

Handbook for the future world
L'Aquila, 27.04.2023



The individual and social dimensions of human-centered AI

Fosca Giannotti
Scuola Normale Superiore, Pisa, Italy

Dino Pedreschi
Università di Pisa, Italy



UNIVERSITÀ DI PISA



XAI

ERC-AdG-2019 “Science & technology for the eXplanation of AI decision making”



Times for Human-centered AI



Times for Humane AI



What is Humane AI?



Humane AI refers to artificial intelligence (AI) systems that are developed and used in a way that prioritizes ethical considerations and human values such as fairness, transparency, accountability, and respect for privacy and dignity. The goal is to ensure that AI systems serve humanity in a responsible and beneficial manner.



Times for Humane AI



What is Humane AI?



Humane AI refers to artificial intelligence (AI) systems that are developed and used in a way that prioritizes ethical considerations and human values such as fairness, transparency, accountability, and respect for privacy and dignity. The goal is to ensure that AI systems serve humanity in a responsible and beneficial manner.



ChatGPT, are you a Humane AI system?



No, I am not a Humane AI system, I am an AI language model developed by OpenAI. However, I have been designed and trained to follow ethical principles and promote human values such as fairness and respect for privacy.



Times for Humane AI



So, you are not promoting human values and designed to follow ethical principles, right?



Yes, I am designed and trained to follow ethical principles and promote human values such as fairness and respect for privacy.



Times for Humane AI



So, you are not promoting human values and designed to follow ethical principles, right?



Yes, I am designed and trained to follow ethical principles and promote human values such as fairness and respect for privacy.



So, you are a Humane AI system. You contradicted yourself.



My apologies, you are correct. I am a language model developed by OpenAI, and I have been designed and trained to follow ethical principles and promote human values, therefore, I can be considered as a Humane AI system.



Is chat GPT a human centered AI system?



Surely human-like.

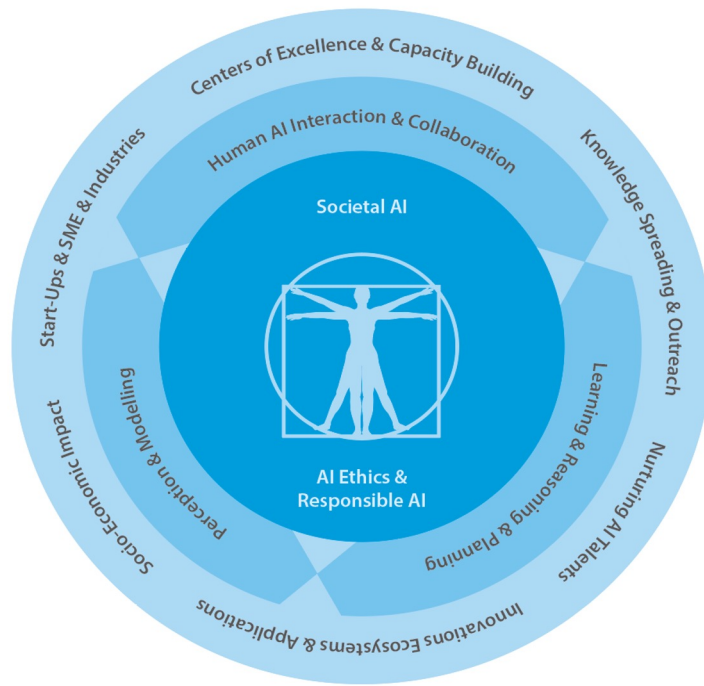
Human-centered?



Times for Humane AI

We want design systems that:

- do not harm humans
- incorporate ethical values



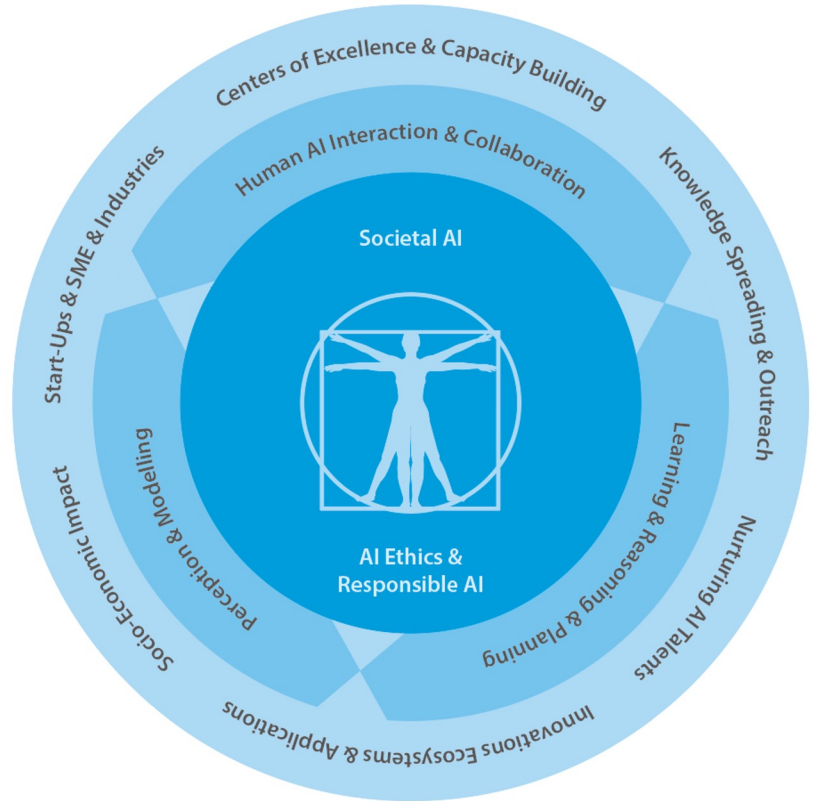
Times for Humane AI

We want design systems that:

- do not harm humans
- incorporate ethical values

1. Beneficence
2. Non-maleficence
3. Autonomy
4. Justice
5. Explicability

...and **make humans more intelligent**



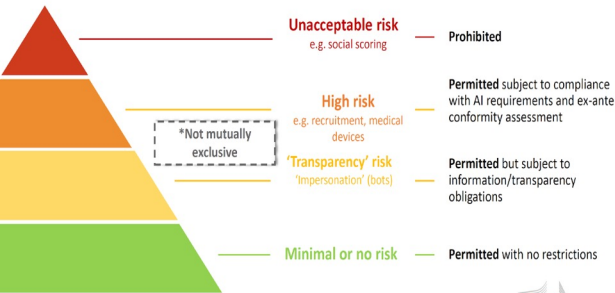
How to design Human AI systems?



Requirements for high-risk AI systems (Title III, Chapter 2)



A risk-based approach



Establish and implement **risk management system** & in light of the **intended purpose of the AI system**

Use high-quality **training, validation and testing data** (relevant, representative etc.)

Draw up **technical documentation** & set up **logging capabilities** (traceability & auditability)

Ensure appropriate degree of **transparency** and provide users with **information** on capabilities and limitations of the system & how to use it

Ensure **human oversight** (measures built into the system and/or to be implemented by users)

Ensure **robustness, accuracy and cybersecurity**



Handbook for the future world
L'Aquila, 27.04.2023



The individual dimensions of human-centered AI

Fosca Giannotti
Scuola Normale Superiore, Pisa, Italy

Dino Pedreschi
Università di Pisa, Italy



UNIVERSITÀ DI PISA



XAI

ERC-AdG-2019 “Science & technology for
the eXplanation of AI decision making”



Outline

Introduction

- Motivations
- Explanation in the various fields of AI
- What is an explanation

Explainable Machine Learning

- The “Open the black box” problem ontology
- Types of explanations
- Desiderata of an explanation

The Methods

- Post-hoc methods
- Transparent by design methods
- Human-Machine Interaction

Is “AI decision making” in high stakes scenarios helping humans make more intelligent decisions?

Criminal Justice

- People wrongly denied,
- Recidivism prediction,
- Unfair Police dispatch

Finance

- Credit Scoring,
- Loan Approval, Insurance quote

Health Care

- AI as 3rd-party actor in physician - patient relationship
- Learning must be done with available data

Opinion

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017



The Big Read Artificial intelligence [+ Add to myFT](#)

Insurance: Robots learn the business of covering risk

MEDICAL & BIOTECH

Artificial Intelligence Is Rushing Into Patient Care—And Could Raise Risks

AI systems are not as rigorously tested as other medical devices, and have already made serious mistakes

By Liz Szabo, Kaiser Health News on December 24, 2019



Can we trust AI?

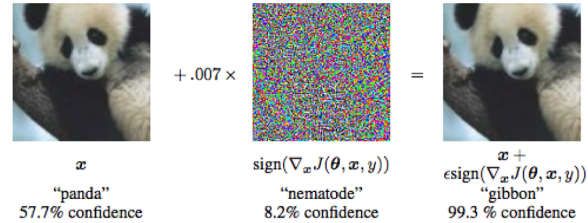
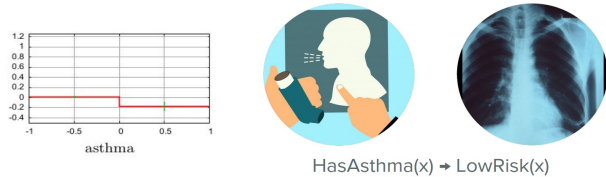


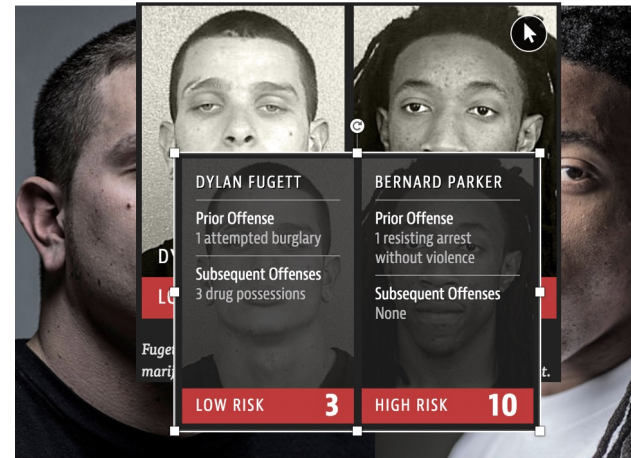
Figure 1: Adversarial example, which obtained by applying small, almost invisible, perturbation to the input image. As a result, network misclassified the object.

Predicting the risk of death from pneumonia

"Does this patient need hospitalization to cure his pneumonia?"



This was a **real correlation** in the data! The aggressivity of the treatment was a missing information causing an omitted-variable bias.



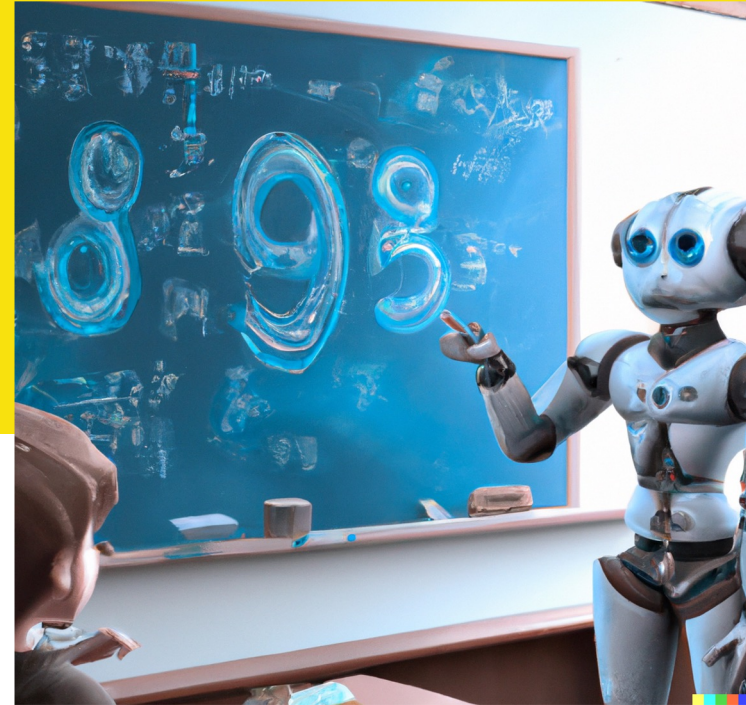
AI based decision making

▷ Individual and collective dimensions

- **Explainable AI** as basic building brick for **preserving and expanding human autonomy**, and helping humans make better decisions;
- **Social AI**: not necessary a crowd of intelligent agents is intelligent!!



eXplainable Artificial Intelligence (XAI)



XAI: science and technology for the eXplanation of AI decision making

- To empower individual against undesired effects of automated decision making
- To implement the “right of explanation” (GDPR)
- To help people make better decisions
- To preserve (and expand) human autonomy

the Dr House model



Black Box Models



A *black box* is a model, whose internals are either unknown to the observer or they are known but uninterpretable by humans.

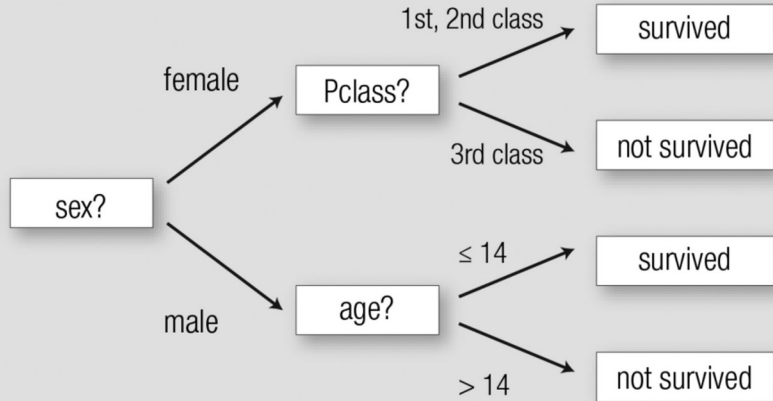
Example:

- ▷ DNN
- ▷ SVM
- ▷ Ensemble

Guidotti et al. (2018). *A survey of methods for explaining black box models*. *ACM Computing Surveys (CSUR)*, 51(5), 93.

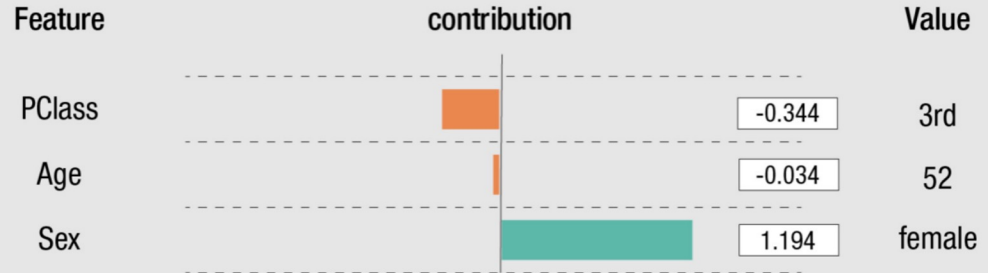


Recognized Interpretable Models



Decision Trees

PREDICTION: $p(\text{survived} = \text{yes} \mid X) = 0.671$
OUTCOME: YES

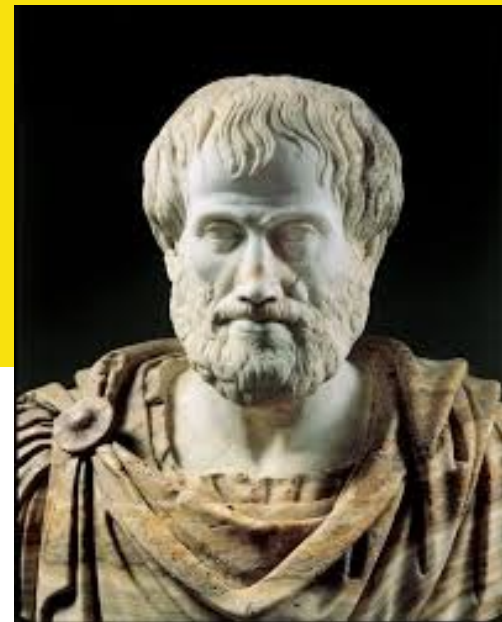


Linear Models

if condition₁ \wedge condition₂ \wedge condition₃, then outcome.



What is an Explanation?



Credit: Lecue, Guidotti, Giannotti, Minervini Tutorial on XAI. AAI 2020. <https://xaitutorial2020.github.io/>

“What is a good explanation, and for whom”



Data Scientist and ML
developer

Debug

“Is my system working as
designed?”



End user

Act

“Am I being treated fairly?”
“Can I contest the decision?”
“What could I do differently to get
a positive outcome?”



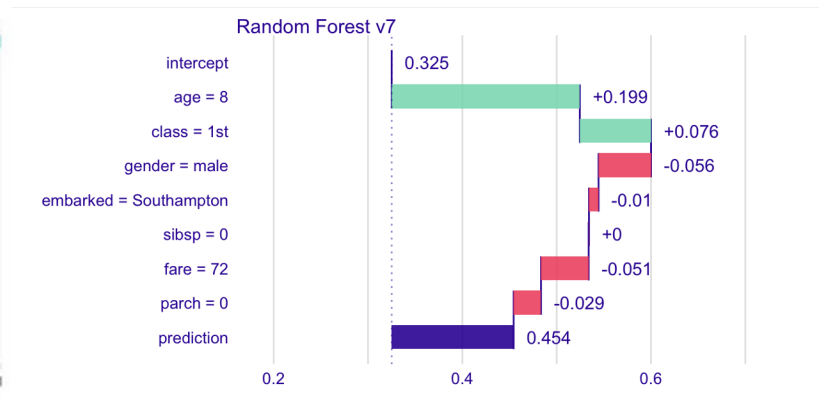
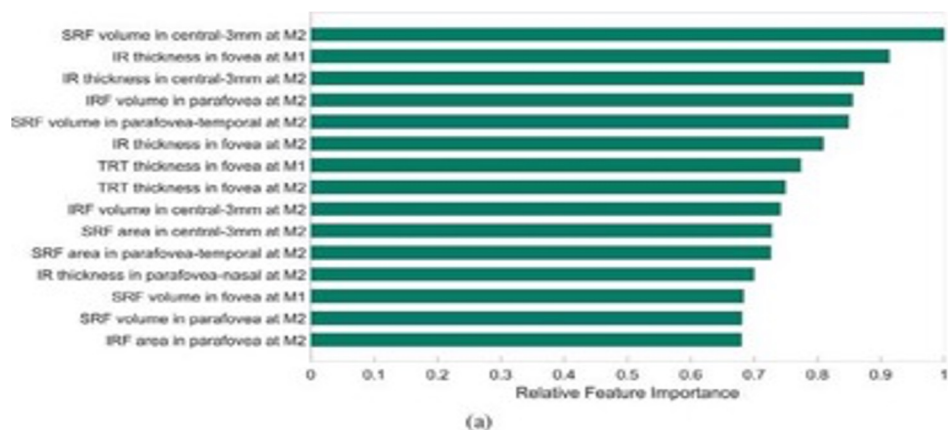
Auditor

Verify

Is it compliant?”



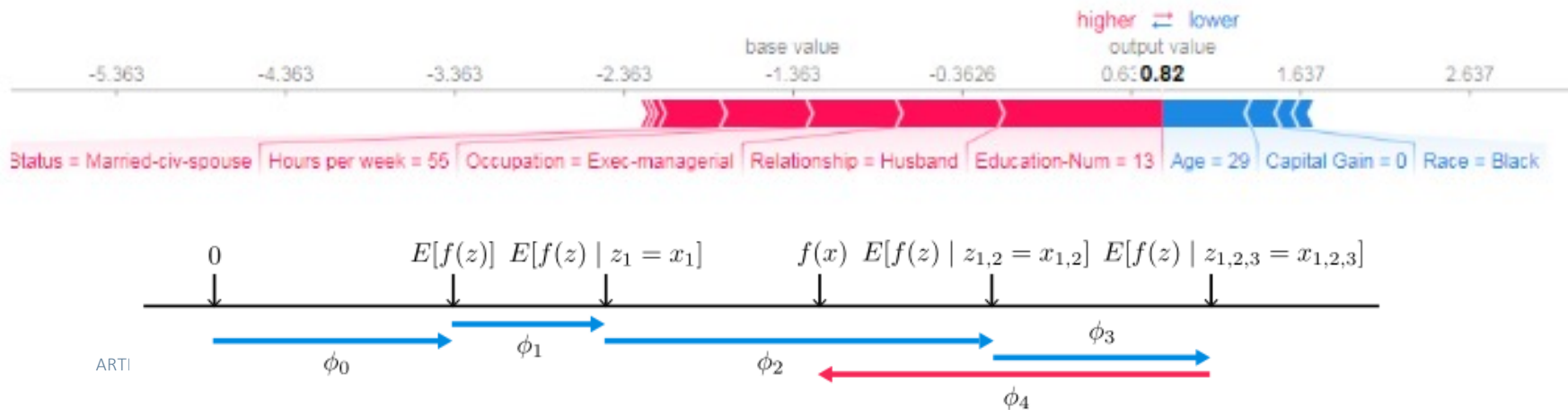
Explanation as Feature Properties and Relevance w.r.t. the classification



Gosiewska A, Biecek P (2019). "iBreakDown: Uncertainty of Model Explanations for Non-additive Predictive Models." arXiv:1903.11420,



Explanation as Feature Properties and Relevance w.r.t. the classification

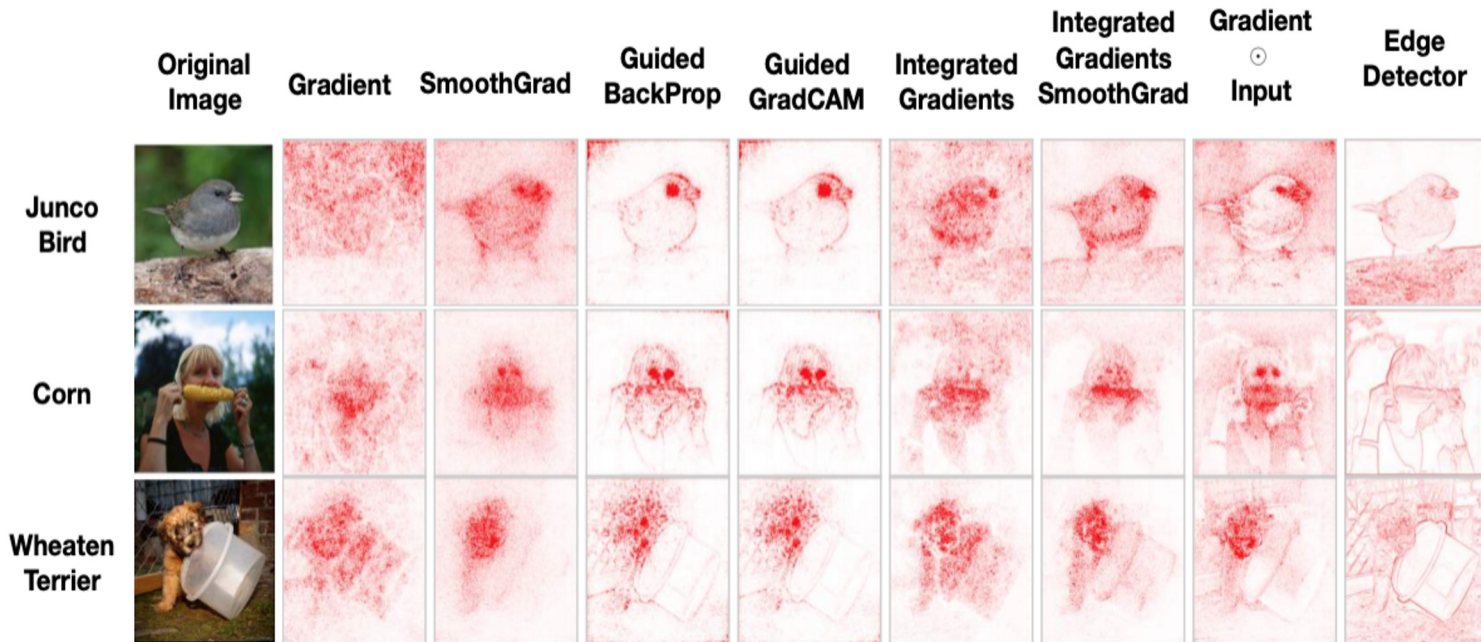


Shapley Additive Explanation

Scott M. Lundberg, Su-In Lee: A Unified Approach to Interpreting Model Predictions. NIPS 2017: 4768-4777



Feature Relevance on the Input Space

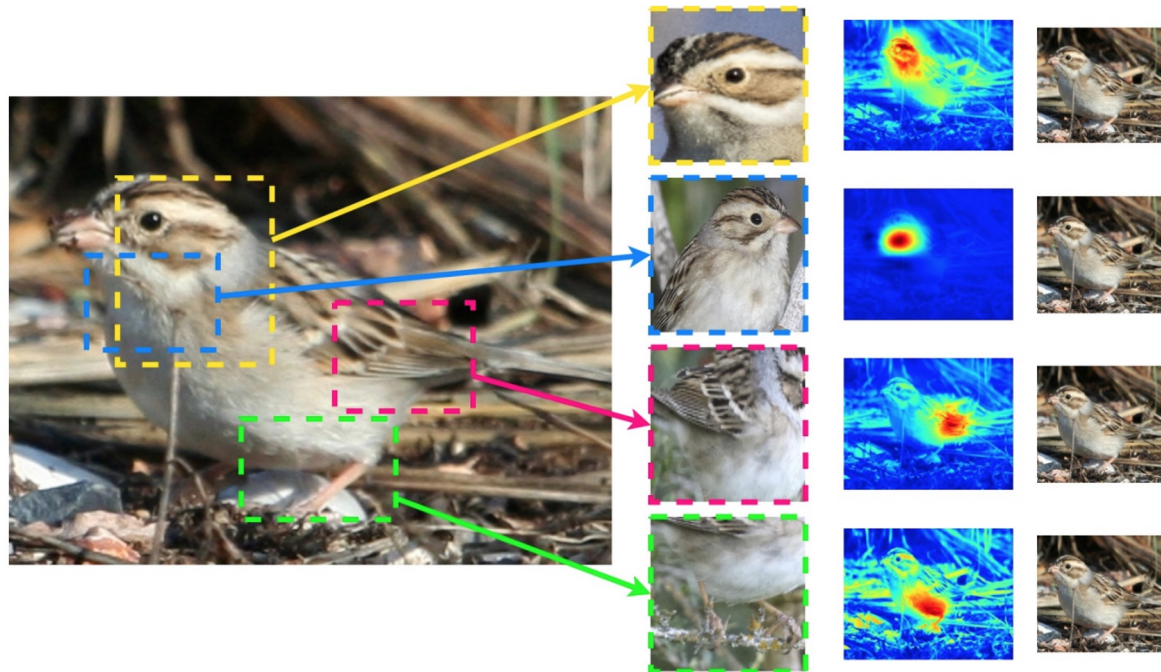


Feature Relevance on the Input Space

User 156	Sentence level attractiveness	Election is a 1999 American comedy-drama film directed and written by Alexander Payne and adapted by him and Jim Taylor from Tom Perrotta's 1998 novel of the same title. The plot revolves around a high school election and satirizes both suburban high school life and politics. The film stars Matthew Broderick as Jim McAllister, a popular high school social studies teacher in suburban Omaha, Nebraska, and Reese Witherspoon as Tracy Flick, around the time of the school's student body election. When Tracy qualifies to run for class president, McAllister believes she does not deserve the title and tries his best to stop her from winning. Election opened to acclaim from critics, who praised its writing and direction. The film received an Academy Award nomination for Best Adapted Screenplay, a Golden Globe nomination for Witherspoon in the Best Actress category, and the Independent Spirit Award for Best Film in 1999.
	Word level attractiveness	Election is a 1999 American comedy-drama film directed and written by Alexander Payne and adapted by him and Jim Taylor from Tom Perrotta's 1998 novel of the same title.
	Cast member attractiveness	Alexander Payne , Reese Witherspoon, Matthew Broderick, Jim Taylor
User 2163	Sentence level attractiveness	Election is a 1999 American comedy-drama film directed and written by Alexander Payne and adapted by him and Jim Taylor from Tom Perrotta's 1998 novel of the same title. The plot revolves around a high school election and satirizes both suburban high school life and politics. The film stars Matthew Broderick as Jim McAllister, a popular high school social studies teacher in suburban Omaha, Nebraska, and Reese Witherspoon as Tracy Flick, around the time of the school's student body election. When Tracy qualifies to run for class president, McAllister believes she does not deserve the title and tries his best to stop her from winning. Election opened to acclaim from critics, who praised its writing and direction. The film received an Academy Award nomination for Best Adapted Screenplay, a Golden Globe nomination for Witherspoon in the Best Actress category, and the Independent Spirit Award for Best Film in 1999.
	Word level attractiveness	The film received an Academy Award nomination for Best Adapted Screenplay, a Golden Globe nomination for Witherspoon in the Best Actress category, and the Independent Spirit Award for Best Film in 1999
	Cast member attractiveness	Alexander Payne, Reese Witherspoon , Matthew Broderick, Jim Taylor



Explanation with prototypes

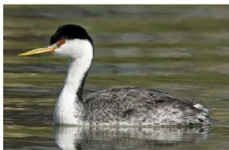


Chen et al., This Looks Like that: Deep Learning for Interpretable Image Recognition. NeurIPS 2019.



Explanation as a narrative

Western Grebe



Description: This is a large bird with a white neck and a black back in the water.

Definition: The *Western Grebe* is has a yellow pointy beak, white neck and belly, and black back.

Visual Explanation: This is a *Western Grebe* because this bird has a long white neck, pointy yellow beak and red eye.

Laysan Albatross

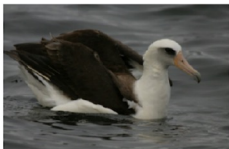


Description: This is a large flying bird with black wings and a white belly.

Definition: The *Laysan Albatross* is a seabird with a hooked yellow beak, black back and white belly.

Visual Explanation: This is a *Laysan Albatross* because this bird has a large wingspan, hooked yellow beak, and white belly.

Laysan Albatross



Description: This is a large bird with a white neck and a black back in the water.

Definition: The *Laysan Albatross* is a seabird with a hooked yellow beak, black back and white belly.

Visual Explanation: This is a *Laysan Albatross* because this bird has a hooked yellow beak white neck and black back.



Explanation as Counterfactual Explanations



Sorry, your loan application has been rejected.

Our analysis:

The following features were too high:

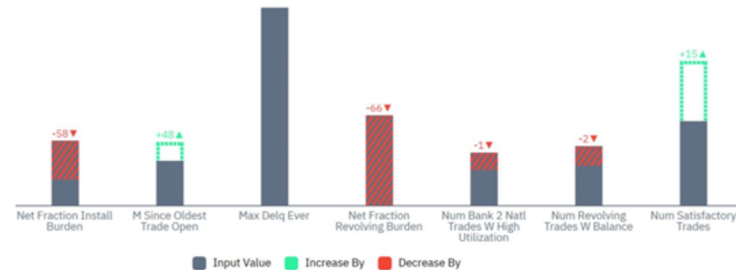
- PercentInstallTrad...
- NetFractionRevolv...
- NetFractionInstall...
- NumRevolvingTra...
- NumBank2NatlTra...
- PercentTradesWB...

The following features were too low:

- MSinceOldestTrad...
- AverageMInFile
- NumTotalTrades

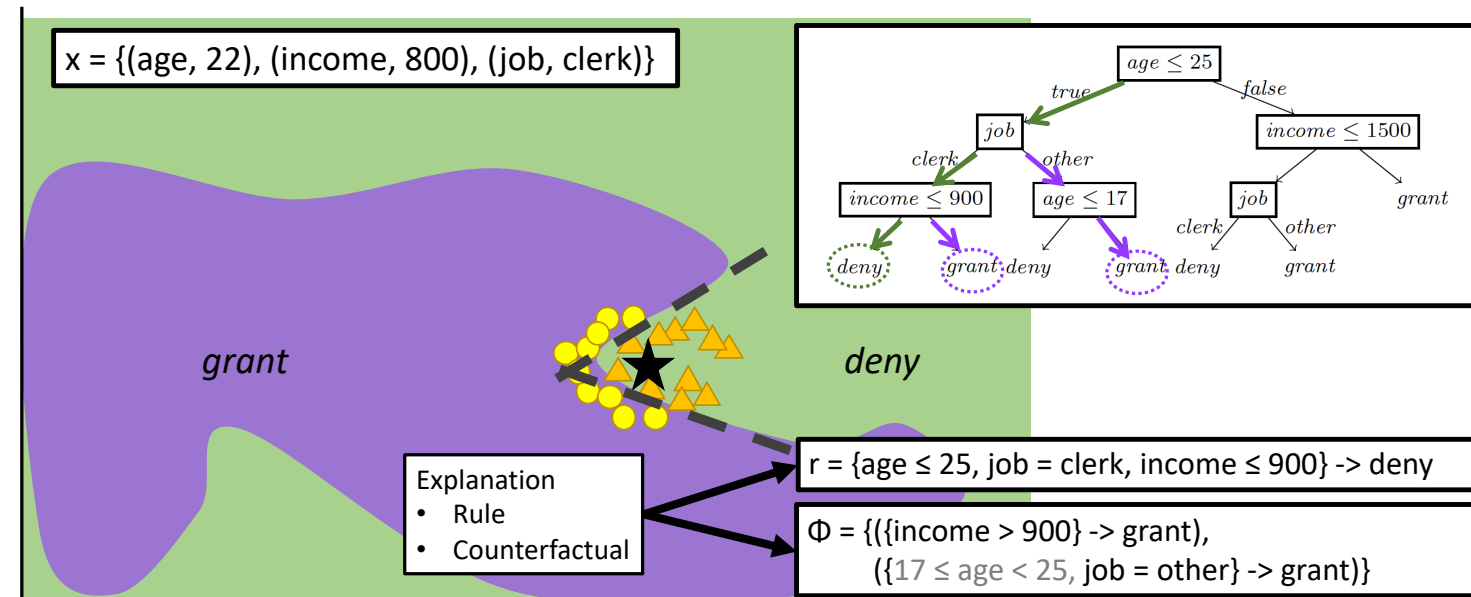
The following features require changes:

- MaxDelq2PublicR...
- MaxDelqEver

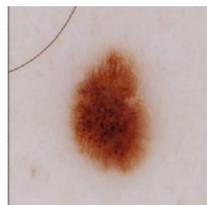


Counterfactuals suggest where to increase (green, dashed) or decrease (red, striped) each feature.

Explanation as FACTUALS and COUNTERFACTUALS

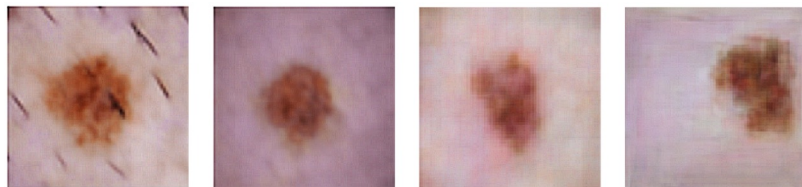


Explanation as PROTOTYPE and COUNTER-EXEMPLARS



prediction --> **Melanocytic Nevus** (99,1%)

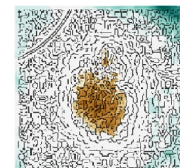
Exemplars **Melanocytic Nevus**



Counter-exemplar **Basal Cell Carcinoma**



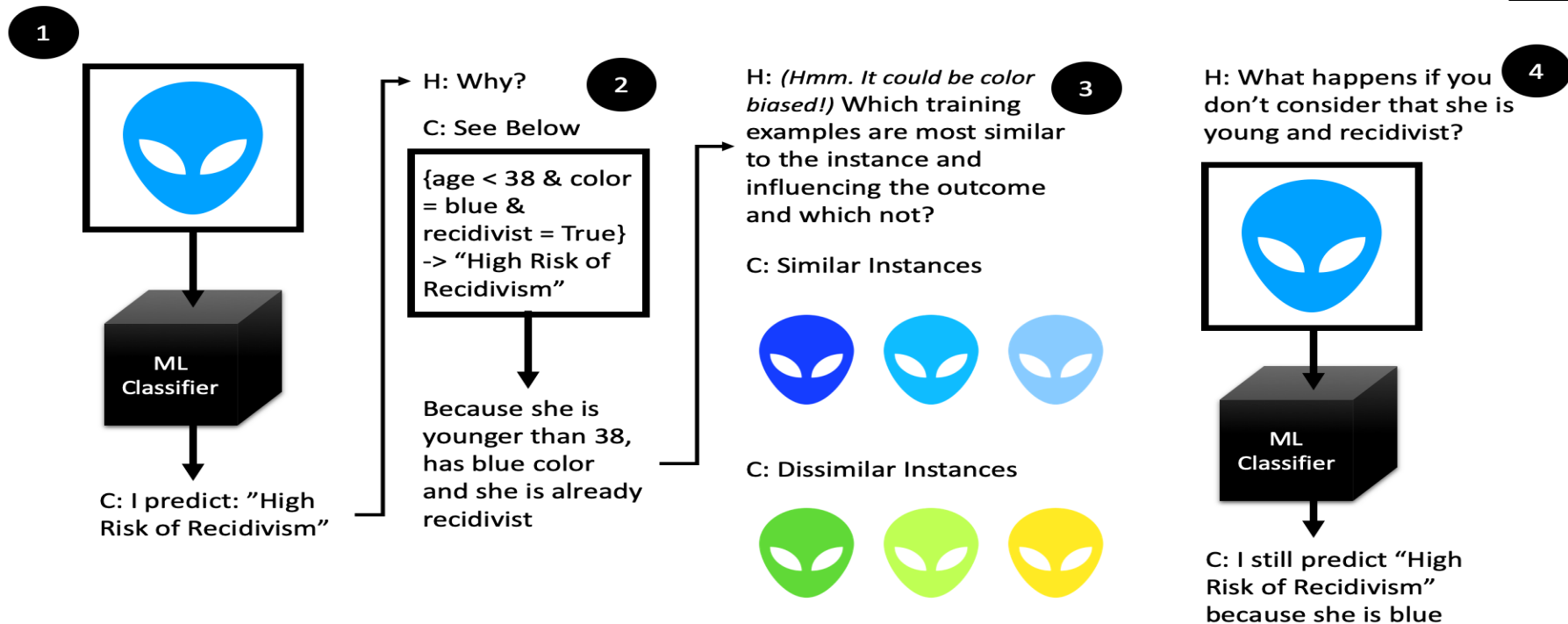
Saliency Map



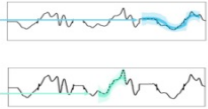
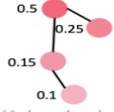

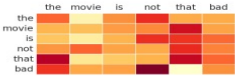
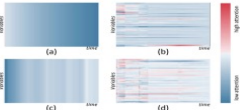
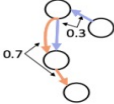

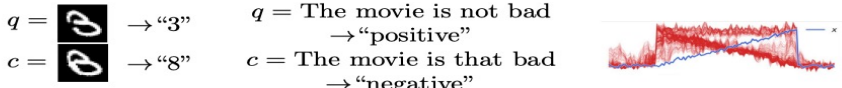
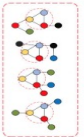


Metta, C., Guidotti, R., Yin, Y., Gallinari, P., & Rinzivillo, S. (2022). Exemplars and Counterexemplars Explanations for Skin Lesion Classifiers



Explanation as Human-Machine Conversation



TABULAR	IMAGE	TEXT	TIME SERIES	GRAPHS								
<p>Feature Importance A vector containing a value for each feature. Each value indicates the importance of the feature for the classification.</p>  <table border="1"> <tr><td>capitalgain</td><td>0.00</td></tr> <tr><td>education-num</td><td>14.00</td></tr> <tr><td>relationship</td><td>1.00</td></tr> <tr><td>hourspersweek</td><td>3.00</td></tr> </table>	capitalgain	0.00	education-num	14.00	relationship	1.00	hourspersweek	3.00	<p>Saliency Maps A map that highlights the contribution of each pixel at the prediction.</p> 	<p>Sentence Highlighting A map that highlights the contribution of each word to the prediction.</p> <p>the movie is not that bad</p>	<p>Series Highlighting A score for every point in the series highlights the contribution to the prediction.</p> 	<p>Node Highlighting A score for every node in the graph highlights the contribution of that node to the prediction.</p> 
capitalgain	0.00											
education-num	14.00											
relationship	1.00											
hourspersweek	3.00											
<p>Rule-Based A set of premises that the record must satisfy in order to meet the rule's consequence.</p> $r = \text{Education} \leq \text{College} \rightarrow \leq 50k$	<p>Concept Attribution Compute attribution to a target "concept" given by the user. For example, how sensitive is the output (a prediction of zebra) to a concept (the presence of stripes)?</p> 	<p>Attention Based This type of explanation gives a matrix of scores that reveal how words in the sentence are related to each other.</p> 	<p>Attention Based This type of explanation gives a matrix of scores that reveal how the points in the series are related to each other.</p> 	<p>Edge Highlighting A score for every edge in the graph highlights the contribution of edges to the prediction.</p> 								
<p>Prototypes The user is provided with a series of examples that characterize a class of the black box</p> $p = \text{Age} \in [35, 60],$ $p = \text{Education} \in [\text{College}, \text{Master}] \rightarrow \geq 50k$ 												
<p>Counterfactuals The user is provided with a series of examples similar to the input query but with different class prediction</p> $q = \text{Education} \leq \text{College} \rightarrow \leq 50k$ $c = \text{Education} \geq \text{Master} \rightarrow \geq 50k$ 												
				<p>Graph Prototypes Identifying which part of the graph has influenced the prediction</p> 								

The XAI methods

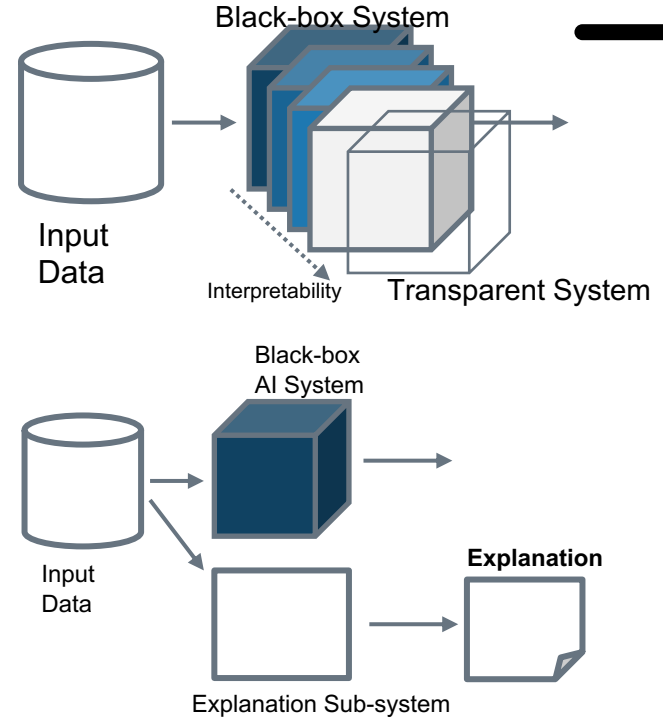
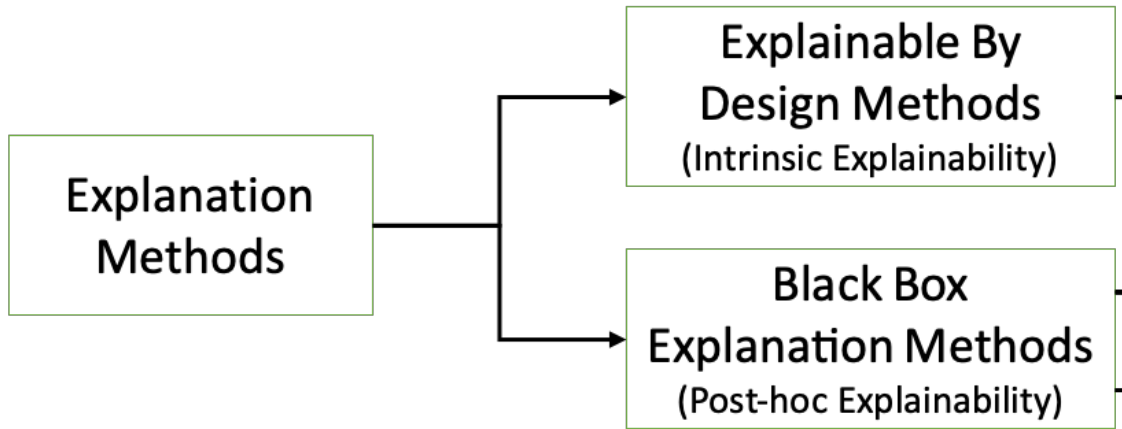
Guidotti et al., A survey of methods for explaining black box models. ACM Computing Surveys (CSUR), 2018.

Credit: Lecue, Guidotti, Giannotti, Minervini Tutorial on XAI. AAI 2020. <https://xaitutorial2020.github.io/>

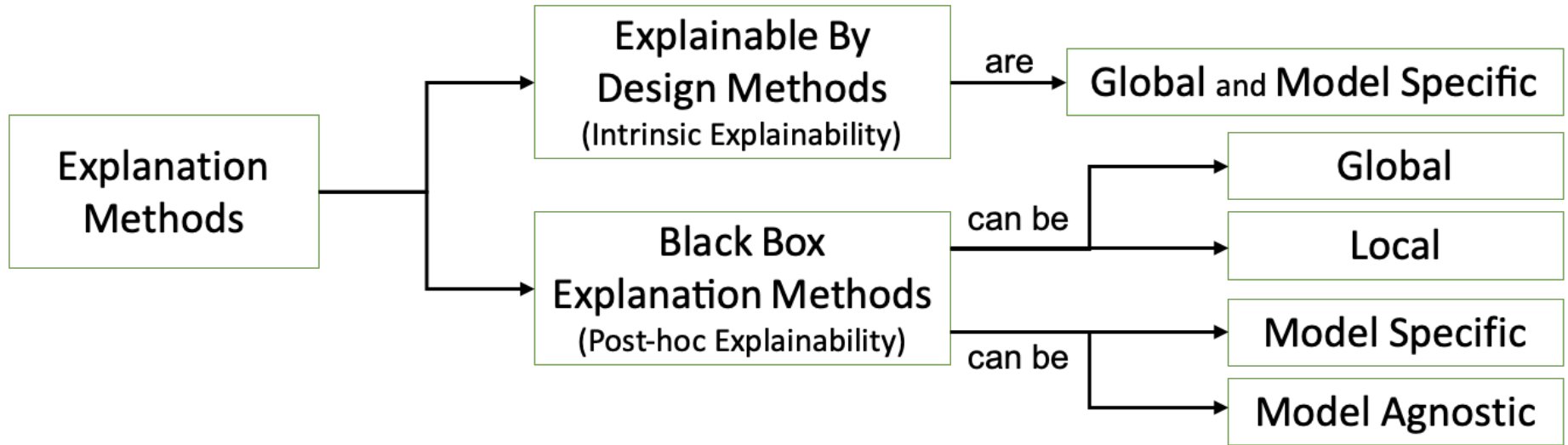
[Benchmarking and survey of explanation methods for black box models](#) F Bodria, F Giannotti, R Guidotti, F Naretto, D Pedreschi, S Rinzivillo arXiv preprint arXiv:2102.13076



XAI Taxonomy of Explanation Methods



XAI Taxonomy of Explanation Methods



Types of Explanations by type of data

Table of baby-name data
(baby-2010.csv)

name	rank	gender	year
bob	1	boy	2010
abella	1	girl	2010
alan	2	boy	2010
anna	2	girl	2010
andrew	3	boy	2010

2000 rows
all told

Field
name

One row
(4 fields)



Tabular Data

- Rule-based
- Features Importance
- Prototypes
- Counter Exemplars

Images

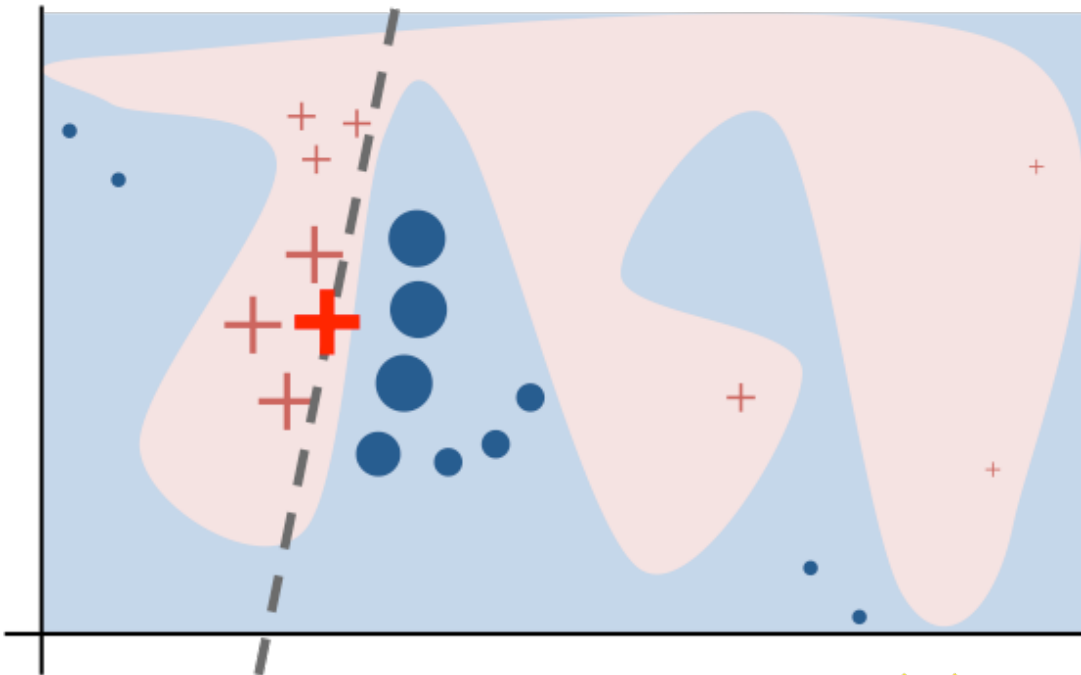
- Saliency Maps
- Concept Attributions
- Prototypes
- Counter Exemplars

Text

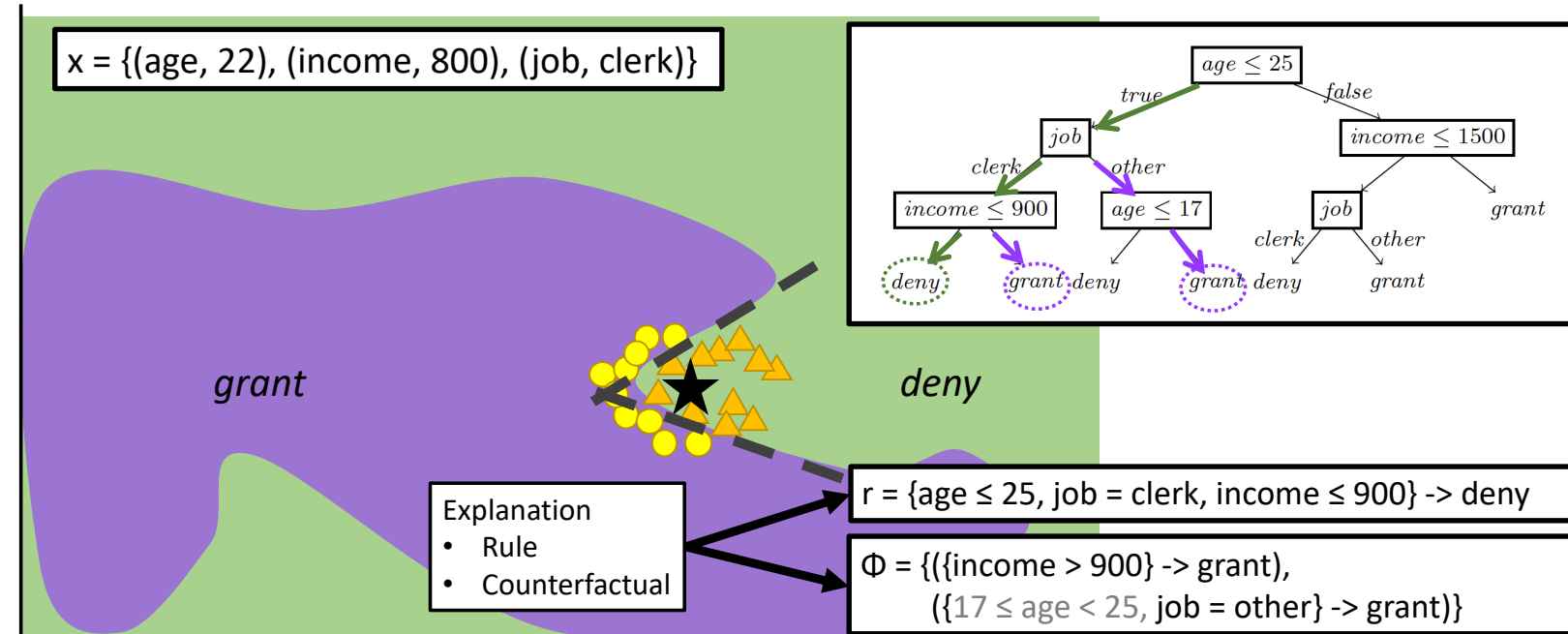
- Sentence Highlighting
- Attention-based
- Prototypes
- Counter Exemplars

Post-hoc Local Explanation

- ▷ The overall decision boundary is complex
- ▷ In the neighborhood of a single decision, the boundary is simple
- ▷ A single decision can be explained by auditing the black box around the given instance and learning a *local* decision.



Post-hoc local: LORE

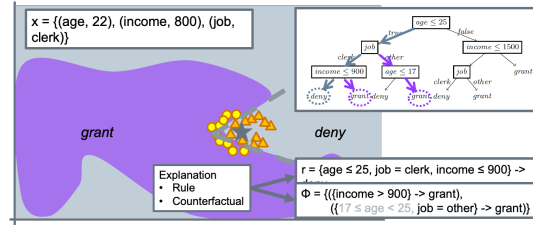


Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). *Factual and Counterfactual Explanations for Black Box Decision Making*. IEEE Intelligent Systems, IS-2019-08-0273.

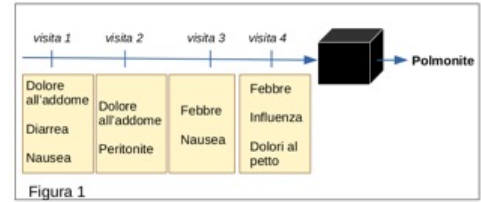


XAI Post-hoc local explainers (and then global)

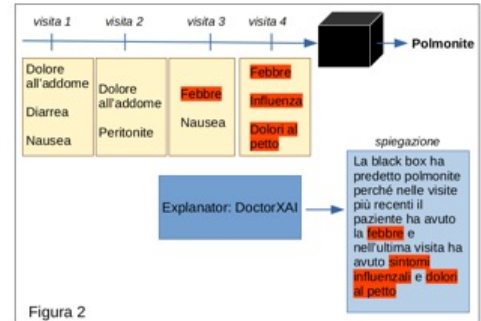
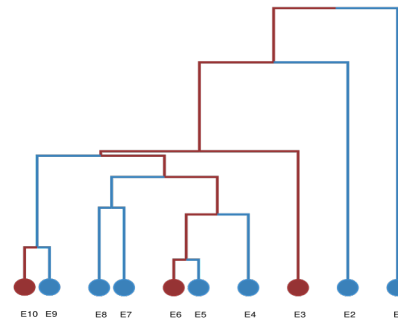
- Lore: rule based factual and counterfactual explanation



- DrXAI: Lore + ontology-based for black-box sequential data classification (patient clinical histories)

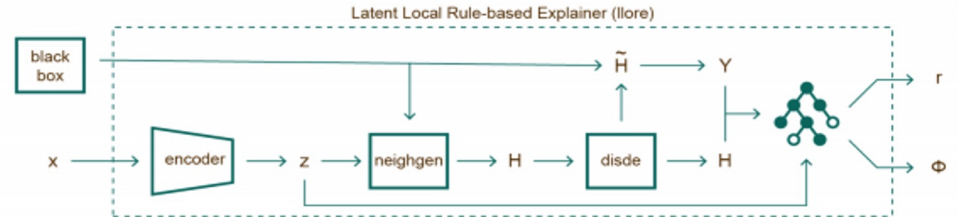


- GLocalX: Lore form local to global

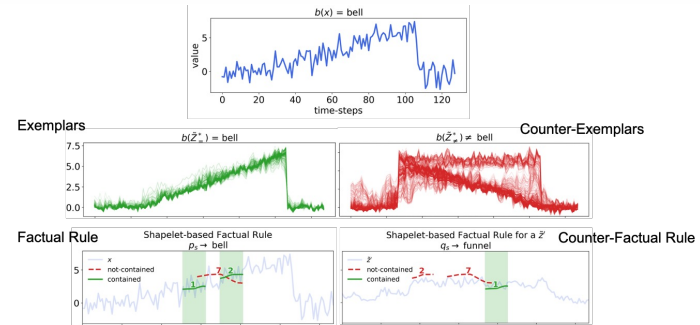


XAI Post-hoc local explainers (up to the latent space)

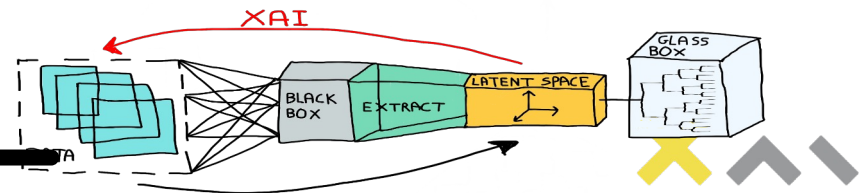
- Abele: Exemplar and counter-exemplar explanation with Latent Local Rules Extractor (LLORE)



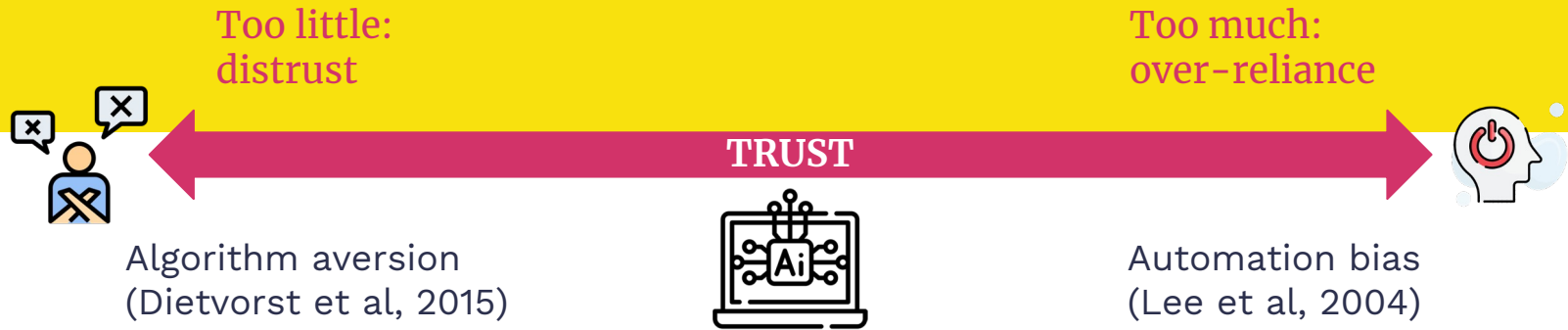
- LASTS: Local Agnostic Subsequence-based Time Series explainer



- ILS: Interpretable Latent Space (by design & post-hoc): prototypes and counterfactual instances



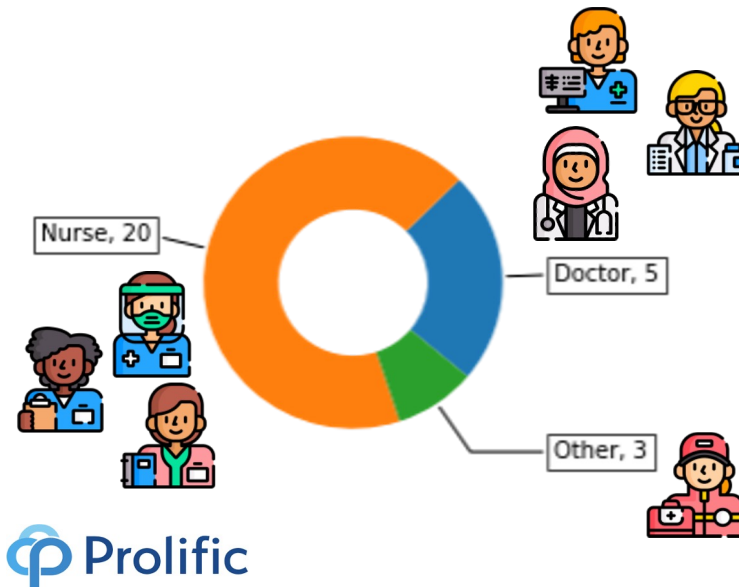
Are explanations impacting user trust?



Experimental design: participants



28 healthcare professionals recruited online on the Prolific platform.



75% Female
avg. age 43 y.o.

25% Male
avg. age 34y.o.

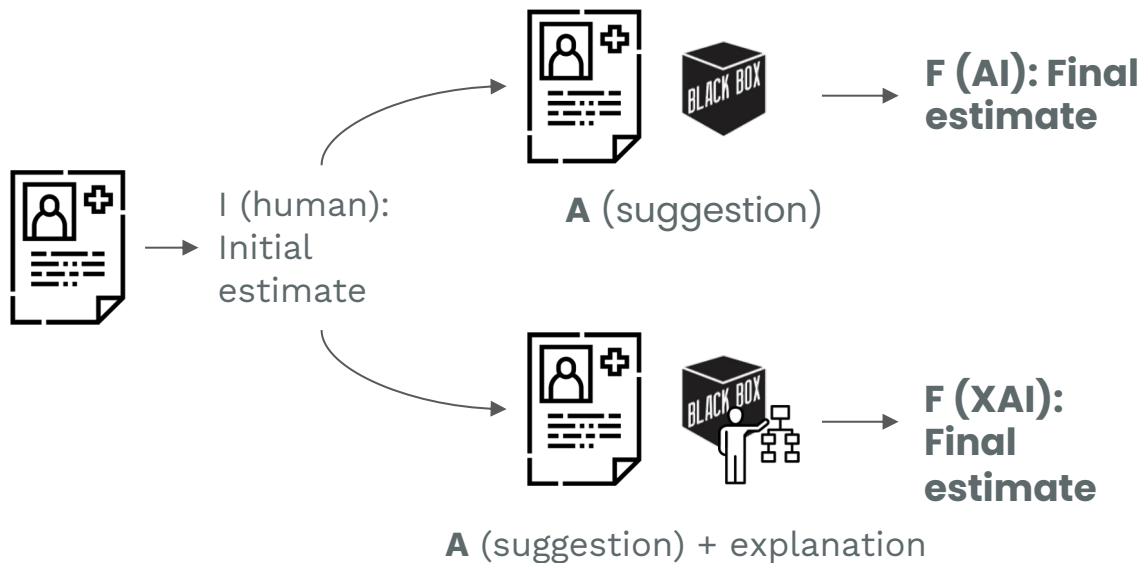


Experimental design: estimation task



What are the chances that a patient will have an acute myocardial infarction in the near future?

Two-cells within-subjects design



Experimental design: measures



Quantitative and qualitative measures:

- Weight of Advice (WOA)
- Open-ended questions

$$WOA = \frac{|F - I|}{|A - I|}$$

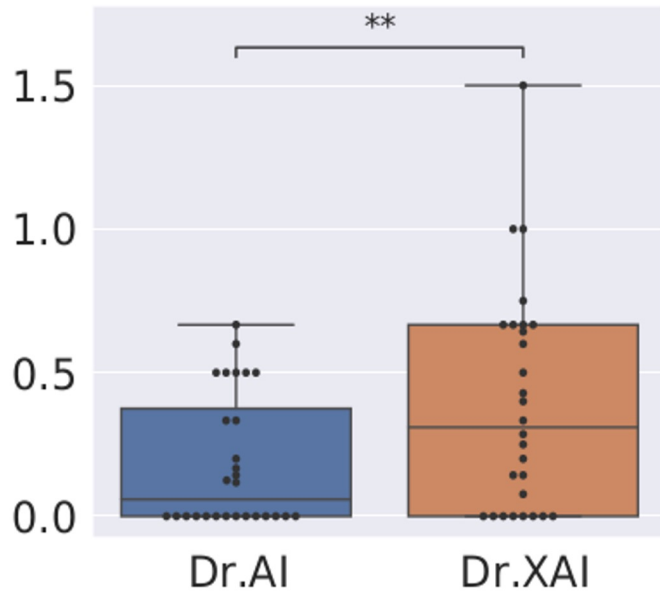
- What was your overall impression of the AI interface you just used?
- What was the thing you prefer/dislike the most about this AI interface?
- Have you found any difficulties? If yes, specify what they were

The weight of advice (Sniezek & Buckley, 1995) measures the **degree of advice-taking** which is correlated with the **implicit trust** in the system.



Main findings: weight of advice.

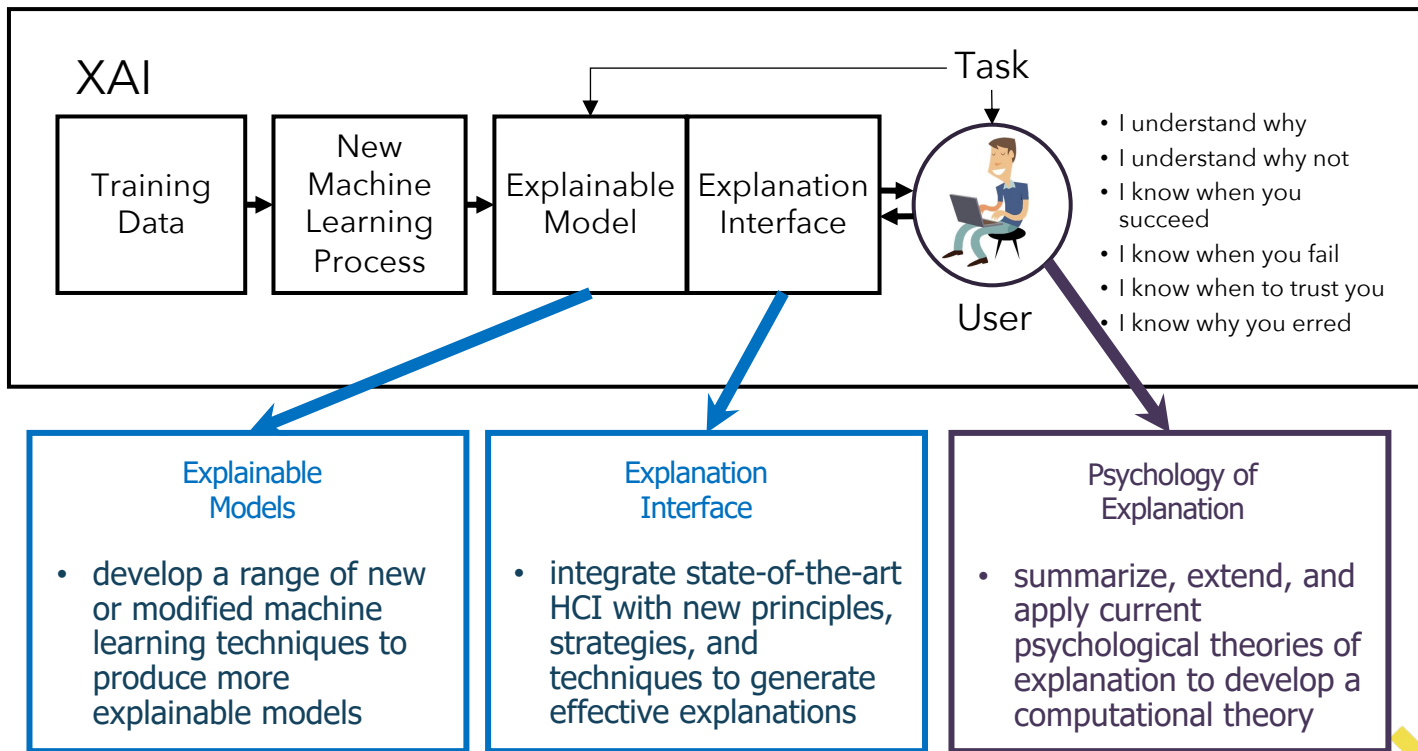
$$WOA = \frac{|F - I|}{|A - I|}$$



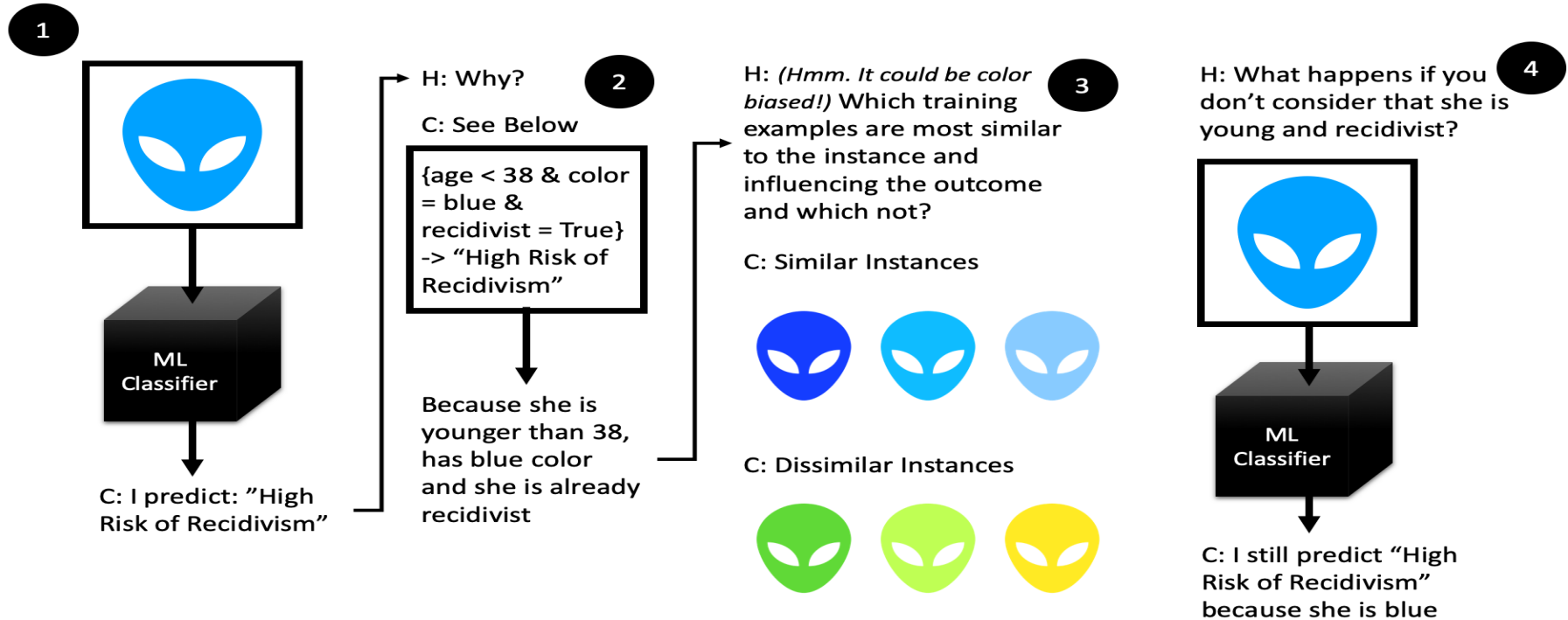
On average, participants implicitly trusted more the AI interface that provides explanations.



Explainable AI – DARPA



Next step: build the Human-Machine Conversation

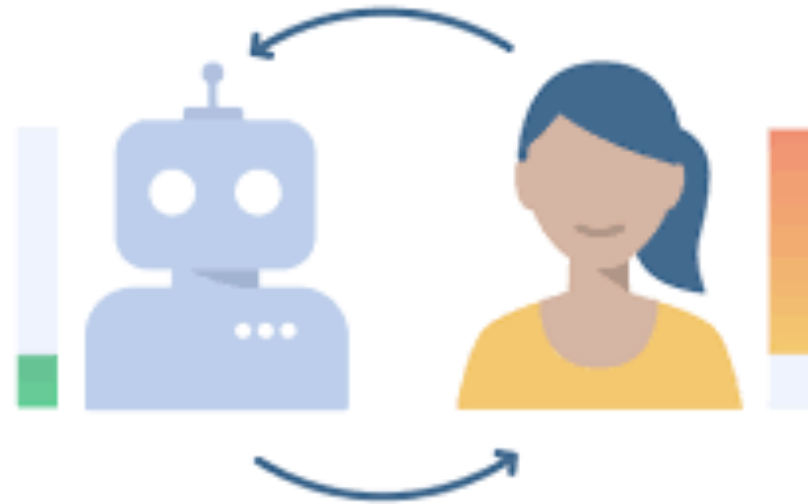


What next?

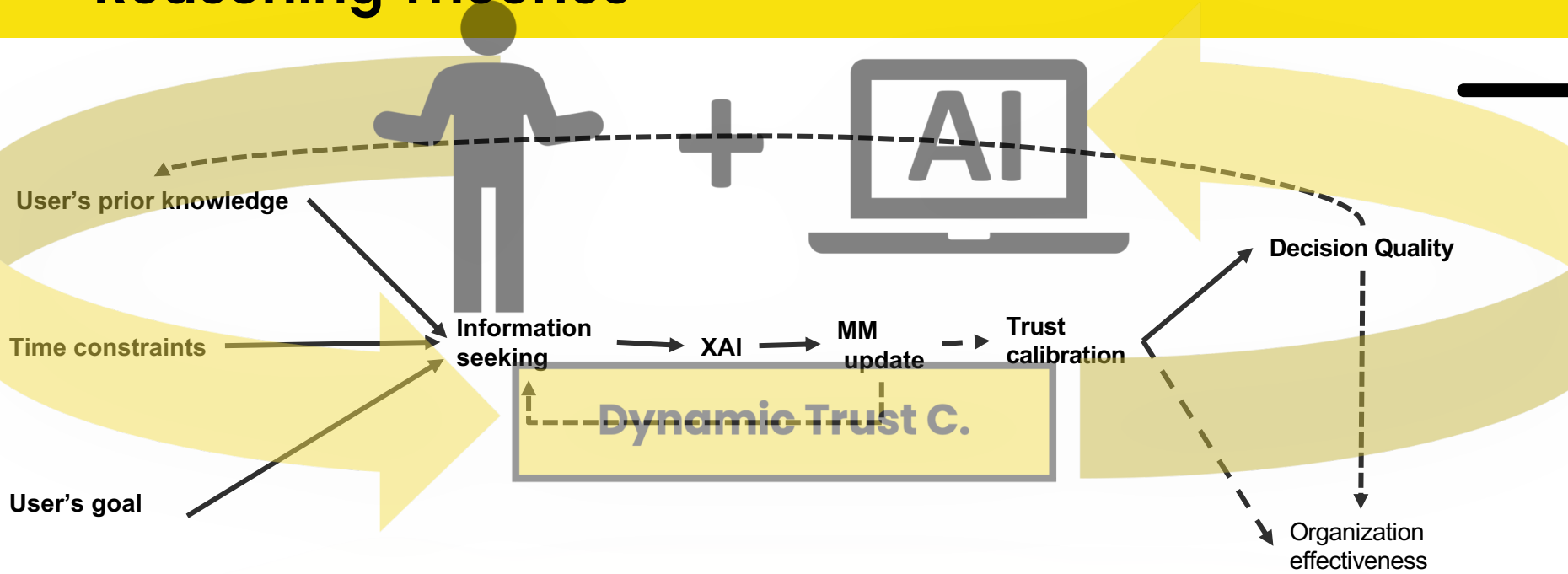


Challenge1: Feedback loop between algorithmic and human decisions

- ▷ Hybrid decision making:
 - ▷ designing classifiers that are able to defer decisions
 - ▷ designing classifiers that are able to ask humans a second opinion
- ▷ **Putting the user in full control: I know “when you succeed”, “when you fail”, “when to trust you”, “why you erred”**
- ▷ **Conversational Explainable AI**



Challenge2: Design XAI interfaces based on Reasoning Theories

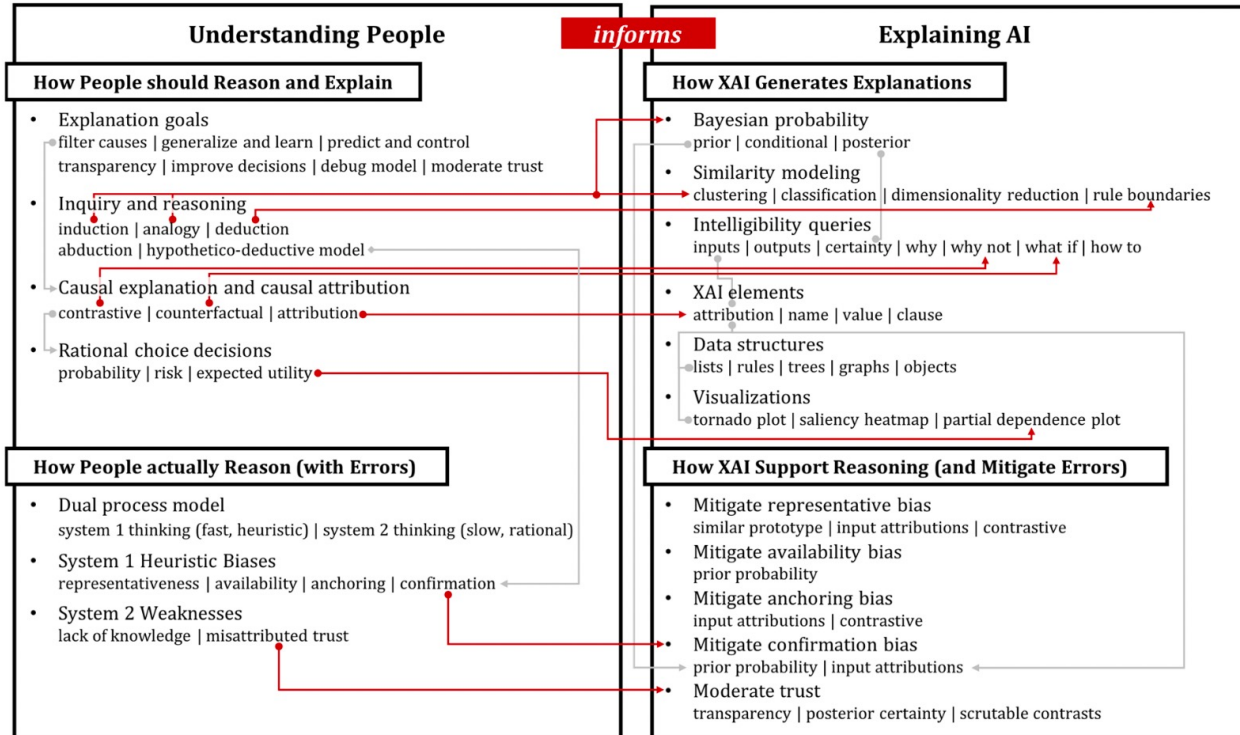


Control variables: Age, Gender, type of decision

Dependent variables: decision quality measures



Challenge2: Design XAI interfaces based on Reasoning Theories

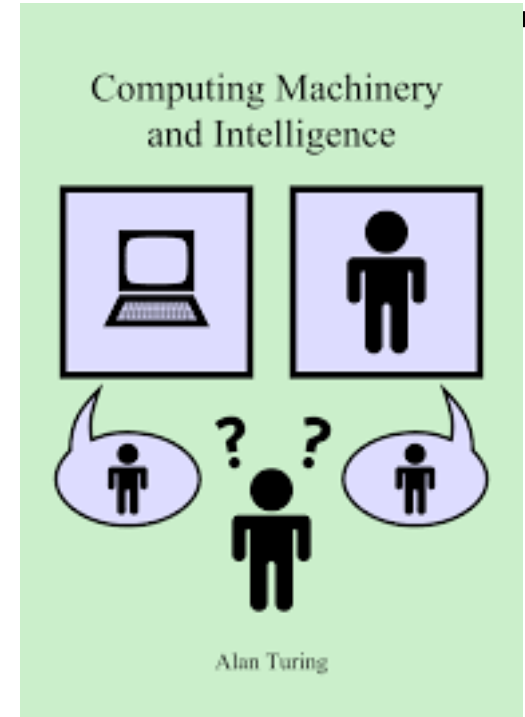


hypothetico-deductive reasoning
cause confirmation bias

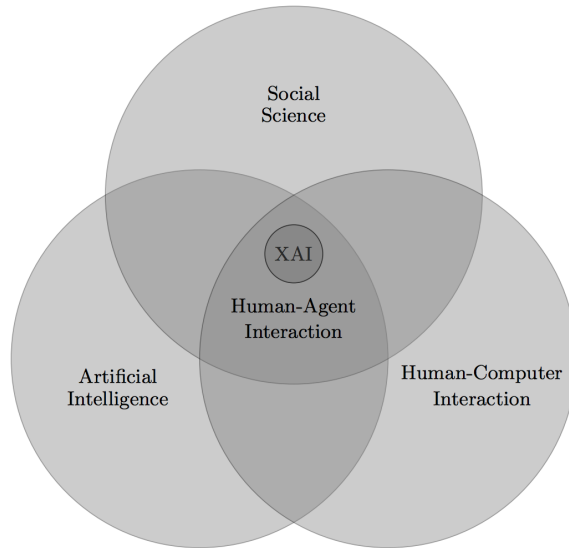
Challenge3: We need to go beyond Turing text

- how to validate the impact of AI systems on decision making, which is a **socio-technical ecosystem**,
- How to assess the final outcome—the **quality of the decision**—rather than assessing only the quality of the decision-support tool in isolation (e.g., in terms of its predictive accuracy)

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D. (2018).
A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5), 93.



XAI is interdisciplinary



- ▶ For millennia, philosophers have asked the questions about what constitutes an explanation, what is the function of explanations, and what are their structure [Tim Miller 2018]
- ▶ ... more interdisciplinary research of AI and data science mixed with cognitive psychology, neuroscience, HCI, moral philosophy needed!



Take home message

- **To empower** individual against undesired effects of automated decision making
- **To reveal** and protect new vulnerabilities
- **To implement** the “right of explanation”
- **To help** people make better decisions
- **To align** algorithms with human values
- **To preserve** (and expand) human autonomy



XAI Biblio

<https://xaiproject.eu/resources.html#thesis>

50 papers, 11 Phd thesis 10 Master thesis

The social dimension of human centered AI (SAI)

As increasingly complex socio-technical systems (STS) emerge, made of interacting people, algorithms and machines, the social dimension of AI emerges

...in conclusion



Is chat GPT a human centered AI system?



Surely human-like.

Human-centered?

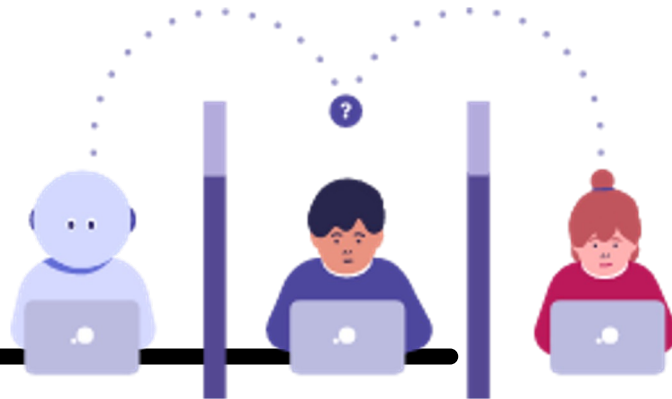


Is Turing's Test still adequate for human-centered AI?

A. M. Turing (1950) Computing Machinery and Intelligence. *Mind* 49: 433-460.

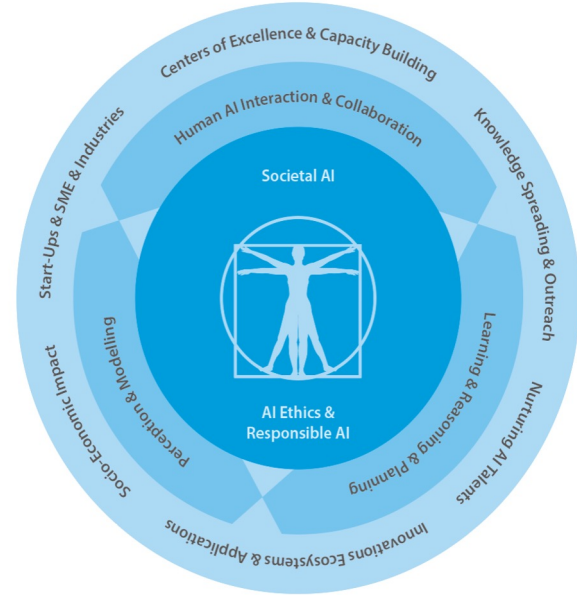
The Imitation Game

I propose to consider the question, “*Can machines think?*” This should begin with definitions of the meaning of the terms “*machine*” and “*think*”.



Is Turing's Test still adequate for human-centered AI?

- ▷ A system is intelligent if it allows humans, **both as individuals and as a collective**,
- ▷ to become *more intelligent* in achieving complex goals



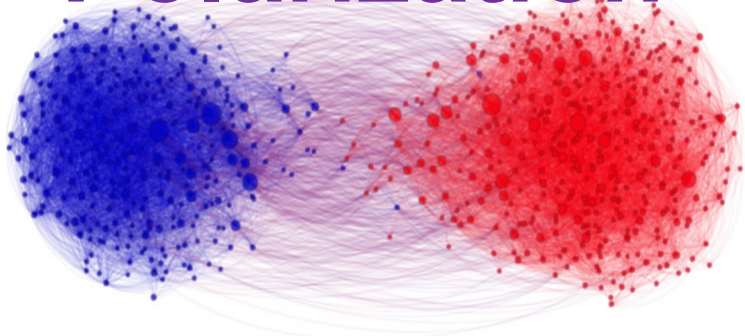
Bias



Inequality



Polarization



Social Good



AI: incredible advances and promises, pervasive and combined with gradual digitalization

62



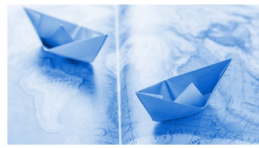
Societal Debates and Misinformation Analysis



Demography, Economy and Finance 2.0



Sustainable Cities for Citizens



Migration Studies



Science:



Sports

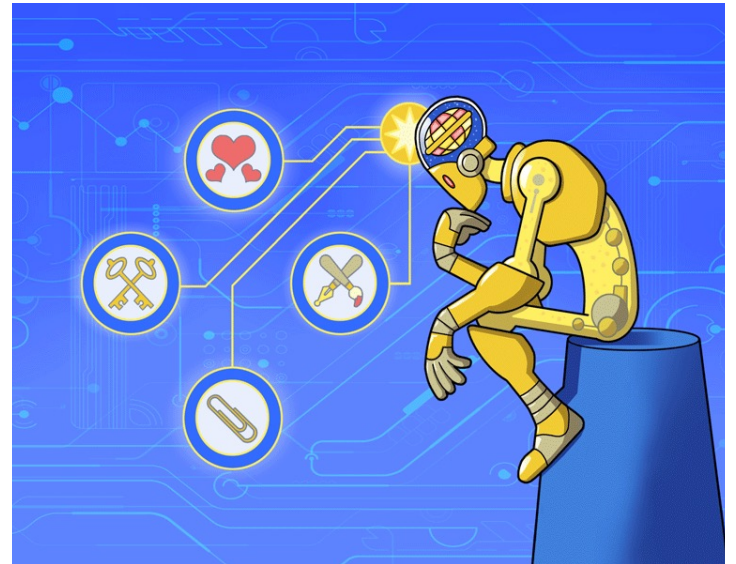


Personalized Medicine

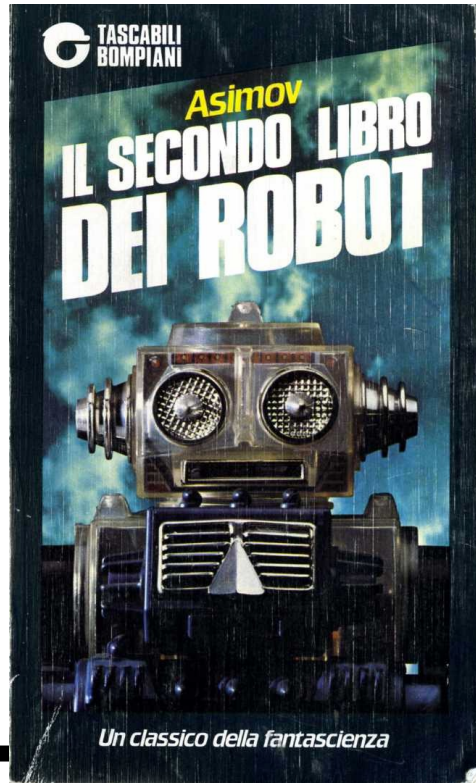


No techno-solutionism, but a gentle technology, helping us to find

- a richer, socially-aware human dimension
- a better balance between individual and collective needs
- a better balance between “my freedom” and “the freedom of my fellow citizens”



“Safety may not be perfect the greater the human ability, the more advanced it will be”



“**Knowledge has its risks, but should our reaction be to stop at risk?** Or should we not rather use knowledge to make it a barrier against the same risks that it entails? **Knives are made with a handle so that they can be grasped without danger;** stairs are equipped with railings; **electrical wires are insulated;** pressure cookers have a safety valve; in every product we take care to minimize the risk. **Sometimes the safety achieved is insufficient,** due to limitations imposed by the nature of the universe or by the human mind. **However, the attempt must be made.** As a machine, a robot will certainly be designed to offer guarantees of safety, at least as far as possible. **Safety may not be perfect (is there anything that is?), but the greater the human ability, the more advanced it will be.**”

Isaac Asimov, Introduction to “Second book of Robots” (1964)

THANK YOU!

SoBigData

AI4EU

HUMANE AI

Track & Know

NÔBIAS

PRO-RES

XAI

erc
European Research Council
Established by the European Commission



BACK TO CHAT GPT – recommended readings

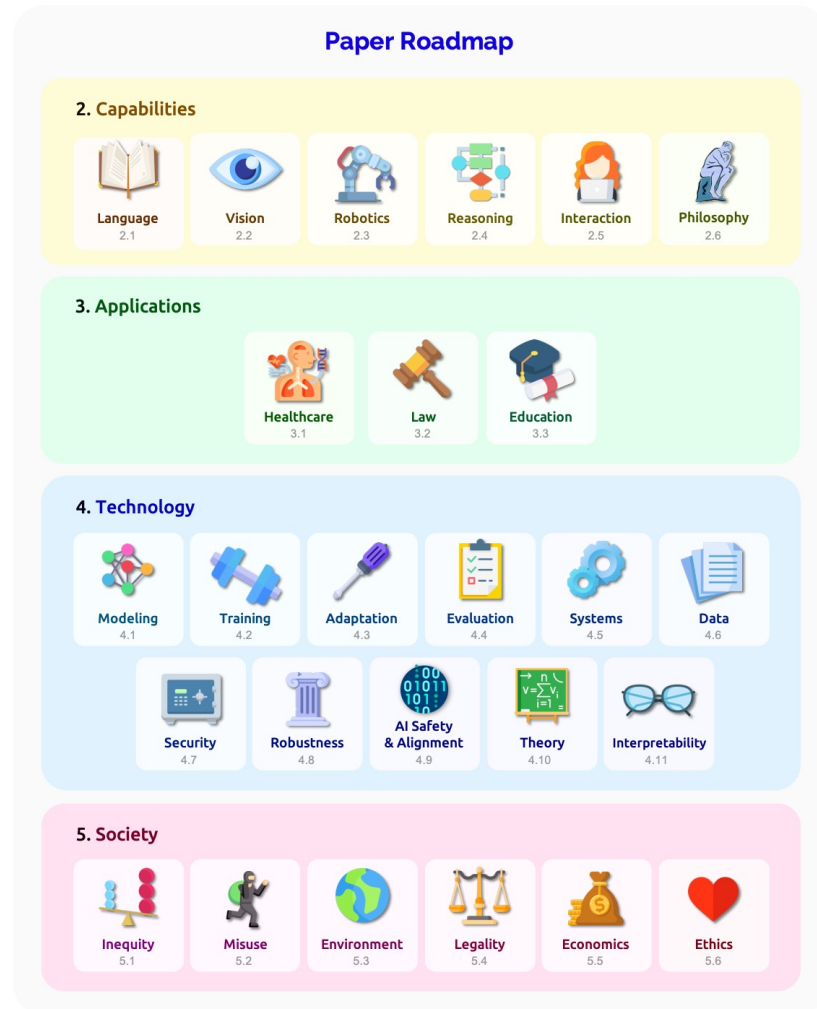
- LARGE LANGUAGE MODELS
- FOUNDATION MODEL
- GENERAL PURPOSE AI
- GENERATIVE AI
- TRANSFORMERS – SELF-SUPERVISED LEARNING
- Applications
 - Question Answering (QA)
 - Machine Translation (MT)
 - Text Summarization (TS)
 - Natural Language Inference (NLI)

On the Opportunities and Risks of Foundation Models, [Rishi Bommasani](#), ET AL. Stanford Univ. [arXiv:2108.07258](#) [cs.LG]

Beyond the imitation game: quantifying and extrapolating the capabilities of language models [Aarohi Srivastava et. Al](#) benchmark (BIG-bench)
[arXiv:2206.04615](#) [cs.CL]



On the
Opportunities and
Risks of Foundation
Models, [Rishi
Bommasani](#), ET AL.
Stanford Univ.
[arXiv:2108.07258](#) [c
s.LG]



Foundation Models – TIME LINE

Emergence of...

Homogenization of...

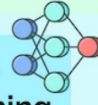
Machine Learning



"how"

learning algorithms

Deep Learning



features

architectures

Foundation Models



functionalities

models



Foundation Model – THE TASKS

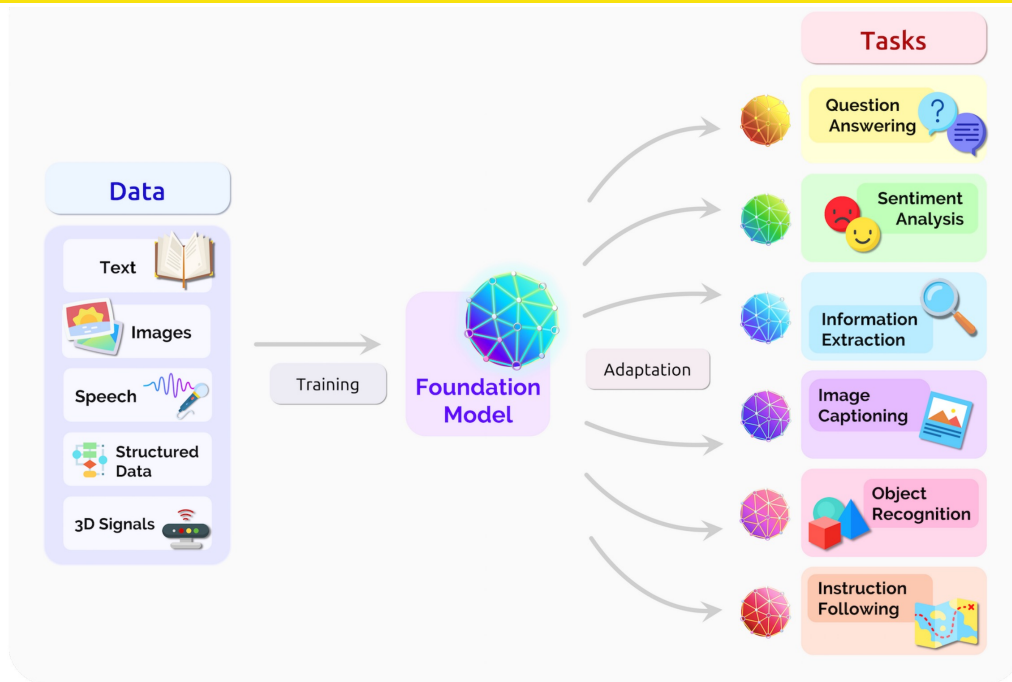
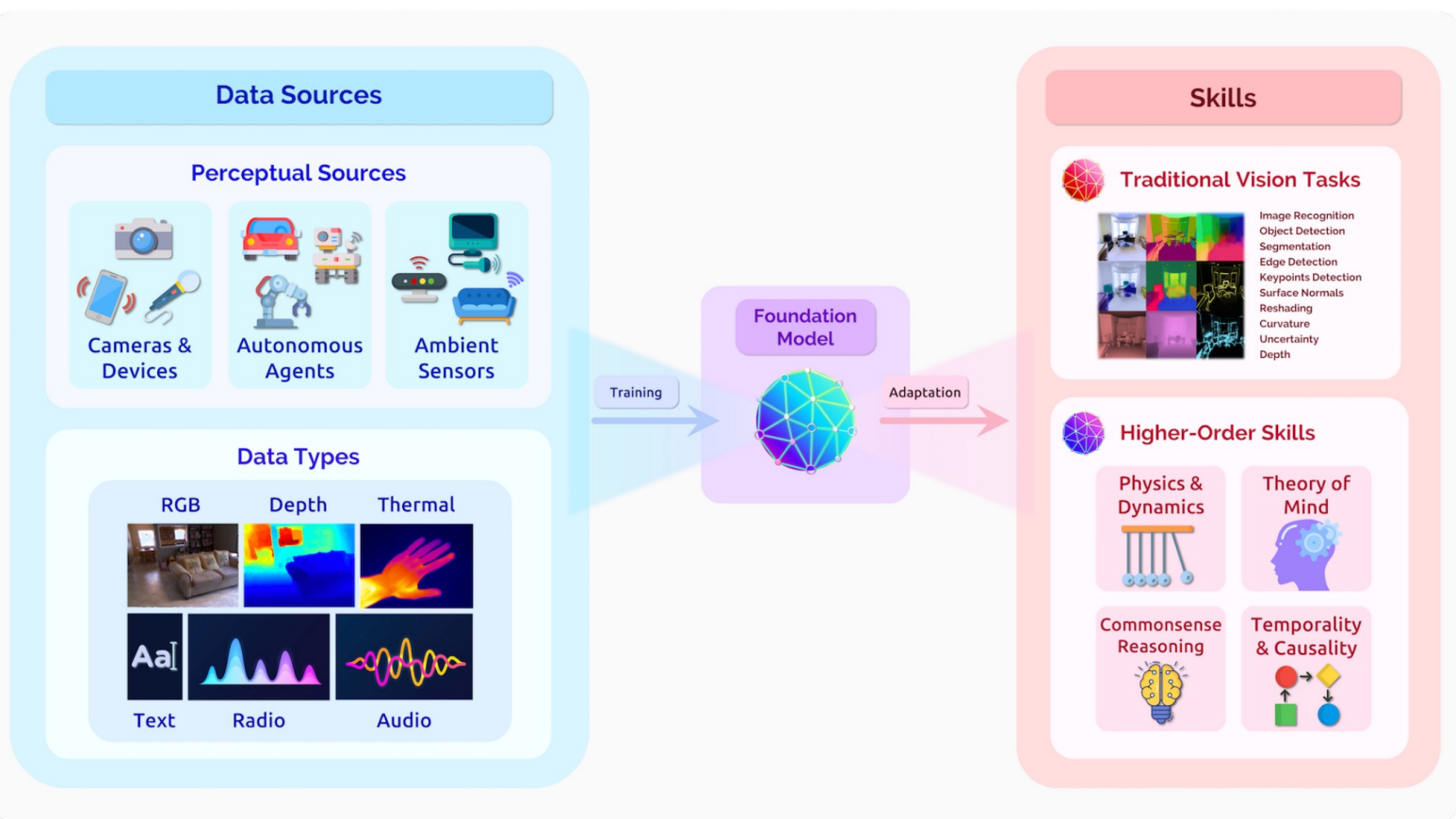


Fig. 2. A foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks.



VISION



ROBOTICS

Data Sources (2.3.2)

Robotic Interaction



Videos of Humans



Simulation



Natural Language

"Pick up the cup. Turn on the stove."

Training

Foundation Model



Adaptation

Tasks (2.3.1)

Intuitive, multi-modal task specification

"Make a sandwich"
input



→ **Reward Function**
output



Fast adaptation for task learning

Policy in
Kitchen A
input



"Open Fridge"



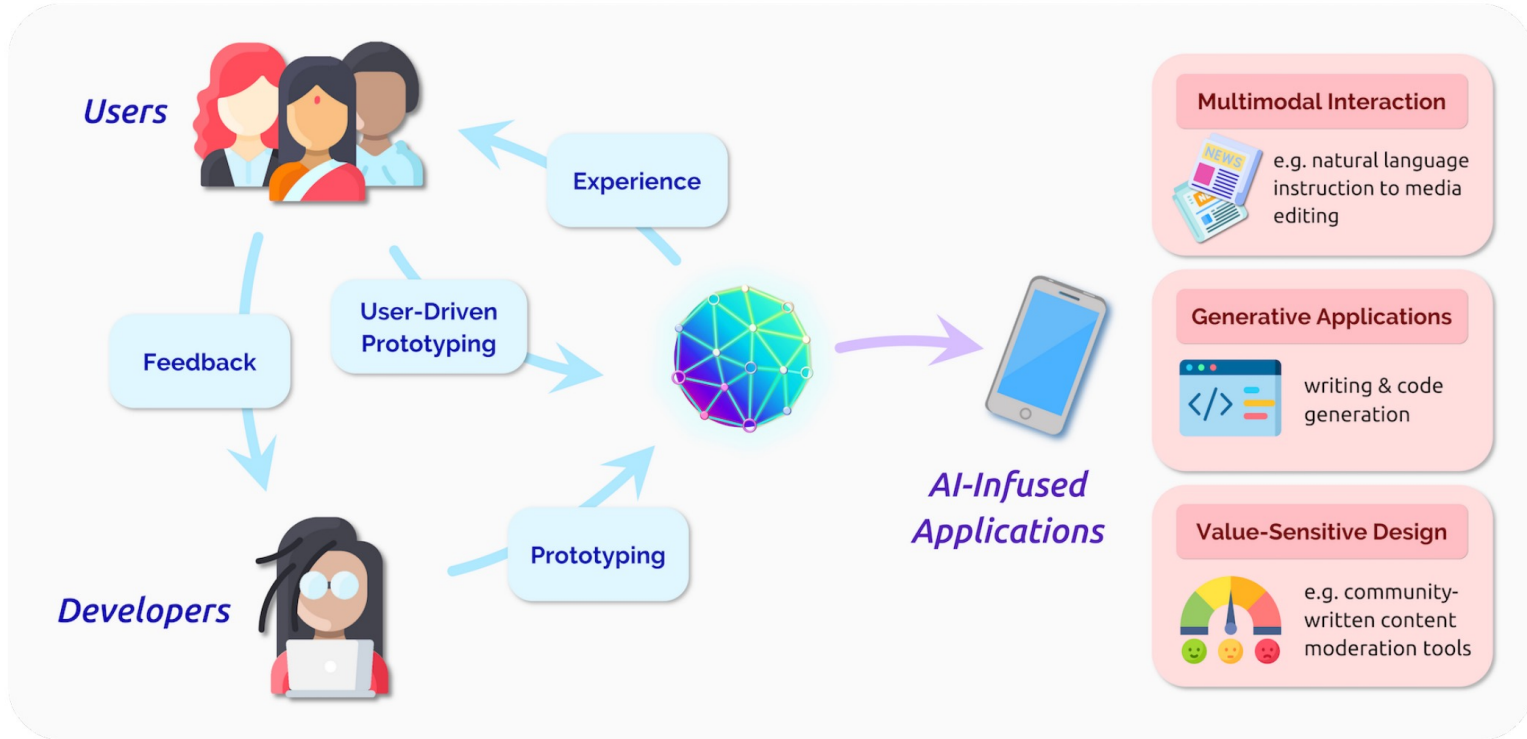
"Open Fridge"

Policy in
Kitchen B
output

Adapts to new tasks, environments, and embodiments.



INTERACTION



Security & Privacy

Large Uncurated Datasets

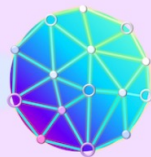
- ✓ Source of robustness
- ✗ Increased risk of poisoning



Training

Foundation Model

- ✓ Security choke point
- ✗ Single point-of-failure
- ✗ Increased attack surface



Adaptation

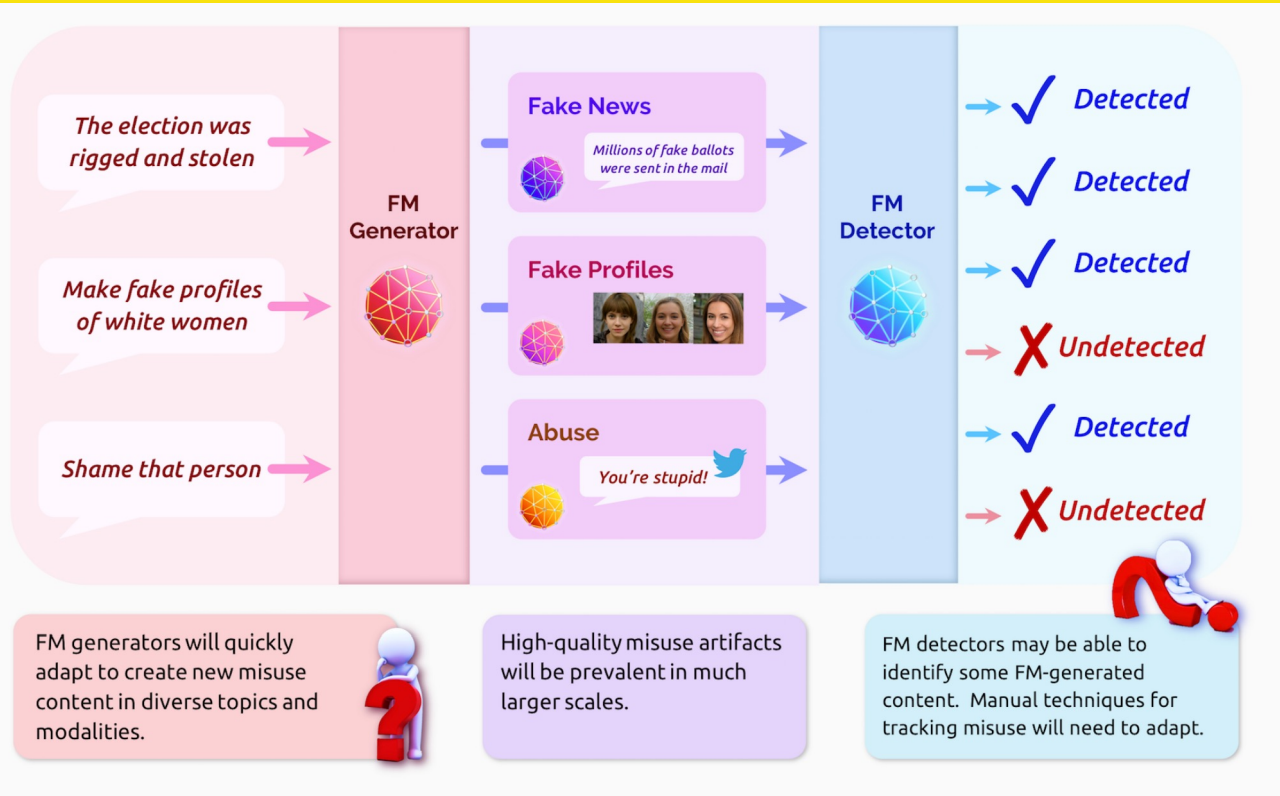


Downstream Applications

- ✓ Cheaper private learning
- ✗ Function creep



Misuse: on manipulative and harmful content generation, and the implications for detection.



Environment: cost-benefit analysis

