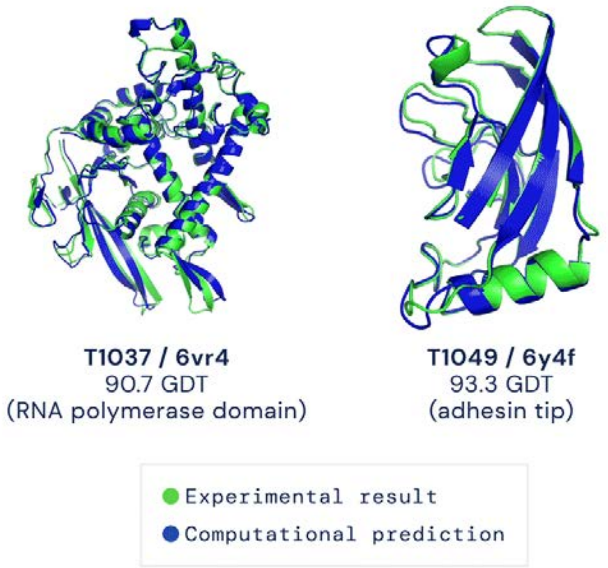
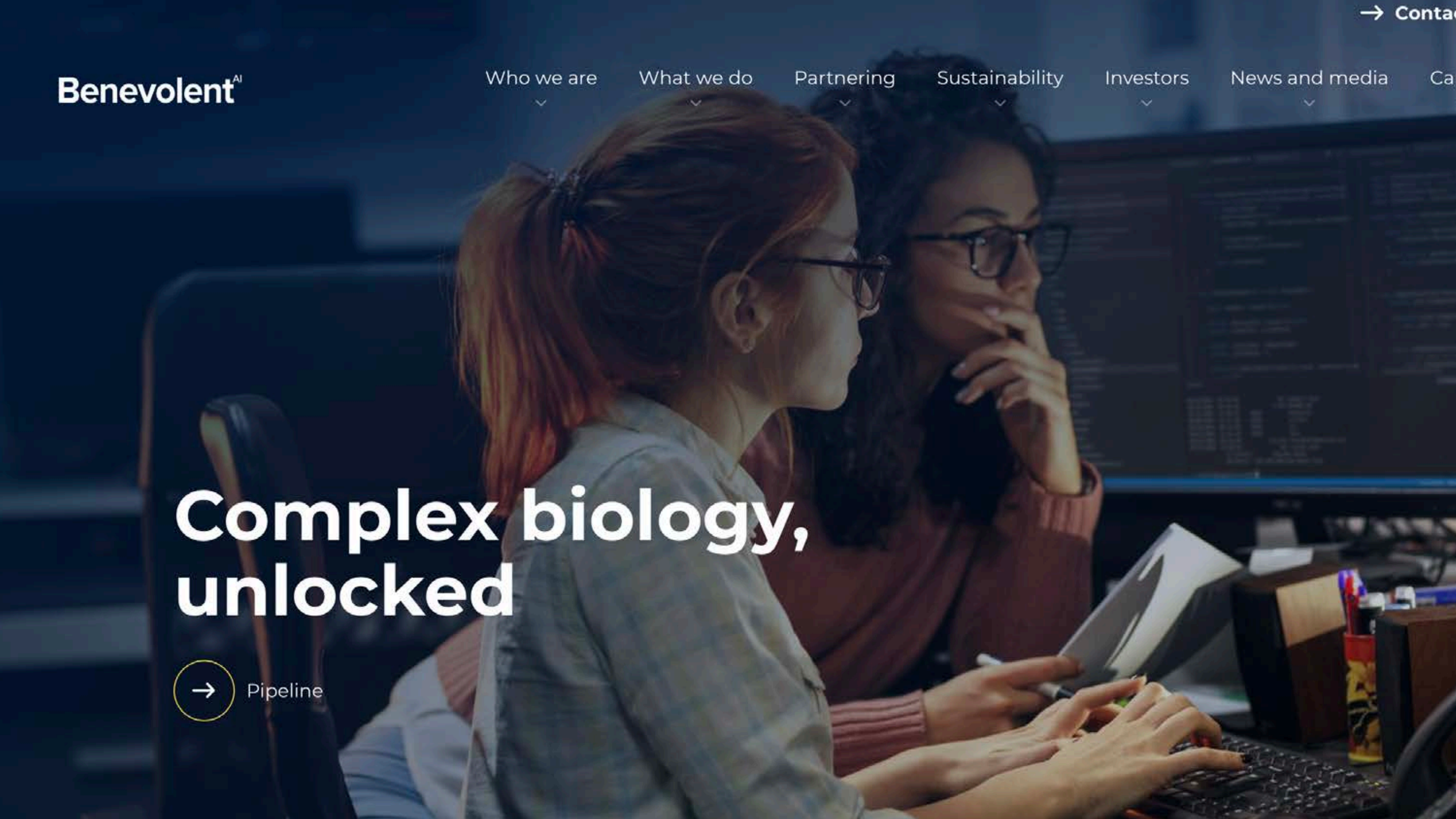
Hacia una epistemología de algoritmos: fiabilismo computational y sus límites

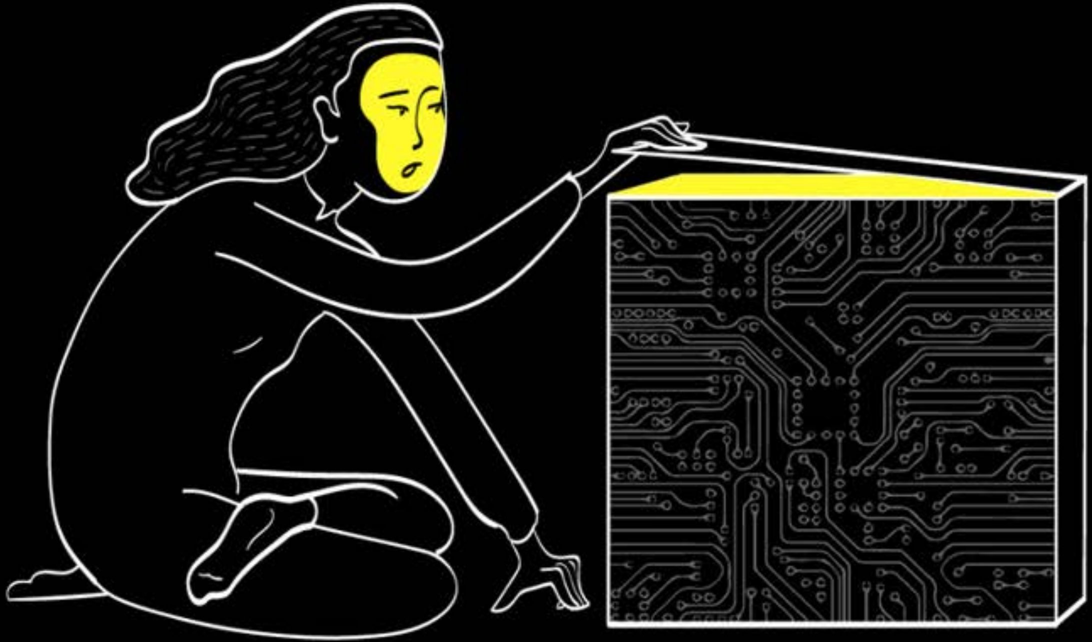
TU Delft

Juan M. Durán



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



Opacidad epistémica

[A] process is epistemically opaque relative to a cognitive agent *S* at time *t* just in case *S* does not know at *t* all of the epistemically relevant elements of the process (Humphreys 2009, 618)

Opacidad epistémica

[A] process is epistemically opaque relative to a cognitive agent *S* at time *t* just in case *S* does not know at *t* all of the epistemically relevant elements of the process (Humphreys 2009, 618)

"For a mathematical proof, one agent may consider a particular step in the proof to be an epistemically relevant part of the justification of the theorem, whereas to another, the step is sufficiently trivial to be eliminable. […] Within the hybrid scenario [i.e., algorithms], no human can examine and justify every element of the computational processes that produce the output of a computer simulation or other artifacts of computational science […] Many, perhaps all, of the features that are special to simulations are a result of this inability of human cognitive abilities to know and understand the details of the computational process" (2009, 618)

Opacidad epistémica

[A] process is epistemically opaque relative to a cognitive agent *S* at time *t* just in case *S* does not know at *t* all of the epistemically relevant elements of the process (Humphreys 2009, 618)

"For a mathematical proof, one agent may consider a particular step in the proof to be an epistemically relevant part of the justification of the theorem, whereas to another, the step is sufficiently trivial to be eliminable. […] Within the hybrid scenario [i.e., algorithms], no human can examine and justify every element of the computational processes that produce the output of a computer simulation or other artifacts of computational science […] Many, perhaps all, of the features that are special to simulations are a result of this inability of human cognitive abilities to know and understand the details of the computational process" (2009, 618)

TU Delft

Opacidad epistémica

[A] process is epistemically opaque relative to a cognitive agent *S* at time *t* just in case *S* does not know at *t* all of the epistemically relevant elements of the process (Humphreys 2009, 618)

"For a mathematical proof, one agent may consider a particular step in the proof to be an epistemically relevant part of the justification of the theorem, whereas to another, the step is sufficiently trivial to be eliminable. […] Within the hybrid scenario [i.e., algorithms], no human can examine and justify every element of the computational processes that produce the output of a computer simulation or other artifacts of computational science […] Many, perhaps all, of the features that are special to simulations are a result of this inability of human cognitive abilities to know and understand the details of the computational process" (2009, 618)

Opacidad epistémica

[A] process is epistemically opaque relative to a cognitive agent *S* at time *t* just in case *S* does not know at *t* all of the epistemically relevant elements of the process (Humphreys 2009, 618)

"For a mathematical proof, one agent may consider a particular step in the proof to be an epistemically relevant part of the justification of the theorem, whereas to another, the step is sufficiently trivial to be eliminable. […] Within the hybrid scenario [i.e., algorithms], no human can examine and justify every element of the computational processes that produce the output of a computer simulation or other artifacts of computational science […] Many, perhaps all, of the features that are special to simulations are a result of this inability of human cognitive abilities to know and understand the details of the computational process" (2009, 618)

Opacidad epistémica

- Naturaleza del proceso y elementos del proceso: instanciación de variables, llamadas a funciones, sentencias condicionales, operaciones aritméticas y lógicas, manejo de errores, estructuras de datos, pero también prácticas, métricas; “one may have excellent reasons for holding that a particular parametric family of models is applicable to the case at hand, yet have only empirical methods available for deciding which parametric values are the right ones” (2004, 150)
- Falta de capacidad de inspección de S: los algoritmos y los procesos computacionales son cajas negras.
- Pertinencia de los elementos epistémicos relevantes son con el propósitos de justificar el resultado

De opacidad a transparencia

“If we think in terms of such a process [i.e., algorithms] and imagine that its stepwise computation was slowed down to the point where, in principle, a human could examine each step in the process, the computationally irreducible process would become epistemically transparent. What this indicates is that the practical constraints we have previously stressed, primarily the need for computational speed, are the root cause of all epistemic opacity in this area. Because those constraints cannot be circumvented by humans, we must abandon the insistence on epistemic transparency for computational science. What replaces it would require an extended work in itself, but the prospects for success are not hopeless.” (Humphreys, 2004, 150)

Transparencia/interpretabilidad

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

Springer

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

568

.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

Springer

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).

Springer

¿Cómo obtenemos transparencia?

Table 2. Summary of Methods for Opening Black Boxes Solving the *Model Explanation* Problem

Name	Ref.	Authors	Year	Explanator	Black Box	Data Type	General	Random	Examples	Code	Dataset
Trepan	[22]	Craven et al.	1996	DT	NN	TAB	✓				✓
–	[57]	Krishnan et al.	1999	DT	NN	TAB	✓		✓		✓
DecText	[12]	Boz	2002	DT	NN	TAB	✓	✓			
GPDT	[46]	Johansson et al.	2009	DT	NN	TAB	✓	✓			
Tree Metrics	[17]	Chipman et al.	1998	DT	TE	TAB					
CCM	[26]	Domingos et al.	1998	DT	TE	TAB	✓	✓			
–	[34]	Gibbons et al.	2013	DT	TE	TAB	✓	✓			
STA	[140]	Zhou et al.	2016	DT	TE	TAB		✓			
CDT	[104]	Schetinin et al.	2007	DT	TE	TAB					
–	[38]	Hara et al.	2016	DT	TE	TAB		✓			
TSP	[117]	Tan et al.	2016	DT	TE	TAB					
Conj Rules	[21]	Craven et al.	1994	DR	NN	TAB		✓			
G-REX	[44]	Johansson et al.	2003	DR	NN	TAB	✓	✓			
REFNE	[141]	Zhou et al.	2003	DR	NN	TAB	✓	✓			
RxREN	[6]	Augasta et al.	2012	DR	NN	TAB		✓			
SVM+P	[82]	Nunez et al.	2002	DR	SVM	TAB					
–	[33]	Fung et al.	2005	DR	SVM	TAB					
inTrees	[25]	Deng	2014	DR	TE	TAB					
–	[70]	Lou et al.	2013	FI	AGN	TAB	✓				
GoldenEye	[40]	Henelius et al.	2014	FI	AGN	TAB	✓	✓			
PALM	[58]	Krishnan et al.	2017	DT	AGN	ANY	✓				
FIRM	[142]	Zien et al.	2009	FI	AGN	TAB	✓	✓			
MFI	[124]	Vidovic et al.	2016	FI	AGN	TAB	✓	✓			
–	[121]	Tolomei et al.	2017	FI	TE	TAB					
POIMs	[111]	Sonnenburg et al.	2007	FI	SVM	TAB					

Table 3. Summary of Methods for Opening Black Boxes Solving the *Outcome Explanation* Problem

Name	Ref.	Authors	Year	Explanator	Black Box	Data Type	General	Random	Examples	Code	Dataset
–	[134]	Xu et al.	2015	SM	DNN	IMG			✓	✓	✓
–	[30]	Fong et al.	2017	SM	DNN	IMG			✓		
CAM	[139]	Zhou et al.	2016	SM	DNN	IMG					
Grad-CAM	[106]	Selvaraju et al.	2016	SM	DNN	IMG					
–	[109]	Simonian et al.	2013	SM	DNN	IMG					
PWD	[7]	Bach et al.	2015	SM	DNN	IMG					
–	[113]	Sturm et al.	2016	SM	DNN	IMG					
DTD	[78]	Montavon et al.	2017	SM	DNN	IMG					
DeapLIFT	[107]	Shrikumar et al.	2017	FI	DNN	ANY					
CP	[64]	Landecker et al.	2013	SM	NN	IMG					
–	[143]	Zintgraf et al.	2017	SM	DNN	IMG					
VBP	[11]	Bojarski et al.	2016	SM	DNN	IMG					
–	[65]	Lei et al.	2016	SM	DNN	TXT					
ExplainD	[89]	Poulin et al.	2006	FI	SVM	TAB					
–	[29]	Strumbelj et al.	2010	FI	AGN	TAB	✓				
LIME	[98]	Ribeiro et al.	2016	FI	AGN	ANY	✓				
MES	[122]	Turner et al.	2016	DR	AGN	ANY	✓				
Anchors	[99]	Ribeiro et al.	2018	DR	AGN	ANY	✓				
–	[110]	Singh et al.	2016	DT	AGN	TAB	✓				
LORE	[37]	Guidotti et al.	2018	DR	AGN	TAB	✓				
MFI	[124]	Vidovic et al.	2016	FI	AGN	TAB	✓				
–	[39]	Haufe et al.	2014	FI	NLM	TAB					

Table 4. Summary of Methods for Opening Black Boxes Solving the *Model Inspection* Problem

Name	Ref.	Authors	Year	Explanator	Black Box	Data Type	General	Random	Examples	Code	Dataset
NID	[83]	Olden et al.	2002	SA	NN	TAB			✓		
GDP	[8]	Baehrens	2010	SA	AGN	TAB	✓		✓		✓
QII	[24]	Datta et al	2016	SA	AGN	TAB	✓		✓		✓
IG	[115]	Sundararajan	2017	SA	DNN	ANY			✓		✓
VEC	[18]	Cortez et al.	2011	SA	AGN	TAB	✓		✓		✓
VIN	[42]	Hooker	2004	PDP	AGN	TAB	✓		✓		✓
ICE	[35]	Goldstein et al.	2015	PDP	AGN	TAB	✓		✓	✓	✓
Prospector	[55]	Krause et al.	2016	PDP	AGN	TAB	✓		✓		✓
Auditing	[2]	Adler et al.	2016	PDP	AGN	TAB	✓		✓	✓	✓
OPIA	[1]	Adebayo et al.	2016	PDP	AGN	TAB	✓		✓		
–	[136]	Yosinski et al.	2015	AM	DNN	IMG			✓		✓
IP	[108]	Shwartz et al.	2017	AM	DNN	TAB			✓		
–	[137]	Zeiler et al.	2014	AM	DNN	IMG		✓		✓	
–	[112]	Springenberg et al.	2014	AM	DNN	IMG			✓		✓
DGN-AM	[80]	Nguyen et al.	2016	AM	DNN	IMG			✓	✓	✓
–	[72]	Mahendran et al.	2016	AM	DNN	IMG			✓	✓	✓
–	[95]	Radford	2017	AM	DNN	TXT			✓		
–	[143]	Zintgraf et al.	2017	SM	DNN	IMG			✓	✓	✓
VBP	[11]	Bojarski et al.	2016	SM	DNN	IMG			✓		✓
TreeView	[119]	Thiagarajan et al.	2016	DT	DNN	TAB			✓		✓

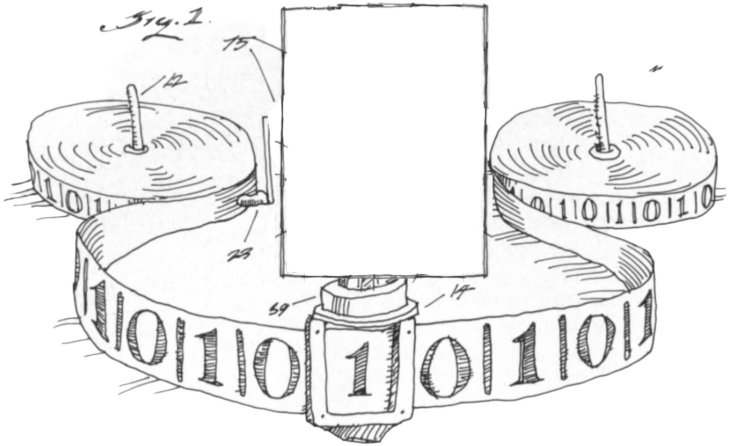
Guidotti, Monreale, Ruggieri, Turini, Giannotti, Pedreschi, (2018)

Transparencia/interpretabilidad

- Requiere “abrir” el algoritmo
  - i.e., rastrear el “path-dependency” de un resultado (Durán, 2021)
- La justification es asegurada mediante un third-party algorithm:
  - Interpretable predictors
  - Algunas formas de XAI (e.g., *post-hoc* explanation)
  - Transparency reports (e.g., Qualitative Input Influence)

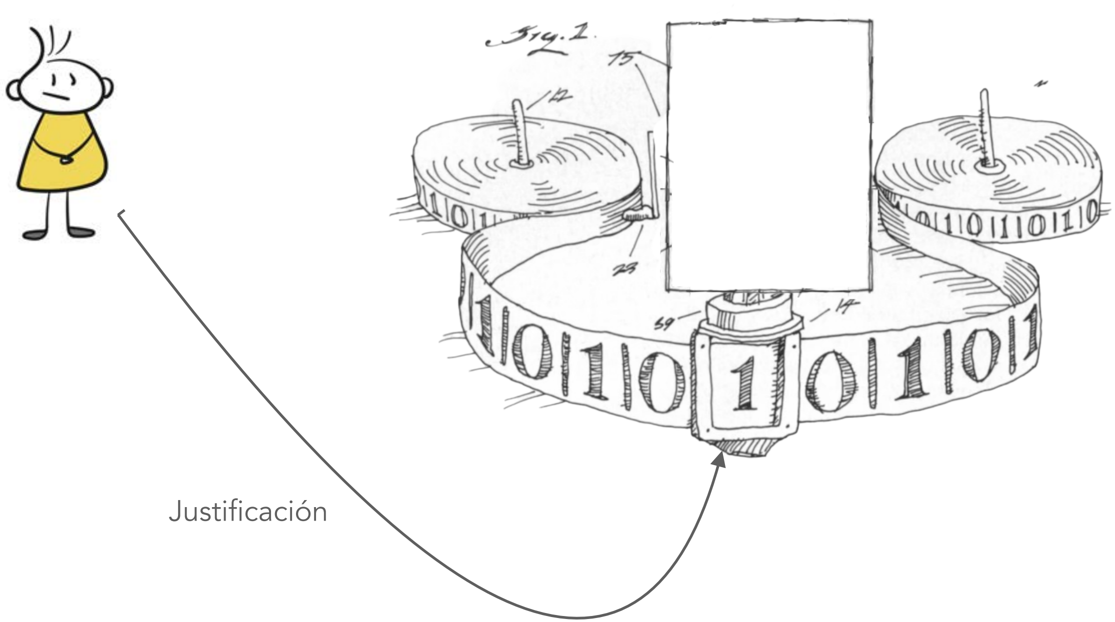
¿Cómo se justifica via transparencia?

Justificación via Transparencia (a sketch)



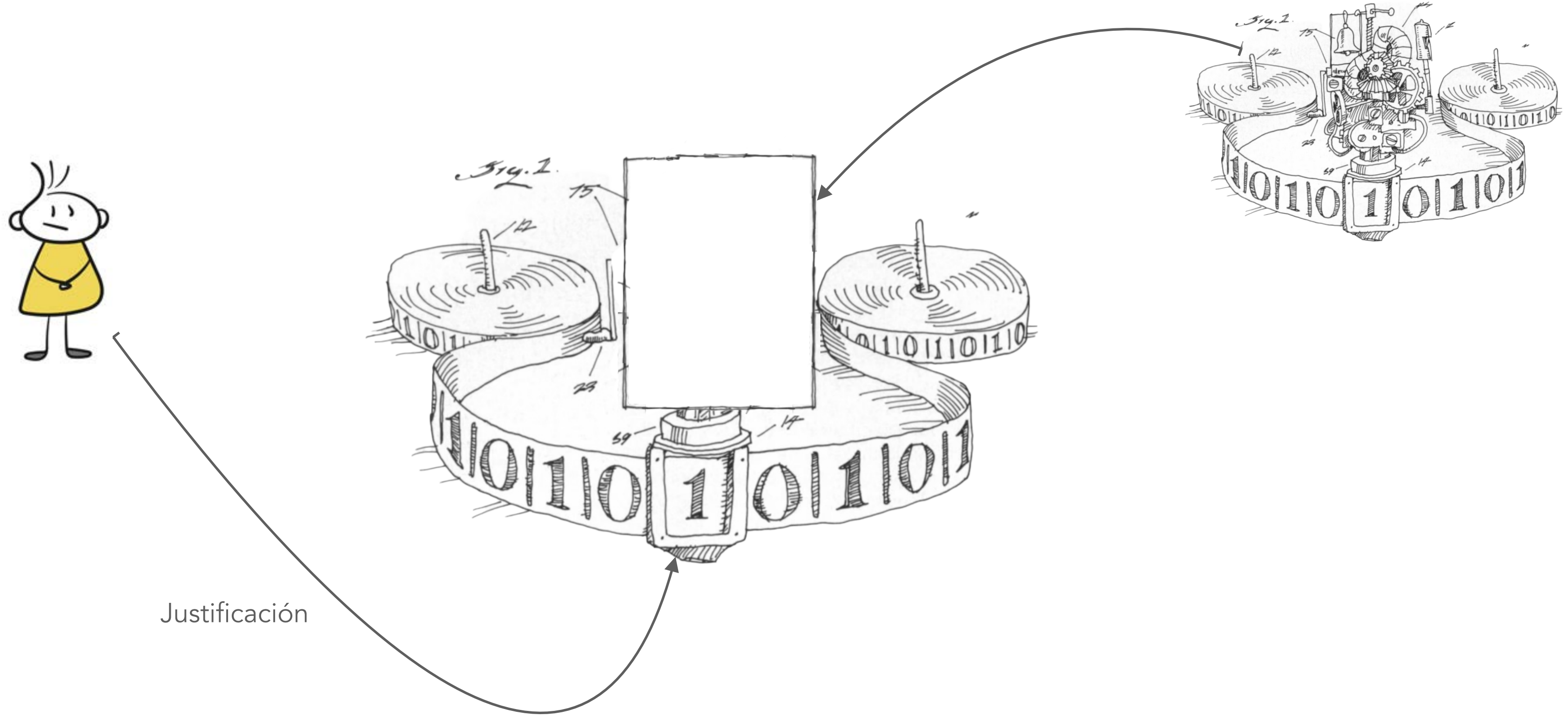
http://bridgethevoid.blogspot.com/

Justificación via Transparencia (a sketch)

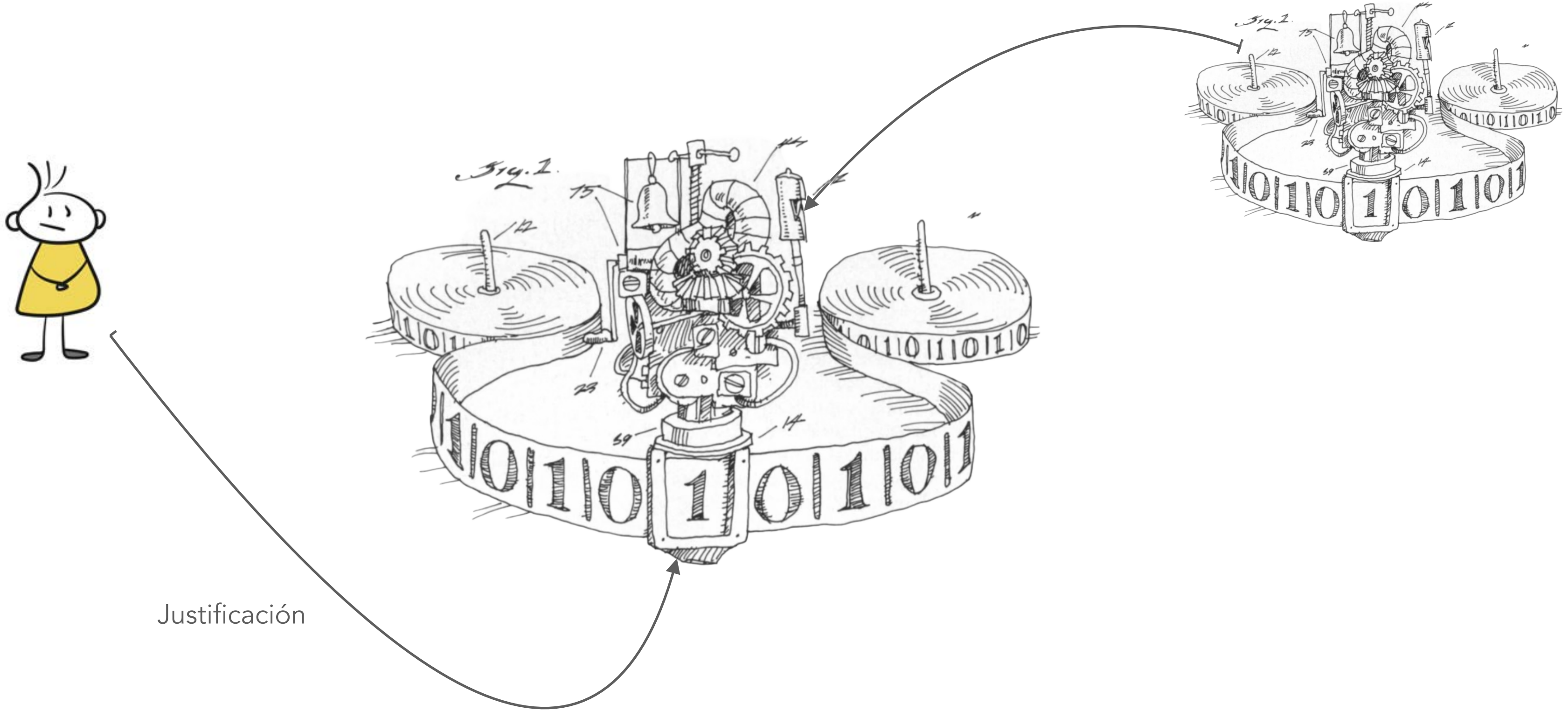


http://bridgethevoid.blogspot.com/

Justificación via Transparencia (a sketch)

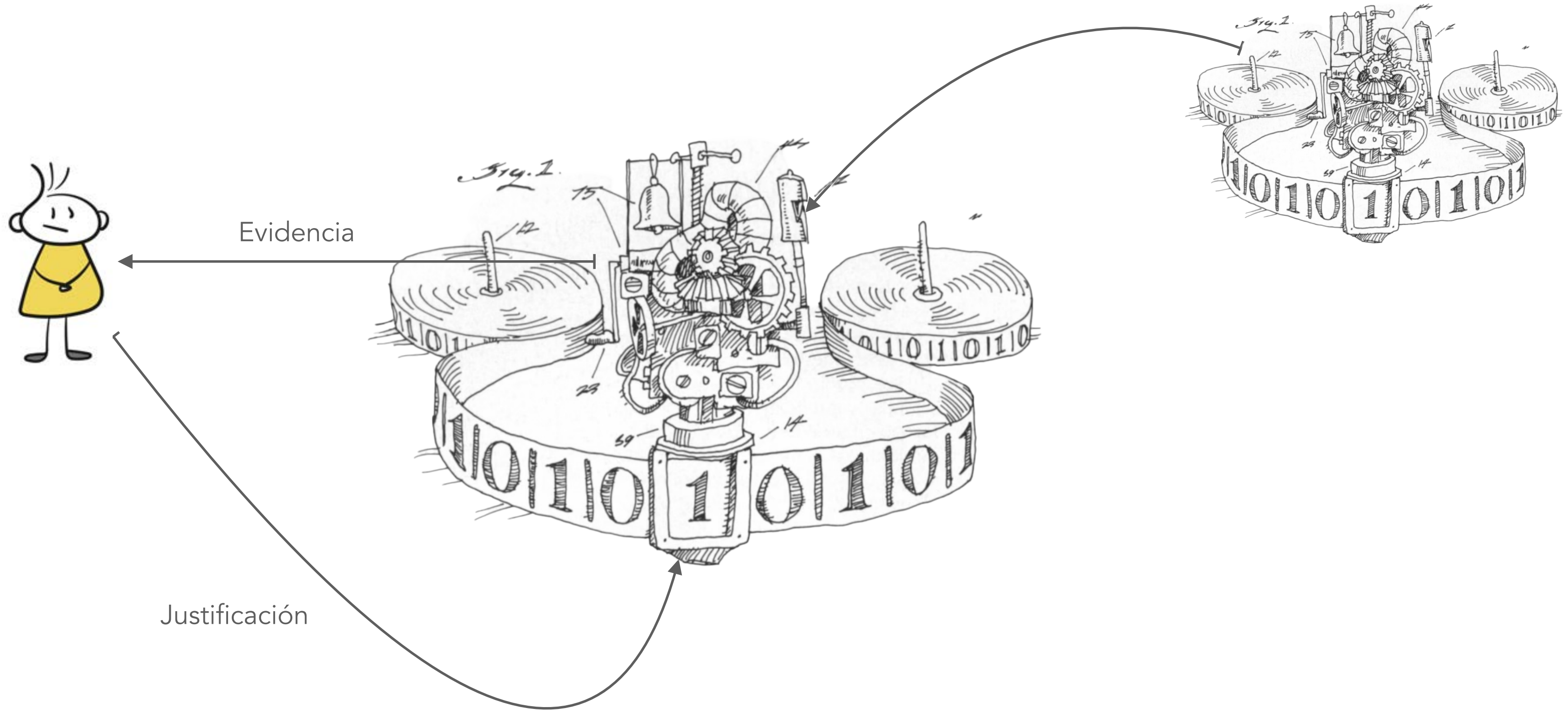


Justificación via Transparencia (a sketch)

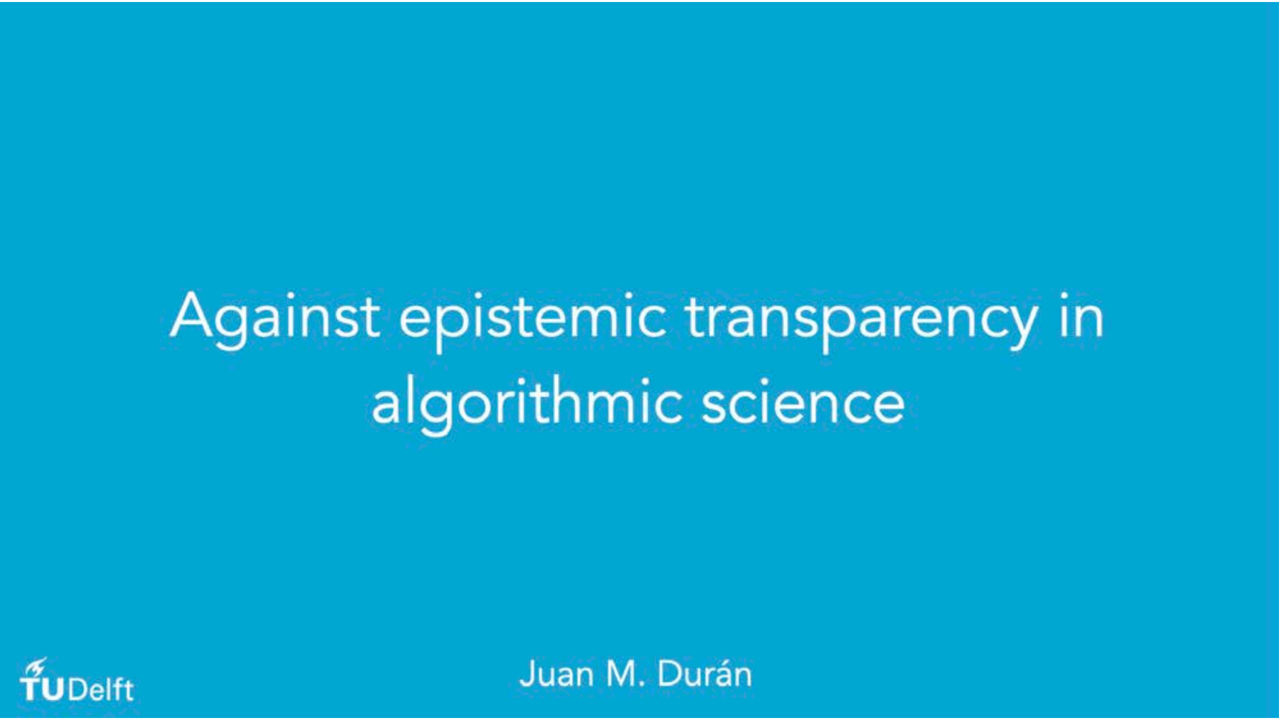


http://bridgethevoid.blogspot.com/

Justificación via Transparencia (a sketch)



http://bridgethevoid.blogspot.com/



Against epistemic transparency of algorithms

Juan M. Durán

Abstract

Epistemic transparency is often proposed as a solution to algorithmic opacity, wherein revealing the inner logic of an algorithm provides reasons or supporting evidence for the justification of its outputs. I argue that transparency is defective in belief formation and thus inadequate as an epistemology of algorithms. Two objections are developed: *transparency regress* and *bootstrapping*, both rooted in transparency's status as a time-sliced, 'outsourced' inferentialist epistemology. While weaker forms—such as contextual transparency—are briefly considered, the central argument remains that transparency fails to provide justification and should be abandoned as an epistemology of algorithms.

1 Introduction

It is widely accepted that algorithms are epistemically opaque. Opacity is here understood in line with Humphreys' definition, whereby the cognitive limitations of human agents prevent them from knowing the epistemically relevant elements of an algorithm that would justify belief in a given output [22, 618]. Now, there would be no special concern about opaque algorithms if it were not for the fact that they are central to many scientific endeavors, progressively displacing humans from the center of knowledge production. The opacity of algorithms, then, triggers significant epistemic anxiety.

To address opacity, many have sought to counteract this lack of knowledge. It is in this context that *transparency* gains traction. Creel, for instance, takes this idea in its strictest form. According to this author, opacity and transparency are "two sides of the same coin: opacity is a lack of transparency and vice versa" [28, 569, footnote 1]. As we will see, many have folded under this view. And while transparency encompasses a range of methodologies and strategies, there is a common justificatory aim, that is, to provide reasons or supporting evidence for believing the algorithm's output. From where is this justification drawn? From identifying the functions, variables, values, and similar epistemically relevant elements within the algorithm that generate the output. This justificatory aim has been collectively referred to as "showing the inner logic of the algorithm" [7, 843]. For example, BenevolentAI identified *baricitinib* as an effective drug to combat COVID-19 symptoms. For a researcher to be justified in this belief, transparency shows how BenevolentAI internally operates: how it identifies drug-signal blocking, maps causal relations between viruses and drugs,

1

TU Delft

De transparencia a fiabilismo (reliabilism)

"If we think in terms of such a process [i.e., algorithms] and imagine that its stepwise computation was slowed down to the point where, in principle, a human could examine each step in the process, the computationally irreducible process would become epistemically transparent. What this indicates is that the practical constraints we have previously stressed, primarily the need for computational speed, are the root cause of all epistemic opacity in this area. Because those constraints cannot be circumvented by humans, we must abandon the insistence on epistemic transparency for computational science. What replaces it would require an extended work in itself, but the prospects for success are not hopeless." (Humphreys, 2004, 150)

¿Qué es process reliabilism? (≈ Goldman 1979)

Process reliabilism: la creencia de Diego está justificada en caso que sea producida por un proceso (o secuencia de procesos) de formación de creencia fiable

Un proceso de formación de creencia fiable tiene la tendencia de producir crencias que son *verdaderas* antes que *falsas*





¿Qué es process reliabilism? (≈ Goldman 1979)

Process reliabilism: la creencia de Diego está justificada en caso que sea producida por un proceso (o secuencia de procesos) de formación de creencia fiable

Un proceso de formación de creencia fiable tiene la tendencia de producir crencias que son *verdaderas* antes que *falsas*

CR + Instrumental reliabilism

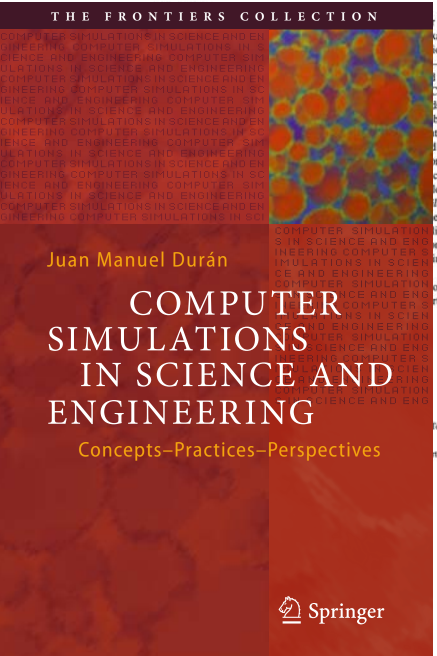
Minds and Machines (2018) 28:645–666
https://doi.org/10.1007/s11023-018-9481-6

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



simulations rel
amples of these
ons (Parker in
-57, 2009), the
(eds) Compute
Cambridge Schol
s) Computer sim
bridge Scholars
ter simulations (
Philos Sci 31(1
rs are right in a
computer simula
em of *epistemic*
lism, a specified
e center. The art
ism, namely, ve
ns, a history of
n simulations.

opacity · Verific
t knowledge

Technology, Jaffalaan

t, Nobelstrasse 19,

Beyond transparency: computational reliabilism as an externalist epistemology of algorithms

forthcoming in

Philosophy of Science for Machine Learning: Core Issues and New Perspectives - Juan M. Durán and Giorgia Pozzi (eds.)

Synthese Library

Juan M. Durán

Abstract This chapter examines the epistemology of algorithms, framing the discussion as a question of epistemic justification. Current approaches emphasize algorithmic transparency, which involves elucidating internal mechanisms—such as functions and variables—and demonstrating how (or that) these compute outputs. Thus, the mode of justification through transparency is contingent on what can be shown about the algorithm and, in this sense, is *internal* to ... contrast, I propose an *externalist* epistemology of algorithms call... *reliabilism* (CR). While I have previously developed CR in the co... simulations ([60, 74, 12]), this chapter extends the framework to a... algorithms used across scientific disciplines, particularly in mac... deep neural networks. At its core, CR posits that an algorithm's ... if it is generated by a reliable algorithm, where reliability is det... bility indicators. These indicators arise from formal methods, alg... expert competencies, research cultures, and other scientific practic... primary objectives are to delineate the foundations of CR, expla... mechanisms, and outline its potential as an externalist epistemolog...

1 Introduction

The use of algorithms for scientific purposes is delivering remarkal... ple of examples will suffice to illustrate this. In molecular biolog...

Juan M. Durán
Department of Values, Technology and Innovation
Faculty of Technology, Policy and Management
Delft University of Technology
Jaffalaan 5
2628 BX Delft
The Netherlands, e-mail: j.m.duran@tudelft.nl

Epistemic Opacity and Epistemic Inaccessibility

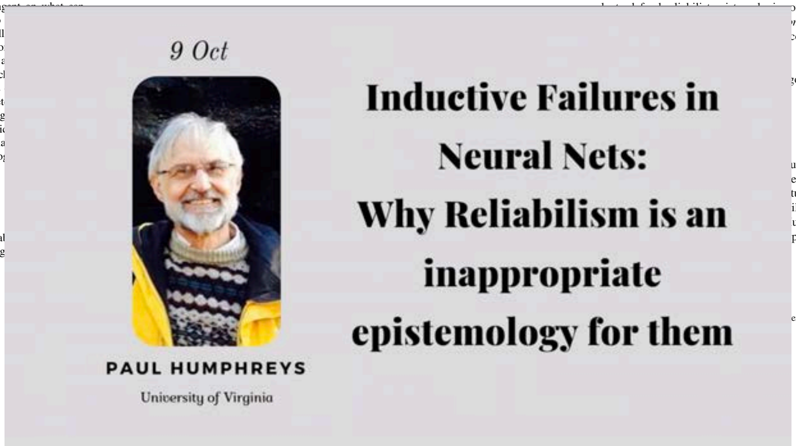
In this paper I shall revisit the concepts of epistemic opacity and essential epistemic opacity with the hope of clarifying and elaborating those concepts. I shall begin with a clarification of the definitions and relate the concept of epistemic opacity to that of epistemic inaccessibility. I then introduce and describe representational opacity, including three distinctions between types of representation and argue that this is an important source of opacity in some deep neural networks. Next, I proceed to an examination of other sources of epistemic opacity, some of which follow from differences between applied and pure mathematics. Then, after looking at some related concepts, I conclude with some ways in which opacity can be ameliorated.

1. Introduction.

As a reminder, here are the two definitions of epistemic opacity as formulated in Humphreys 2009:

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.

A process is *essentially epistemically opaque* to X if and only if it is impossible, given the nature of X, for X to know all of the epistemically relevant elements of the process.



European Journal for Philosophy of Science (2025) 15:37
https://doi.org/10.1007/s13194-025-00664-2

PAPER IN GENERAL PHILOSOPHY OF SCIENCE

Check for updates

In defense of reliabilist epistemology of algorithms

Juan M. Durán[1]

Received: 19 August 2024 / Accepted: 3 June 2025

Abstract
In a reliabilist epistemology of algorithms, a high frequency of accurate output representations is indicative of the algorithm's reliability. Recently, Humphreys challenged this assumption, arguing that reliability depends not only on frequency but also on the quality of outputs. Specifically, he contends that radical and egregious misrepresentations have a distinct epistemic impact on our assessment of an algorithm's reliability, regardless of the frequency of their occurrence. He terms these *statistically insignificant but serious errors* (SIS-Errors) and maintains that their occurrence warrants revoking our epistemic attitude towards the algorithm's reliability. This article ... of algorithms against the challenge posed ... *mputational reliabilism* as a foundational ... conditions designed to prevent SIS-Errors

...orithms · Computational reliabilism ·

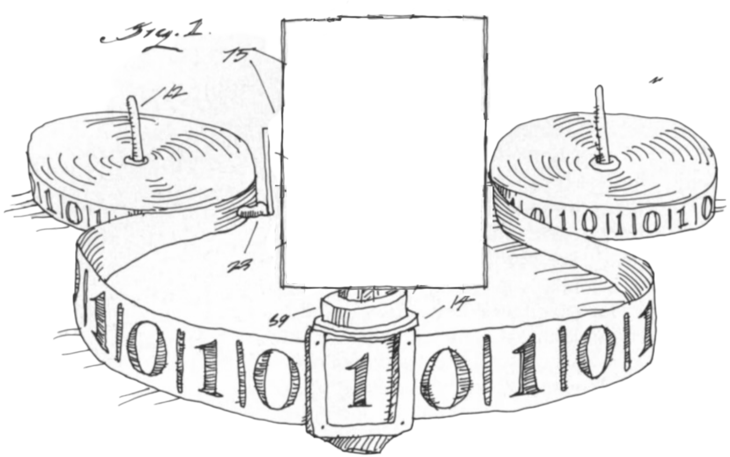
...ut the alleged novelty of computer simula... ...eiss argued that, although technologically ...tute a philosophical novelty nor "[a] revo... ...ilosophers were worried about in the past" ...umphreys highlighted four specific issues ... philosophical novelty (Humphreys, 2009).

...ersity of Technology, Jaffalaan 5, 2628 BX Delft,

Springer

¿Cómo se justifica via CR?

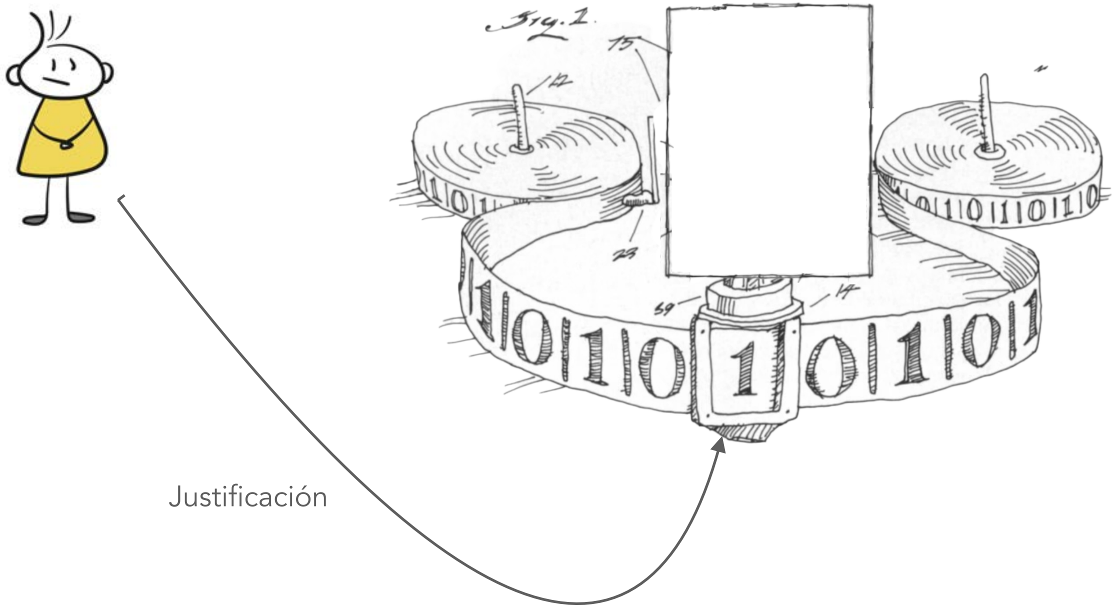
Justificación via CR (a sketch)



TU Delft

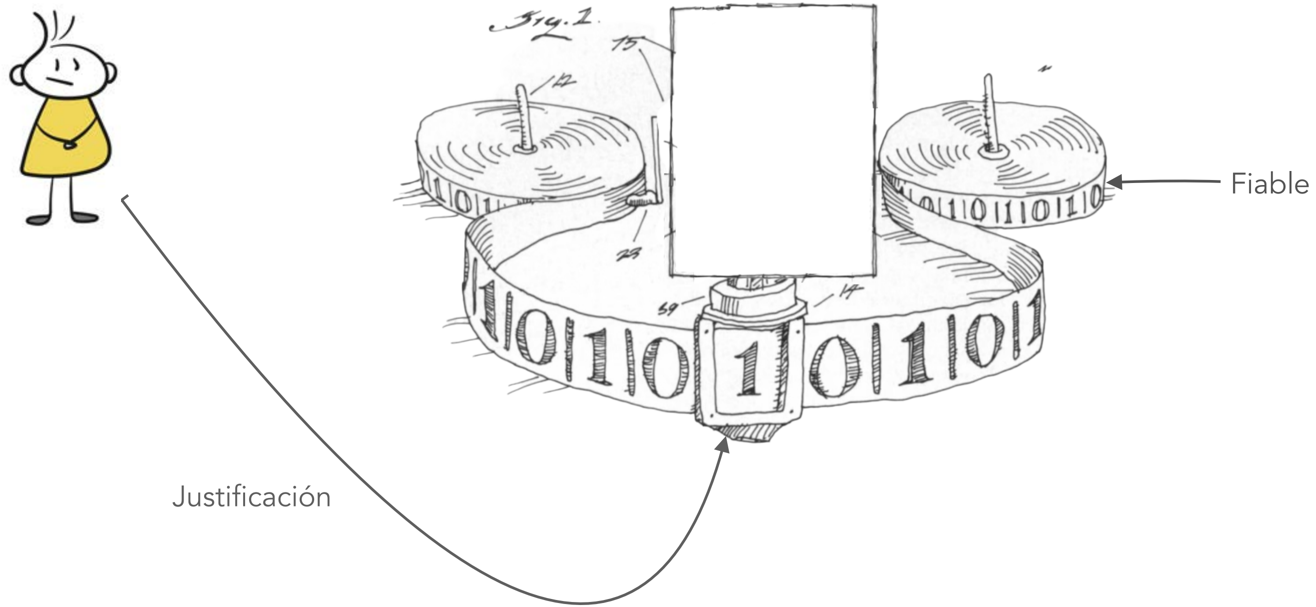
http://bridgethevoid.blogspot.com/

Justificación via CR (a sketch)



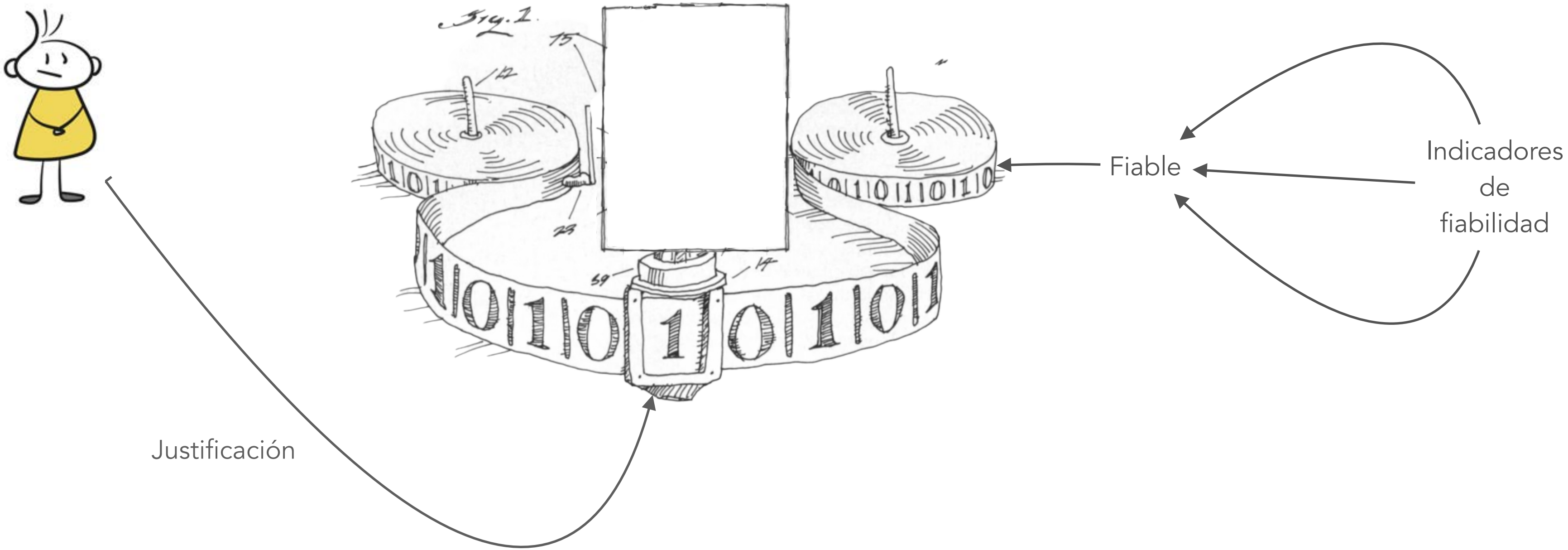
http://bridgethevoid.blogspot.com/

Justificación via CR (a sketch)



http://bridgethevoid.blogspot.com/

Justificación via CR (a sketch)



TUDelft

http://bridgethevoid.blogspot.com/

Fiabilismo computacional

- Aceptamos algoritmos de “caja negra”
- La justificación viene de asegurarse la fiabilidad del algoritmo a través de indicadores de fiabilidad

TUDelft

Fiabilismo computacional

- Aceptamos algoritmos de “caja negra”
- La justificación viene de asegurarse la fiabilidad del algoritmo a través de indicadores de fiabilidad
    - RI_1 Performance de los algorithmos
    - RI_2 Práctica científica basada en algoritmos
    - RI_3 Construcción social de la fiabilidad

Somes obvious sources of reliabilty

- RI_1 Performance de los algoritmos
- RI_2 Práctica científica basada en algoritmos
- RI_3 Construcción social de la fiabilidad

Somes obvious sources of reliabilty

- RI_1 Performance de los algoritmos

TU Delft

Somes obvious sources of reliabilty

- RI_1 Performance de los algoritmos
  - Métodos de verificación y validación (→ accuracy)
  - Elecciones de datos y bases de datos
  - Parametrización e hyper-parametrización
  - Prácticas de entrenamiento de modelos
  - Una historia de éxitos y fracasos en las implementaciones
  - Tratamiento de errores (e.g., uncertainty quantification)

(Durán 2018; Durán & Formanek 2018)

Somes obvious sources of reliabilty

- RI_1 Performance de los algoritmos
- RI_2 Práctica científica basada en algoritmos

Somes obvious sources of reliabilty

- RI_1 Performance de los algoritmos
- RI_2 Práctica científica basada en algoritmos
  - Causalidad y otros mechanismos implementados
  - Valores epistémicos, morales, etc.
  - Representación, idealizaciones, abstracciones, etc
  - Terías, leyes, principios, etc
  - Algún grade de coherencia teórica, conceptual, etc
    - Conocimiento de fondo (background knowledge, e.g., knowledge of the learned weights)
    - Algún morfismo entre hipótesis, modelos, etc y al sintaxis del algoritmo (e.g., al usal natural kinds, cut-off-values que son aceptados, definiciones — e.g., cuándo un melanoma es cancerígeno)

TU Delft

Somes obvious sources of reliabilty

- RI_1 Performance de los algoritmos
- RI_2 Práctica científica basada en algoritmos
- RI_3 Construcción social de la fiabilidad

TUDelft

Somes obvious sources of reliabilty

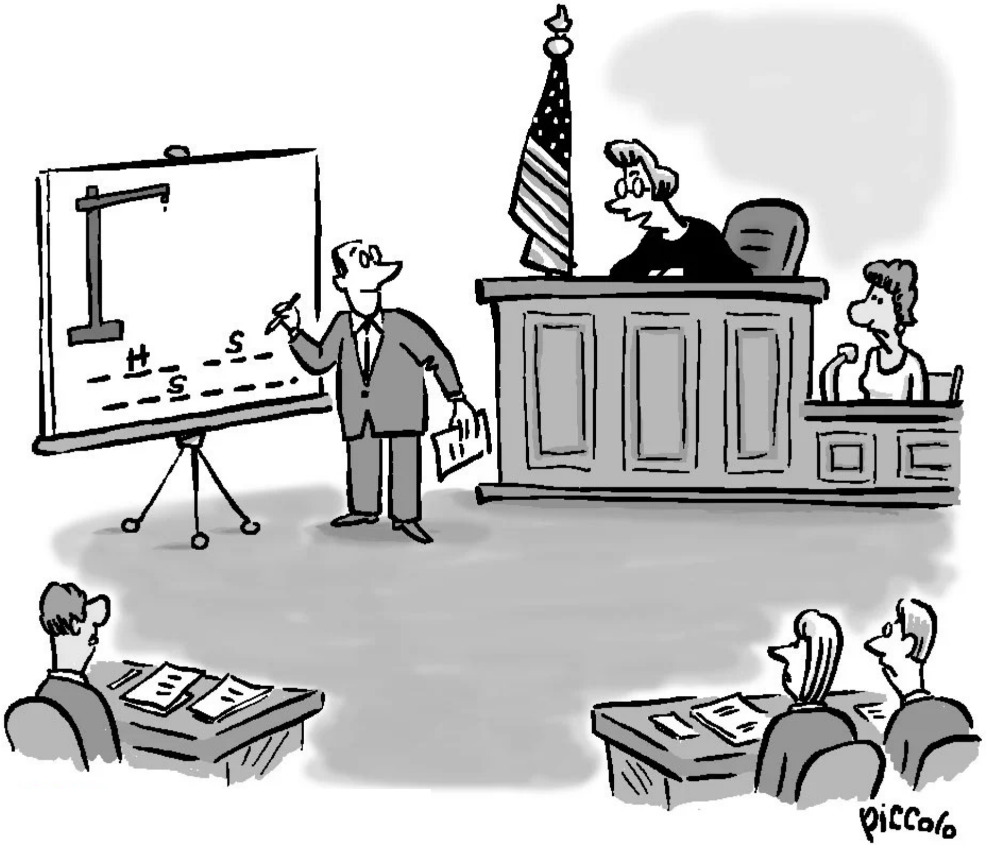
- RI_1 Performance de los algoritmos
- RI_2 Práctica científica basada en algoritmos
- RI_3 Construcción social de la fiabilidad
    - Aceptabilidad social del resultado del algoritmo
    - Favorable, exitoso/fracaso, usos (in)válidos del algoritmo y sus resultados (e.g., cuando se desarrollan nuevas hipótesis o se expande en nuevas estrategias)
    - Coherencia con un cuerpo de conocimiento estable
    - Coherencia con los compromisos epistémicos, éticos, etc de los científicos y sus comunidades
    - Realización de los valores epistémicos, éticos, etc
    - Culturas de investigación (científico, modelaje, comunidades…)

TUDelft

¿Cómo funciona CR?
Un caso en ciencia forense
Parte I

El experto, el abogado, y el juez

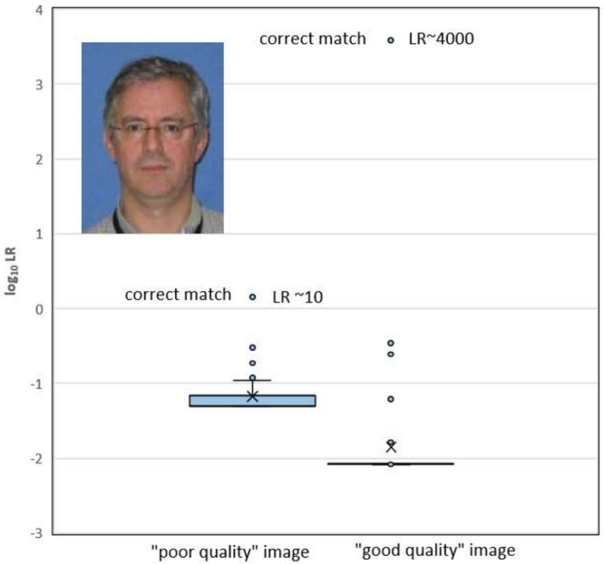
Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid



TU Delft

Comparación de rostros en ciencia forense

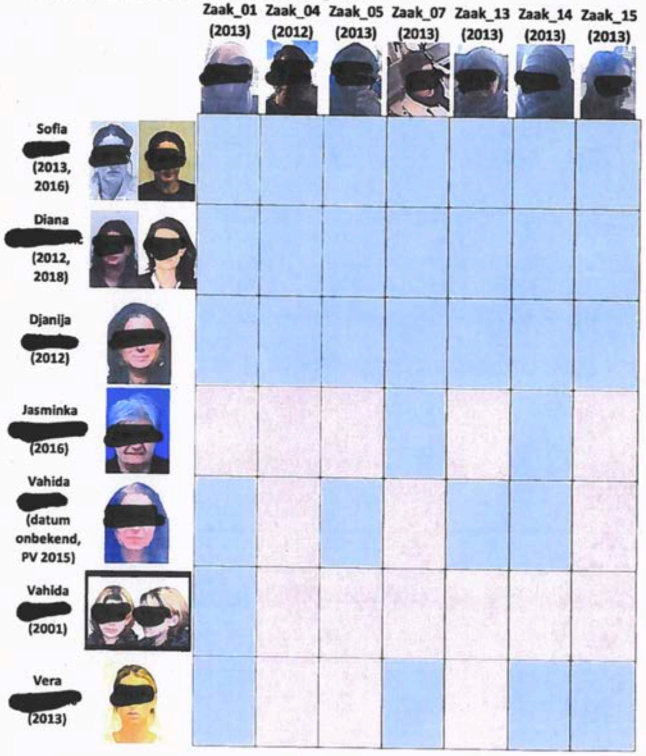
"Poor quality" image



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

Zaak_01 (2013) | Zaak_04 (2012) | Zaak_05 (2013) | Zaak_07 (2013) | Zaak_13 (2013) | Zaak_14 (2013) | Zaak_15 (2013)

Sofia [redacted] (2013, 2016)

Diana [redacted] (2012, 2018)

Djanija [redacted] (2012)

Jasminka [redacted] (2016)

Vahida [redacted] (datum onbekend, PV 2015)

Vahida [redacted] (2001)

Vera [redacted] (2013)

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

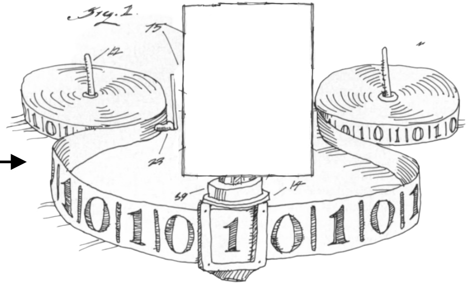
Zaaknr.	Rapportage aanvraag 002 (25 maart 2016)	Huidige aanvraag 003
Zaak 01	Waarschijnlijker	Waarschijnlijker
Zaak 04	Zeer veel waarschijnlijker	Zeer veel waarschijnlijker
Zaak 05	Zeer veel waarschijnlijker	Zeer veel waarschijnlijker
Zaak 07	--	Iets waarschijnlijker
Zaak 13	Veel waarschijnlijker	Veel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

...nderzoek zijn:
...r H2 waar is
...r H2 waar is
...r H2 waar is
...r H2 waar is
...r H2 waar is
... H2 waar is

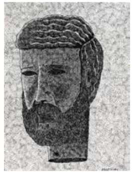
Cijfer	De bevindingen van het onderzoek zijn:
-1	iets waarschijnlijker wanneer H2 waar is dan wanneer H1 waar is
-2	waarschijnlijker wanneer H2 waar is dan wanneer H1 waar is
-3	veel waarschijnlijker wanneer H2 waar is dan wanneer H1 waar is
-4	zeer veel waarschijnlijker wanneer H2 waar is dan wanneer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

Ciencia forense y AI

Evidence

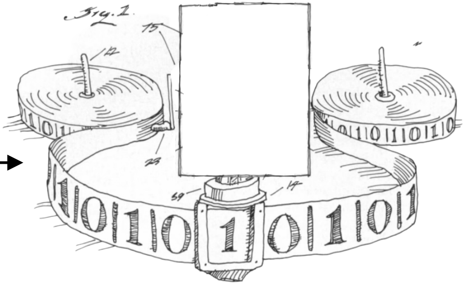


Suspects



Ciencia forense y AI

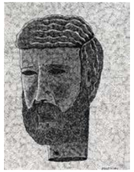
Evidence



Suspects

ψ_1

ψ_2

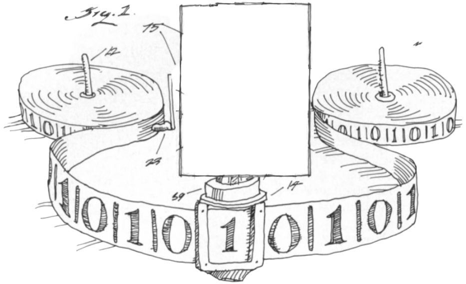


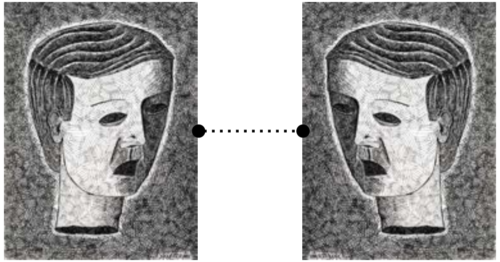
ψ_3

Probability: $\psi_3 > \psi_2 > \psi_1$

TU Delft

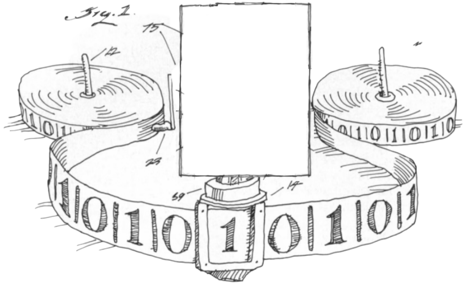
Ciencia forense y CR

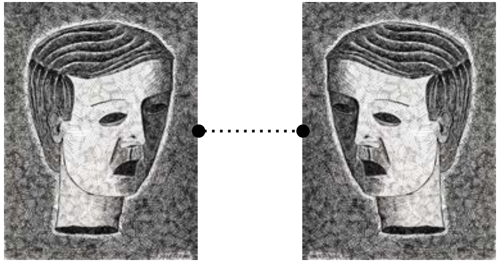




TU Delft

Ciencia forense y CR

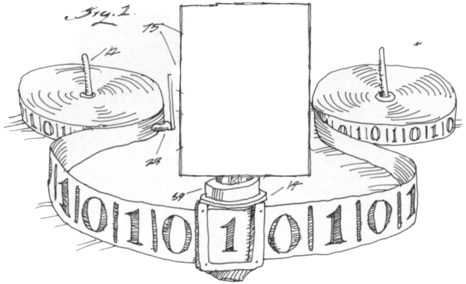


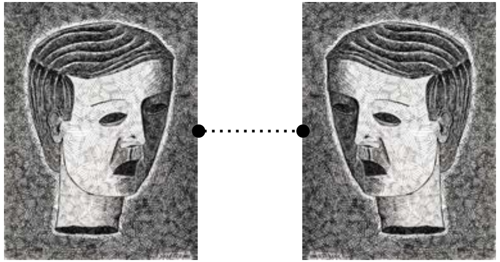


RI$_1$ Performance de los algoritmos

- Verificación y Validación
- Análisis de robustez
- Redundancia
- "Calidad" de los datos

Ciencia forense y CR





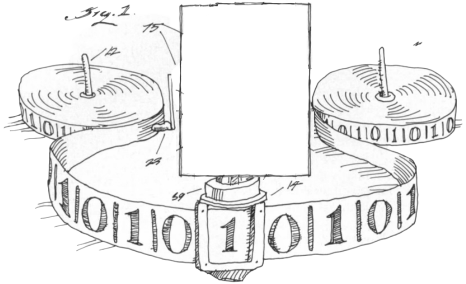
RI$_1$ Performance de los algoritmos

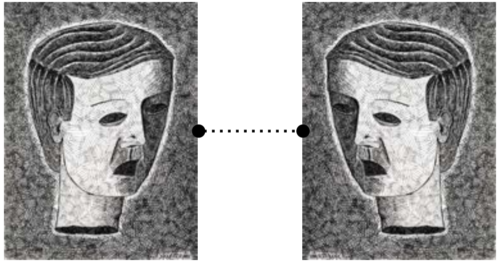
- Verificación y Validación
- Análisis de robustez
- Redundancia
- "Calidad" de los datos

RI$_2$ Práctica científica basada en algoritmos

- Implementación de "quality assurance methods"
- "We implement well-known and tested libraries"
- Técnicas biométricas

Ciencia forense y CR





RI_1 Performance de los algoritmos

- Verificación y Validación
- Análisis de robustez
- Redundancia
- "Calidad" de los datos

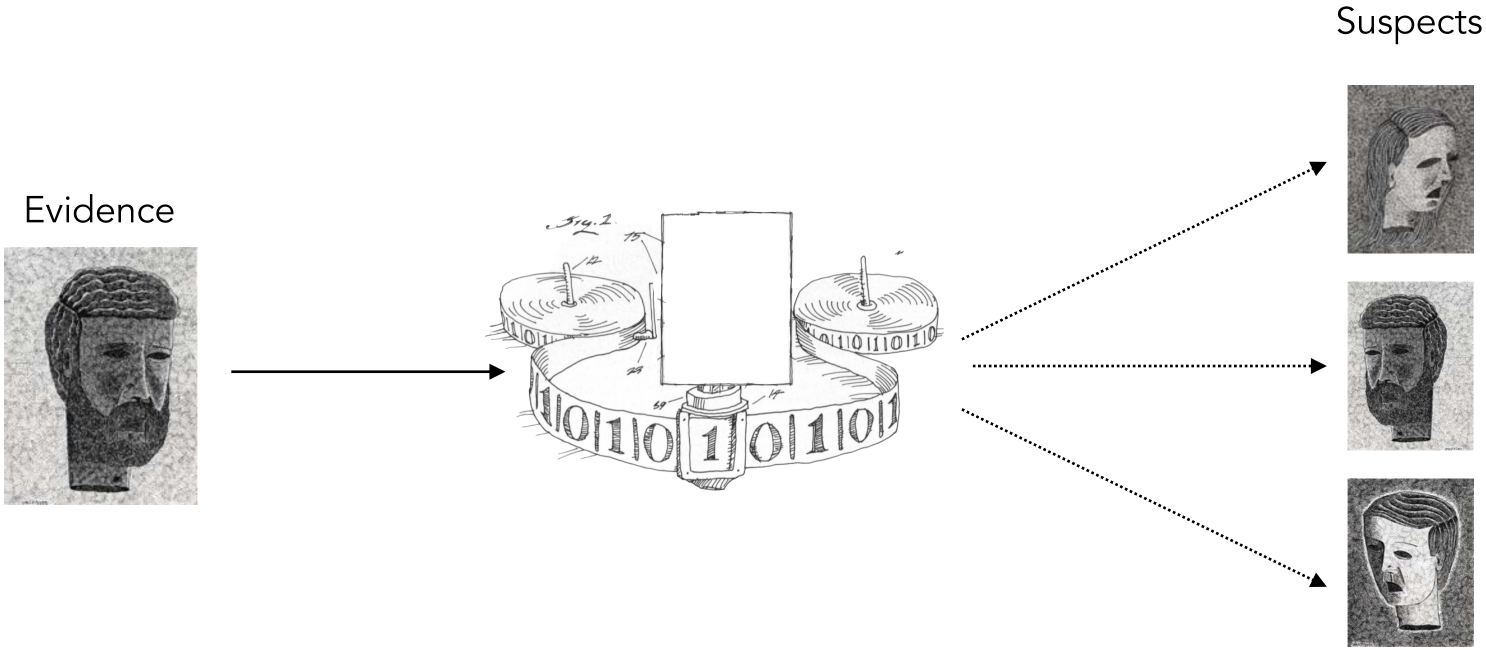
RI_2 Práctica científica basada en algoritmos

- Implementación de "quality assurance methods"
- "We implement well-known and tested libraries"
- Técnicas biométricas

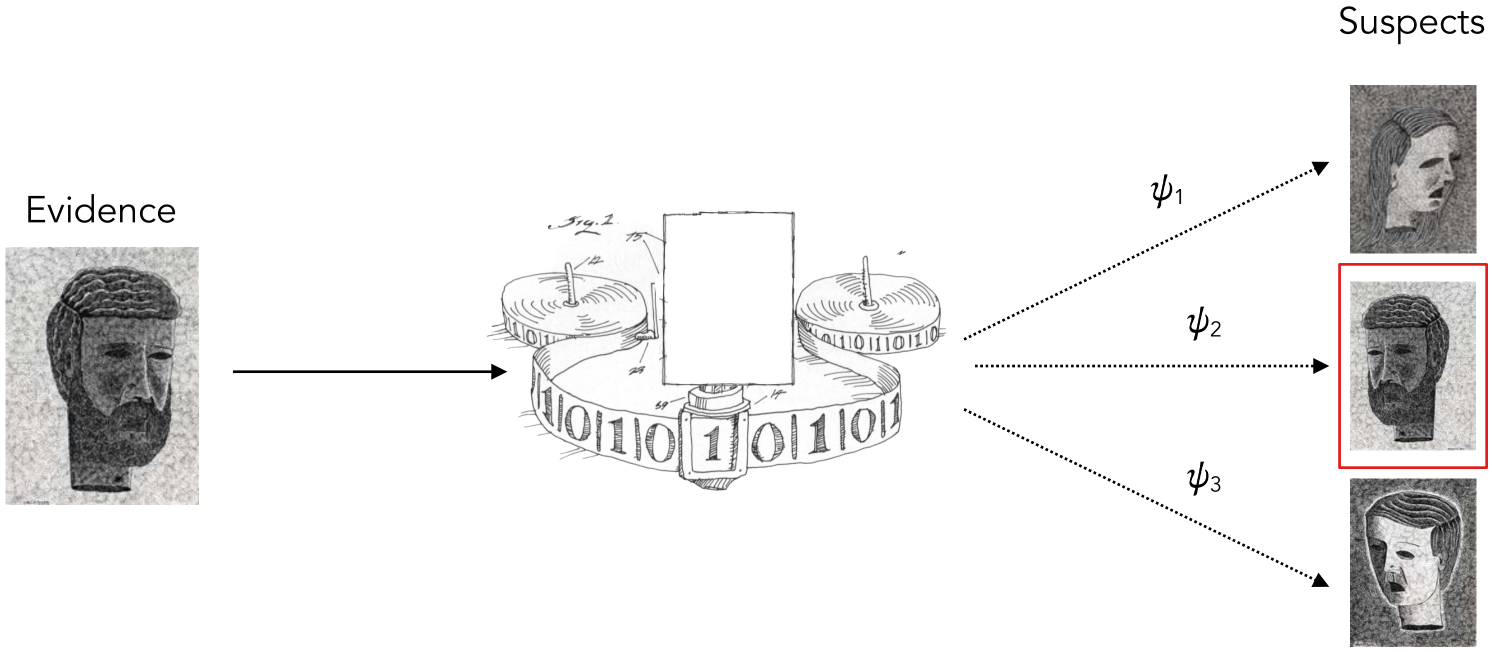
RI_3 Construcción social de la fiabilidad

- Debates sobre la aplicabilidad de los resultados (e.g., soundness, legal, etc.)
- Contraste con otros métodos no-computacionales (e.g., uso de revisiones de código, reportes de expertos)

Ciencia forense y CR



Ciencia forense y CR



Probability: $\psi_2 > \psi_3 > \psi_1$

¿Cómo funciona CR?
Un caso en ciencia forense
Parte II

TU Delft

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.

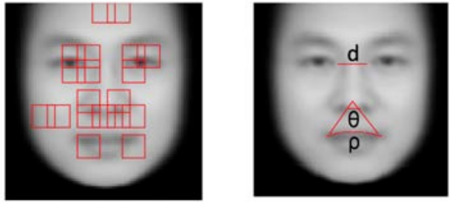
TU Delft

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.

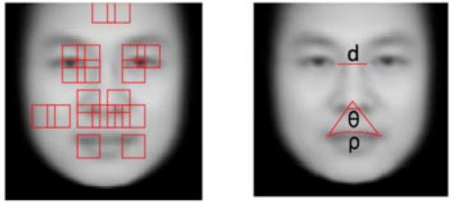
Wu & Zhang (2016) *Automated Inference on Criminality using Face Images* - arXiv

Automated Inference on Criminality Using Face Images

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

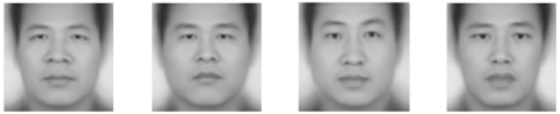
(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.



(a) (b)

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



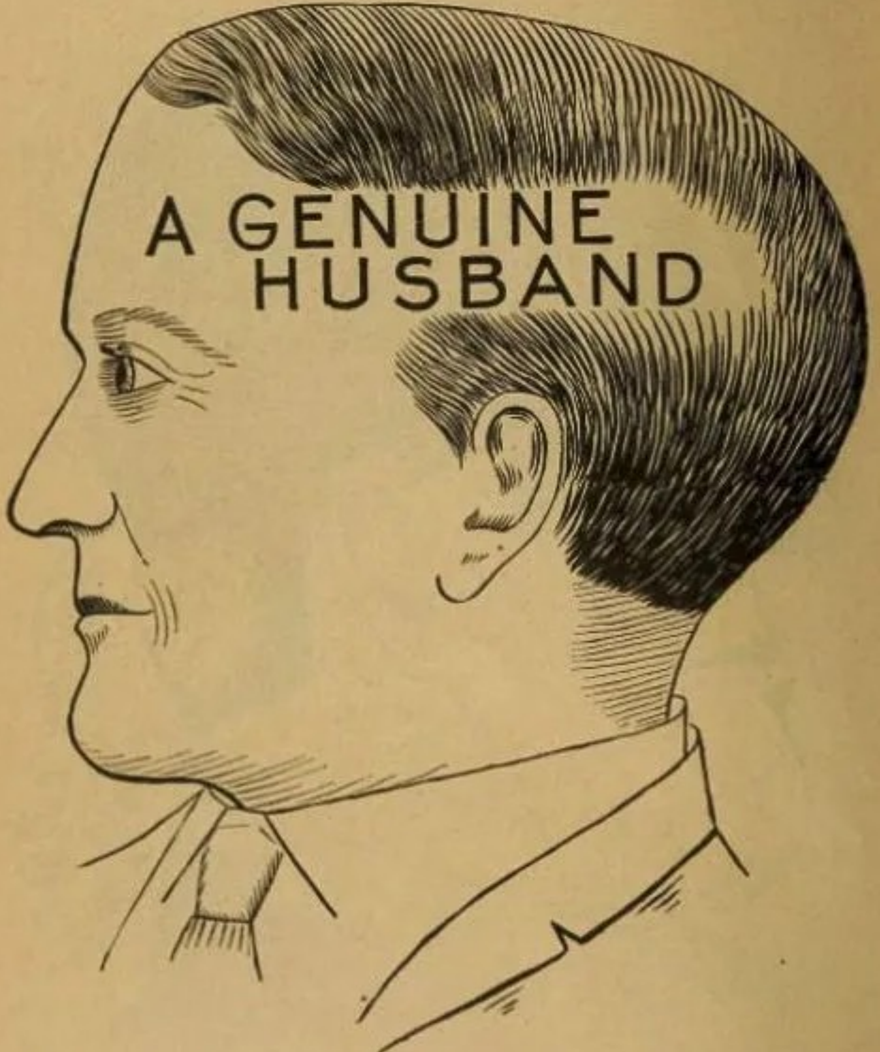
(a)

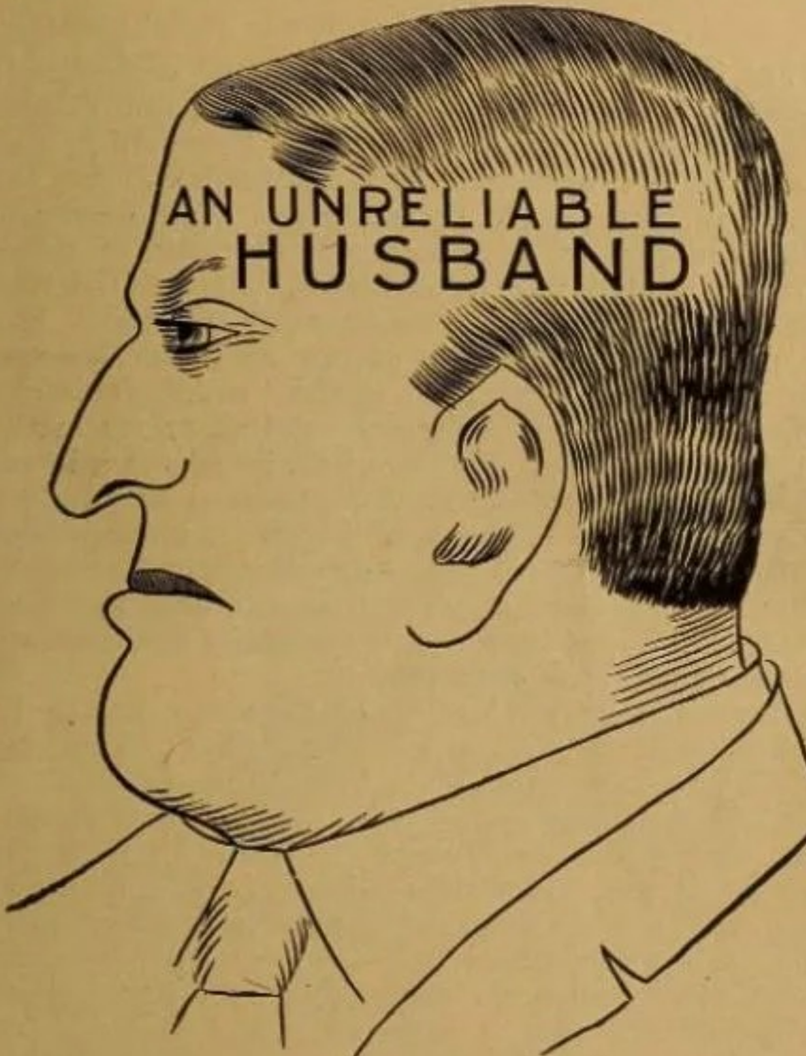
(b)

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

TU Delft

Wu & Zhang (2016) *Automated Inference on Criminality using Face Images* - arXiv


A GENUINE
HUSBAND


AN UNRELIABLE
HUSBAND

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.

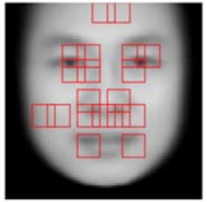
"This newly discovered knowledge suggests a law of normality for faces of non-criminals: Given the race, gender and age, the faces of general law-biding public have a greater degree of resemblance compared with the faces of criminals. In other words, criminals have a significantly higher degree of dissimilarity in facial appearance than normal population" (2016, 2)

Automated Inference on Criminality Using Face Images

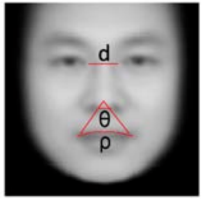
(a) Three samples in criminal ID photo set S_c.

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

Figure 1. Sample ID photos in our data set.

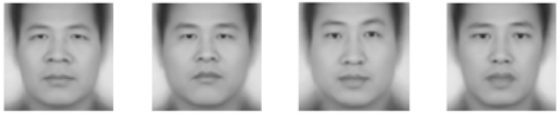


(a)



(b)

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



(a)

(b)

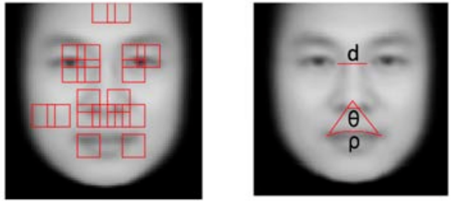
Figure 9. (a) The four subtypes of criminal faces; (b) The three subtypes of non-criminal faces.

Automated Inference on Criminality Using Face Images

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

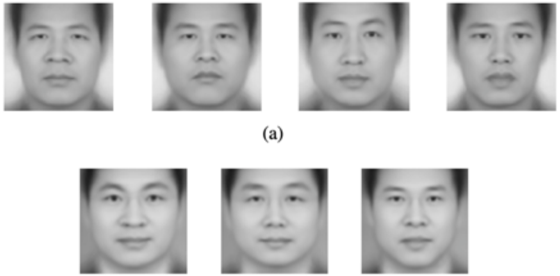
(b) Three samples in non-criminal ID photo set S_n

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.

RI_1 Performance de los algoritmos

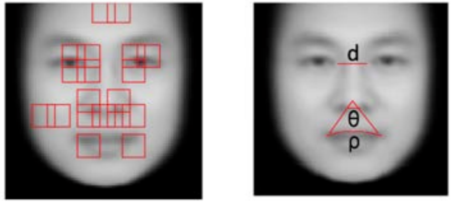
- Verificación y Validación (?)
- "Calidad" de los datos
- El sistema es forzado a elegir entre does clases: {criminal}/{no-criminal}

Automated Inference on Criminality Using Face Images

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

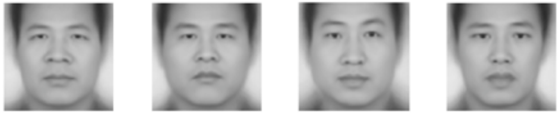
(b) Three samples in non-criminal ID photo set S_n

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.

RI_1 Performance de los algoritmos

- Verificación y Validación (?)
- "Calidad" de los datos
- El sistema es forzado a elegir entre does clases: {criminal}/{no-criminal}

RI_2 Práctica científica basada en algoritmos

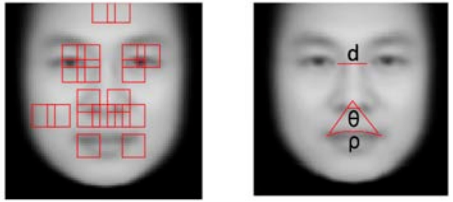
- Fosiliza conceptos tales como "criminal"
- Está desconectada de un cuerpo de conocimiento:
  - La construcción social de la criminalidad
  - Las bases socio-económicas
  - La psicología de los autores de crímenes

Automated Inference on Criminality Using Face Images

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

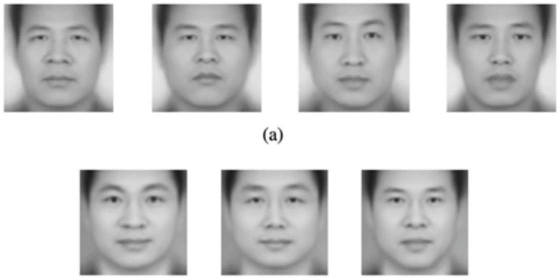
(b) Three samples in non-criminal ID photo set S_n

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.

RI$_1$ Performance de los algoritmos

- Verificación y Validación (?)
- "Calidad" de los datos
- El sistema es forzado a elegir entre does clases: {criminal}/{no-criminal}

RI$_2$ Práctica científica basada en algoritmos

- Fosiliza conceptos tales como "criminal"
- Está desconectada de un cuerpo de conocimiento:
  - La construcción social de la criminalidad
  - Las bases socio-económicas
  - La psicología de los autores de crímenes

RI$_3$ Construcción social de la fiabilidad

- Falla en dar cuenta de, o incluir valores epistémicos, éticos, etc (e.g., la presunción de inocencia)

Un límite para CR

SISE-errors

SISE-errors

SIS-Errors: Statistically insignificant, but serious errors pueden debilitar la fidelidad del algoritmo (warrant revoking the reliability of the algorithm)

SIS-Errors: the algorithm's output does not match, to the degree permissible by the relevant community, a given representation (Humphreys, 2021)

Tipos de Errores
(y de cómo los aborda el fiabilista)

SIS-Errors: Errores aleatorios

Errores aleatorios:

Son errores que no pueden anticiparse, que son impredecibles, y que son usualmente ejecuciones fallidas del algoritmo o los datos que se procesan. Estos errores no son sistemáticos ni consistentes, sino que ocurren esporádicamente y son muy difíciles de reproducir (e.g., bit flips in memory due to cosmic rays; randomised initiations ⇢ non-deterministic ML outputs)

TU Delft

SIS-Errors: Errores aleatorios

Anti "epistemic bad luck" condition: Permítase que un algoritmo se ejecute al menos dos veces bajo condiciones similares (i.e., variables de entrada, parámetros, configuraciones de datos, metaparámetros, etc.). Si el algoritmo produce SIS-Errors idénticos (o considerados lo suficientemente idénticos) en ejecuciones independientes, entonces la probabilidad de que estos resultados sean casos de "epistemic bad luck" es insignificante. Por lo tanto, tal recurrencia proporciona un fuerte respaldo a que el error es sistemático y no aleatorio.

TU Delft

SIS-Errors: Errores sistemáticos

A[1] A[2] A[3] A[4] A[5] … A[n]

Errores sistemáticos:

Son errores que no son arbitrarios, y que se reproducen a lo largo de distintas ejecuciones del mismo algoritmo. Con las herramientas apropiadas, errores sistemáticos se pueden identificar y medir (e.g., rounding off errors)

TUDelft

La estrategia

I. Hacer una estrategia epistemológica, entendida como condiciones de posibilidad para la fiabilidad. No necesariamente tiene correlato metodológico o aplicable en la práctica.

La estrategia

I. Hacer una estrategia epistemológica, entendida como condiciones de posibilidad para la fiabilidad. No necesariamente tiene correlato metodológico o aplicable en la práctica.

II. Conectar errores de representación con tipos de indicadores de fidelidad.
    i. Para poder evitar tener que revocar la fiabilidad de un algoritmo, nuestra mejor estrategia es identificar IF que han fallado (inadecuados, incorrectos, o ausentes)

TUDelft

Clases de errores sistemáticos

"Once you unleash it on large data, deep learning has its own dynamics, it does its own repair and its own optimisations, and it gives you the right results most of the time. But when it doesn't, you don't have a clue abut what went wrong and what should be fixed. In particular, you d not now if the fault is in the program, in the method, or because things have changed in the environment." (Pearl, 2019, 15)

Clases de errores sistemáticos

"Once you unleash it on large data, deep learning has its own dynamics, it does its own repair and its own optimisations, and it gives you the right results most of the time. But when it doesn't, you don't have a clue abut what went wrong and what should be fixed. In particular, you d not now if the fault is in the program, in the method, or because things have changed in the environment." (Pearl, 2019, 15)

- Errores de clase$_1$: errores que ocurren cuando el algoritmo produces resultados incorrectos dado calcular mal (e.g., rounding off errors, overflow, etc)

Clases de errores sistemáticos

"Once you unleash it on large data, deep learning has its own dynamics, it does its own repair and its own optimisations, and it gives you the right results most of the time. But when it doesn't, you don't have a clue abut what went wrong and what should be fixed. In particular, you d not now if the fault is in the program, in the method, or because things have changed in the environment." (Pearl, 2019, 15)

- Errores de clase$_1$: errores que ocurren cuando el algoritmo produces resultados incorrectos dado calcular mal (e.g., rounding off errors, overflow, etc)
- Errores de clase$_2$: errores que surgen cuando los métodos o técnicas implementadas en el algoritmo son inapropiadas para un objetivo en particular (e.g., usar el "sorting algorithm" incorrecto, o la implementación de conceptos núcleo {criminal/no-criminal})

Clases de errores sistemáticos

"Once you unleash it on large data, deep learning has its own dynamics, it does its own repair and its own optimisations, and it gives you the right results most of the time. But when it doesn't, you don't have a clue abut what went wrong and what should be fixed. In particular, you d not now if the fault is in the program, in the method, or because things have changed in the environment." (Pearl, 2019, 15)

- Errores de $clase_1$: errores que ocurren cuando el algoritmo produces resultados incorrectos dado calcular mal (e.g., rounding off errors, overflow, etc)
- Errores de $clase_2$: errores que surgen cuando los métodos o técnicas implementadas en el algoritmo son inapropiadas para un objetivo en particular (e.g., usar el "sorting algorithm" incorrecto, o la implementación de conceptos núcleo {criminal/no-criminal})
- Errores de $clase_3$: errores que ocurren cuando el entorno hace el algoritmo y sus resultados irrelevantes (e.g., detección facial en los tiempos del COVID-19)

Errores e Indicadores de Fiabilidad

Errores de clase$_1$ → RI_1 Performance de los algoritmos

Errores de clase$_2$ → RI_2 Práctica científica basada en algoritmos

Errores de clase$_3$ → RI_3 Construcción social de la fiabilidad

Errores e Indicadores de Fiabilidad

Errores de clase$_1$

Errores de clase$_2$

Errores de clase$_3$

RI$_1$ Performance de los algoritmos

RI$_2$ Práctica científica basada en algoritmos

RI$_3$ Construcción social de la fiabilidad

Errores e Indicadores de Fiabilidad

Errores de clase$_1$	inadecuados	RI_1 Performance de los algoritmos
Errores de clase$_2$	incorrectos	RI_2 Práctica científica basada en algoritmos
Errores de clase$_3$	ausentes	RI_3 Construcción social de la fiabilidad

IF inadecuados

SPRINGER NATURE Link

Find a journal　Publish with us　Track your research　Search

Home > Advances in Biometrics > Conference paper

Why Is Facial Occlusion a Challenging Problem?

Conference paper

pp 299–308 | Cite this conference paper

Advances in Biometrics

(ICB 2009)

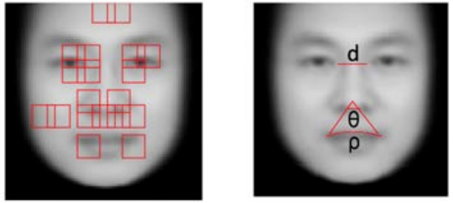
Download book PDF

IF incorrectos

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

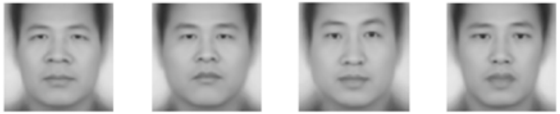
(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.



(a) (b)

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

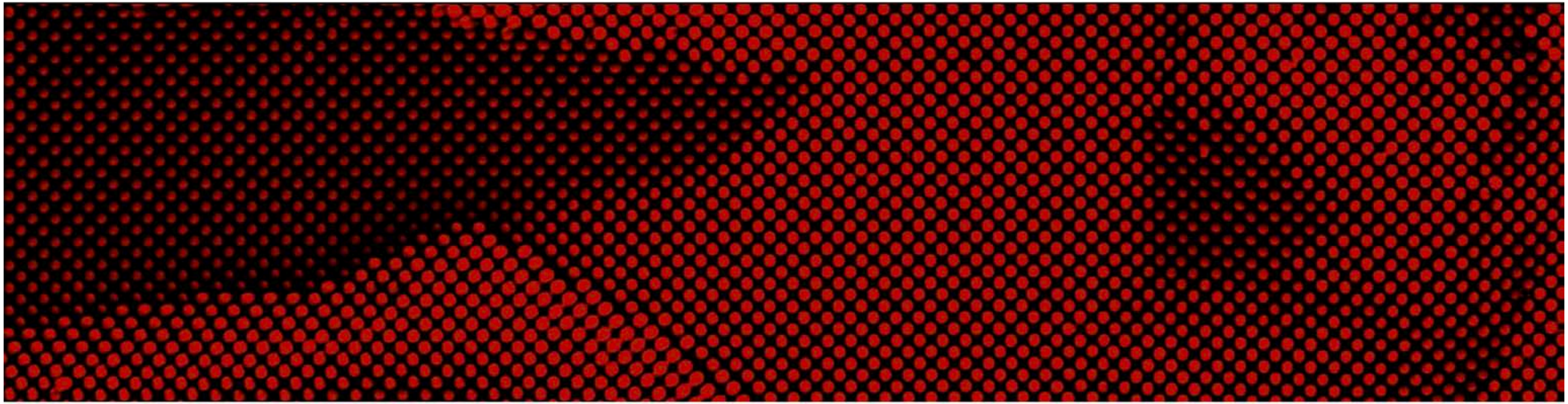


(a)

(b)

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

IF ausentes



VIDEO: SAM CANNON

MAIA SZALAVITZ THE BIG STORY AUG 11, 2021 6:00 AM

The Pain Was Unbearable. So Why Did Doctors Turn Her Away?

A sweeping drug addiction risk algorithm has become central to how the US handles the opioid crisis. It may only be making the crisis worse.

La estrategia

I. Hacer una estrategia epistemológica, entendida como condiciones de posibilidad para la fiabilidad. No necesariamente tiene correlato metodológico o aplicable en la práctica.

II. Conectar errores de representación con tipos de indicadores de fidelidad.

    i. Para poder evitar tener que revocar la fiabilidad de un algoritmo, nuestra mejor estrategia es identificar IF que han fallado (inadecuados, incorrectos, o ausentes)

III. Establezco cláusulas de fiabilidad que, de violarse, nos obligan a suspender nuestras creencias —y así resguardar la fiabilidad (i.e., no tener que revocarla)

SIS-Errors

Errores de $clase_1$

SIS-Errors $\dashrightarrow$ Errores de $clase_2$

Errores de $clase_3$

TU Delft

Errores de clase$_1$ → RI$_1$ Robustez técnica de los algoritmos

SIS-Errors → Errores de clase$_2$ → RI$_2$ Práctica científica basada en algoritmos

Errores de clase$_3$ → RI$_3$ Construcción social de la fiabilidad

SIS-Errors → Errores de clase$_2$

Errores de clase$_1$ ··· inadecuados → RI$_1$ Robustez técnica de los algoritmos

Errores de clase$_2$ ··· incorrectos → RI$_2$ Práctica científica basada en algoritmos

Errores de clase$_3$ ··· ausentes → RI$_3$ Construcción social de la fiabilidad

SIS-Errors

Errores de $clase_1$ inadecuados RI_1 Robustez técnica de los algoritmos

Errores de $clase_2$ incorrectos RI_2 Práctica científica basada en algoritmos

Errores de $clase_3$ ausentes RI_3 Construcción social de la fiabilidad

Mantener fiabilidad?

condiciones para suspender, revisar o anular la formación de creencia

SIS-Errors: Errores sistemáticos clase$_1$

A[1] A[2] A[3] A[4] A[5] A[n]

- **Conditional reliable token-RI**: is one that confers reliability [to the new context] if the methods, standards, and breadth of application are based on are also reliable [for the new context]

SIS-Errors: Errores sistemáticos clase$_1$

- **Conditional reliable token-RI**: is one that confers reliability [to the new context] if the methods, standards, and breadth of application are based on are also reliable [for the new context]

Idea: Realiza un seguimiento de la representación evaluando la probabilidad de que el token-RI mantenga la confiabilidad en todos los contextos.

Nos permite mantener nuestras creencia en la fiabilidad del algoritmo en tanto y en cuanto las condiciones iniciales que dieron fiabilidad todavía se aplican en contextos nuevos

SIS-Errors: Errores sistemáticos clase$_1$

- **Conditional reliable token-RI**: is one that confers reliability [to the new context] if the methods, standards, and breadth of application are based on are also reliable [for the new context]

Ejemplo (IF inadecuado - suspendo creencia): un algoritmo que fue diseñado para detección de rostros aplicado en un nuevo contexto (e.g., durante el uso de COVID-19) necesita una actualización de la base de datos de entrenamiento, un nuevo módulo que trate con oclusión facial, etc

SIS-Errors: Errores sistemáticos clase$_2$

- **Anti-defeat clause**: *S* is epistemically warranted in maintaining token-RI as a suitable reliability indicator unless there is a defeater token-RI* epistemically available to *S* such that, if *S* were to use toke-RI*, *S* would no longer hold the belief that the output represents a fact *F*

SIS-Errors: Errores sistemáticos clase$_2$

- **Anti-defeat clause**: *S* is epistemically warranted in maintaining token-RI as a suitable reliability indicator unless there is a defeater token-RI* epistemically available to *S* such that, if *S* were to use toke-RI*, *S* would no longer hold the belief that the output represents a fact *F*

Idea: El estatus justificado de una creencia se preserva mientras que la IF siga valiendo frente a alternativas nuevas y que potencialmente remueven credibilidad.

Mantenemos la fiabilidad en tanto y en cuanto no encontremos un defeater. En ese caso, estamos epistémicamente obligados a adoptarlo

SIS-Errors: Errores sistemáticos clase$_2$

A[1] A[2] A[3] A[4] A[5] ... A[n]

- **Anti-defeat clause**: *S* is epistemically warranted in maintaining token-RI as a suitable reliability indicator unless there is a defeater token-RI* epistemically available to *S* such that, if *S* were to use toke-RI*, *S* would no longer hold the belief that the output represents a fact *F*

Ejemplo (IF incorrecto - reviso de la creencia): Un algoritmo que selecciona {criminal, no criminal} basado en características faciales: token-IF = distancia entre los ojos, curvatura boca, etc.

Defeater, más robusto: token-IF* = implementación de conceptos socio-económicos de criminalidad

SIS-Errors: Errores sistemáticos clase₃

- **Supercharging type-RI$_3$ indicator**: Maintaining algorithmic reliability depends on subjecting their outputs to debate and other forms of scientific engagement. In this sense, the social construction of belief plays a crucial role in determining reliability and can, at times, take precedence over other indicators

SIS-Errors: Errores sistemáticos clase$_3$

A[1] A[2] A[3] A[4] A[5] … A[n]

- **Supercharging type-RI$_3$ indicator**: Maintaining algorithmic reliability depends on subjecting their outputs to debate and other forms of scientific engagement. In this sense, the social construction of belief plays a crucial role in determining reliability and can, at times, take precedence over other indicators

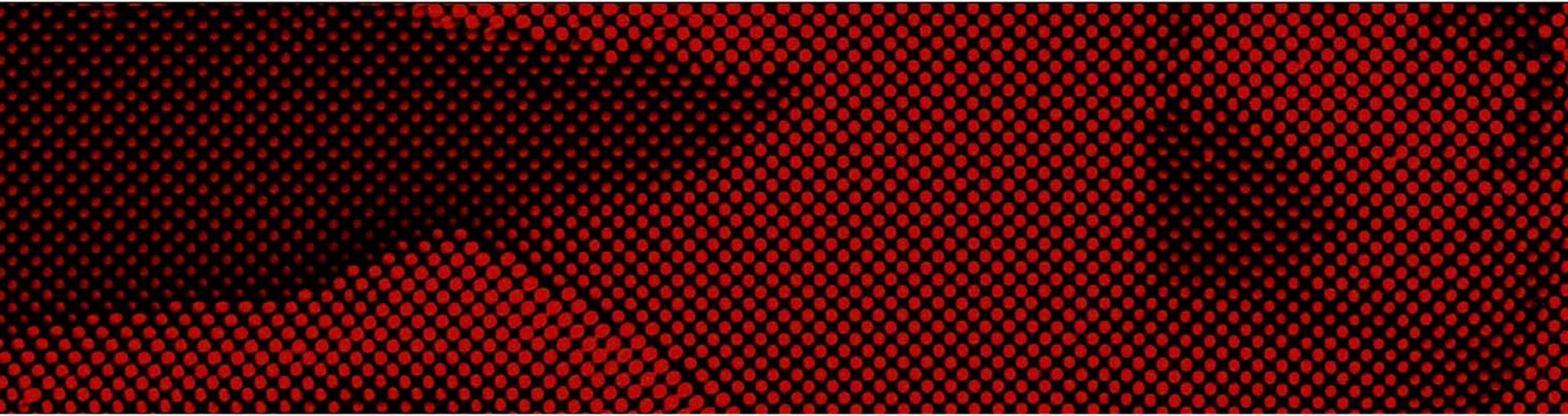
Idea: principio preventivo epistémico, donde la comunidad relevante evalúa los méritos de los resultados del algoritmo (tiene poder de veto!)

TU Delft

SIS-Errors: Errores sistemáticos clase$_3$

- **Supercharging type-RI$_3$ indicator**: Maintaining algorithmic reliability depends on subjecting their outputs to debate and other forms of scientific engagement. In this sense, the social construction of belief plays a crucial role in determining reliability and can, at times, take precedence over other indicators

Ejemplo (ausencia de IF - anulo creencia): Kathryn!



VIDEO: SAM CANNON

MAIA SZALAVITZ THE BIG STORY AUG 11, 2021 6:00 AM

The Pain Was Unbearable. So Why Did Doctors Turn Her Away?

A sweeping drug addiction risk algorithm has become central to how the US handles the opioid crisis. It may only be making the crisis worse.

https://www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/

Recapitulando

TU Delft

Recapitulando

- Externalista:
    - La justificación no depende de un "third-party algorithms" y es independiente de nuestras capacidad por entender algoritmos
- Cara indicador de fiabilidad está justificado independientemente
- Hay una variedad de RI (e.g., es una epistemología "no-centralizada")
- En principio se pueden detectar indicadores de resultados no creíbles/falsos
- Se acomoda a diferentes necesidades epistémicas/éticas: en algunos sistemas, los métodos de verificación y validación son más importantes, mientras que en otros es la coherencia teórica, mientras que en otros es el alineamiento con leyes vigentes

Recapitulando

- SIS-Errors contituyen un problema para el fiabilista (para toda epistemología)
  - Condiciones para suspender nuestras creencias—y mantener fiabilidad?
  - Problemas de orden práctico
- Más problemas:
  - No es claro la precedencia, el orden, el peso, etc de cada indicador.
  - No es claro cómo se procede cuando hay conflicto entre los indicadores
  - *La tiranía de unos pocos indicadores*: unos pocos indicatores pueden ser suficiente para desbalancear la fiabilidad/no-fiabilidad de un algoritmo

Muchas gracias!

j.m.duran@tudelft.nl

Johannes Vermeer, *Girl with a Pearl Earring*, c. 1665 - https://www.mauritshuis.nl/