From Epistemic Opacity to Trustworthy Medical AI: is transparency the pathway?

Mariana Vitti Rodrigues

Yes!

Philosophy & Technology (2019) 32:661–683
https://doi.org/10.1007/s13347-018-0330-6

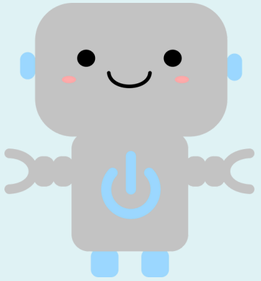
RESEARCH ARTICLE

Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?

John Zerilli [1] · Alistair Knott [2] · James Maclaurin [1] · Colin Gavaghan [3]

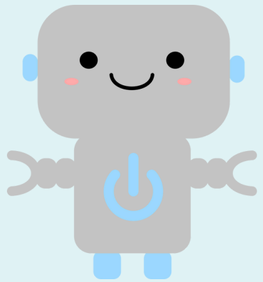
Received: 9 May 2018 / Accepted: 28 August 2018 / Published online: 5 September 2018

“While we do not deny that transparency and explainability are important desiderata in algorithmic governance, we worry that automated decision-making is being held to an unrealistically high standard here, possibly owing to an unrealistically high estimate of the degree of transparency attainable from human decision-makers”. (Zerilli et al. 2019, p. 662)



Objective

To investigate kinds of ***epistemic opacity*** and propose an account of ***transparency*** grounded on the notion of ***epistemic risk***.



Summary

- Introduction
- Typology of Opacity
- Epistemic Opacity, Interpretability, Explainability and Explanatory Adequacy
- Rethinking transparency
- Final Remarks

Introduction

Epistemic Opacity

A process is ***epistemically opaque*** relative to a ***cognitive agent X*** at time t just in case X does not know at t all of the ***epistemically relevant elements*** of the process. (Humphreys 2016).

Epistemic Opacity might create *Trust*, *Justificatory* and *Responsibility* **GAPS** when an agent needs to act upon *uncertainty*, *risk*, and *ignorance*.

GAPS: { *easy gaps (**in practice** opacity)*; *hard gaps (**in principle** opacity)*

(For trust gaps cf. Gouveia & Malik 2024; Gouveia, 2025)

Epistemic in principle opacity of Machine Learning: *trade-offs dilemma*

More performance, accuracy, efficiency (in terms of time/money), less human mistakes and biases.
Less transparency, understanding, interpretability (susceptible to ‘machine biases’).

Alternatives

- Restrict the use of high-risk AI in medicine and healthcare;
- Accept the risks of using epistemic opaque systems;
- ***Search for strategies*** to enhance the reliability of epistemically opaque models.

Search for strategies

Search for strategies

Explainable AI: global and local interpretability (model-of-the-model).

Search for strategies

Explainable AI: global and local interpretability (model-of-the-model).

Human in the loop: requiring humans to be part of computer-assisted decision-making processes (Crootof, Kaminski, Price, 2023).

Search for strategies

Explainable AI: global and local interpretability (model-of-the-model).

Human in the loop: requiring humans to be part of computer-assisted decision-making processes (Crootof, Kaminski, Price, 2023).

Machine-in-the-loop: propose ‘Evaluative AI’ designed to present cons and pros of human generated hypotheses.(Miller 2023)

Search for strategies

Explainable AI: global and local interpretability (model-of-the-model).

Human in the loop: requiring humans to be part of computer-assisted decision-making processes (Crootof, Kaminski, Price, 2023).

Machine-in-the-loop: propose ‘Evaluative AI’ designed to present cons and pros of human generated hypotheses (Miller 2023)

External Validation: formulating reliability criteria to assess the model’s performance without opening the Blackbox (Durán 2021).

.

Search for strategies

Explainable AI: global and local interpretability (model-of-the-model).

Human in the loop: requiring humans to be part of computer-assisted decision-making processes (Crootof, Kaminski, Price, 2023).

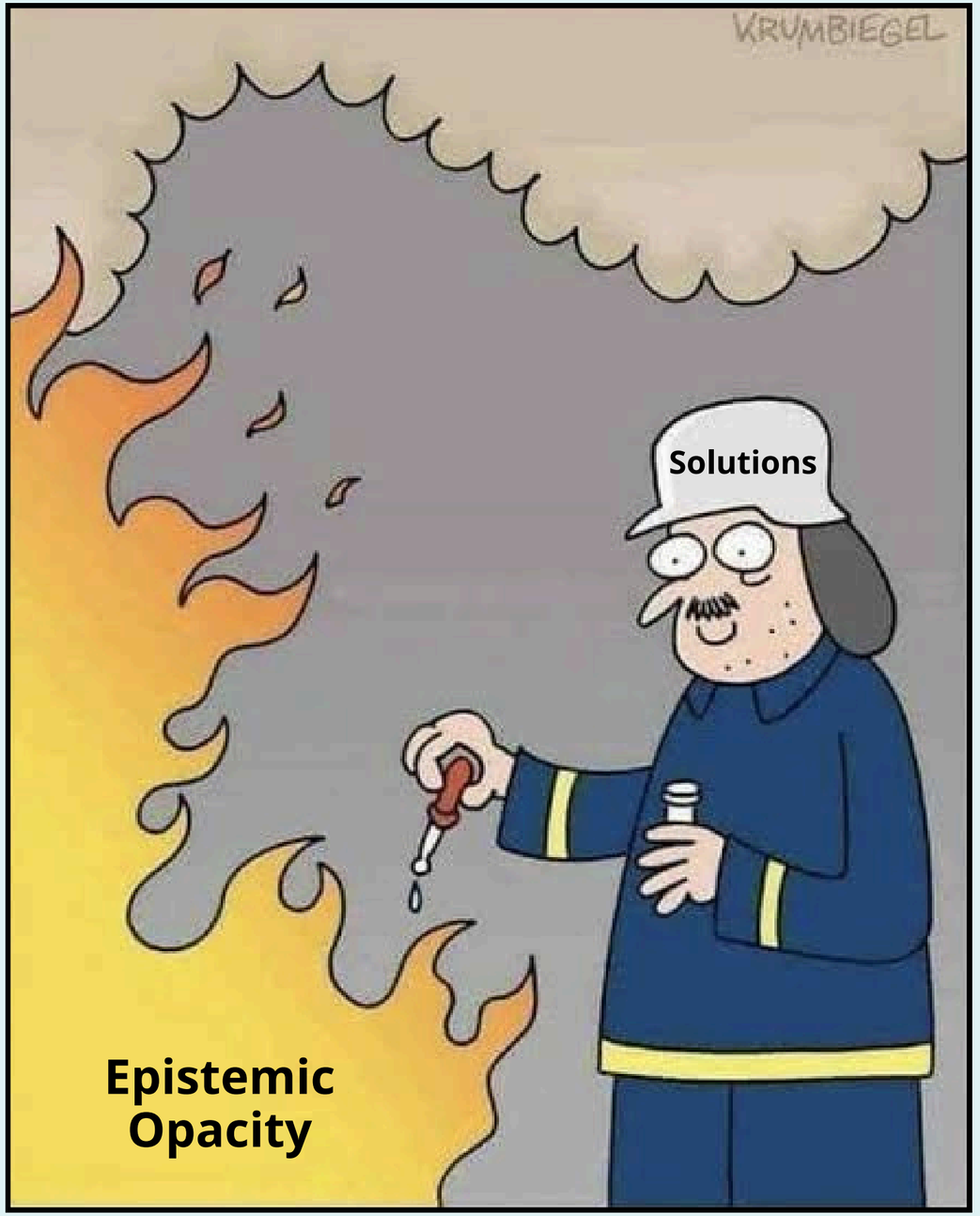
Machine-in-the-loop: propose ‘Evaluative AI’ designed to present cons and pros of human generated hypotheses (Miller 2023)

External Validation: formulating reliability criteria to assess the model’s performance without opening the Blackbox (Durán 2021).

Automating abduction: create a model-of-the-model that automatically generate explanations of the blackbox’s output. (Gouveia & Malik 2024; Gouveia, 2025).

Taking a step back....

... to investigate kinds and sources of Epistemic Opacity



Typology of Opacity

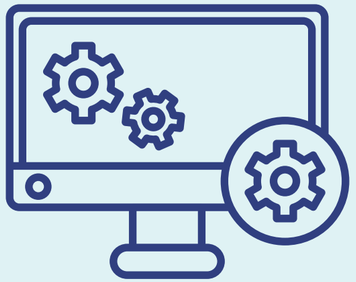
TYPOLOGY OF OPACITY

- State or corporation secrecy;
- Technical illiteracy;
- Algorithmic Opacity;
- Model Opacity;
- Hardware Opacity;
- Cognitive Opacity;
- Data Opacity;
- Methodological Opacity.

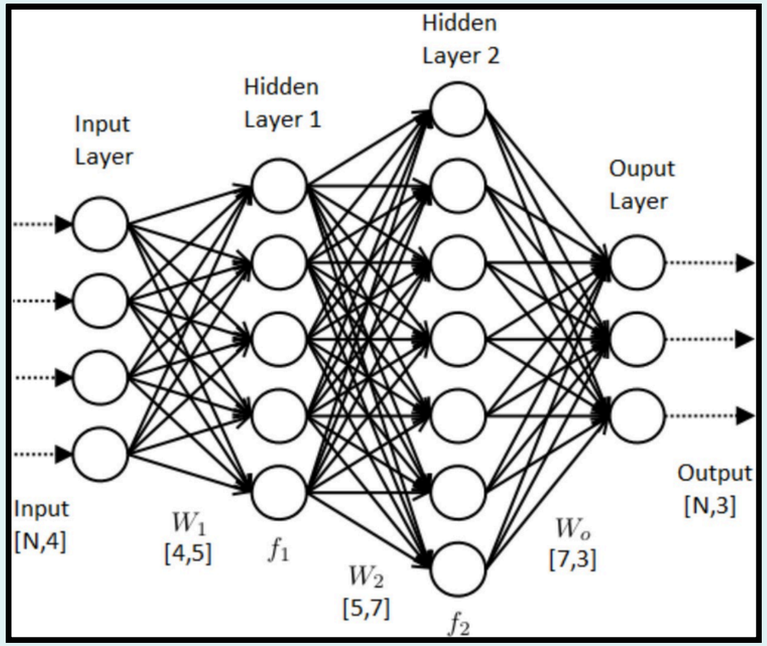
TYPOLOGY OF OPACITY

- State or corporation secrecy;
- Technical illiteracy;
- **Algorithmic Opacity;**
- **Model Opacity;**
- Hardware Opacity;
- **Cognitive Opacity**;
- **Data Opacity**;
- Methodological Opacity.

Machine Learning Opacity and BlackBox Medicine



Complexity



Blackbox medicine:
"the use of opaque computational models to make decisions related to health care."
(Price 2015, p. 421)

Cognitive Opacity and Abductive Reasoning

Human Cognition

Reasoning from Surprise to Inquiry

The surprising fact, C, is observed.
But if A were true, C would be a matter of course.
Hence, there is reason to suspect that A is true.
(EP2: 231, CP 5.189).

Cognitive Opacity, Medical Reasoning, and Tacit Knowledge

Human Cognition

Tacit knowledge underlies decision-making and reasoning processes encompassing embodied and embedded biases, values, and assumptions, etc., that are difficult to express propositionally and, thus, be translated in formal languages.

Data Opacity

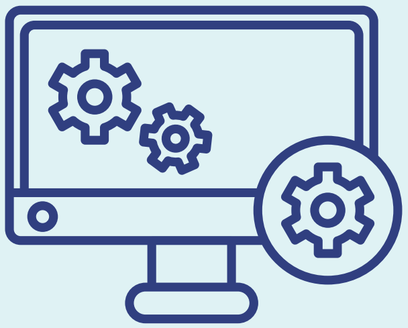
Data Infrastructures

“The paradox consists of the observation that, despite their epistemic value as ‘given’, data are clearly made.”(Leonelli 2015, p. 813)

Data Opacity and Data Journeys

(i) the process of collecting, mining, integrating and storing data that configures the ‘datafication' of the object of study, enabling the mechanical/automated processing of data.
(ii) access to data, “cleaned” and mechanically structured, for analysis and interpretation.
(iii) De-contextualization, re-contextualization, retrieval, reuse, reanalysis, repurposing, repositioning.

In practice or in principle epistemic opacity?



Complexity ***Human cognition*** ***Data Infrastructures***

What difference does this difference make?

Epistemic Opacity, Interpretability, Explainability and Explanatory Adequacy

....Rethinking Transparency....

Epistemic Opacity, Interpretability, Explainability and Explanatory Adequacy

"In machine learning, the concept of interpretability is both important and slippery"
(Lipton 2018)

(For a very good discussion on Interpretability, Explainability and Explanatory adequacy see Mittelstadt 2021)

Epistemic Opacity and Interpretability

A process is ***interpretable*** relative to a ***cognitive agent X*** at time t just in case X ***has access*** at t to ***relevant elements*** of the process.

Cognitive agent X - Artificial agent (model-of-the-model approaches)

Epistemically relevant elements - system´s features responsible for an output given a specific input (local interpretability), model's rationale (global interpretability).

Epistemic Opacity and Explainability

A process is ***explainable*** relative to a ***cognitive agent X*** at time t just in case X knows at t all of the ***epistemically relevant elements*** of the process.

Cognitive agent X - Human expert.

Epistemically relevant elements - decision criteria involving the internal functioning and external behaviour of the system (information, meaning and context adequacy)

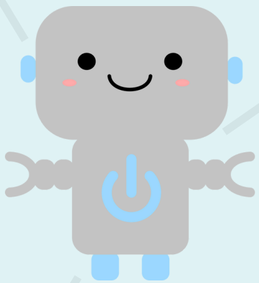
Epistemic Opacity and Explanatory Adequacy

A process is ***explainable*** relative to a ***cognitive agent X*** at time t just in case X knows at t all of the ***epistemically relevant elements*** of the process.

Cognitive agent X - Human from different expertises and non-expert agents.

Epistemically relevant elements - decision criteria expressed in user-friendly vocabulary

Rethinking transparency

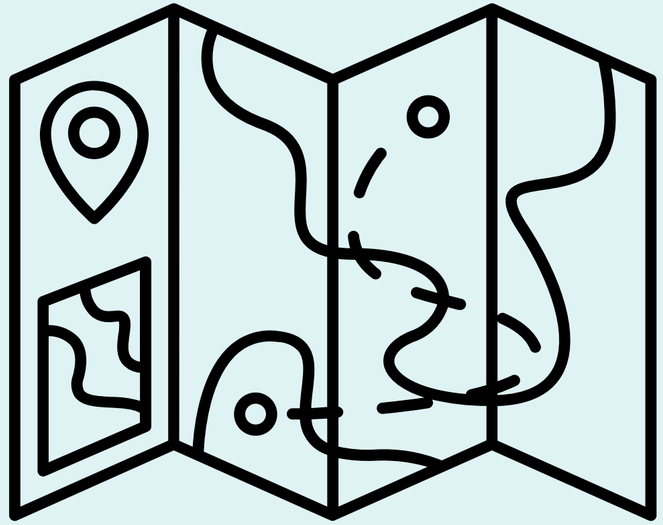


Transparency: navigating islands of opacity

- Identify ***sources of epistemic opacity*** in order to propose targeted strategies;

- Different opacities require different strategies - in terms of interpretability, explainability, and explanatory adequacy - to build the pathway - ***transparent data journeys*** - towards Trustworthy Medical AI;

- But, what to do when opacity is inevitable?



Transparency as risk documentation

Transparency as risk documentation

Inductive risk [strict def., inspired by Hempel 1965]:

"[...] is the chance that one will be wrong in accepting (or rejecting) a scientific hypothesis" (Douglas 2000).

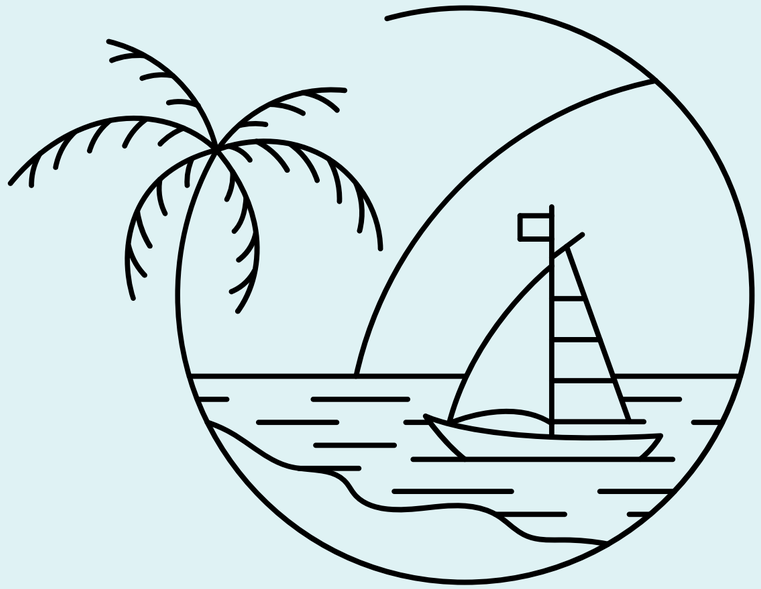
"[…] the decision of how much evidence is enough to accept or reject a hypothesis". (Biddle 2016, p. 192).

Transparency as risk documentation

Inductive risk [strict def., inspired by Hempel 1965]:
"[...] is the chance that one will be wrong in accepting (or rejecting) a scientific hypothesis" (Douglas 2000).
"[…] the decision of how much evidence is enough to accept or reject a hypothesis". (Biddle 2016, p. 192).

Epistemic risk [broad def.; cf. *inductive risk* in Douglas 2000]:
there is inductive risk in several points of the research process (choice of methodology, gathering and characterization of evidence, interpretation of data; background assumptions, set of tested subjects, policy, conceptual definitions). (Biddle 2016)

Final Remarks



Transparency as risk documentation

- Transparency should be understood in degrees;
- Interpretability, explainability, and explanatory adequacy are steps towards transparency and, thus, Trustworthy Medical AI;
- Transparency is achieved by building ***reliable data journeys*** to support 'islands of opacity' in a given pipeline:
- Document ***epistemic risk*** involved in decision-making processes: ***assumptions, judgements, reasons, decisions, and values involved in reasoning processes in the training, test, use, and maintainance of algorithmic models for data analysis.***

Transparency as risk documentation

- ***Question:*** What kind of epistemic opacity could be considered epistemologically justifiable and, therefore, ethically acceptable in the implementation of algorithmic systems for data analysis?
- ***Hypothesis*:** The solution is local: achieving transparency means to build reliable data journeys that support islands of [in practice/in principle] epistemic opacity;
- ***Challenges***: openness, generalizability, initiatives of standardization (FAIR, CARE, TRUST), up-to-date ethical guidelines.

Collaborations and Feedback!

Suggestions? Comments? Ideas?

Acknowledgements

Claus Emmeche, Henrik Nielsen, João Cortese, Natale Cavaçana, Adamantios Koupis, Vincent Müller, Sascha Fink, Juan Durán, Samantha Copeland, Maria Eunice Quilici Gonzalez, Mariana Cláudia Broens, Ettore Bresciani.

By Shane Rounce - unsplash

REFERENCES

Biddle, J. Inductive Risk, Epistemic Risk, and Overdiagnosis of Disease. Perspectives on Science 2016; 24 (2): 192–205.

Crootof, Kaminski, Price II (2023) Humans in the Loop, In: Vanderbilt Law Review, 76(2) 429.

Douglas, H. (2000) Inductive Risk and Values in Science. In: Philosophy of Science 67: 559-579.

Durán, J. (2021) Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare. In: Artificial Intelligence, v. 297.

Gouveia, S.S. (2025) The Ethics of Artificial Intelligence in Medicine: Preliminary Remarks. glob. Philosophy 35, 4 (2025).

Gouveia, S.S., Malík, J (2024). Crossing the Trust Gap in Medical AI: Building an Abductive Bridge for xAI. Philos. Technol. 37, 105.

Hansson, Sven Ove, "Risk", *The Stanford Encyclopedia of Philosophy* (Summer 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.)

Hempel, C. (1965) *Science and Human Values.* In: Aspects of Scientific Explanation and other Essays in the Philosophy of Science, 81-96. New York: The Free Press.

Humphreys, P. (2016). Epistemic Opacity and Epistemic Inaccessibility. https:// wordpress.its.virginia.edu/ Paul_Humphreys_Home_Page/ files/ 2016/ 02/ epistemic-opacity-and-epistemic-inaccessibility. pdf. Accessed 15 April 2024.

Leonelli, S. (2015). What counts as scientific data? A relational framework. In: Philosophy of Science, vol. 82:810-821.

Miller, T. (2023). Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai. In Proceedings of the 2023 ACM conference on fairness, accountability, and transparency (p. 333-342).

Mittelstadt, Brent, Interpretability and Transparency in Artificial Intelligence (2021). in Carissa Véliz (ed.), The Oxford Handbook of Digital Ethics (online edn, Oxford Academic, 10 Nov. 2021).

Price II, W, N. (2015) Black-Box Medicine (September 22, 2014). 28 Harv. J.L. & Tech. 419.

Zerilli, J., Knott, A., Maclaurin, J. et al. (2019) Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? Philos. Technol. 32, 661–683.

Thank you for your attention!!

Mariana Vitti Rodrigues

mvittirodrigues@gmail.com