Fairness

1) Similar predictions to similar individuals;
2) Treat different groups equally

Potential harms

1) Extend or withhold opportunities, resources, or information unfairly.
   Example: [Determining whether a client is eligible for receiving a loan;]
2) Provide unequal quality of service.
   Example: [Twitter cropped images to fit on mobile screen favouring white faces over black ones]

Technical bias
E.g. a search engine showing only the first three results on the first page

Emergent (also called interaction) bias
Learning from data that comes from interactions with an environment (including humans) that does not accurately reflect the real world. (Examples: predictive police patrolling, the shoe example)

Latent (also called pre-existing) bias
Incorrect predictions because of pre-existing bias in the training data set, e.g. when training on historical data that contains a societal stereotype. (Example: Amazon recruitment)

Selection bias
Learning from a set of training data that does not include sufficient data on *all* possible instances in the test data set (does not reflect the entire population). (Example: facial recognition that fails in recognizing faces of people of color)

A. Tools for fairness and mitigating bias

Fairlearn: navigate any trade-offs between fairness and performance, and select the mitigation strategy that best fits your needs.
Two components:

1) An assessment dashboard for assessing which groups are negatively impacted.
2) A set of strategies for mitigating fairness issues.

White paper (Microsoft): Fairlearn: A toolkit for assessing and improving fairness in AI

ML-Fairness Gym: exploring algorithmic feedback loops in evolving machine learning systems.

- Adaptation of Google's OpenAI Gym used for analysis of reinforcement learning algorithms.
- Collection of test problems and environments.

Paper: "Fairness is not static", Example: Determining whether a client is eligible for receiving a loan.

What-if Tool: Visually probe the behavior of trained machine learning models. WIT allows you to test performance in hypothetical situations, analyze the importance of different data features, and visualize model behavior across multiple models and subsets of input data, and for different ML fairness metrics.

Also see this tutorial which combines the What-if tool with SHAP (see this slide)

Amazon SageMaker Clarify
Detect bias in ML models and understand model predictions

IBM Research Trusted AI
AI Fairness 360

B. Misunderstanding / misuse of algorithms

B. Misunderstanding / misuse of algorithms

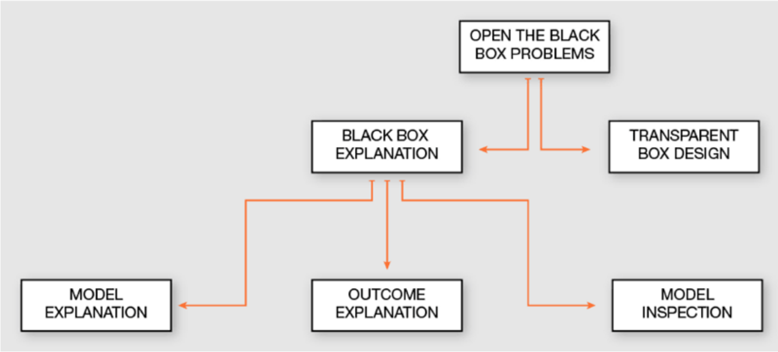
*A model takes input data and turns it into output data. The **black box** is used as a metaphor for models that make it difficult for a human to sufficiently understand the input-output relation. What is sufficient in any given situation depends on the context of use and the purpose of the model.*

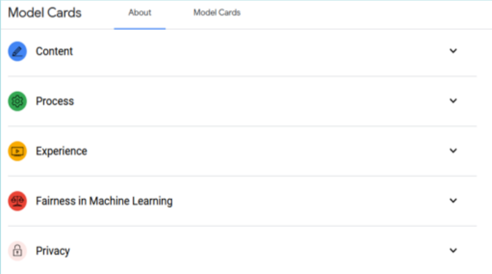
Two causes for "black box" behaviour:

1) Complexity of machine learning models.

2) Limited access to inner workings for users, customers, and other stakeholders (often due to IP protection).

B. Tools to open the black box

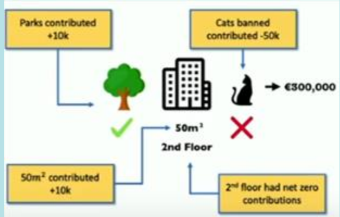




Model cards (Google research)

- Glass box approaches, e.g. Explainable Boosting Machines (EBM)
- Intuitive Confidence Measure: computes the probability that a given output in a new situation is correct.
- Contrastive explanations: 'Why'-questions of the form "Why P rather than Q?"where Q is an expected foil case instead of true outcome P.
- Causal & counterfactual explanations: what-if reasoning (e.g. what-if tool)
- InterpretML: Combines many of the tools listed here

SHAP:
Shapley Additive Explanations



LIME:
Local Interpretable Model-Agnostic Explanations

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

B. Tools to increase reproducibility

Abstracting computational steps:

https://www.youtube.com/watch?v=eOzl-LFqYFM

https://www.seldon.io/

Adopting open standards:

https://onnx.ai/

https://www.khronos.org/nnef

http://dmg.org/

C. Analysis of stakeholders and potential impacts

C. Analysis of stakeholders and potential impacts

Stakeholder analysis

Ensuring that **all stakeholders** are represented in evaluations and discussions. Identifying various characteristics and representativeness within stakeholder groups as well as a wide collection of both direct and indirect stakeholder groups is crucial.

Impact analysis

Looking at **a wide angle of potential impact** when envisioning the technological outcome. Critical thinking is highly important in this stage, and inviting a variety of stakeholders can help collect the relevant input.

C. Tools for stakeholder and impact analysis

Value-sensitive design:

- Include important stakeholders affected by envisioned solution:
  - Who (else) is affected by your solution?
  - Are those affected included in the design conversation?
  - What are the values and needs of different stakeholders?
  - What value tensions may exist?
  - How can value tensions be mitigated (by design)?

Ethical Toolkit for development of AI Applications:

- Create awareness within organisation to prevent overlooking of potential (detrimental) side effects
  - Ethics workshop
    - Trolley problem
    - Ethical dilemmas
  - AI Project checklist (workshop)
    - Stakeholders impacted
    - Type of impact
    - Ethical principles checklist
  - Responsible AI Deck (workshop)

Judgment call - the game:

- A card game to practice value sensitive design
- Introduces various scenarios and stakeholders in fictional use cases
- Players write reviews from the perspective of a fictional character from the stakeholder group and evaluate among one another

C. Security and Privacy by Design

Privacy:

Protect personal data wherever you can (e.g. differential privacy kit)

Be aware of what information may be given away by metadata about your users / clients

Security:

Mitigate basic loopholes for adversarial patch tricking models (also here)

For true security, it is important that technologists take into consideration the whole lifecycle of the machine learning algorithm. The process and infrastructure to store the training data, accuracy, documentation, trained model, orchestration of the model, inference results and beyond.

D. Limited human-system feedback for system improvement

D. Limited human-system feedback for system improvement

D. Tools enabling human-system feedback

E. Unsustainable energy use by (re)training and data use

E. Unsustainable energy use by (re)training and data use

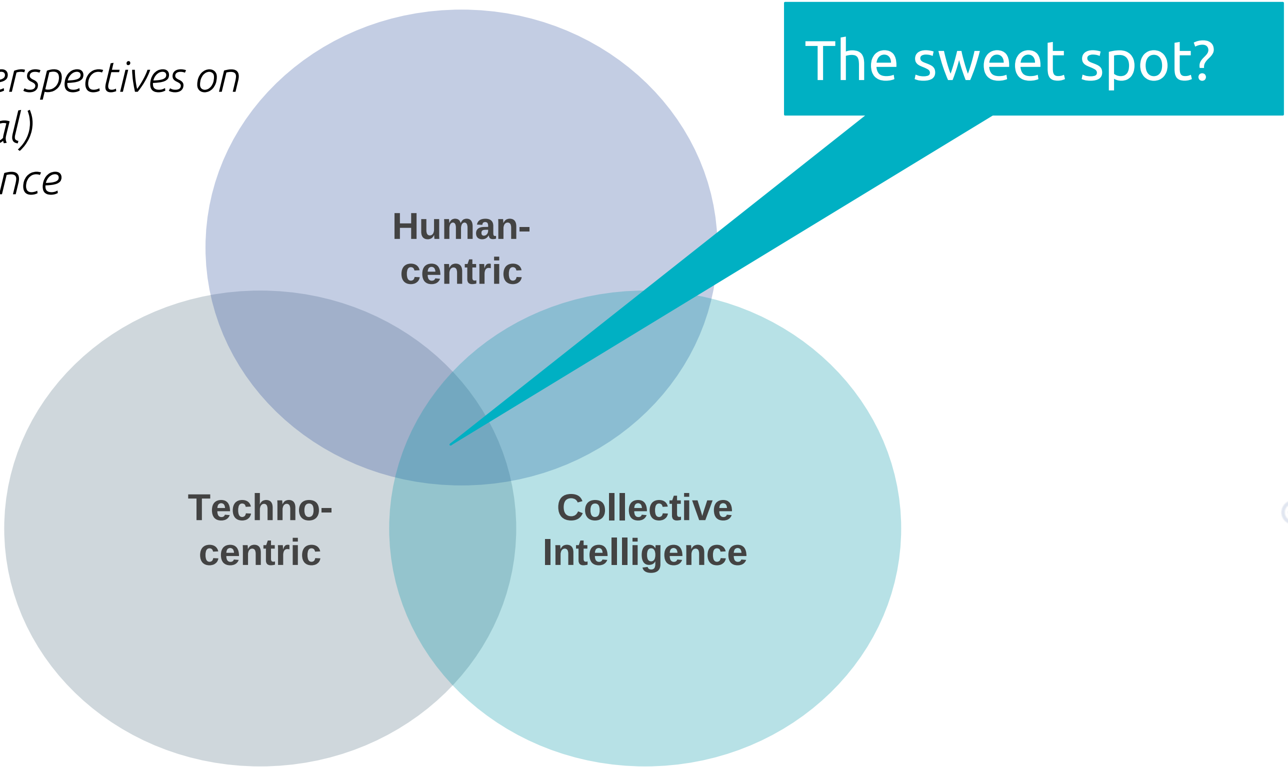
E. Tools for sustainable energy use

2. Responsible AI - design method and approach

Towards responsible AI

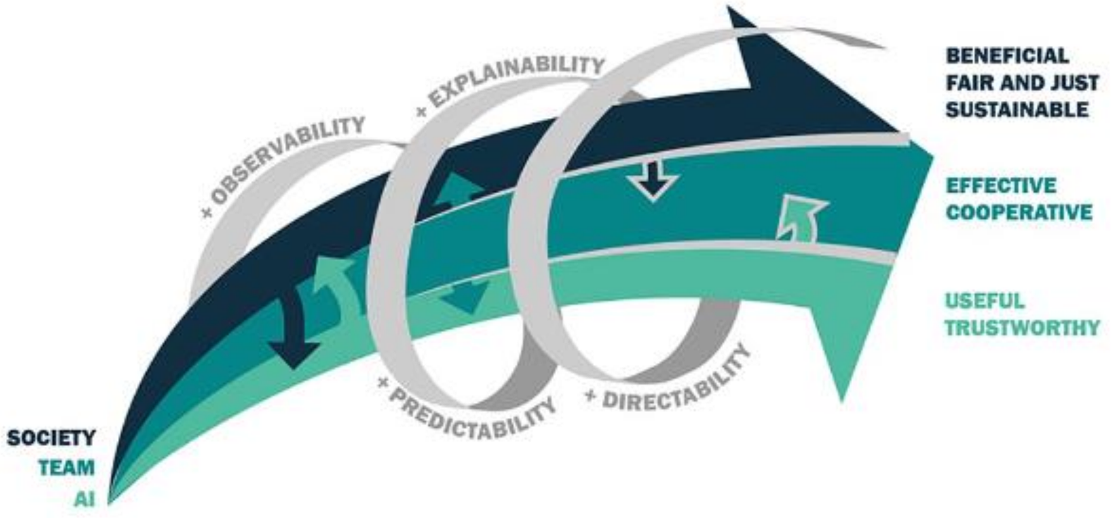
True intelligence is found in...?

Three perspectives on (Artificial) Intelligence



Original publication: Peeters, M. M. M., van Diggelen, J., Van Den Bosch, K., Bronkhorst, A., Neerincx, M. A., Schraagen, J. M., & Raaijmakers, S. (2020). Hybrid collective intelligence in a human–AI society. *AI & SOCIETY*, 1-22.

Towards the design and development of responsible AI

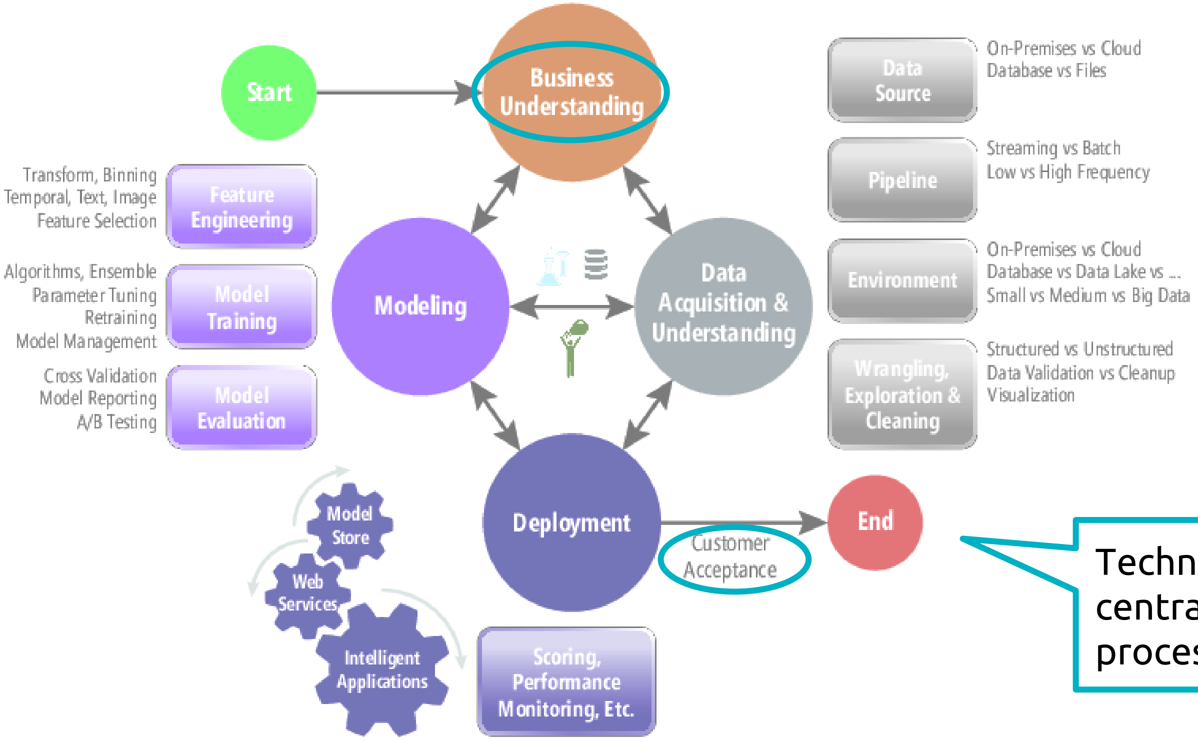


Include additional KPIs for ML/AI solutions. Make these *measurable*.

Original publication: Peeters, M. M. M., van Diggelen, J., Van Den Bosch, K., Bronkhorst, A., Neerincx, M. A., Schraagen, J. M., & Raaijmakers, S. (2020). Hybrid collective intelligence in a human–AI society. *AI & SOCIETY*, 1-22.

Three design methods

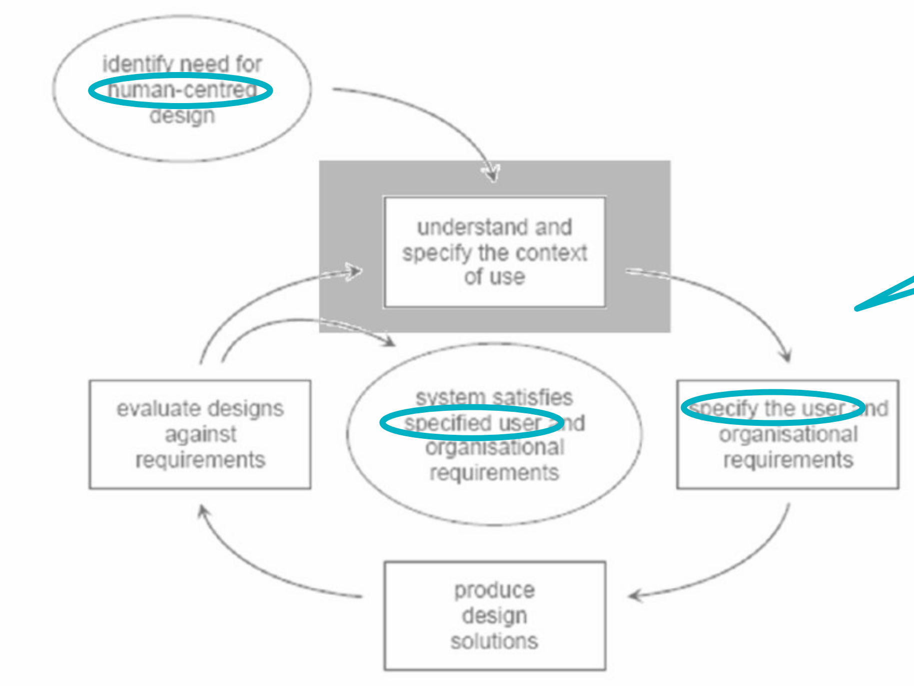
When and how are humans involved in the design of AI?



Technology is central to design process

Example of a techno-centric design framework

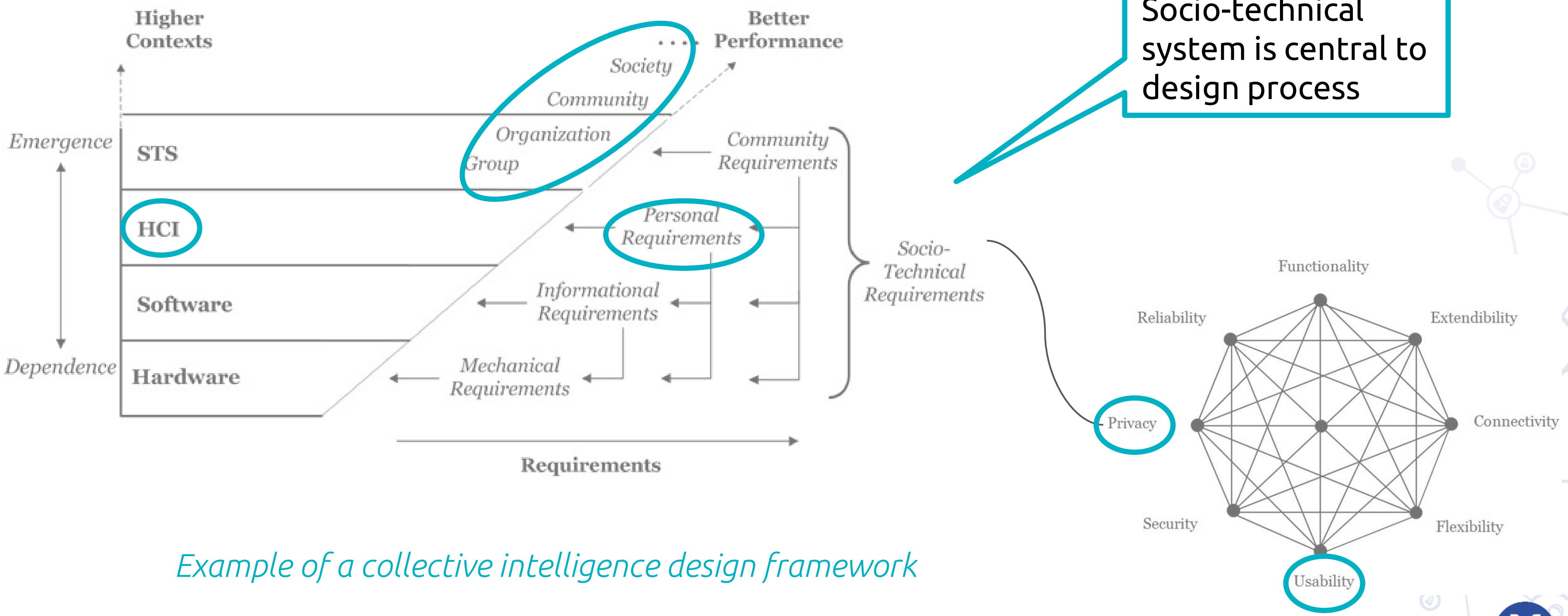
When and how are humans involved in the design of AI?



Human is central to design process

Example of a human-centric design framework

When and how are humans involved in the design of AI?



Socio-technical system is central to design process

Example of a collective intelligence design framework

Role-specific Responsible AI

Data Engineer

- Add human-in-the-loop feedback loops
- Security & privacy by design
- Increase sustainability
- Avoid technical bias
- Ensure reproducibility

Data Scientist

- SHAP
- LIME
- Model Cards
- Fairlearn / ML-Fairness-gym
- Intuitive Confidence Measure
- What-if Tool
- Glass box approaches
- Contrastive Explanations
- InterpretML

Analytics Translator

- Finding “the right” datasets
- Ethical toolkit for AI
- Value-sensitive design
- Judgment call the game