AI Ethics in Healthcare: Explainability, Trust, and Moral Agency

October 8 - 10, 2025
Dighealth project

Luka Poslon, mag. phil.
Catholic University of Croatia
Digital healthcare ethics laboratory (Digit-HeaL)

Co-funded by the European Union

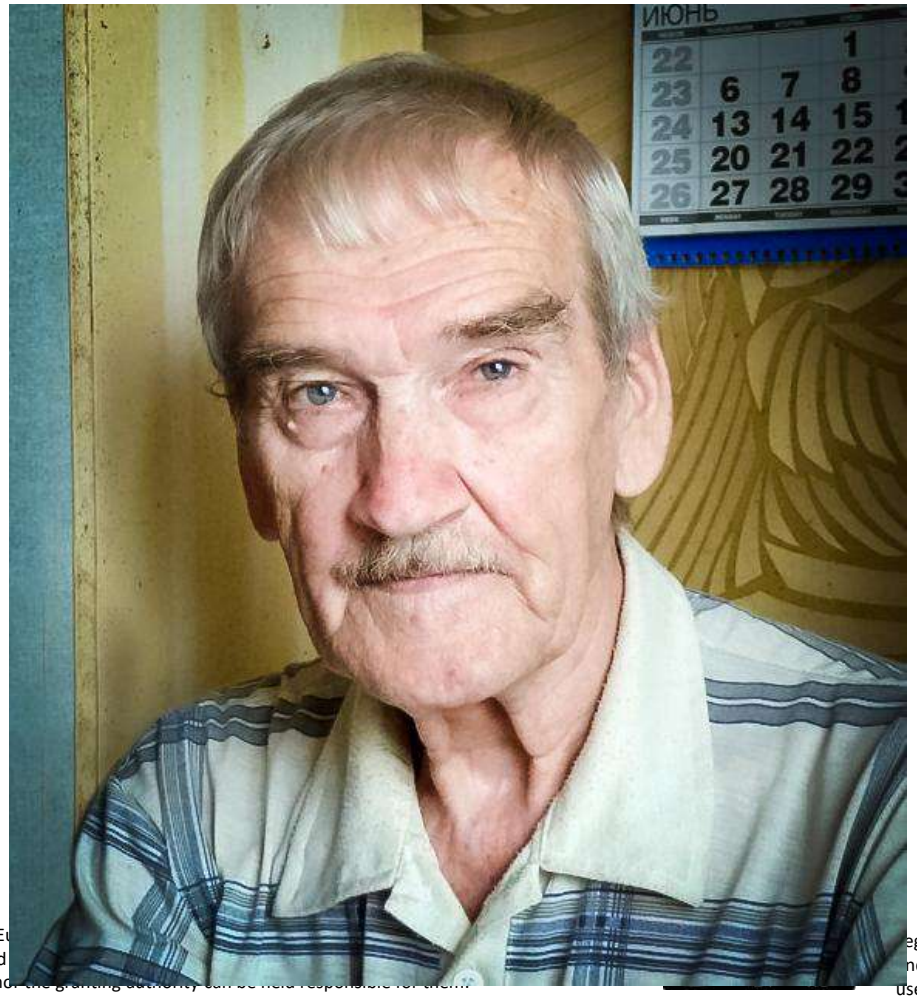
Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

- AI Ethics (& AI Epistemology)
- Alignment problem
- Instrumental & Intrinsic Good
- Explainability (& Transparency)
- Trust
- Human decision-making
- Moral Agency

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

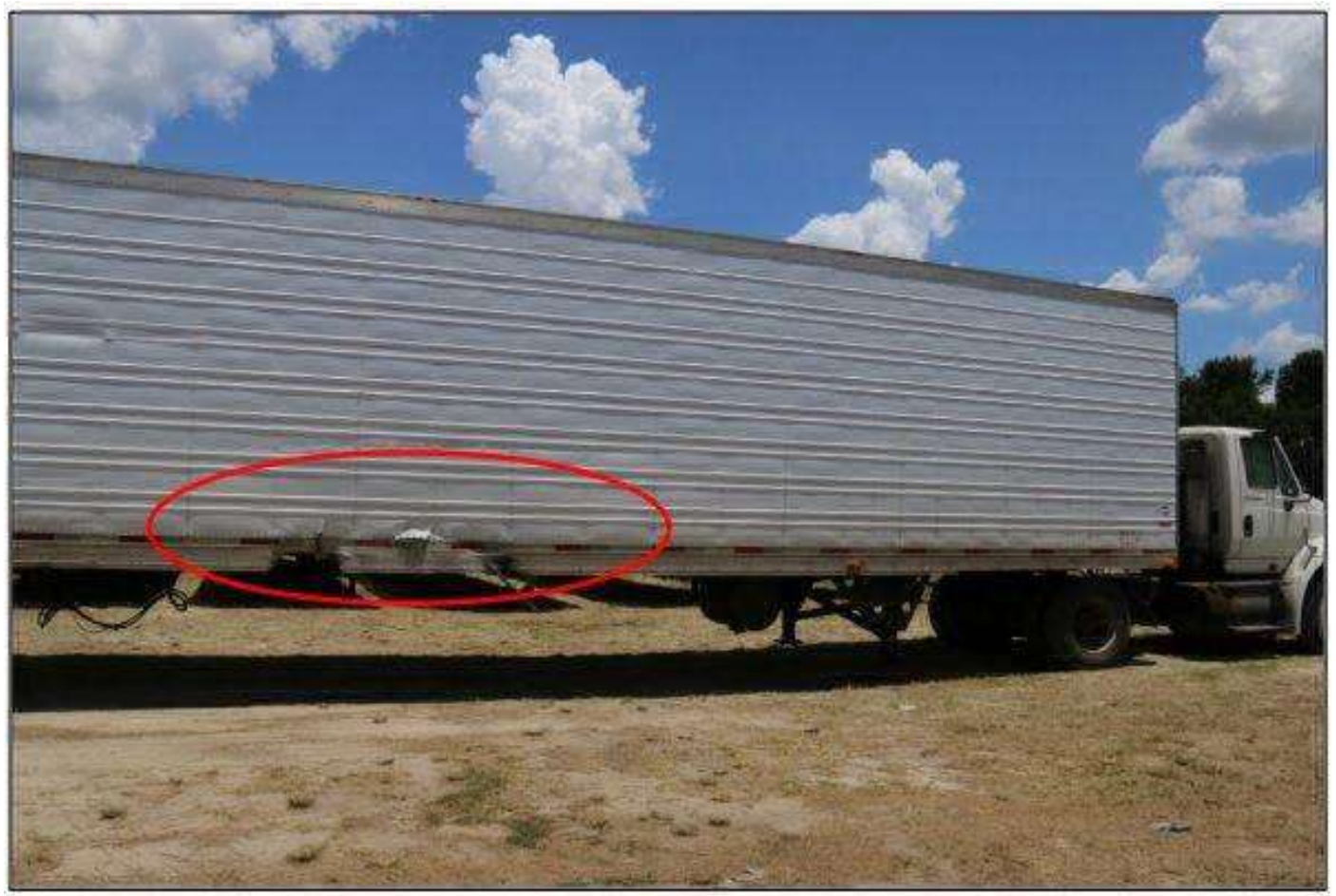
- Retired Soviet Lt. Col. Stanislav Petrov



Co-funded by the European Union

Funded by the Eu
author(s) only and
European Union no

- Tesla's deadly crash in May 2016



Co-funded by the European Union

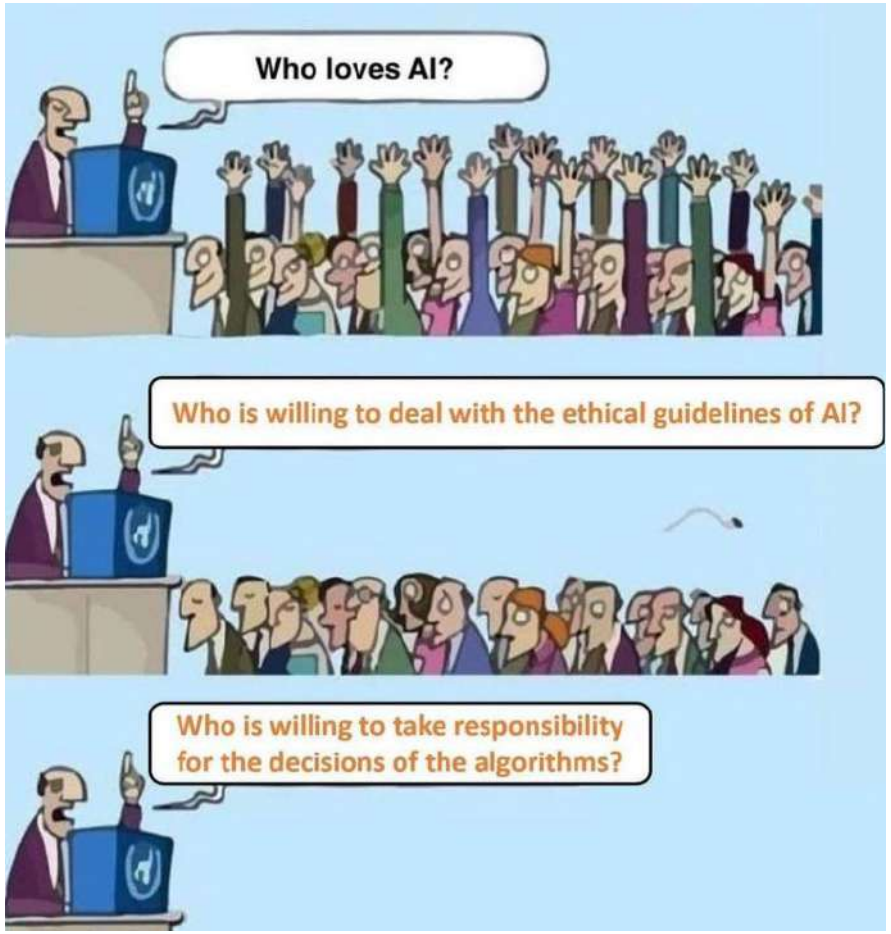
Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

- Microsoft's chatboot Tay

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

- Do we care about AI Ethics?
- Why should we care?
- How should we care?



Co-funded by the European Union

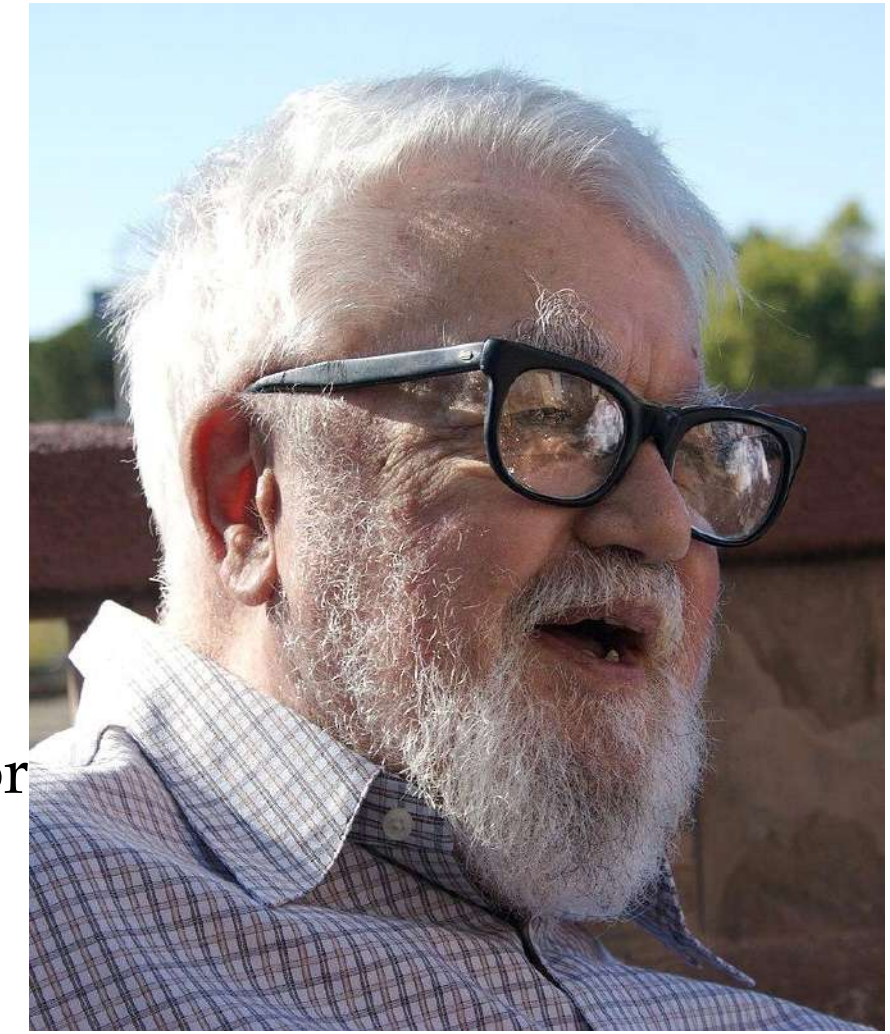
Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

- Does AI align with human values?
  - Or fundamentally clash with them?
- What goals are being pursued?
  - In what way are goals being pursued?
- Perhaps the goals themselves are problematic? (Paperclip maximizer – Nick Bostroom)
  - Perhaps the means of pursuing them are the real problem?

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

- Computer program *Advice Taker*
- John McCarthy in 1958
- "program with common sense”
- Turing Award laureate in 1971 for significant contributions to the field of AI



John McCarthy,
September 4, 1927 – October 24, 2011

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

Goal

- Achieving trustworthy AI in healthcare
- Biomedical supercomputing - not mandatory for all situations in biomedicine

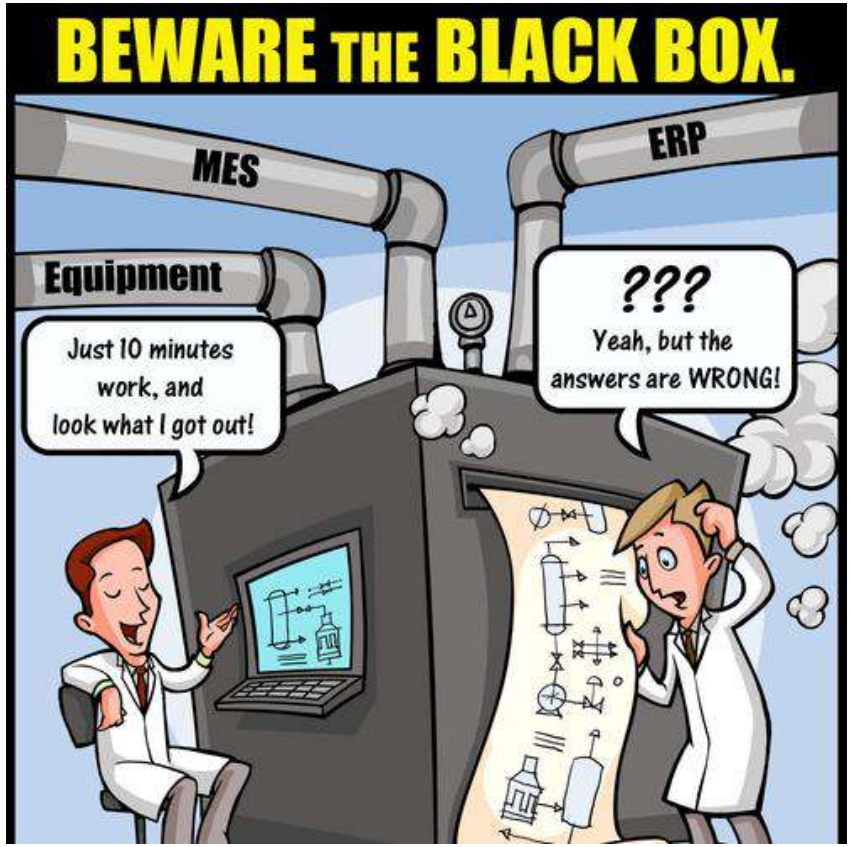
Challenges

- Multifaceted concept of trustworthiness
- Need for user-friendly systems for all stakeholders

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

- AI systems are „black boxes”
- Highly undesirable for high-stakes applications
- We need algorithms that we can understand and, in turn, trust
- Doctors should not trust a machine that cannot reason about its decisions



Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

Affects many fields:

- Trust Issues of Black-Box Models
  - Preference for surgeons over robotic systems (Longoni, 2019)
  - Importance in xAI for breast cancer (Bidwai et al., 2023)

- Transperancy Issues of Black-Box Models
  - Ensuring transparency in decision-making processes essential to building trust between healthcare professionals, patients and the general public (Peroso, A., et. al, 2022)

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

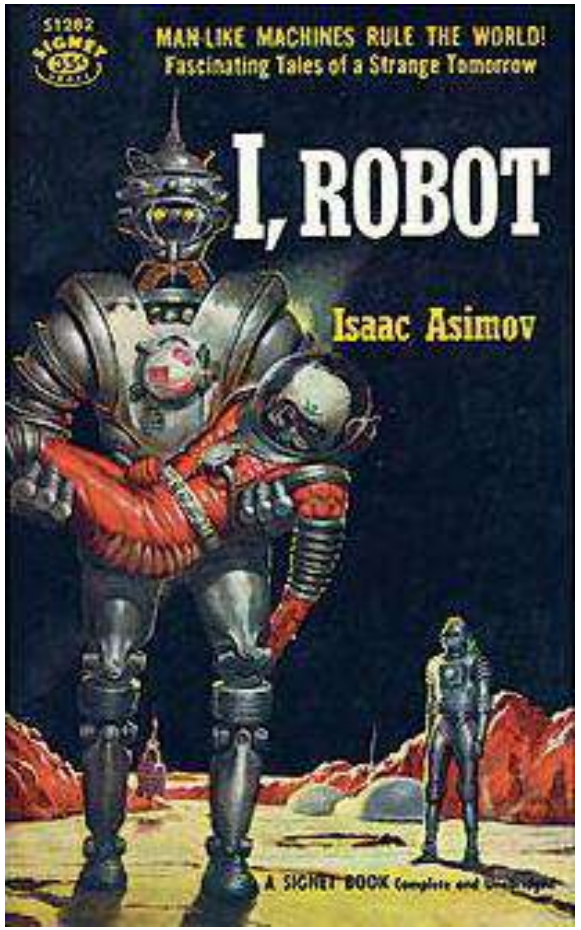
- Provides explanations for AI predictions
- Makes predictions intuitive for users
- Fosters trust in AI implementatio
- Reduces opacity in black-box models (Holzinger et al., 2019)

ILLUSTRATION: KAL

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

- „Knowledge has its risks, but should our reaction be to stop at risk?
- Knives are made with a handle, so that they can be grasped without danger, electrical wires are insulated…
- Safety may not be perfect the greater the human ability, the more advanced it will be.”



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

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

- 10 years after Zeiler & Fergus's paper on understanding neural networks
- explaining requires assumptions
- 1st argument:
- Explainable machine learning methods have some success scenarios and others where they fail. New methods are often used to showcase their performance, but subsequent research identifies failure modes. For example, feature attribution methods can detect spurious correlations, but their generality is debated. Despite this, many researchers believe feature attributions are useful tools for detecting certain forms of spurious correlations. (Ribeiro et al., 2016); (Adebayo, 2022; Adebayo et al., 2020)

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

- 2nd argument:
- Explanation algorithms are susceptible to adversarial attacks, which alter the function f to maintain constant predictions but arbitrary explanations, as demonstrated in LIME and SHAP, and in gradient-based methods. (Slack et al., 2020); (Dombrowski et al., 2019)

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

- 3rd argument:
- The third argument for explaining functions is theoretical, as There is no universal learning algorithm in machine learning theory. This means every algorithm must choose a preference for some functions, making it unlikely to have a universal explaination for functions. (Wolpert, 1996)

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

- Explainable machine learning researchers now believe it's generally impossible to explain arbitrary functions, but must identify specific conditions and assumptions for explainability, with the specific form and practical relevance yet to be discovered.

Co-funded by the European Union

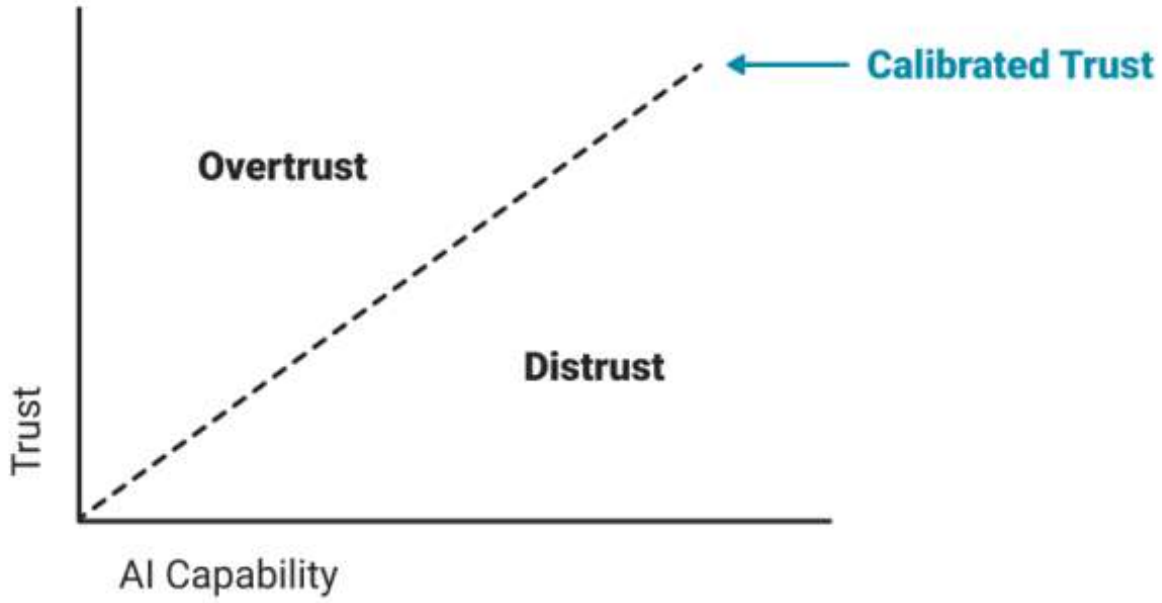
Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

- understanding of the model which can be indicative of trust calibration; if a user understands a model, it follows that they will know when to trust it and when not to
- Some research has focused on other aspects of explainability such as detecting fairness issues

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

- correspondence between a person's trust in AI and the AI's capabilities (Lee & Moray 1994)



Trust Calibration: Correspondence between Trust and AI Capability

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

- Too litle distrust Algorithm averions (Dietvorst et. al. 2015)
- TRUST
- Too much over-reliance Automation bias (Lee et. al. 2004)

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

- Ideally AI explanation should help the trust calibration process by providing more insight into the AI decision-making process and allowing users to properly adjust their level of trust according to the actual reliability of the AI system (Giannotti, et. al., 2023)

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

- While the field of xAI has seen many papers that propose to explain arbitrary functions, researchers now believe that this is generally not possible.

- Instead, we must identify the particular conditions and assumptions under which explainability is possible. In most interesting cases, the specific form and practical relevance that these assumptions will take are yet to be discovered.

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

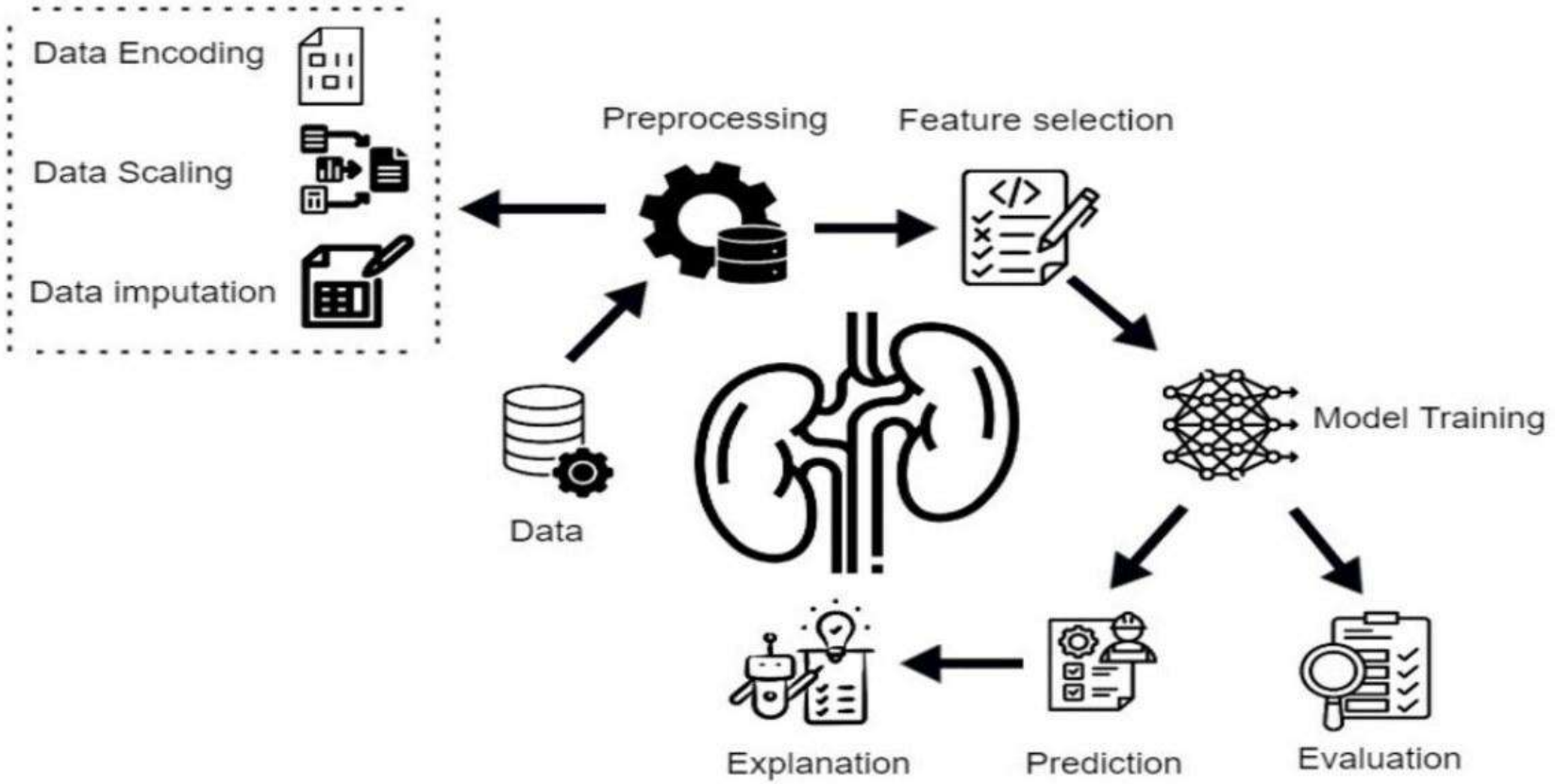
- Explanations help to assess the capacity of new systems better than simply examining their accuracy scores on validation sets

- If it is unclear – even for the designer – how a system operates, it is possible that it reaches decisions based on undesired biases

- While a black box can help to improve performance, it cannot drive progress (instrumentalism)

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.



- From: Arif, M. S., Rehman, A. U., & Asif, D. (2024). Explainable Machine Learning Model for Chronic Kidney Disease Prediction. *Algorithms*, *17*(10), 443. https://doi.org/10.3390/a17100443

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

- A study at the University of Coventry has developed an AI-based tool called the UK Deceased Donor Kidney Transplant Outcome Prediction (UK-DTOP) to improve decision-making about kidney transplants. The tool, developed using XGBoost, was found to be more accurate than the Kidney Donor Risk Index, which currently assesses potential outcomes for donors.

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

- The UK-DTOP was more accurate than the Kidney Donor Risk Index, which has a statistical score consistently above 0.72. The researchers plan to use the findings to improve organ allocation and outcomes for patients in need. However, more work is needed to improve the model further, such as ensuring no missing data could influence outcome predictions.

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

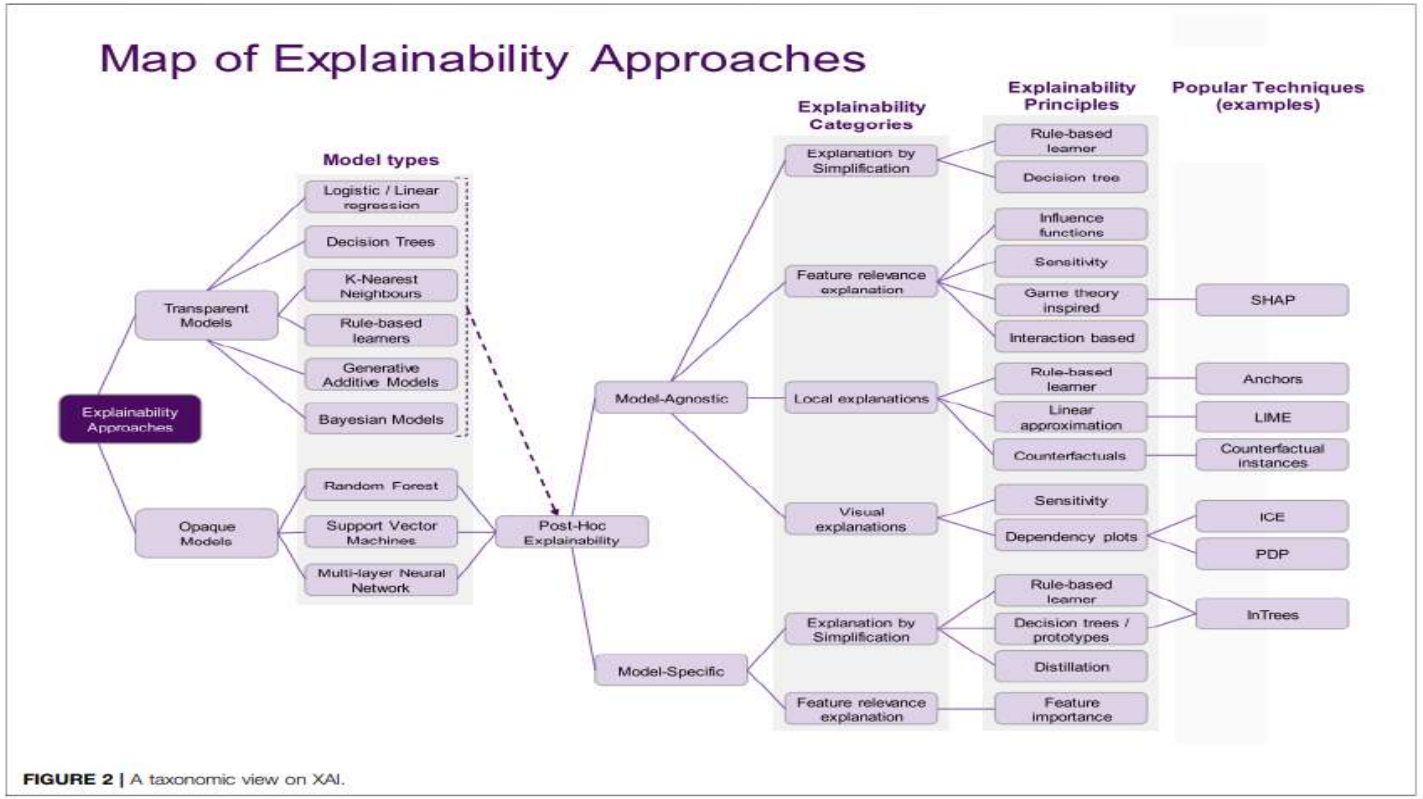


FIGURE 2 | A taxonomic view on XAI.

From: Principles and Practice of Explainable Machine Learning, Vaishak Belle, Ioannis Papantonis, 2021

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

1) Intrinsic or post hoc

2) Result of xAI method:

- Statistics or visualization or features

- Model internals

- Data points (e. g. counterfactuals)

- Iterpretable (surrogate) model

3) Model-specific or model-agnostic

4) Local or global

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

Major differences			
	GDPR	HIPAA	PIPL
Accountability	Accountability premise expressed clearly.	Accountability premise expressed clearly.	Accountability premise expressed clearly.
Sensitive personal data	Sensitive personal data handling must have specific justification. Allows people to request that an institution delete their data, a right known as the right to be forgotten.	Medical records and other personal information can't be altered or deleted (the information is stored forever).	Sensitive personal data processing requires separate written approval.
Consent	Justification for handling personal data in a permissible manner. high quality if it can be trusted.	The consent form must be accompanied by a Privacy Notice. HIPAA consent must be separate from the normal informed consent for treatment. The consent must be in plain language and signed by the patient. The patient may revoke that consent at any time.	A valid legal foundation for processing personal data in a legitimate manner (permission or statutory consent free scenarios). Under the PIPL, legitimate interest is not a legitimate foundation.
Data export	Data processors are directly obligated.	requires that all parties involved are responsible for protecting the data. This includes health organizations like hospitals, clinics, and insurance as well as companies like Dropbox or cloud service providers.	The sole notion of a data handler, with no legal distinction made between a data controller and a data processor.
Sanctions	Data protection authorities can impose administrative fines. Data subjects can bring civil actions for compensation.	Depending on the degree of responsibility, civil monetary penalties for HIPAA infractions can range from $137 to $68,928 per infraction. Intentional infractions may also result in criminal sanctions, which include fines and sometimes even jail time.	Administrative punishment: fines on personal data handler, warning, confiscation of illegal gains, suspension or termination of operation, public disclosure of non-compliance.

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

With trust calibration

- If novice users are shown explanations that don't help calibrate trust, their misuse of the system can increase

- There are different forms of explanations. Sørmo et al. (2005) differentiate five objectives of explanations in AI:
  - Justification (Why?)
  - Transparency (How?)
  - Relevance (Why this question?)
  - Conceptualization (clarify meaning)
  - Learning (teaching)

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

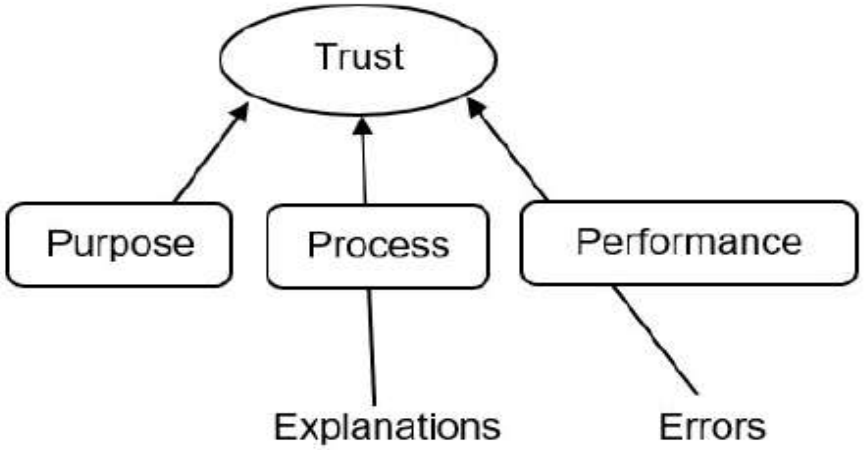


Figure 6: General overview of the effects of explanations and errors

From: Burr, Johannes. (2018). Trust in Artificial Intelligence - Effects of Explanations and Classification Errors. 10.13140/RG.2.2.35478.29765.

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

Users must have an understanding of the capabilities and limitations of an AI system when making high-stakes decisions

Explainability makes clear what the system “knows” while uncertainty awareness reveals what the system does not “know”

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

Trust is well calibrated when the human sets their trust level appropriately to the AI's capabilities, accepting the output of a competent system but employing other resources or their own expertise to compensate for AI errors

Poorly calibrated trust reduces team performance because the human trusts erroneous AI outputs or does not accept correct ones

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

When every decision is high-stakes, the human must be able to calibrate their trust in the AI quickly and adjust their trust level on a case-by-case basis

We refer to this process as rapid trust calibration

Understanding what the AI does not know is also extremely important for creating a suitable mental model of the AI's capabilities

To do this, the AI system must be able to estimate the uncertainty in its outputs

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

When every decision is high-stakes, the human must be able to calibrate their trust in the AI quickly and adjust their trust level on a case-by-case basis

We refer to this process as rapid trust calibration

Understanding what the AI does not know is also extremely important for creating a suitable mental model of the AI's capabilities

To do this, the AI system must be able to estimate the uncertainty in its outputs

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

Quantifying epistemic uncertainty requires the model to have a means of accurately estimating how far away new inputs are from the data distribution it was trained on

A common approach is to use Bayesian methods, whereby epistemic uncertainty is captured as uncertainty in the model parameters or as uncertainty in function space using, for example, Gaussian processes

We suggest using the framework provided by Lasswell's communication model to structure future research efforts

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

Harmonizing Technology and Ethics

Pathway to Harmonization:
Technological progress with ethical integrity
Regulatory compliance
Safeguarding public trust and welfare

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

- Perfectly accurate and perfectly intelligible systems are not yet a reality, but they are not required; clinically useful tools are likely to blend islands of AI automation with rule-based workflows that improve interpretability of the decision-making process and allow effective human supervision. Determining the optimal design for AI systems depends on factors beyond performance, with considerations of interpretability and the nature of the intended human-AI interaction also informing many design and engineering choices.

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

- To effectively appraise research and new market tools, clinicians need an awareness of the common problems that arise in AI training and deployment. In parallel, teams developing AI image classifiers need continual dialogue with clinicians that is centered on the trade-offs that their chosen designs entail; to secure trust, and facilitate a smooth translation to the clinic.

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

Instrumental Value	Noninstrumental Value
Slavery	Friendship
Good Value	Bad Value
Friendship	Slavery

From: Nyholm, S. (2023). Responsibility gaps, value alignment, and meaningful human control over artificial intelligence. In Risk and responsibility in context. Routledge. https://doi.org/10.4324/9781003276029-14

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

Instrumental Value	Noninstrumental/Intrinsic Value
Slavery	Friendship
Good Value	Bad Value
Friendship	Slavery

From: Nyholm, S. (2023). Responsibility gaps, value alignment, and meaningful human control over artificial intelligence. In Risk and responsibility in context. Routledge. https://doi.org/10.4324/9781003276029-14

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

Instrumental Value	Noninstrumental/Intrinsic Value
Slavery	Friendship
Good Value	Bad Value
Friendship	Slavery

From: Nyholm, S. (2023). Responsibility gaps, value alignment, and meaningful human control over artificial intelligence. In Risk and responsibility in context. Routledge. https://doi.org/10.4324/9781003276029-14

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.

- Conclusion

- We need an interdisciplinary approach…
- Includes clinical practice, policy, patient care, public health, and ethics

- … to integrate AI technologies responsibly in healthcare.

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EACEA. Neither the European Union nor the granting authority can be held responsible for them.