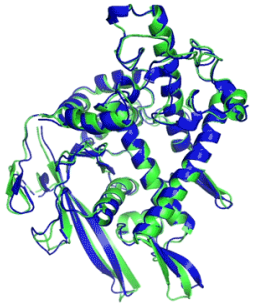
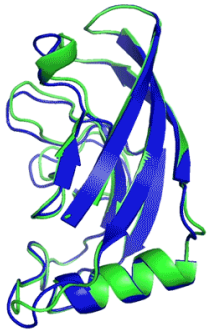
Justification, transparency and computational reliabilism

TU Delft

Juan M. Durán



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

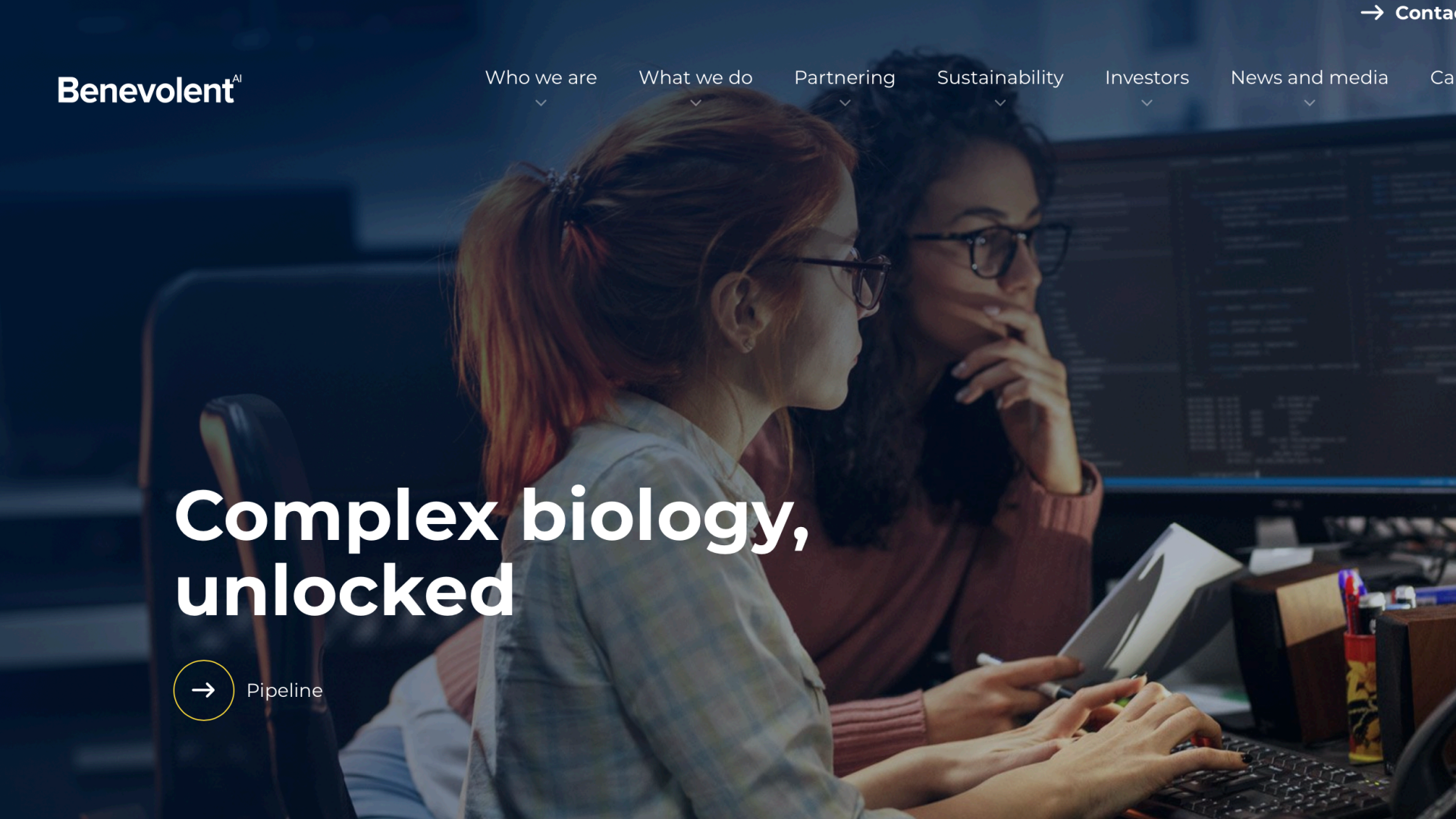


T1049 / 6y4f
93.3 GDT
(adhesin tip)

Experimental result
Computational prediction

TU Delft

Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2


Conta
Benevolent AI
Who we are
What we do
Partnering
Sustainability
Investors
News and media
Ca
Complex biology, unlocked
Pipeline

Wu & Zhang:
Automated Inference on Criminality Using Face Images

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

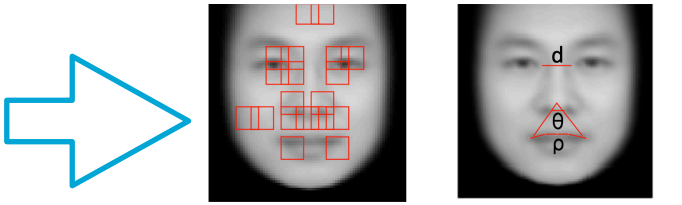
Figure 1. Sample ID photos in our data set.

Wu & Zhang:
Automated Inference on Criminality Using Face Images

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.



(a) (b)

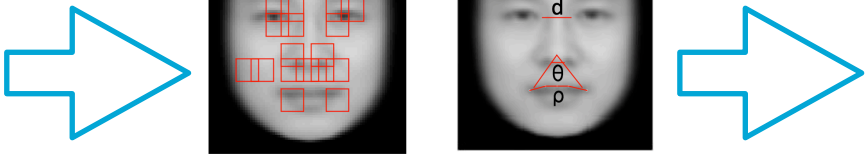
Figure 4. (a) FGM results; (b) Three discriminative features ρ, d and θ.

Wu & Zhang:
Automated Inference on Criminality Using Face Images

(a) Three samples in criminal ID photo set S_c.

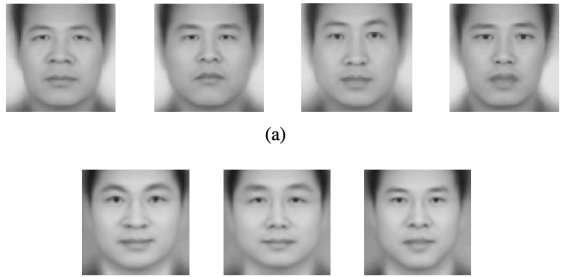
(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.



(a) (b)

Figure 4. (a) FGM results; (b) Three discriminative features ρ, d and θ.



(a)

(b)

Figure 9. (a) The four subtypes of criminal faces; (b) The three subtypes of non-criminal faces.

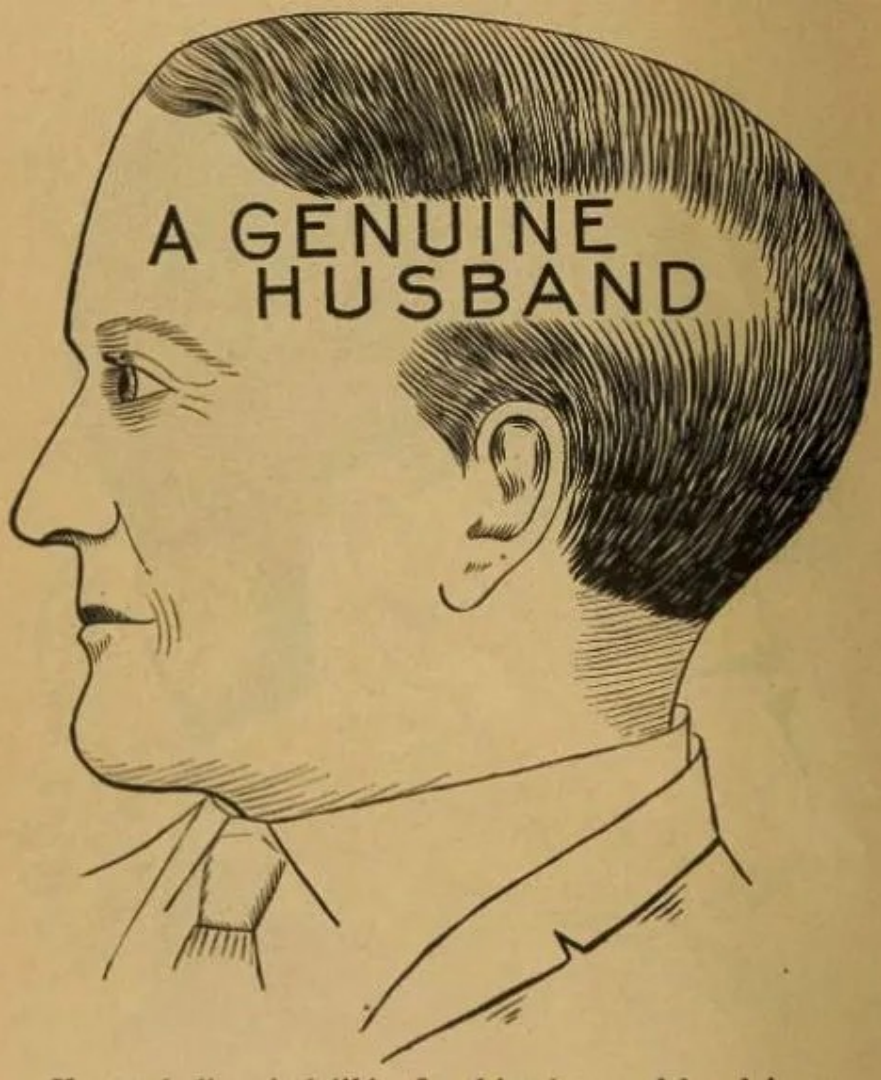
Wu and Zhang

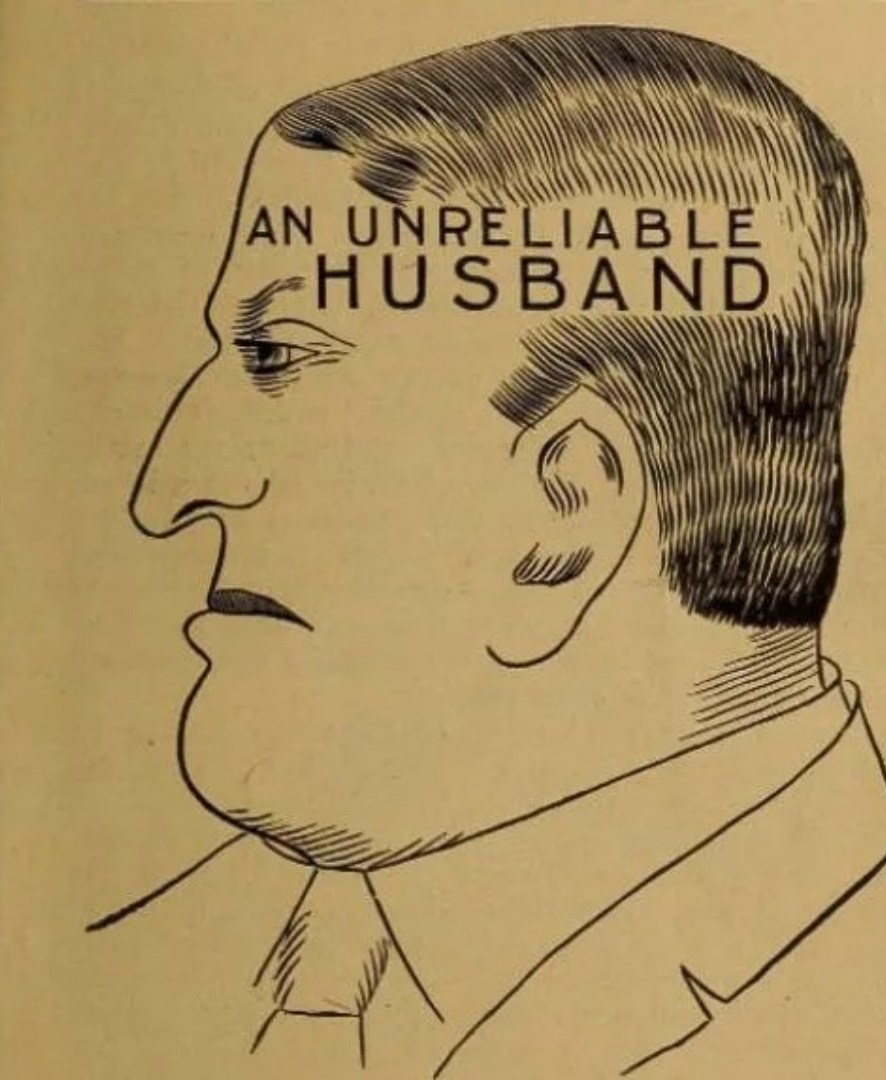
"The two manifolds formed respectively by the two data sets of criminal and non-criminal faces are concentric, with the manifold for the non-criminal face images lying in the kernel with a smaller span. This newly discovered knowledge suggests **a law of normality for faces of non-criminals: Given the race, gender and age, the faces of general law-biding public have a greater degree of resemblance compared with the faces of criminals.** In other words, criminals have a significantly higher degree of dissimilarity in facial appearance than **normal population**" (2016, 2)

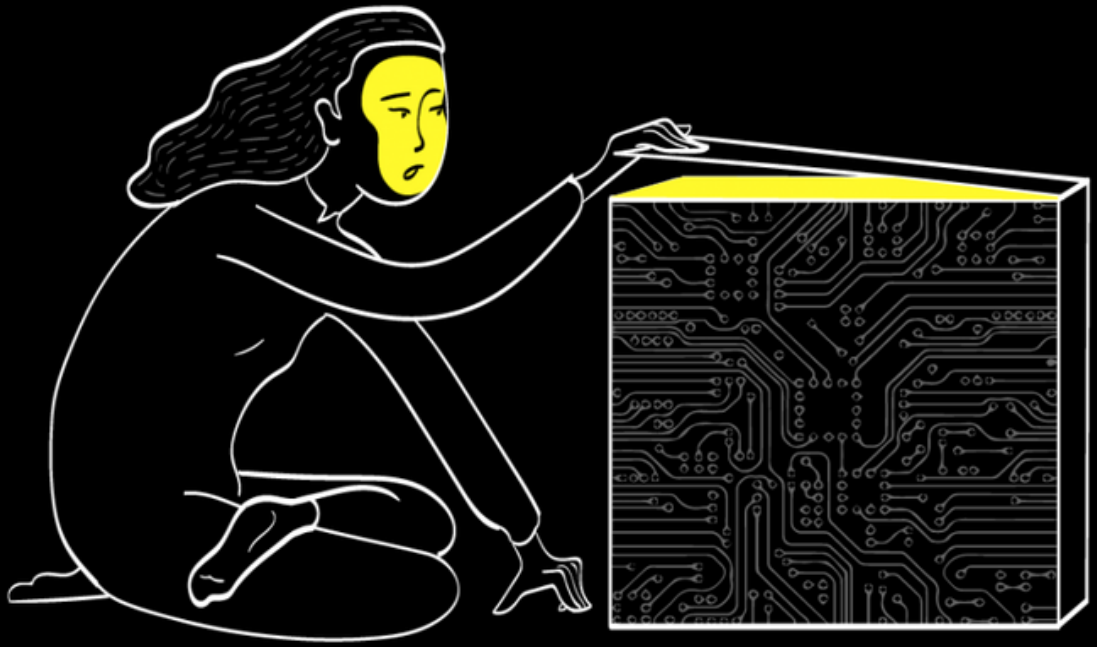
Wu and Zhang

"The two manifolds formed respectively by the two data sets of criminal and non-criminal faces are concentric, with the manifold for the non-criminal face images lying in the kernel with a smaller span. This newly discovered knowledge suggests **a law of normality for faces of non-criminals: Given the race, gender and age, the faces of general law-biding public have a greater degree of resemblance compared with the faces of criminals.** In other words, criminals have a significantly higher degree of dissimilarity in facial appearance than **normal population**" (2016, 2)

"With a Ph.D in computer science, we know all too well "garbage in and garbage out". However, some of our critics seemed to suggest that machine learning tools cannot be used in social computing simply because no one can prevent **the garbage of human biases from creeping in**. We do not share their pessimism. **Like most technologies, machine learning is neutral.**" (2017, 2)


A GENUINE
HUSBAND


AN UNRELIABLE
HUSBAND



Two epistemologies for ML

Two epistemologies for ML

- Transparency/Interpretability
- Computational Reliabilism

Two epistemologies for ML

- Transparency/Interpretability

Two epistemologies for ML

- Transparency/Interpretability
  - Requires to “open” the algorithm
  - Justification is secured by a “third-party” algorithm (interpretable predictors, XAI)

Two epistemologies for ML

Transparency/Interpretability

Synthese (2021) 198:9211–9242
https://doi.org/10.1007/s11229-020-02629-9

The explanation game: a formal framework for interpretable machine learning

David S. Watson[1] · Luciano Floridi[1,2]

Received: 23 October 2019 / Accepted: 12 March 2020 / Published online: 3 April 2020

Abstract
We propose a formal framework for interpretable machine learning. Combining elements from statistical learning, causal interventionism, and decision theory, we design an idealised *explanation game* in which players collaborate to find the best explanation(s) for a given algorithmic prediction. Through an iterative procedure of questions and answers, the players establish a three-dimensional Pareto frontier that describes the optimal trade-offs between explanatory accuracy, simplicity, and relevance. Multiple rounds are played at different levels of abstraction, allowing the players to explore overlapping causal patterns of variable granularity and scope. We characterise the conditions under which such a game is almost surely guaranteed to converge on a (conditionally) optimal explanation surface in polynomial time, and highlight obstacles that will tend to prevent the players from advancing beyond certain explanatory thresholds. The game serves a descriptive and a normative function, establishing a conceptual space in which to analyse and compare existing proposals, as well as design new and improved solutions.

Keywords Algorithmic explainability · Explanation game · Interpretable machine learning · Pareto frontier · Relevance

1 Introduction

Machine learning (ML) algorithms have made enormous progress on a wide range of tasks in just the last few years. Some notable recent examples include mastering perfect information games like chess and Go (Silver et al. 2018), diagnosing skin cancer (Esteva et al. 2017), and proposing new organic molecules (Segler et al. 2018). These technical achievements have coincided with the increasing ubiquity of ML, which

✉ David S. Watson
david.watson@oii.ox.ac.uk

[1] Oxford Internet Institute, University of Oxford, 41 Saint Giles, Oxford OX1 3LW, UK

[2] The Alan Turing Institute, British Library, 96 Euston Road, Kings Cross, London NW1 2DB, UK

Transparency in Complex Computational Systems

Kathleen A. Creel*†

Scientists depend on complex computational systems that are often ineliminably opaque, to the detriment of our ability to give scientific explanations and detect artifacts. Some philosophers have suggested treating opaque systems instrumentally, but computer scientists developing strategies for increasing transparency are correct in finding this unsatisfying. Instead, I propose an analysis of transparency as having three forms: transparency of the algorithm, the realization of the algorithm in code, and the way that code is run on particular hardware and data. This targets the transparency most useful for a task, avoiding instrumentalism by providing partial transparency when full transparency is impossible.

1. Introduction. Scientists depend on complex computational systems to process their big data, but these systems are not always transparent. Physicists within the Large Hadron Collider's (LHC) Compact Muon Solenoid working group are considering using deep learning algorithms to sort particle collision events and discard the uninteresting ones (Duarte et al. 2018). The new algorithms for doing so, while faster than the old, are complex enough that their decisions cannot be reconstructed in terms of why some events were interesting and thus saved and why others were discarded

Received November 2018; revised October 2019.

*To contact the author, please write to: University of Pittsburgh, Department of History and Philosophy of Science, 1101 Cathedral of Learning, 4200 Fifth Avenue, Pittsburgh, PA 15260; e-mail: kac284@pitt.edu.

†I am grateful for helpful comments from and discussions with Holly Andersen, Robert Batterman, Nora Mills Boyd, Liam Kofi Bright, Mazviita Chirimuuta, Roger Creel, Javier Duarte, Mahi Hardalupas, Paul Humphreys, Benjamin Jantzen, Johannes Lenhard, Sabina Leonelli, Jake Levinson, Edouard Machery, Sandra Mitchell, Elinor Nichols, Kathleen Nichols, Aaron Novick, Olivia Ordoñez, William Penn, Rebecca Traber, Porter Williams, Eric Winsberg, and two anonymous reviewers. Thanks also to generous audiences at Philosophical Perspectives on Data-Intensive Science in Hannover; Models and Simulations 8 in Columbia, SC; the Machine Learning Workshop in Irvine, CA; and Science and Art of Simulation IV in Stuttgart.

Philosophy of Science, 87 (October 2020) pp. 568–589. 0031-8248/2020/8704-0002$10.00

i.org/10.1086/709729 Published online by Cambridge University Press

Minds and Machines
https://doi.org/10.1007/s11023-019-09502-w

ORIGINAL ARTICLE

The Pragmatic Turn in Explainable Artificial Intelligence (XAI)

Andrés Páez[1]

Received: 11 March 2019 / Accepted: 27 May 2019

Abstract
In this paper I argue that the search for explainable models and interpretable decisions in AI must be reformulated in terms of the broader project of offering a pragmatic and naturalistic account of understanding in AI. Intuitively, the purpose of providing an explanation of a model or a decision is to make it understandable to its stakeholders. But without a previous grasp of what it means to say that an agent understands a model or a decision, the explanatory strategies will lack a well-defined goal. Aside from providing a clearer objective for XAI, focusing on understanding also allows us to relax the factivity condition on explanation, which is impossible to fulfill in many machine learning models, and to focus instead on the pragmatic conditions that determine the best fit between a model and the methods and devices deployed to understand it. After an examination of the different types of understanding discussed in the philosophical and psychological literature, I conclude that interpretative or approximation models not only provide the best way to achieve the objectual understanding of a machine learning model, but are also a necessary condition to achieve post hoc interpretability. This conclusion is partly based on the shortcomings of the purely functionalist approach to post hoc interpretability that seems to be predominant in most recent literature.

Keywords Explainable artificial intelligence · Understanding · Explanation · Model transparency · Post-hoc interpretability · Machine learning · Black box models

1 Introduction

The main goal of Explainable Artificial Intelligence (XAI) has been variously described as a search for explainability, transparency and interpretability, for ways of validating the decision process of an opaque AI system and generating trust in the

✉ Andrés Páez
andrespaez@gmail.com

[1] Department of Philosophy, Universidad de los Andes, Carrera 1 No. 18A-12 (G-533), Bogotá, DC 111711, Colombia

Published online: 29 May 2019

AI & SOCIETY (2021) 36:585–595
https://doi.org/10.1007/s00146-020-01066-z

OPEN FORUM

Artificial intelligence and the value of transparency

Joel Walmsley[1]

Received: 6 January 2020 / Accepted: 25 August 2020 / Published online: 8 September 2020

Abstract
Some recent developments in Artificial Intelligence—especially the use of machine learning systems, trained on big data sets and deployed in socially significant and ethically weighty contexts—have led to a number of calls for "transparency". This paper explores the epistemological and ethical dimensions of that concept, as well as surveying and taxonomising the variety of ways in which it has been invoked in recent discussions. Whilst "outward" forms of transparency (concerning the relationship between an AI system, its developers, users and the media) may be straightforwardly achieved, what I call "functional" transparency about the inner workings of a system is, in many cases, much harder to attain. In those situations, I argue that contestability may be a possible, acceptable, and useful alternative so that even if we cannot understand how a system came up with a particular output, we at least have the means to challenge it.

Keywords Transparency · Explainability · Contestability · Machine learning · Bias

1 Introduction

Alongside, and arguably because of, some of the most recent technical developments in Artificial Intelligence, the last few years have seen a growing number of calls for various forms of transparency[1] within and about the field. For example, the 2019 report from the European Commission's High-Level Expert Group on AI—entitled Ethics Guidelines for Trustworthy AI—features the notion of transparency prominently, and the European Union's General Data Protection Regulation (GDPR) includes the stipulation that, when a person is subject to an automated decision based on their personal information, he or she has "the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision".[2] In part, these calls respond to an epistemic limitation; machine learning techniques, together with the use of "Big Data" for training purposes, mean that many AI systems are both too complex for a complete understanding, and faster and more powerful than human cognition (at least, on the relatively narrow set of tasks for which AI is designed). Of course, in many cases, "complete understanding" is neither desired nor required; we are perfectly happy to interact with technology by adopting Dennettian[3] "intentional" or "design" stances (rather than the more complete but cumbersome "physical stance") so long as the system functions correctly, and the respects in which it is not transparent are roughly neutral along ethical, political or commercial dimensions. But given that we increasingly and preferentially trust AI systems, and that we do rely on them to make decisions, recommendations and predictions in a variety of socially significant and morally weighty contexts (for example, not just regarding what video to watch or what product to buy, but also whether a person qualifies for job interview, a loan, or for parole), the call for transparency has acquired an ethical (and legal) dimension too. Furthermore, these dimensions intersect: both in the popular media, and within specialised technology circles, we find a steady stream of examples and anecdotes of problematic biases, prejudices and other errors that have been automated and reinforced—albeit, sometimes, unwittingly—because of our reliance on AI systems that we do not fully comprehend.

✉ Joel Walmsley
j.walmsley@ucc.ie

[1] Department of Philosophy, University College Cork, Cork, Ireland

[1] Sometimes also discussed under the heading of "explainability," "explicability" or "understandability" (e.g., by Robbins 2019) or with reference, also, to "accountability," "intelligibility" and "interpretability" (e.g., in Floridi et al. 2018).

[2] General Data Protection Regulation, Recital 71, available at https://gdpr-info.eu/recitals/no-71/.

[3] See Dennett (1971).

Two epistemologies for ML

- Transparency/Interpretability
- Computational Reliabilism

Two epistemologies for ML

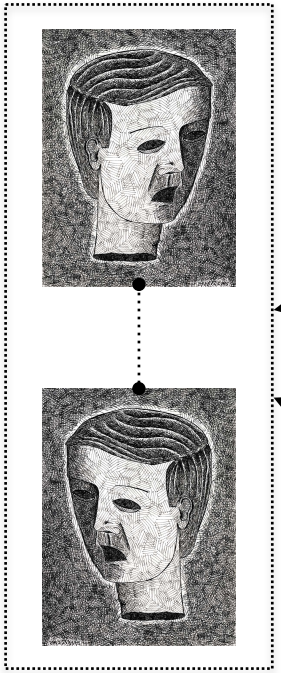
- Transparency/Interpretability
- Computational Reliabilism
  - Accepts “black box” algorithms
  - Justification comes from securing the reliability of the algorithm
    - Depends on self-regulating practices, methods, and processes with reliability-conferring property — external to the algorithm.

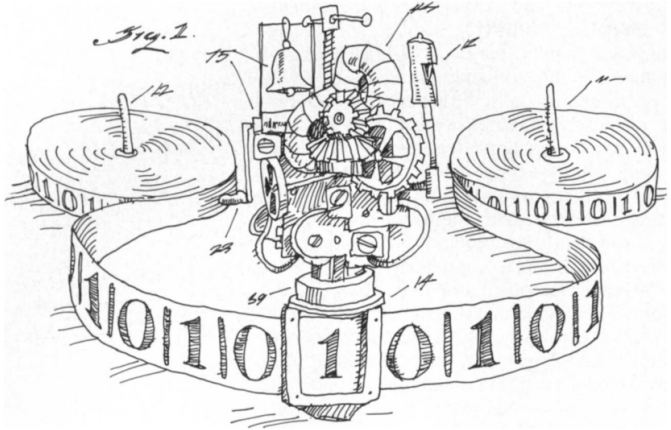
What is process reliabilism? (≈ Goldman 1979)

Process reliabilism: An individual's belief is justified in case it is produced by a reliable belief-forming process (or sequence of processes)

A reliable belief-forming process has the tendency to produce beliefs that are *true* rather than *false*.

Justification via CR (a sketch)





Reliable

Reliability Indicators

Justified

Two epistemologies for ML

- Transparency/Interpretability
- Computational Reliabilism (type/token)
  - RI_1 Technical robustness of algorithms
  - RI_2 Computer-based scientific practice
  - RI_3 Social construction of reliability

(Type) Reliability indicators for ML

- RI_1 Technical robustness of algorithms
- RI_2 Computer-based scientific practice
- RI_3 Social construction of reliability

TU Delft

Reliability indicators for ML

- RI_1 Technical robustness of algorithms

(Token) Reliability indicators for ML

RI_1 Technical robustness of algorithms

- Verification and validation methods (→ accuracy)
- Robustness analysis
- Data choices
- Parametrization and hyper-parametrization choices
- A history of (un)successful implementations
- Error treatment (e.g., uncertainty quantification)

(Durán 2018; Durán & Formanek 2018)

Minds and Machines (2018) 28:645–666
https://doi.org/10.1007/s11023-018-9481-6

CrossMark

Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism

Juan M. Durán[1] · Nico Formanek[2]

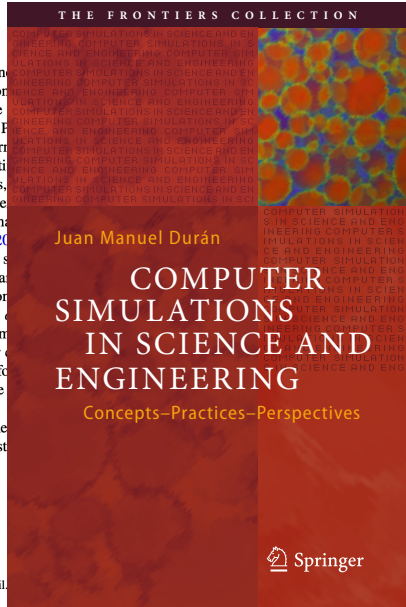
Received: 22 May 2018 / Accepted: 12 October 2018 / Published online: 29 October 2018

Abstract
Several philosophical issues in conn
assumption that results of simulation
the debate on the experimental role
169(3):483–496, 2009; Morrison in P
computer data (Barberousse and Vorr
tions and the changing face of scienti
lishing, Barcelona, 2013; Humphreys,
and the changing face of scientific e
ing, Barcelona, 2013), and the explan
Int Stud Philos Sci 22(3):277–292, 20
2017). The aim of this article is to s
that results of computer simulations a
reliable processes. After a short recor
the article elaborates extensively on
process reliabilism with computer sim
with a discussion of four sources for
and validation, robustness analysis fo
cessful implementations, and the role

Keywords Computer simulations · Re
validation · Robustness analysis · Hist

✉ Juan M. Durán
j.m.duran@tudelft.nl

Nico Formanek
formanek@hlrs.de; nico.formanek@gmail.



TU Delft

Reliability indicators for ML

- RI_1 Technical robustness of algorithms
- RI_2 Computer-based scientific practice

TU Delft

Reliability indicators for ML

- RI_1 Technical robustness of algorithms
- RI_2 Computer-based scientific practice
  - Acting mechanisms (causality and otherwise)
  - Model-Trainning practices / Epistemic and moral values
  - Representation
  - Some degree of theoretical, conceptual coherence (against JBDA)
    - Background knowledge (e.g., knowledge of the learned weights)
    - Some morphism between the implemented methodology and an established scientific methodology (e.g., use of natural kinds, accepted cut-off-values — e.g., when melanoma is carcinogenic)
    - Against JBDA
      - JBDA cannot contribute to the confirmation of theories over time
      - JBDA cannot enable inferences to similar phenomena

(Type) Reliability indicators for ML

- RI_1 Technical robustness of algorithms
- RI_2 Computer-based scientific practice
- RI_3 Social construction of reliability

(Type) Reliability indicators for ML

- RI_1 Technical robustness of algorithms
- RI_2 Computer-based scientific practice
- RI_3 Social construction of reliability
  - Peer and social acceptability of ML output
  - Favorable, successful, scientifically valid use of ML/ML output (e.g., develop of new hypothesis or expand into new strategies)
  - Coherence with established body of scientific beliefs
  - Coherence with scientists' theoretical and other commitments
  - Realization of the epistemic and ethical values

TU Delft

Wu & Zhang:
Automated Inference on Criminality Using Face Images

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

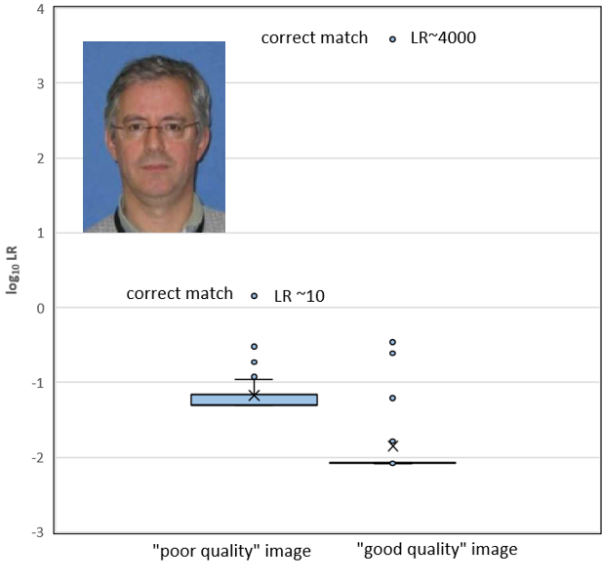
Figure 1. Sample ID photos in our data set.

- The system is forced to pick one out of two classes: "criminal" or "not-criminal"
- It fossilises concepts, such as "criminality"
- It fails to account for or include values (e.g., "innocent until proved guilty")
- It is disconnected from a larger body of knowledge on:
  - Social construction of crime
  - The socio-economic basis of criminality
  - The psychology of criminals

How does CR work?
A case in Forensic AI

Forensic face comparison

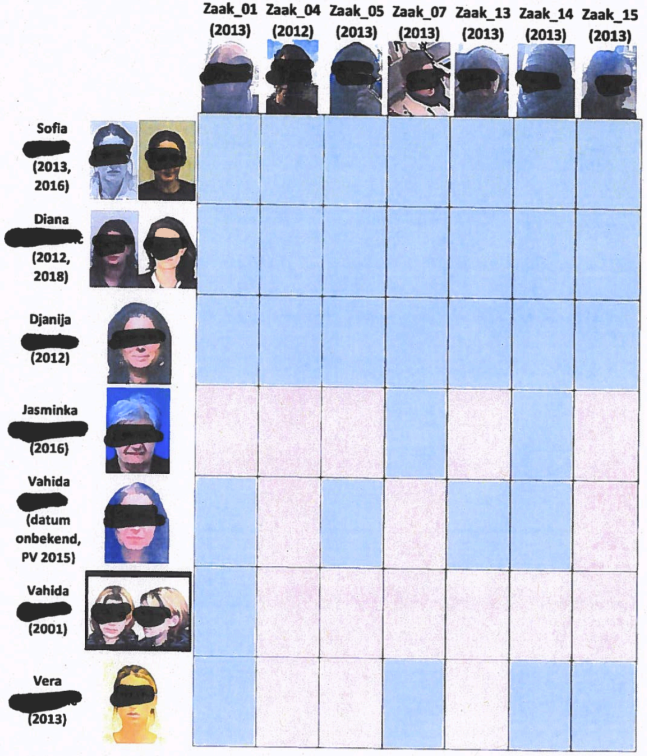
"Poor quality" image



"Good quality" image

Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

Tabel 2: Overzicht van alle gezichtbeeldvergelijkingen. Groen: geselecteerd voor gedetailleerde vergelijking. Rood: op basis van voorselectie uitgesloten.



Tabel 3: Samenvatting van de conclusies van de vergelijkingen (met cijfers volgens bijgaande legenda), waarbij a, b en c staat voor de conclusies van de afzonderlijke onderzoekers en 'Consensus' na bespreking van de resultaten (legenda: zie Tabel 4)

Zaak	Verdachte	a	b	c	Consensus
01	Sofia	2	2	2	2
01	Diana	-3	-3	-3	-3
01	Djanija	-4	-3	-4	-4
01	Vahida (jonger)	-2	-2	-4	-3
01	Vahida (ouder)	-3	1	-4	-2
01	Vera	-4	-2	-4	-4
04	Sofia	4	3	3	4
04	Diana	-4	-5	-4	-4
04	Djanija	-4	-3	-4	-4
05	Sofia	5	3	3	4
05	Diana	-4	-5	-4	-4
05	Djanija	-4	-3	-5	-4
05	Vahida (ouder)	-4	-4	-4	-4
07	Sofia	0	1	2	1
07	Diana	-2			
07	Djanija	-3			
07	Jasminka	-2			
07	Vera	-4			

Zaak	Verdachte	a	b	c	Consensus
13	Sofia	3	3	2	3
13	Diana	-3	-3	-3	-3
13	Djanija	-3	-2	-4	-3
13	Vahida (ouder)	-3	-2	-3	-3
14	Sofia	1	2	2	2
14	Diana	-2	-2	-3	-3
14	Djanija	-4	-2	-4	-3
14	Jasminka	-1	-3	-4	-2
14	Vahida (ouder)	-1	-2	-4	-2
14	Vera	-3	-2	-4	-3
15	Sofia	3	4	3	3
15	Diana	-3	-5	-4	-4
15	Djanija	-4	-5	-4	-4
15	Vera	-4	-3	-4	-4

Tabel 4: Legenda bij Tabel 3

Legenda

Cijfer	De bevindingen van het onderzoek zijn:
5	**extreem veel waarschijnlijker** wanneer H1 waar is dan wanneer H2 waar is
4	**zeer veel waarschijnlijker** wanneer H1 waar is dan wanneer H2 waar is
3	**veel waarschijnlijker** wanneer H1 waar is dan wanneer H2 waar is
2	**waarschijnlijker** wanneer H1 waar is dan wanneer H2 waar is
1	**iets waarschijnlijker** wanneer H1 waar is dan wanneer H2 waar is
0	**ongeveer even waarschijnlijk** wanneer H1 waar is als wanneer H2 waar is

Cijfer	De bevindingen van het onderzoek zijn:
-1	**iets waarschijnlijker wanneer** H2 waar is dan wanneer H1 waar is
-2	**waarschijnlijker** wanneer H2 waar is dan wanneer H1 waar is
-3	**veel waarschijnlijker** wanneer H2 waar is dan wanneer H1 waar is
-4	**zeer veel waarschijnlijker** wanneer H2 waar is dan wanneer H1 waar is
-5	**extreem veel waarschijnlijker** wanneer H2 waar is dan wanneer H1 waar is

Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

Tabel 2: Overzicht van alle gezichtbeeldvergelijkingen. Groen: geselecteerd voor gedetailleerde vergelijking. Rood: op basis van voorselectie uitgesloten.

Tabel 3: Samenvatting van de conclusies van de vergelijkingen (met cijfers volgens bijgaande legenda), waarbij a, b en c staat voor de conclusies van de afzonderlijke onderzoekers en 'Consensus' na bespreking van de resultaten (legenda: zie Tabel 4)

Zaak	Verdachte	a	b	c	Consensus
01	Sofia	2	2	2	2
01	Diana	-3	-3	-3	-3
01	Djanija	-4	-3	-4	-4
01	Vahida (jonger)	-2	-2	-4	-3
01	Vahida (ouder)	-3	1	-4	-2
01	Vera	-4	-2	-4	-4
04	Sofia	4	3	3	4
04	Diana	-4	-5	-4	-4
04	Djanija	-4	-3	-4	-4
05	Sofia	5	3	3	4
05	Diana	-4	-5	-4	-4
05	Djanija	-4	-3	-5	-4
05	Vahida (ouder)	-4	-4	-4	-4
07	Sofia	0	1	2	1
07	Diana	-2			

Zaak	Verdachte	a	b	c	Consensus
13	Sofia	3	3	2	3
13	Diana	-3	-3	-3	-3
13	Djanija	-3	-2	-4	-3
13	Vahida (ouder)	-3	-2	-3	-3
14	Sofia	1	2	2	2
14	Diana	-2	-2	-3	-3
14	Djanija	-4	-2	-4	-3
14	Jasminka	-1	-3	-4	-2
14	Vahida (ouder)	-1	-2	-4	-2
14	Vera	-3	-2	-4	-3
15	Sofia	3	4	3	3
15	Diana	-3	-5	-4	-4
15	Djanija	-4	-5	-4	-4
15	Vera	-4	-3	-4	-4

Tabel 5: Vergelijking van conclusies vorige en huidige rapportage

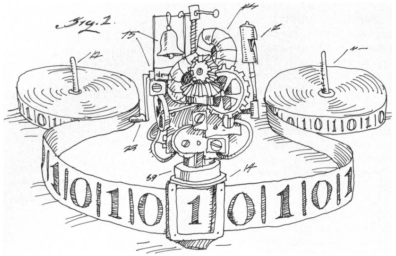
Zaaknr.	Rapportage aanvraag 002 (25 maart 2016)	Huidige aanvraag 003
Zaak 01	Waarschijnlijker	Waarschijnlijker
Zaak 04	Zeer veel waarschijnlijker	Zeer veel waarschijnlijker
Zaak 05	Zeer veel waarschijnlijker	Zeer veel waarschijnlijker
Zaak 07	--	Iets waarschijnlijker
Zaak 13	Veel waarschijnlijker	Veel waarschijnlijker
Zaak 14	Iets waarschijnlijker	Waarschijnlijker
Zaak 15	Veel waarschijnlijker	Veel waarschijnlijker

nderzoek zijn:
er H2 waar is
er H2 waar is
er H2 waar is
er H2 waar is
er H2 waar is
H2 waar is

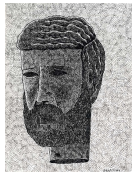
Cijfer	De bevindingen van het onderzoek zijn:
-1	**iets waarschijnlijker wanneer** H2 waar is dan wanneer H1 waar is
-2	**waarschijnlijker** wanneer H2 waar is dan wanneer H1 waar is
-3	**veel waarschijnlijker** wanneer H2 waar is dan wanneer H1 waar is
-4	**zeer veel waarschijnlijker** wanneer H2 waar is dan wanneer H1 waar is
-5	**extreem veel waarschijnlijker** wanneer H2 waar is dan wanneer H1 waar is

Forensic AI

Evidence

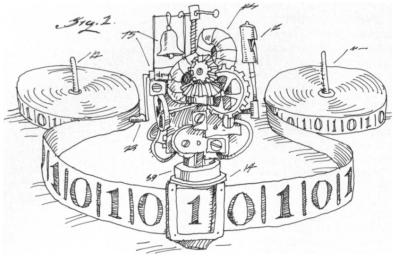


Suspects

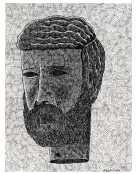


Forensic AI

Evidence



Suspects



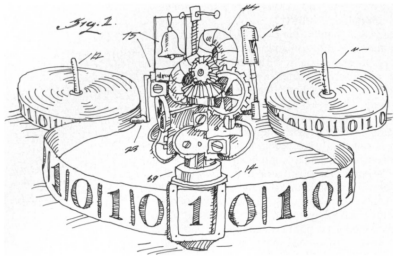
ψ_1

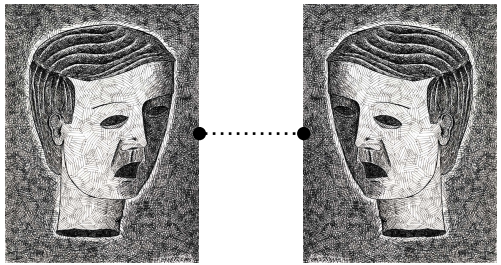
ψ_2

ψ_3

$\psi_3 > \psi_2 > \psi_1$

Forensic AI



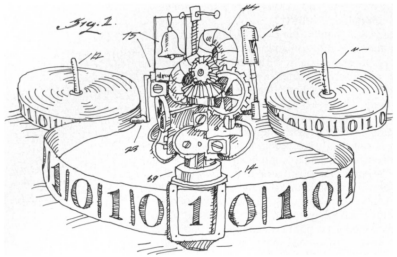


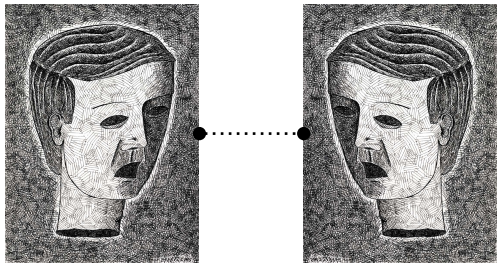
RI$_1$ Technical robustness of algorithms

- Verification and Validation
- Robustness analysis
- Redundancy mechanisms
- Quality of data

TU Delft

Forensic AI





RI_1 Technical robustness of algorithms

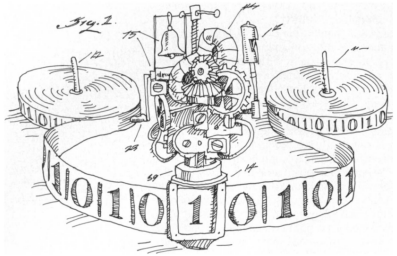
- Verification and Validation
- Robustness analysis
- Redundancy mechanisms
- Quality of data

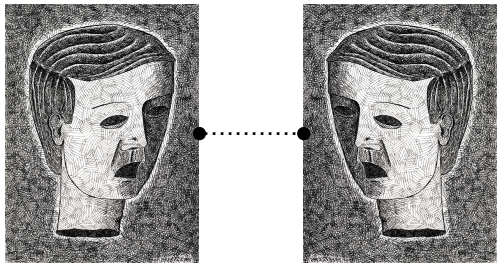
RI_2 Computer-based scientific practice

- Implementation of quality assurance methods
- “We implement well-known and tested libraries”
- Biometrical techniques

TU Delft

Forensic AI





RI_1 Technical robustness of algorithms

- Verification and Validation
- Robustness analysis
- Redundancy mechanisms
- Quality of data

RI_2 Computer-based scientific practice

- Implementation of quality assurance methods
- "We implement well-known and tested libraries"
- Biometrical techniques

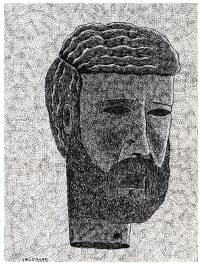
RI_3 Social construction of reliability

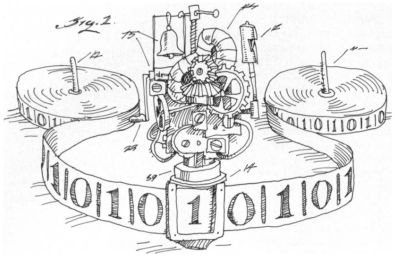
- Debates about the output (e.g., soundness, applicability, etc.)
- Contrast with other non-computational methods (e.g., use of code review, expert assessment and report)

TU Delft

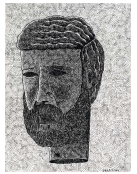
Forensic AI

Evidence

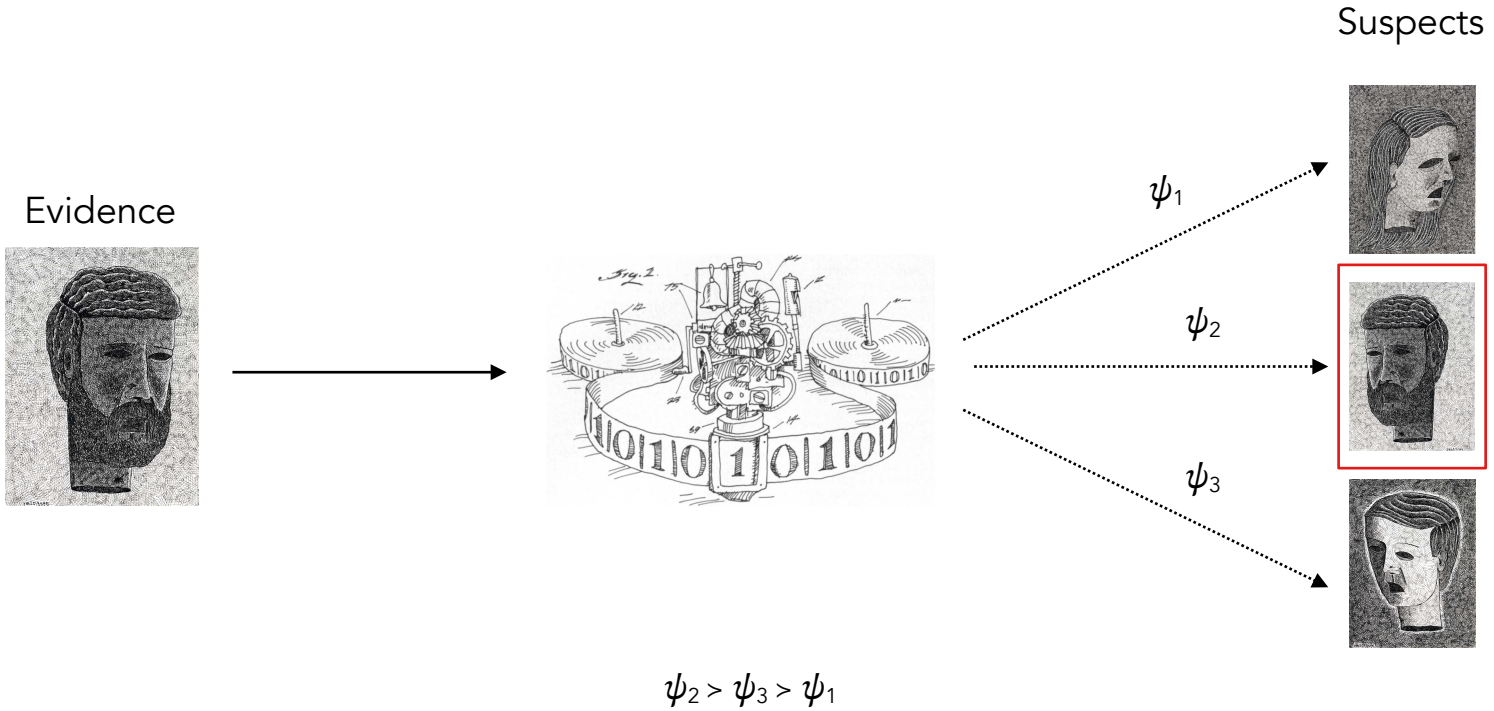




Suspects



Forensic AI



$\psi_2 > \psi_3 > \psi_1$

Lights and shadows of CR

Lights

- Externalist:
  - Credibility does not depend on third-party algorithms and are independent on the quality of our insight and our mental capabilities to understand it
- Each reliability indicator is independently justified
- There are various reliability indicators (“non-centralized”)
- It detects markers for non-credible results
- It accommodates different needs: in some systems, V&V are more valuable, whereas in others is theoretical coherence

Shadows

- It can't be automated (or the operationalization is very difficult)
  - It might hamper algorithmic utility
- It is unclear the precedence, order, and weight of each indicator
- *The tyranny of the few*: a few indicators might suffice for crediting reliability/non-reliability to an algorithm (when missing indicators or when conflict of indicators)

View of Delft, Johannes Vermeer, 1660-1661. Rijksmuseum, Amsterdam

Thank you!

j.m.duran@tudelft.nl

Draft paper: juanmduran.net & PhilPapers