

Buenos Aires, February 8th, 2023

The social dimensions of human-centered Al

Fosca Giannotti Scuola Normale Superiore, Pisa, Italy









 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.



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.





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.



ፊ ⊽



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.



ፊ 🖓



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



We want design systems that:

- do not harm humans
- incorporate ethical values



We want design systems that:

- do not harm humans
- incorporate ethical values

Beneficence
Non-maleficence
Autonomy
Justice
Explicability



How to design Human AI systems?

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





European

General

Regulation

Data Protection

	Use high-quality training, validation and testing data (relevant, representative etc.)
and t risk ient the the d f the m	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



9

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

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

When a Computer Program Keeps You in Jail

By Rebecca We

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?"





 $HasAsthma(x) \rightarrow LowRisk(x)$

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



ERC Advanced Grant 2018 – 834756 <u>https://xai-project.eu/</u>

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

v3



Recognized Interpretable Models



Decision Trees

PREDICTION: p(survived = yes | X) = 0.671OUTCOME: YES



Linear Models

if $condition_1 \wedge condition_2 \wedge condition_3$, then *outcome*.

Rules



What is an Explanation?



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



"What is a good explanation, and for whom"



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





Julius Adebayo et al. Sanity checks for saliency maps. 2018.

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

Hu et al., Interpretable recommendation via attraction modeling: Learning multilevel attractiveness over multimodal movie contents. IJCALECAI, 2018.

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.

Hendricks et al., Generating Visual Explanations. ECCV 2016.



Explanation as Counterfactual Explanations





Explanation as FACTUALS and COUNTERFACTUALS



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.



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



because she is blue



Benchmarking and survey of explana tion methods for black box models F Bodria, F Giannotti, R Guidotti, F Naretto, D Pedreschi, S Rinzivillo arXiv preprint arXiv:2102.13076

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. AAAI 2020. https://xaitutorial2020.github.io/

Benchmarking and survey of explana tion 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





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 <u>Latent Local Rules</u> <u>Extractor (LLORE)</u>

• LASTS: Local Agnostic Subsequence-based Time Series explainer

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



Are explanations impacting user trust?



Cecilia Panigutti, Andrea Beretta, Fosca Giannotti and Dino Pedreschi. "Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems" Proceedings of 2022, ACM CHI Conference on Human Factors in Computing Systems

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



Experimental design: measures



Quantitative and qualitative measures:

 Weight of Advice (WOA)

Open-ended questions

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

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

- 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



Main findings: weight of advice.



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



Explainable AI – DARPA



https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

Still a lot of open research questions

Feedback loop between algorithmic and human decisions

- \triangleright Modeling human decision making:
 - designing classifiers that are able to defer decisions
 - designing classifiers that are able to as 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"







The social dimension of human centered Al (SAI)

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





A crowd of intelligent individuals is not necessarily an "intelligent" crowd

The sum of many individually "optimal" choices may have unintended negative outcomes

Because such choices influence and interfere with each other on top of shared resources





Als determine a **feedback loop**:

- users' preferences determine the training data on which Als are learned
- Al recommendations influence back users' preferences
-which determine future re-training

……and so on…

• Thus, Als determine impactful changes in networks

We largely ignore how



Als determine impactful changes in networks:

- diffusion of opinions and financial decisions shaped by AI recommender systems
 - Recommendations on social media may artificially amplify echo chambers, filter bubbles, and radicalisation
 - Profiling and targeted advertising may further increase inequality and monopolies, perpetuating and amplifying biases, discriminations, and "tragedies of the common



Sîrbu A, Pedreschi D, Giannotti F, Kertész J (2019) Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model.7 PLoS ONE 14(3): e0213246. https://doi.org/10.1371/journal.pone.0213246



Twitter has recently published a paper describing exactly the kind of fact-finding study we envisage

Huszár et al. "Algorithmic Amplification of Politics on Twitter", PNAS, Dec 2021



Als determine impactful changes in networks: case2: impact on sales

- By Dokyun Lee and Kartik Hosanagar, The Wharton School, University of Pennsylvania (2014)
- investigate the **impact of different recommender algorithms** (e.g., Amazon.com's "Consumers who bought this item also bought"), commonly used in ecommerce and online services, on sales volume and **diversity**, using **field experiment data** on movie sales from a top retailer
- discover that collaborative filtering algorithms cause individuals to discover and purchase a greater variety of products but push each individual to the same set of popular titles, leading to concentration bias at the aggregate level.



Als determine impactful changes in networks: case3 Navigation Apps

The New York Times

Navigation Apps Are Turning Quiet Neighborhoods Into Traffic Nightmares



The corner of Fort Lee Road and Broad Avenue in Leonia, N.J. With traffic apps suggesting shortcuts for commuters through the borough, officials have decided to take a stand. Bryan Anselm for The New York Times

Are routing apps impacting urban emissions?

TrafficO2: a simulation framework to compare the impact of navigation apps on CO2 emissions





ACM SIG SPATIAL Cornacchia et al., How routing strategies impact urban emissions, 2022

Rome – 0% vs 100% routed with gmaps



Rome – 0% vs 100% routed with gmaps



Rome – total C02, travel time and entropy with varying fractions



The network effects of AI and their impact on society are not sufficiently addressed by AI research

- We need a step ahead in the trans-disciplinary integration of A with network/complexity science and computational social science
- How to understand and mitigate harmful aggregated outcomes? *
- How to design AI mechanisms that help steering STS towards such agreed collective outcomes, e.g.,
 - sustainable mobility in cities,
 - diversity and pluralism in the public debate,
 - fair distribution of resources?



* Within an ethical and legal framework and public policy that sets the goals

The need of real "interventional" studies

Most studies are:

 observational: they use data gathered from browser loggers, platforms' APIs, bots, and controlled studies

• simulative: mathematical models of STS that introduce interventions probabilistically to study aggregated social effects

Only a few works are interventional: they explicitly examine the causal effects of AI on actual users (e.g., with A/B testing or randomised controlled trials)

- Interventional methods are crucial but hard to conduct
- How to solve this issue/limitation?



...in conclusion



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 humancentered AI?

A system is intelligent if it allows humans, **both as individuals and as a collective**,

▷to become more intelligent in achieving complex goals





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"





XAI: Science and technology for the eXplanation of AI decision making: We are hiring!

- Joint work with: Dino Pedreschi, Anna Monreale, Salvatore Ruggieri, Franco Turini, Riccardo Guidotti, Salvo Rinzivillo, Daniele Fadda, Andrea Beretta, Carlo Metta, Mattia Setzu, Francesca Naretto, Francesco Bodria, Cecilia Panigutti, Francesco Spinnato
- 2 research fellow position open till **17 February**, **2023**
- <u>https://amministrazionetrasparente.sns.it/bando/selezione-di-n-2-ricercatoriricercatrici-tempo-determinato-di-tipo-sc-01b1-ssdinf01-recruitment-2-fixed-terms-researcher</u>
- Write to: fosca.giannotti@sns.it





[1] Setzu, M., Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., & Giannotti, F. (2021). GLocalX-From Local to Global Explanations of Black Box AI Models. Artificial Intelligence, 294, 103457, and PD Thesis

[2] Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., & Turini, F. (2019). Factual and counterfactual explanations for black box decision making. IEEE Intelligent Systems, 34(6), 14-23

[3] Guidotti, R., & Monreale, A. (2020, November). Data-Agnostic Local Neighborhood Generation. In 2020 IEEE International Conference on Data Mining (ICDM) (pp. 1040-1045). IEEE.

[4] Panigutti, C., Perotti, A., & Pedreschi, D. (2020, January). Doctor XAI: an ontology-based approach to black-box sequential data classification explanations. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 629-639), and PD Thesis

[5] Guidotti, R., Monreale, A., Matwin, S., & Pedreschi, D. (2020, April). Explaining Image Classifiers Generating Exemplars and Counter-Exemplars from Latent Representations. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 09, pp. 13665-13668).

[6] Guidotti, R., Monreale, A., Spinnato, F., Pedreschi, D., & Giannotti, F. (2020, October). Explaining Any Time Series Classifier. In 2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI) (pp. 167-176). IEEE.

[7] Francesco Bodria Understanding and Exploiting the Latent Space of Machine Learning Models, Phd Thesis

[8] Francesco Spinnato Explanation Methods for Sequential Data Models, Phd Thesis

[9] Giovanni Puccetti Text Generation from Multimodal Sources, Phd Thesis

[10] M. Cinquini. Boosting Synthetic Data Generation with Effective Nonlinear Causal Discovery. Master's thesis in Data Science and Business Informatics. Pisa, May 2021

[11] Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., & Rinzivillo, S. (2021). Benchmarking and Survey of Explanation Methods for Black Box Models. arXiv preprint arXiv:2102.13076

[12] Cecilia Panigutti, Anna Monreale, Giovanni Comandè, Dino Pedreschi: Ethical, societal and legal issues in deep learning for healthcare. Chapter in Deep Learning in Biology and Medicine. World Scientific Publishing.

[13] Francesca Naretto, Roberto Pellungrini, Anna Monreale, Franco Maria Nardini, Mirco Musolesi: Predicting and Explaining Privacy Risk Exposure in Mobility Data. DS 2020: 403-418

[14] Francesca Naretto, Roberto Pellungrini, Franco Maria Nardini, Fosca Giannotti: Prediction and Explanation of Privacy Risk on Mobility Data with Neural Networks. PKDD/ECML Workshops 2020: 501-516

[15] Panigutti, C., Perotti, A., Panisson, A., Bajardi, P., & Pedreschi, D. (2020). FairLens: Auditing Black-box Clinical Decision Support Systems. arXiv preprint arXiv:2011.04049. An improved version of the manuscript is currently submitted to the Information Processing & Management journal.

[16] Francesca Naretto. The relationship between privacy and explanations, Phd ThesisPanigutti, C., Guidotti, R., Monreale, A., & Pedreschi, D. (2019, January). Explaining multi-label black-box classifiers for health applications. AAAI 2019. Springer, Cham.

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



THANK YOU!



Vous préférez un conseiller qui répond humainement ou une machine qui répond machinalement?

