Identifying the 'right' level of explainability

Beaudouin, V., Bloch, I., Bounie, D., Clémençon, S., d'Alché-Buc, F., Eagan, J., Maxwell, W., Mozharovskyi, P., & Parekh, J. (2020b). *« Identifying the right level of explanability in a given situation »* https://hal.telecom-paris.fr/hal-02507316

Winston Maxwell, Director of Law & Technology Studies, Télécom Paris

Astrid Bertrand, PhD Student in Explainable AI, Télécom Paris

telecom-paris.fr/en/ai-ethics

Operational AI Ethics: five interdisciplinary research pillars

Algorithmic bias and fairness

Governance and regulation

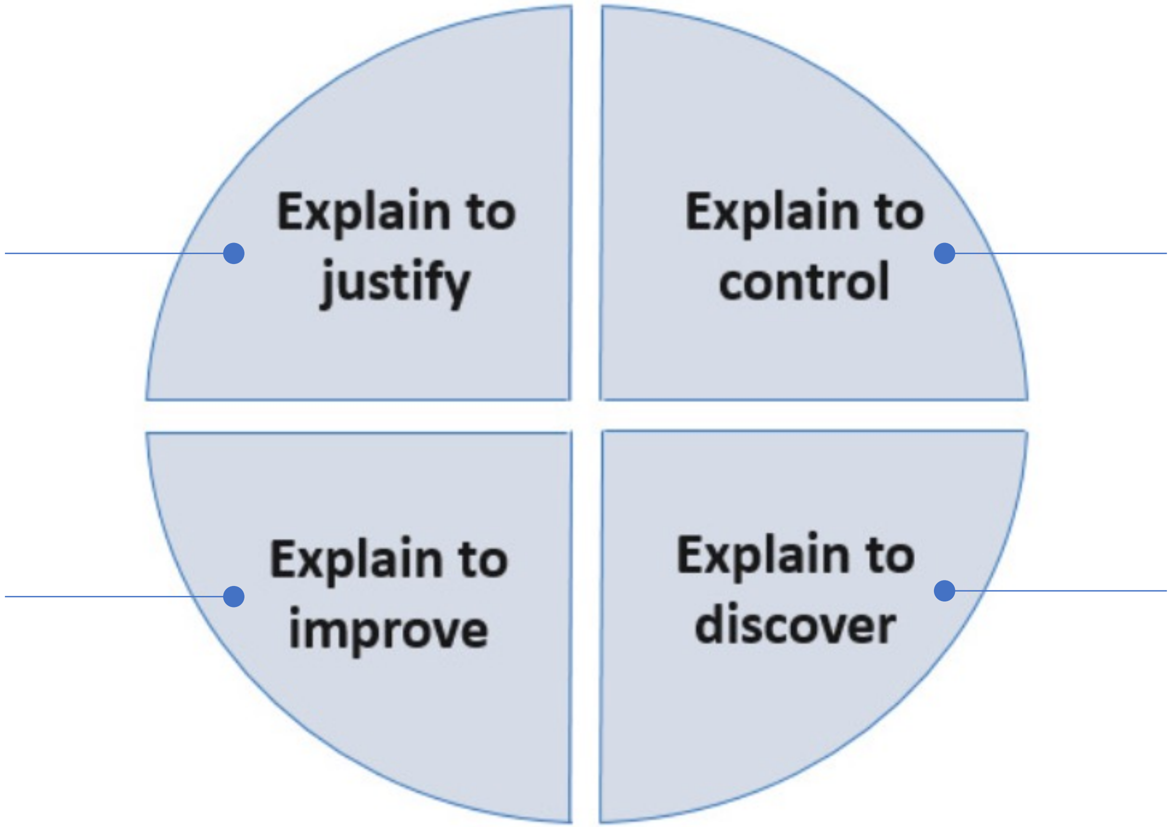
Algorithmic explainability

AI and General Interest

AI liability

Applied math
Statistics
Data science
Computer science
Sociology
Law
Economics

What purposes do explanations serve?



Explain results, defend algorithmic decisions as being fair and ethical, which leads to building trust.

Understand to identify and correct errors, learn vulnerabilities, challenge decisions.

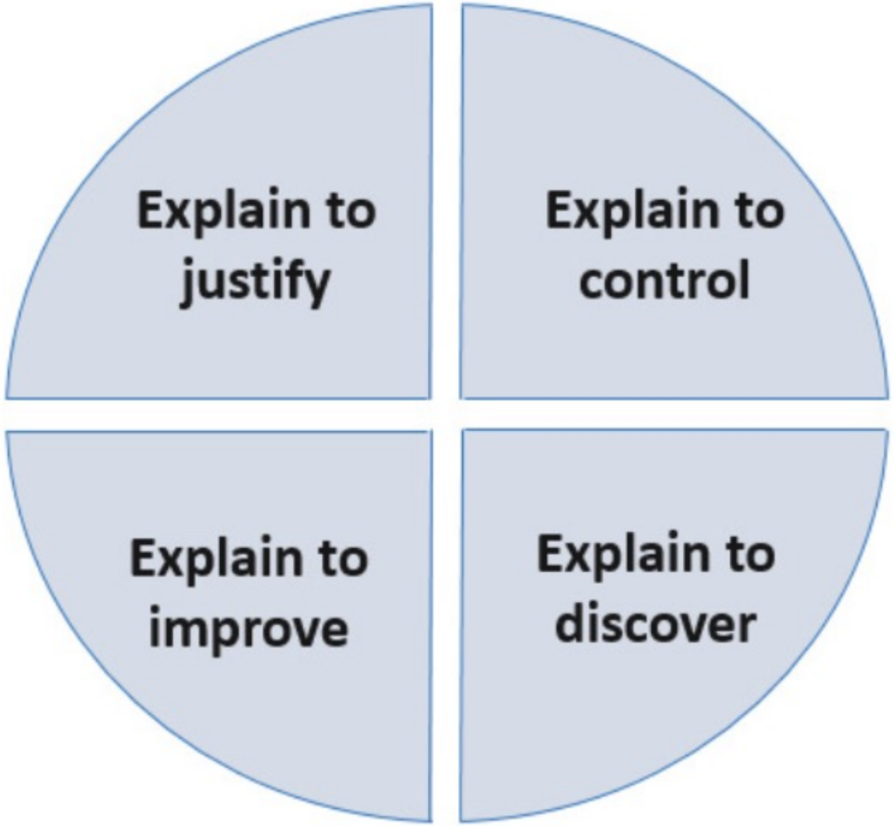
Improve the robustness and reliability of the algorithm.

Can Alpha Go teach us about its learned strategy? Can AI teach us hidden laws in biology, chemistry and physics?

Reasons For XAI

Adadi and Berrada - 2018 - Peeking Inside the Black-Box A Survey on Explainable AI

What purposes do explanations serve?



Explain to justify

Explain to control

Explain to improve

Explain to discover

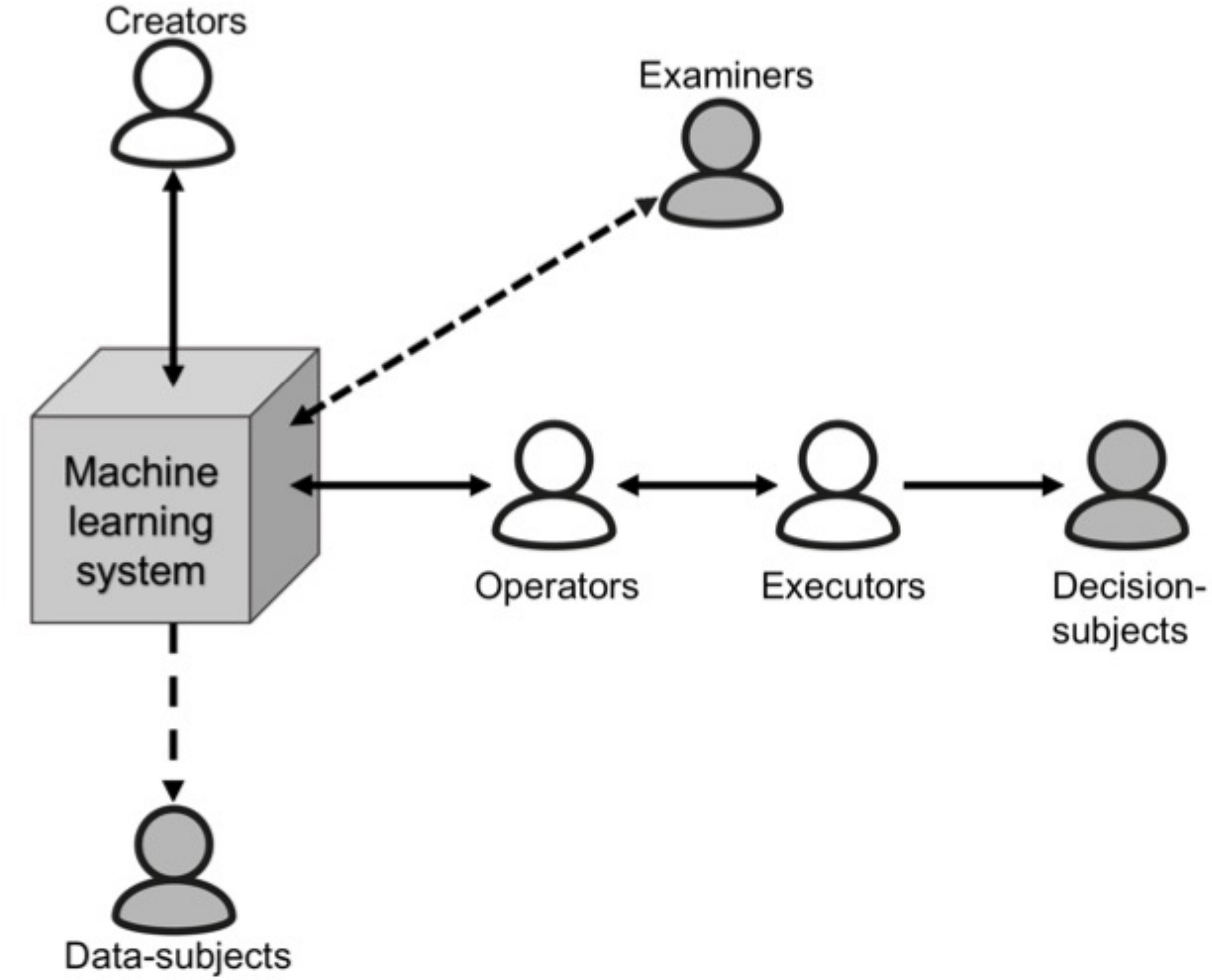
The explanations contribute to:

- ➔ traceability,
- ➔ auditability and
- ➔ accountability.

Reasons For XAI

Adadi and Berrada - 2018 - Peeking Inside the Black-Box A Survey on Explainable AI

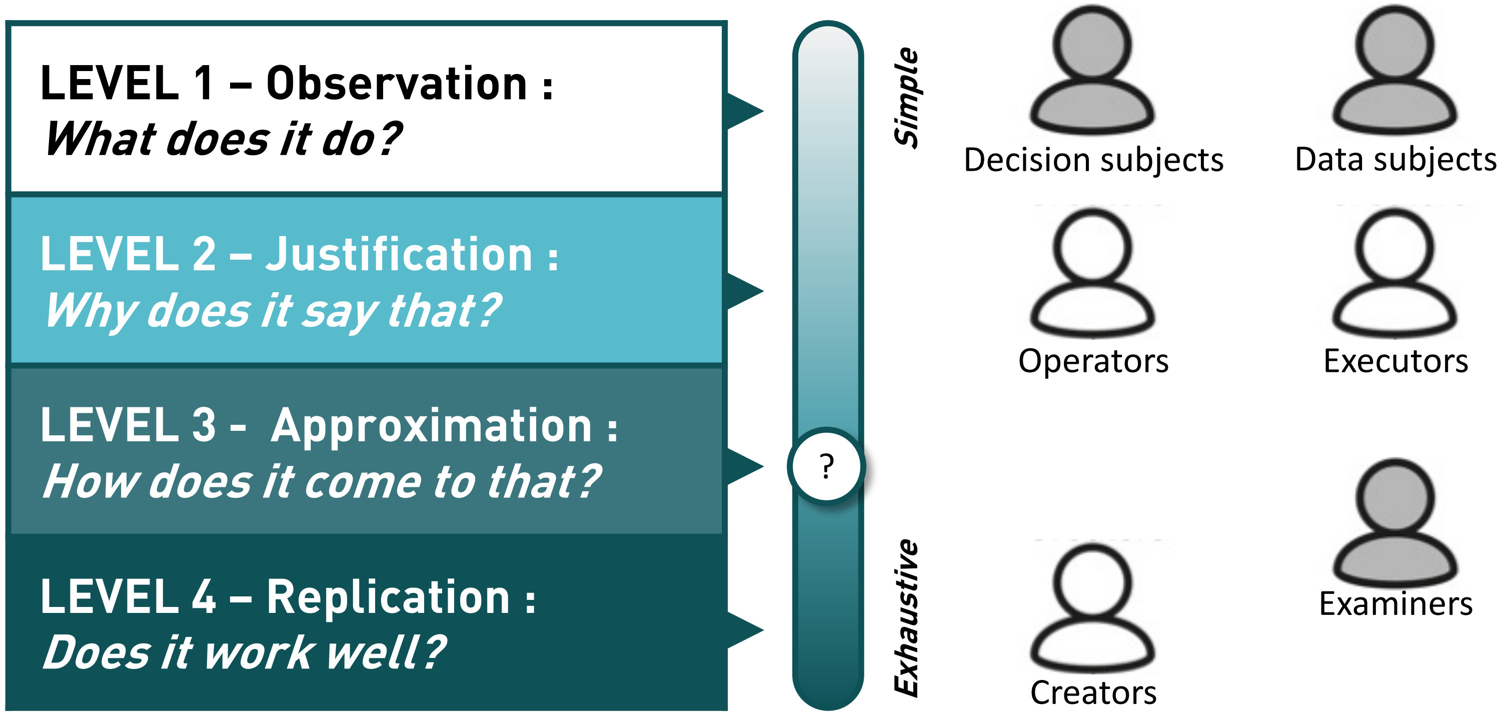
What audiences?



A machine learning ecosystem

Tomsett et al. - 2018 - Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems

In financial applications, the ACPR's guidelines sets 4 XAI levels



LEVEL 1 – Observation :
What does it do?

LEVEL 2 – Justification :
Why does it say that?

LEVEL 3 - Approximation :
How does it come to that?

LEVEL 4 – Replication :
Does it work well?

The level of the explanation depends on the risks and audiences
Dupont et al. - 2020 - 'Governance of Artificial Intelligence in Finance' (ACPR Report)

What costs and benefits, and when are explanations worth the costs? (1/2)

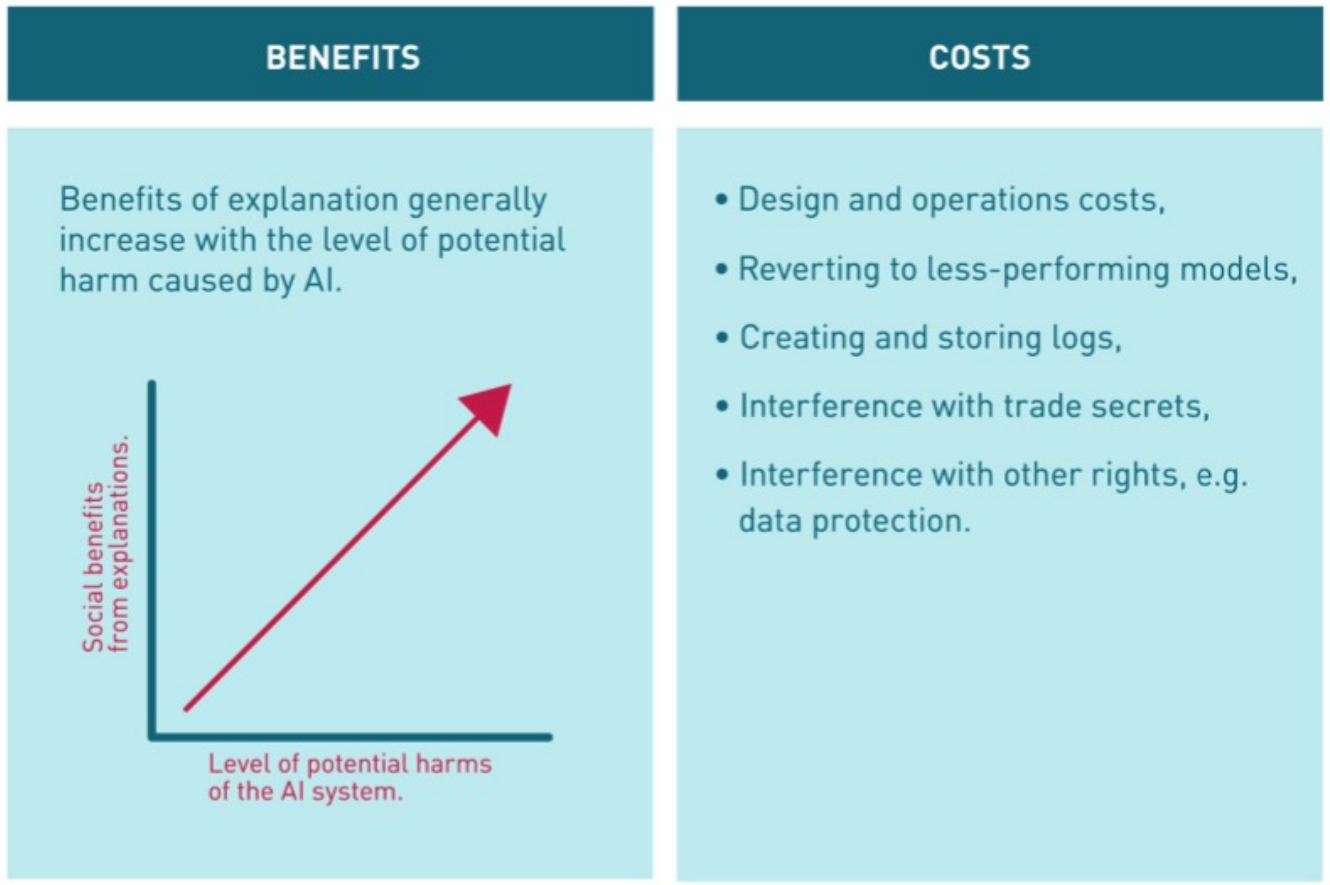
- Not all AI systems need to sacrifice their accuracy for interpretability

> *‘We should be careful when giving up predictive power, that the desire for transparency is justified and isn’t simply a concession to institutional biases against new methods.’*
> (Lipton, 2018)

- Doshi-Velez et al. (2017) illustrate this idea using the example of a smart toaster:

> *‘Requiring every AI system to explain every decision could result in less efficient systems, forced design choices, and a bias towards explainable but suboptimal outcomes. For example, the overhead of forcing a toaster to explain why it thinks the bread is ready might prevent a company from implementing a smart toaster feature – either due to the engineering challenges or concerns about legal ramifications'*

What costs and benefits, and when are explanations worth the costs? (2/2)



BENEFITS

Benefits of explanation generally increase with the level of potential harm caused by AI.

Social benefits from explanations.
Level of potential harms of the AI system.

COSTS

- Design and operations costs,
- Reverting to less-performing models,
- Creating and storing logs,
- Interference with trade secrets,
- Interference with other rights, e.g. data protection.

Summary of benefits and costs of explanations

Beaudouin et al. - 2020 - Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach

Explainability and human rights

The Rule of Law/Etat de droit requires:

- Laws published in advance and understandable
- Predictable application of laws
- Supremacy of laws
- Universality of laws
- Predictable and fair legal procedures to contest decisions
- Citizens should understand how decisions that affect them fit into a coherent, consistent, legal framework

➢ Human dignity requires individual agency and choice

Each individual is an entity open to argument and persuasion, and deserving of reasoned explanations rather than simply objects to be coerced into compliance
Emily Berman, A Government of Laws and Not of Machines, 98 Boston Univ. L. Rev. 1277 (2018)

the rule of law . . . is preferable to that of any individual. Aristotle, *Politics* 1282b

When does the law require explainability?

USA

State v. Loomis (2016)

- COMPAS algorithm
- Judge imposes global explainability

Houston Federation of Teachers (Tex. 2017)

- Scoring algorithm for teacher performance
- Judge imposes replicability of individual scores to test for errors

Regulation B, Fair Credit Reporting Act, Equal Credit Opportunity Act

- Bank must give specific reasons for loan denial
- Regulation B provides for 24 reason codes

Administrative Procedure Act

- Requires statement of findings and conclusions, and the reasons or basis therefor

Washington State Facial Recognition Law

- Requires accountability report
- Meaningful human review
- Testing for bias

Example of credit refusal:

Your application was processed by a credit scoring system that assigns a numerical value to the various items of information we consider in evaluating an application. These numerical values are based upon the results of analyses of repayment histories of large numbers of customers. The information you provided in your application did not score a sufficient number of points for approval of the application. The reasons you did not score well compared with other applicants were:

- *Insufficient bank references*
- *Type of occupation*
- *Insufficient credit experience*
- *Number of recent inquiries on credit bureau report*

When does the law require explainability?

Europe and France

La Quadrature du Net

- Specific and reliable models
- Pre-established models and criteria
- Re-examination by human experts
- Institutional oversight

SyRI decision (Netherlands)

- Transparency required to permit individual challenge of scores
- Verification of lack of discrimination

French Code of relations between the public and the administration

- The degree and method by which the algorithmic calculation contributed to the decision
- The data relied on and their source
- The parameters used, and their ponderation as applied to the situation of the individual
- The operations conducted by the processing

General Data Protection Regulation

- 'meaningful information about the logic involved'
- 'fair and transparent' processing
- data protection impact assessments

European Platform to Business Regulation

- Main parameters determining ranking
- Explanation of relative effects thereof
- Permit "adequate understanding" by users

The challenge of designing appropriate reliance: automation bias

<Automation-Induced Complacency > :

- Evidenced in aviation through analysis of accidents [Parasuraman 1993]
- It’s larger than the lack of vigilance, boredom, or excessive workload. It’s a distinctive Attitude.
- Occurs in monitoring situations that involves multiple tasks.
- The substandard monitoring leads to poorer performance (usually missed errors or delayed response).
- *“self-satisfaction that may result in nonvigilance based on an unjustified assumption of satisfactory system state” [Billings and al. 1976]*
- *“a psychological state characterized by a low index of suspicion”* [Wiener, 1981]

<Automation bias> :

- *‘the tendency to use automated cues as a heuristic replacement for vigilant information seeking and processing’* [Mosier and Skitka, 1996]
- Active bias (over-reliance) towards automation.
- Overlapping phenomenons, with allocation of limited user attention being central to both [Parasuraman]

<(Inappropriate) Trust in automation> :

- Low self-confidence is related to a greater inclination to rely on the automatic controller.
- The opposite is also true: When operators’ self-confidence is high and trust in the system is low, they are more inclined to rely on manual control. [Lee and See, 2004].

How can explanations help mitigate human biases (such as automation bias)?

Context: AI-advised human decision making. AI generates an explanation.

<Explanation> Refers to the causes of an event. Result of an abductive reasoning :
Causal connections ➔ Selection ➔ Evaluation [Miller, Peirce]
Human cognitive biases can be involved in selecting and evaluating explanations.

<Interpretability> Degree to which an observer can understand the cause of a decision [Miller, Biran and Cotton] Latent, subjective property.

Can be measured through [Poursabzi-Sangdeh et al. 2019]:

- Simulatability : *How well can people estimate what a model will predict?*
- Deviation: *To what extent do people follow a model's predictions when it is beneficial to do so?*
- Capacity to detect errors: *How well can people detect when a model has made a sizable mistake?*

Human relies appropriately on the algorithm and takes autonomous decision.
"Meaningful human intervention"

Cognitive biases in XAI

« Misuse » and « Disuse »

Misuse

- ➢ Automation complacency
- ➢ Excessive confidence in the hypothesis explained
- ➢ Anchoring, confirmation, availability [Kahneman]

Risk not to detect algorithm's errors

« Adopting a conditional reference frame might make certain aspects of the problem prominent, so that these aspects become the ones by which hypotheses are evaluated when assessing confidence »

[Griffin & Tversky, 1990]

« People more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake »

[Dietvorst 2015]

Disuse

- ➢ Algorithmic aversion [Dietvorst]
- ➢ Belief perseverance [Koehler]

Risk not to follow the correct algorithm's recommendations

Avoiding misuse

- ➢ Automation bias
- ➢ Anchoring bias

What are the means to enhance the user's ability to detect algorithmic errors?

Koehler « Explanation, Imagination, and Confidence in Judgment »

➡ **Give a counter-explanation**

Lombrozo, Rehder, Miller : The explanation is useful to learn and generalize

➡ **Train users with explanations, then let them disappear as they use them.**

Logg: lay people overestimate the algorithm's answer due to lack of confidence

➡ **Put the user in conditions where he has his own answer before seeing the recommendation**

Poursabzi-Sangdeh

➡ **Don't overload with information**

➡ **Send alert messages when unusual situations/parameters occur**

Avoiding disuse

- ➢ Algorithmic aversion
- ➢ Belief perseverance

How to select the explanation(s) that best satisfies the user's questions, hereby increasing his confidence?

Josephson, Miller: Humans tend to ignore statistical arguments

➡ **Give causes rather than statistical arguments**

Hilton, Lipton, Miller : why questions are contrastive

➡ **Define the underlying opposing hypothesis ('the foil')**

Thagard's theory, Read and Marcus-Newhall, Miller

➡ **Use simple (citing fewer causes) and more generalized (explaining more events) explanations**

Grice, Hilton, Graaf and Malle, Miller :

➡ **Present the explanation as a conversation.**

➡ **Respect Grice's maxims of communication Quality, Quantity, Relation, Manner**

Thank you for your attention!

Feel free to contact us at:
winston.maxwell @telecom-paris.fr
astrid.bertrand@telecom-paris.fr
telecom-paris.fr/en/ai-ethics

Références

- Beaudouin, V., Bloch, I., Bounie, D., Clémençon, S., d'Alché-Buc, F., Eagan, J., Maxwell, W., Mozharovskyi, P., & Parekh, J. (2020b). Identifying the 'Right' Level of Explanation in a Given Situation. https://hal.telecom-paris.fr/hal-02507316
- Beaudouin, V., Bloch, I., Bounie, D., Clémençon, S., d'Alché-Buc, F., Eagan, J., Maxwell, W., Mozharovskyi, P., & Parekh, J. (2020a). Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach. hal-polytechnique.archives-ouvertes.fr.
- Miller, *« Explanation in artificial intelligence: Insights from the social sciences »,* 2019, Journal of Artificial intelligence
- Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, Hanna Wallach, *« Manipulating and Measuring Model Interpretability"*, 2018
- Derek J. Koehler, *"Explanation, Imagination, and Confidence in Judgment"*, 1991
- Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey, *"Algorithm aversion: People erroneously avoid algorithms after seeing them err."*, 2015
- Jennifer-Marie Logg, *"Theory of Machine: When Do People Rely on Algorithms?",* 2017
- Dietvorst, *« Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can Even Slightly Modify Them »*, 2018
- Lord, Pepper and Preston, *« Considering the opposite: a corrective strategy for social judgment »,* 1984
- Miller and Parasuraman, *« Designing for Flexible Interaction Between Humans and Automation: Delegation Interfaces for Supervisory Control »,* 2007
- Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. ArXiv:1702.08608 [Cs, Stat]. http://arxiv.org/abs/1702.08608
- Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36–43. https://doi.org/10.1145/3233231
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. Human Factors, 52(3), 381–410. https://doi.org/10.1177/0018720810376055
- Parasuraman, R., Molloy, R., & Singh, I. (1993). Performance Consequences of Automation Induced Complacency. International Journal of Aviation Psychology, 3. https://doi.org/10.1207/s15327108ijap0301_1

TELECOM Paris

INSTITUT POLYTECHNIQUE DE PARIS