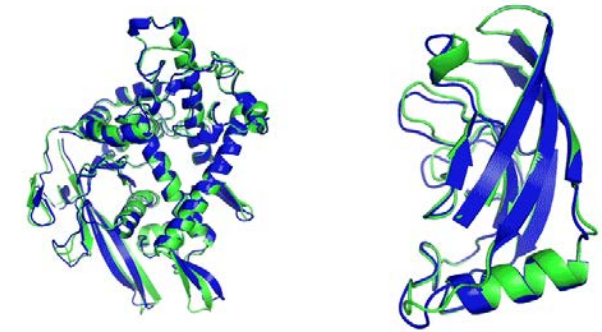
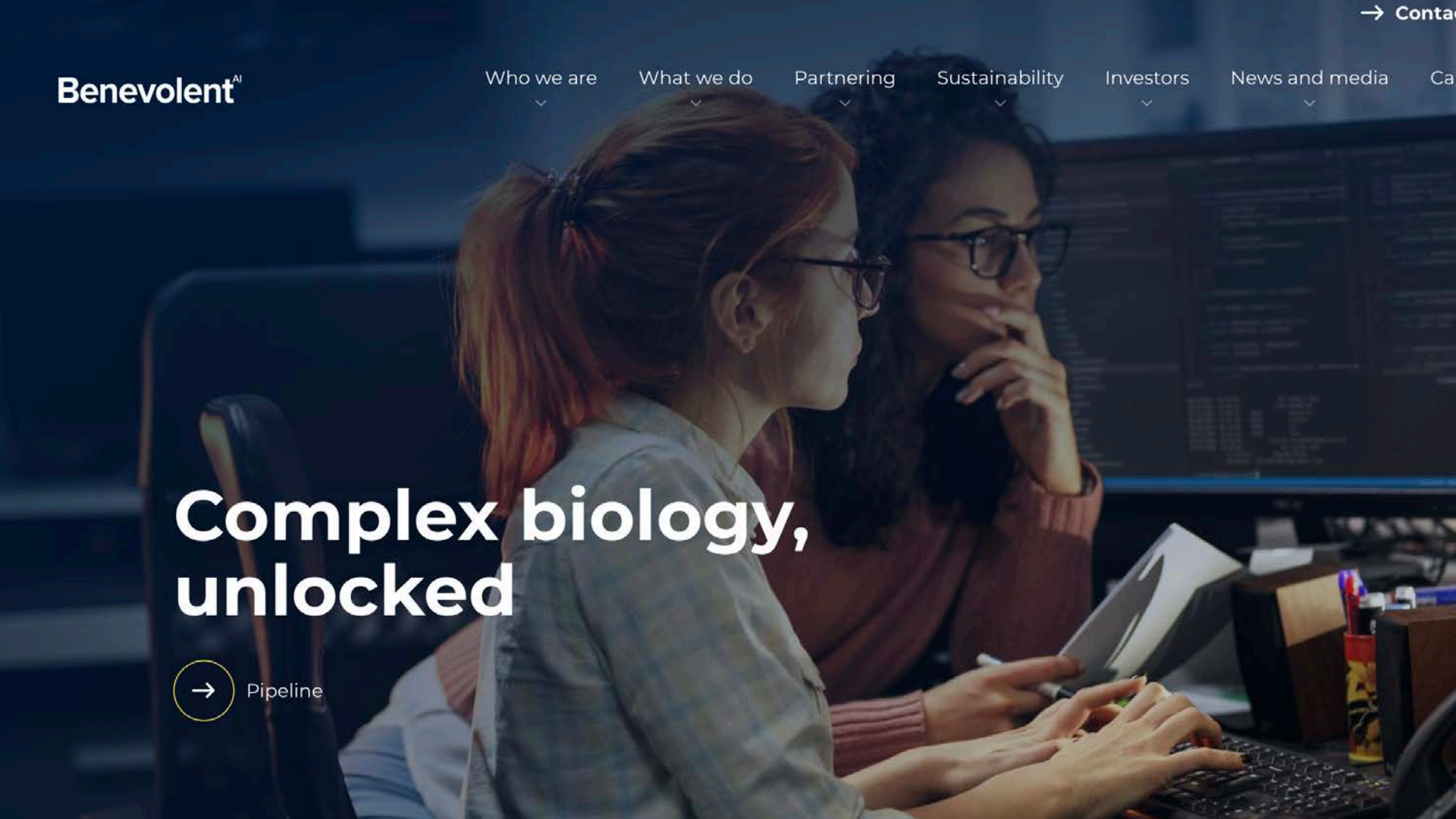
Against epistemic transparency in algorithmic science

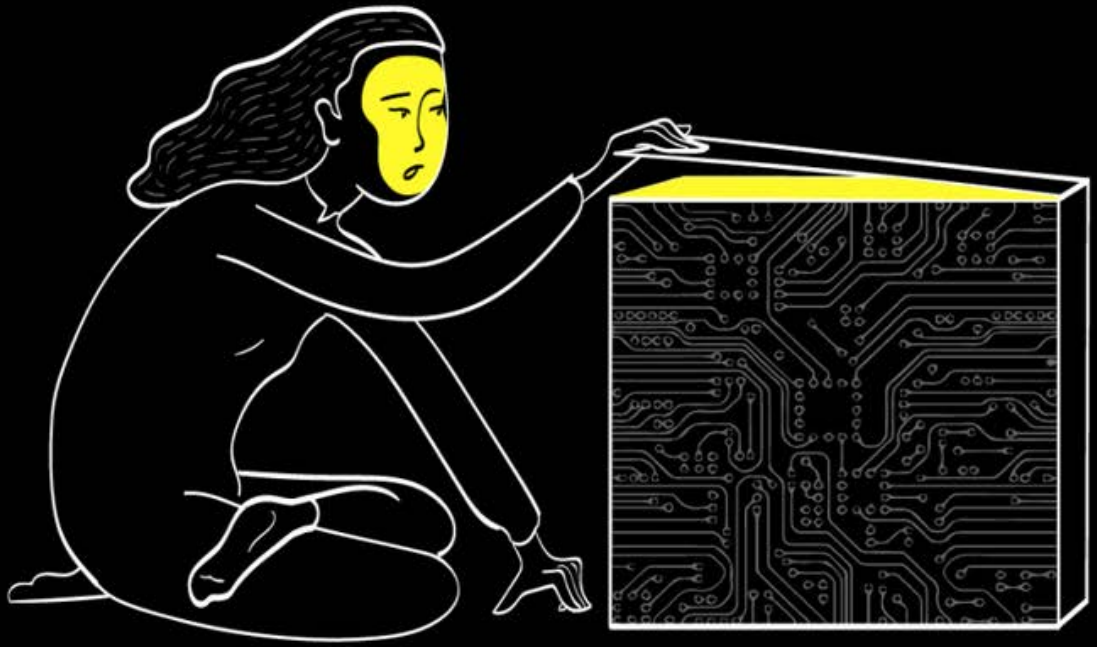
Juan M. Durán

TU Delft



Jumper, J., Evans, R., Pritzel, A. et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589


→ 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)

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

- 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)

Opacidad y Transparencia

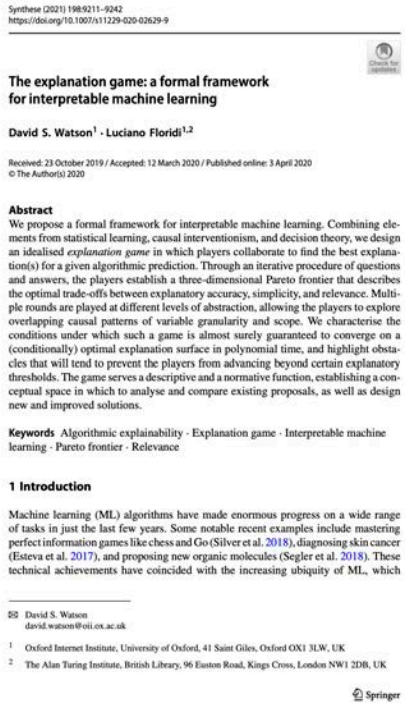
Creel: “opacity and transparency are two sides of the same coin: opacity is a lack of transparency and vice versa” (2020, FN2)

> Boge: “I take it for granted that epistemic opacity is relative to an agent and involves a lack of knowledge […] *h-opacity* may concern all three forms of transparency in complex computational systems identified by Creel” (2022, 15. FN 18)

Lipton: “Informally, transparency is the opposite of opacity or “black-boxness.” It connotes some sense of understanding the mechanism by which the model works” (2017)

> Zerilli’s *fathomability* (2022)

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

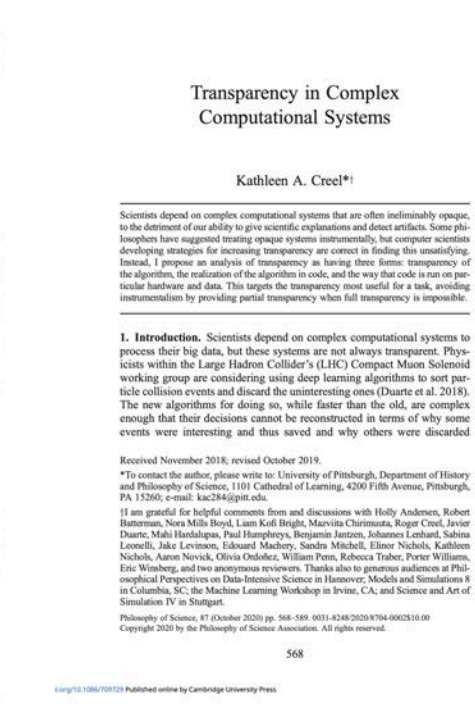
1 Introduction

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



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).

¿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)

TU Delft

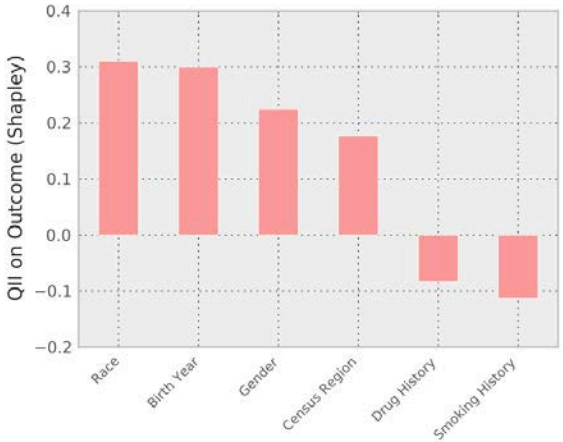
Transparency reports: arrests

Quantitative Input Influence (QII) measures the influence of (sets of) input on a quantity of interest —understood as representing a property of the behavior of the system.

Mr. Z: […] is from the arrests dataset. History of drug use and smoking are both strong indicators of arrests. However, Mr. X received positive classification by this classifier even without any history of drug use or smoking. On examining his classifier, it appears that race, age and gender were most influential in determining his outcome. In other words, the classifier that we train for this dataset (a decision forest) has picked up on the correlations between race (Black), and age (born in 1984) to infer that this individual is likely to engage in criminal activity. Indeed, our interventional approach indicates that this is not a mere correlation effect: race is actively being used by this classifier to determine outcomes (Datta et al, 2016, 609)

Birth Year	1984
Drug History	None
Smoking History	None
Census Region	West
Race	Black
Gender	Male

(a) Mr. Z's profile



(b) Transparency report for Mr. Z's positive classification

Fig. 6: Mr. Z.

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?

Transparencia y evidencia(lismo)

Kroll et al. “In algorithmic systems, **transparency provides the evidence** needed to assess the system’s fairness and accuracy. Without transparency, stakeholders are forced to **accept or reject results** without the **supporting evidence** necessary to form justified beliefs about those outcomes.”

Vallor: “Transparency is key to fostering **epistemic trust** in artificial intelligence. Only when the **reasons behind an algorithm’s decisions are clear** can people form well-supported beliefs about its reliability and fairness. In this way, transparency contributes to justified trust in algorithmic processes”

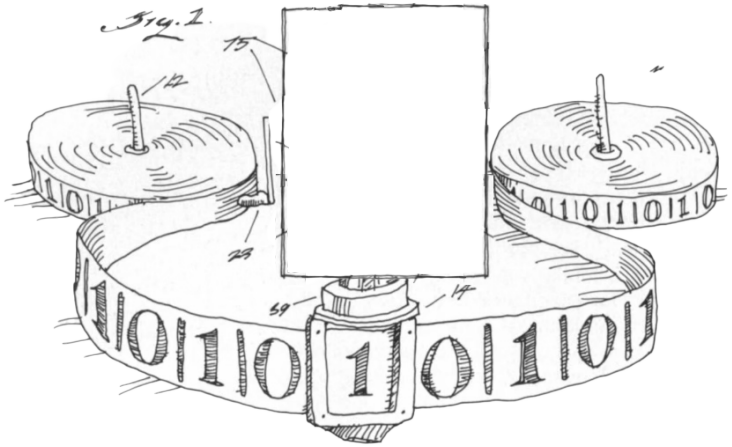
O’Neil: “**Transparency** allows people to see and evaluate the **reasons** embedded in algorithms, supporting a reasoned belief in their fairness—or exposing biases. Without such transparency, beliefs about an algorithm’s fairness are based on faith rather than evidence.”

Transparencia y evidencia(lismo)

Creel about *Post Hoc Explanation and LIME*: "because the diagnostic systems do not give an explanation or **reason** for the diagnosis, doctors often deem them **untrustworthy** and avoid them. Applying LIME to such a system **gives doctors a rationale** for the existing system's diagnoses. Because of the nature of the algorithms used, however, the diagnostic systems did not already contain those **reasons** in human-understandable form."

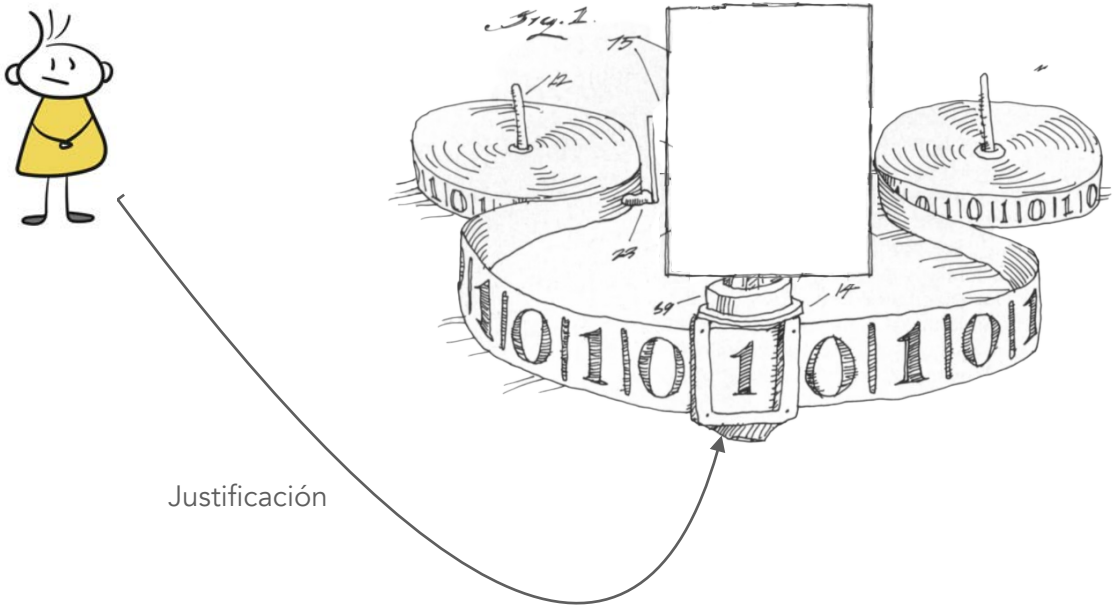
Burrell: "One approach to building more interpretable classifiers is to implement an enduser facing component to provide not only the classification outcome, but also **exposing some of the logic of this classification**. A real-world implementation of this in the domain of spam filtering is found in Google's gmail 'spam' folder. If you select a spam message in this folder, a yellow alert box with the query 'why is this message in Spam?' above the text of the email itself provides one **reason** why it has been placed in this folder"

Justificación via Transparencia (a sketch)



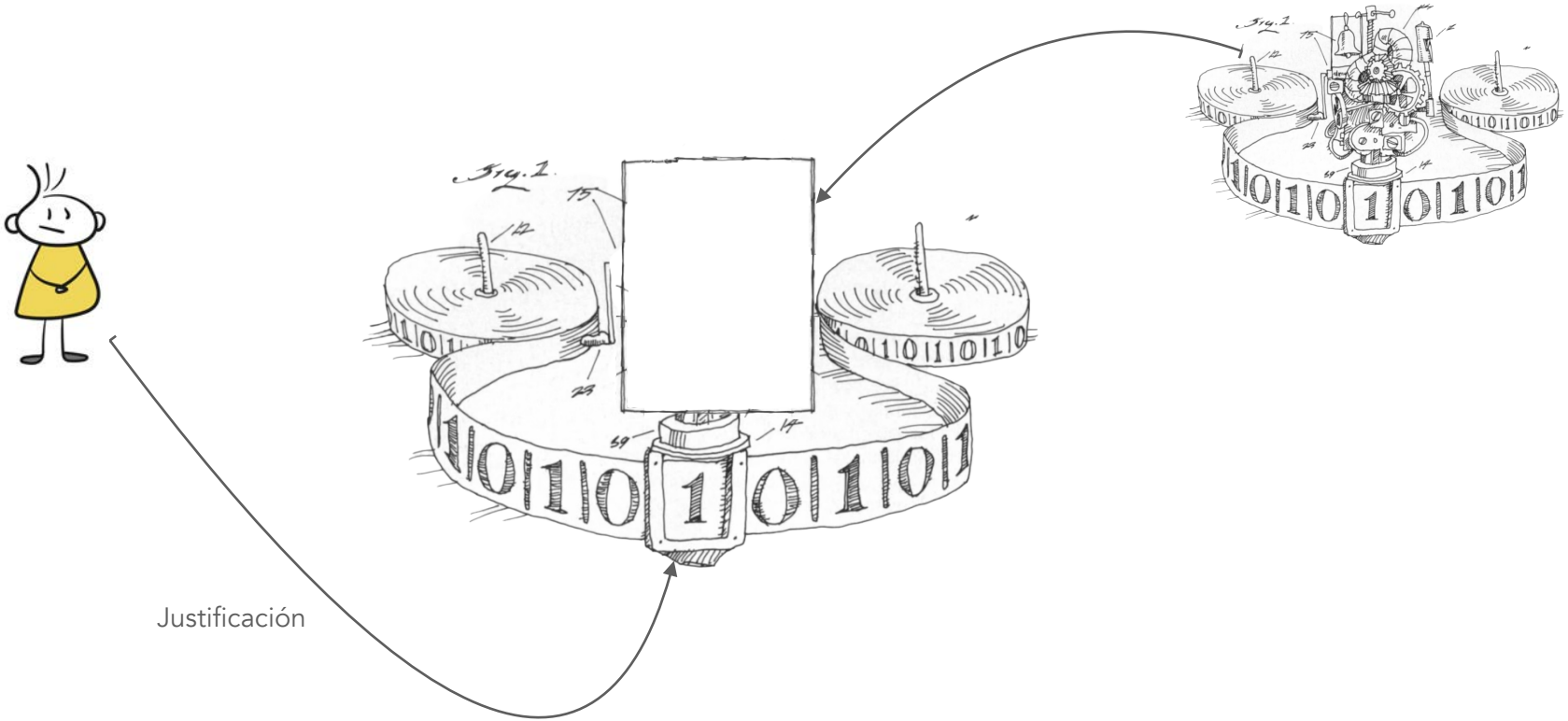
http://bridgethevoid.blogspot.com/

Justificación via Transparencia (a sketch)



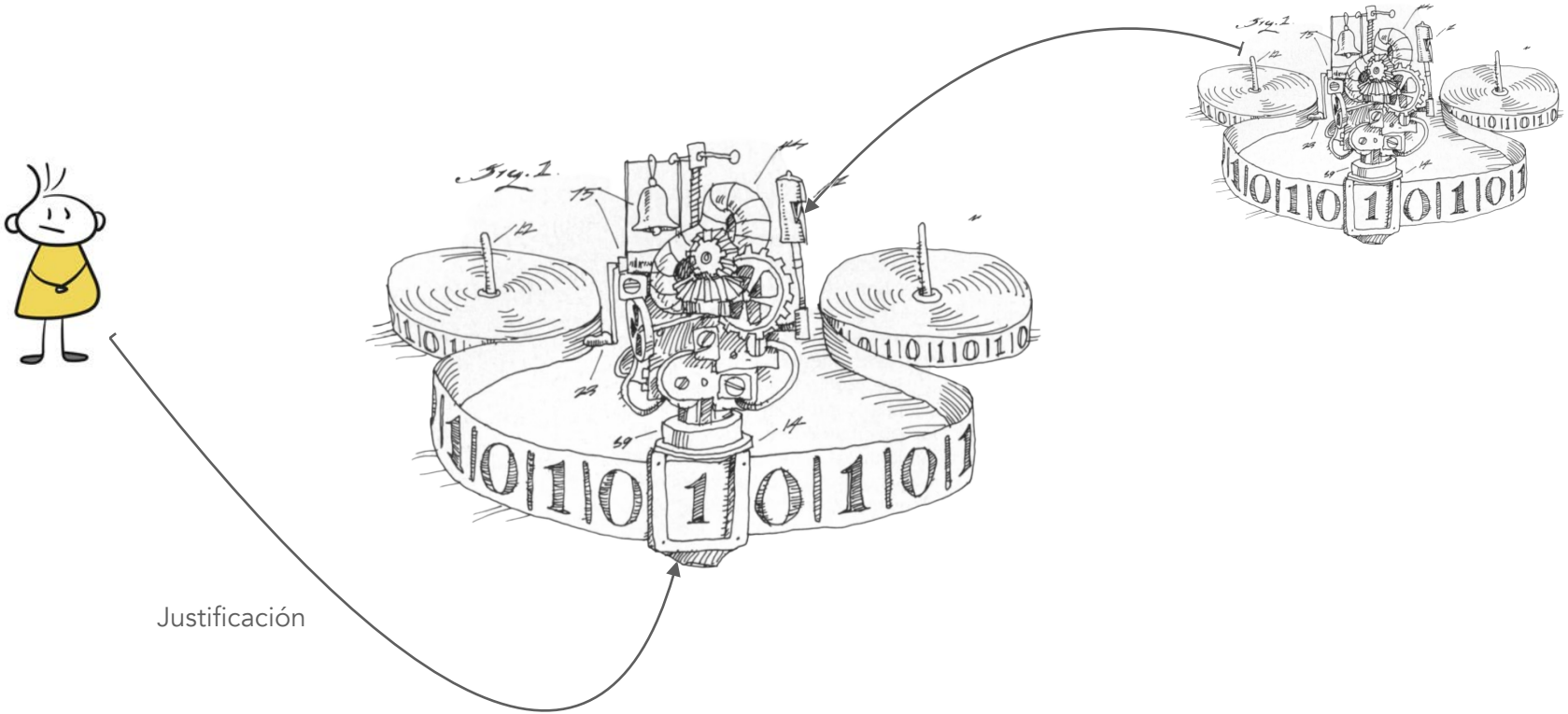
http://bridgethevoid.blogspot.com/

Justificación via Transparencia (a sketch)



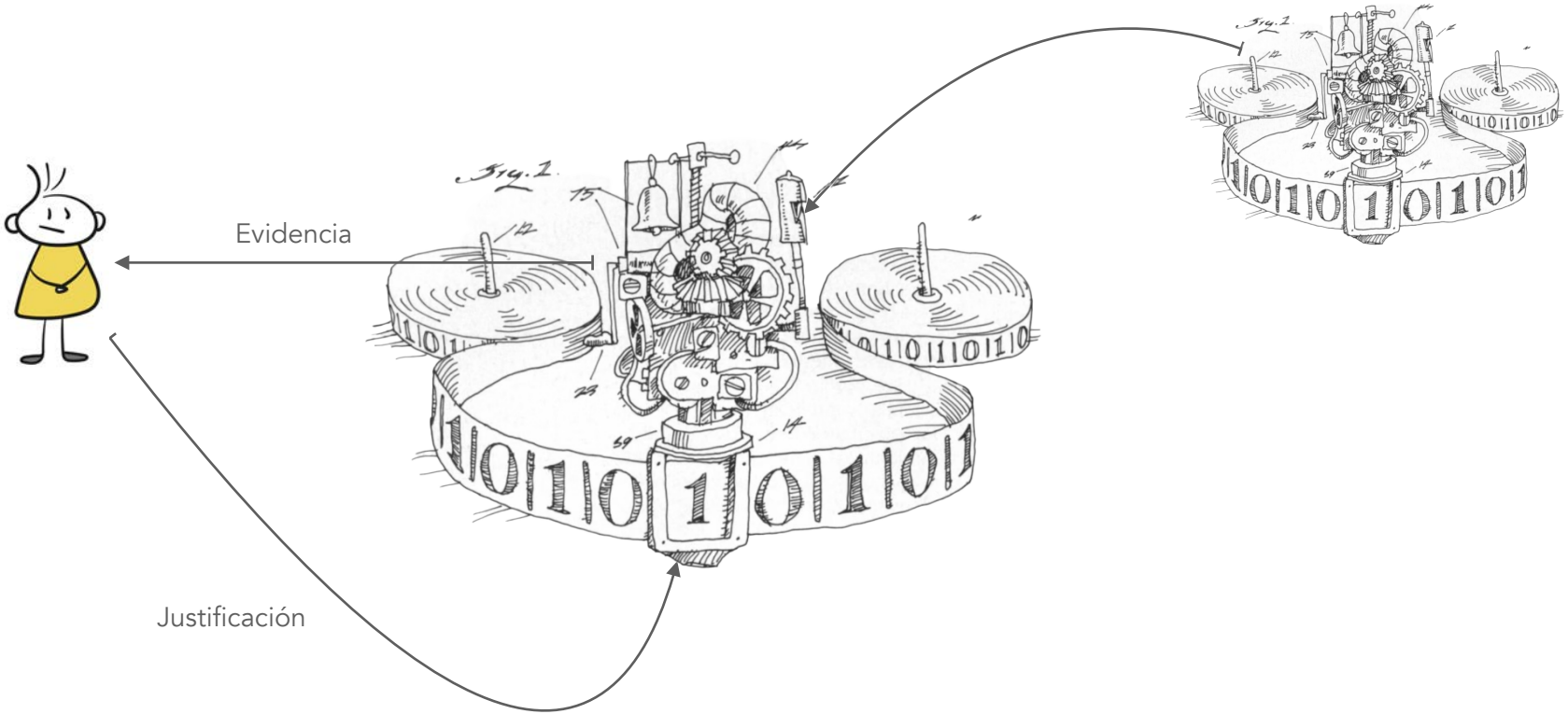
http://bridgethevoid.blogspot.com/

Justificación via Transparencia (a sketch)



http://bridgethevoid.blogspot.com/

Justificación via Transparencia (a sketch)



http://bridgethevoid.blogspot.com/

Puntos clave sobre transparency

- **¿Qué?:** Transparencia se opone a opacidad mostrando los elementos epistémicamente relevantes internos al algoritmo —i.e., reduce la falta de conocimiento de cómo el output se generó
  - NB: transparencia es una epistemología **interna-al-algoritmo**

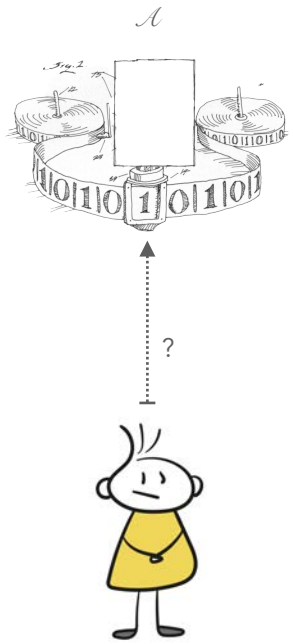
- **¿Cómo?:** XAI, Transparency reports, decision trees, saliency maps, LIME/SHAP

- **¿Por qué?:** Transparencia forma creencias a través de evidencia que el output es:
  - Creíble, confiable, verdadero, científicamente válido, correcto, representa, etc.

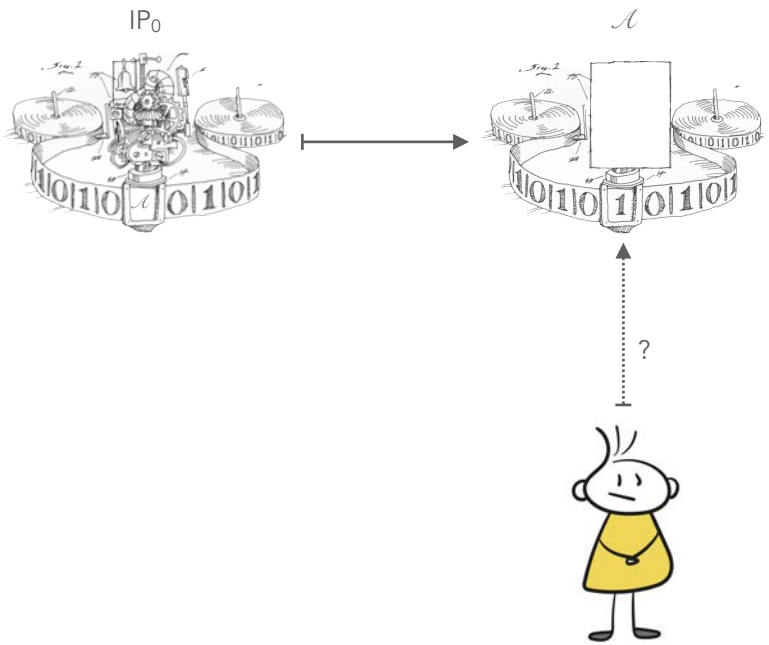
¿Por qué transparencia es inadecuada para justificar?

§1 Transparency regress

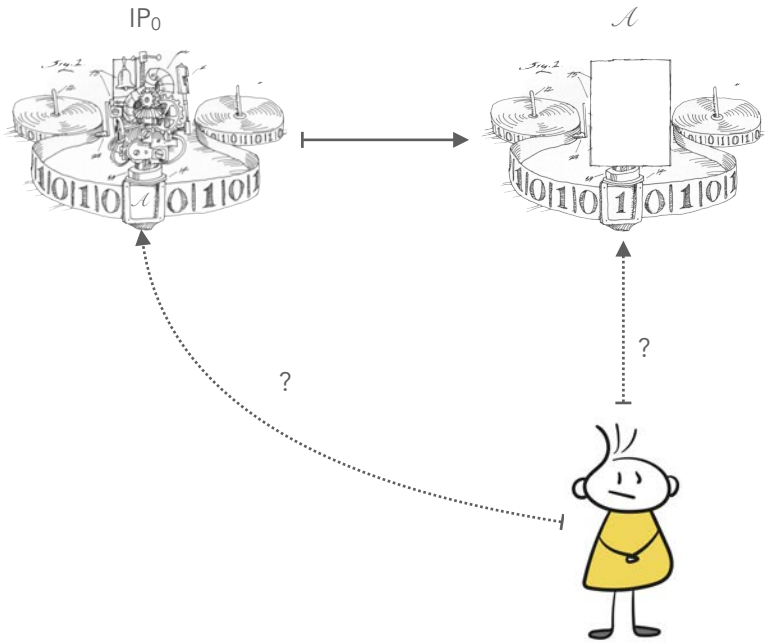
Transparency regress



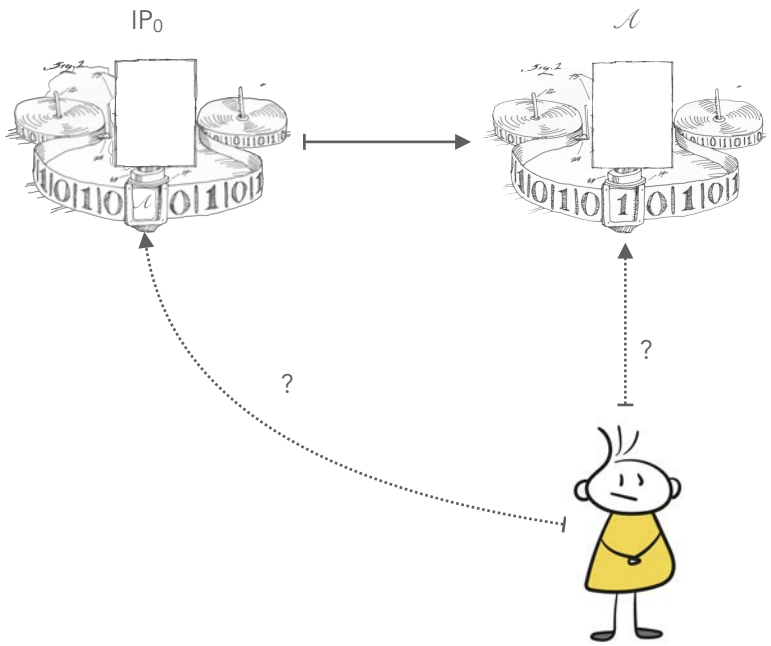
Transparency regress



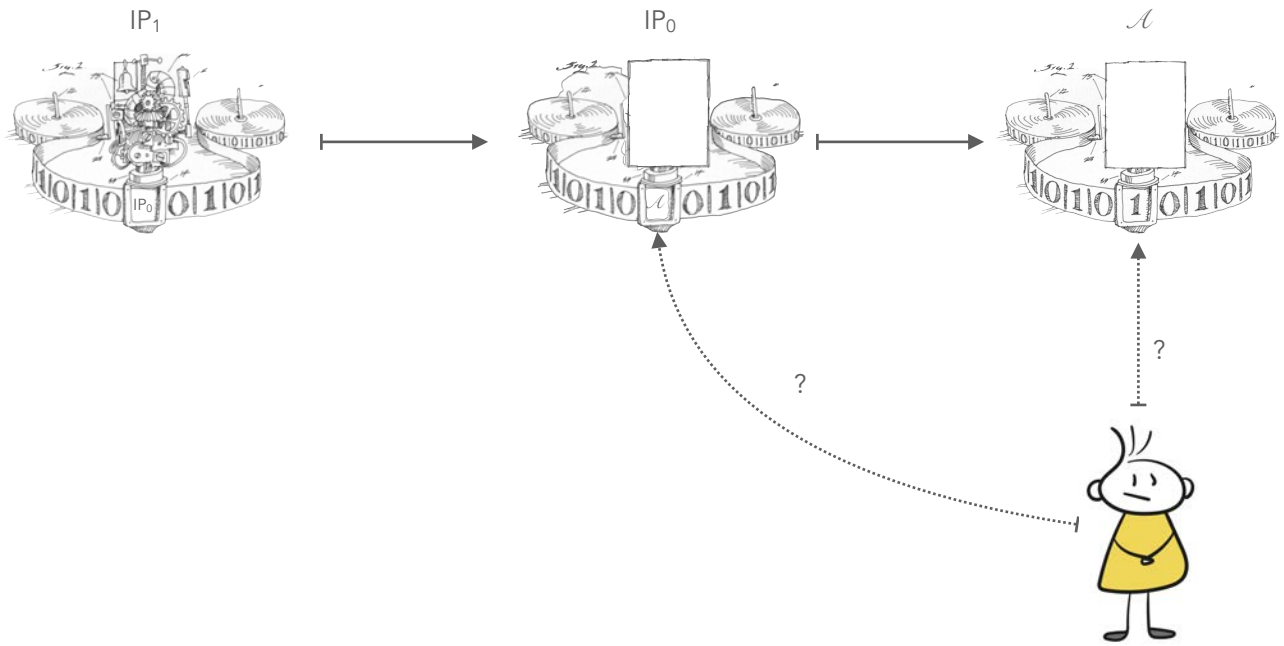
Transparency regress



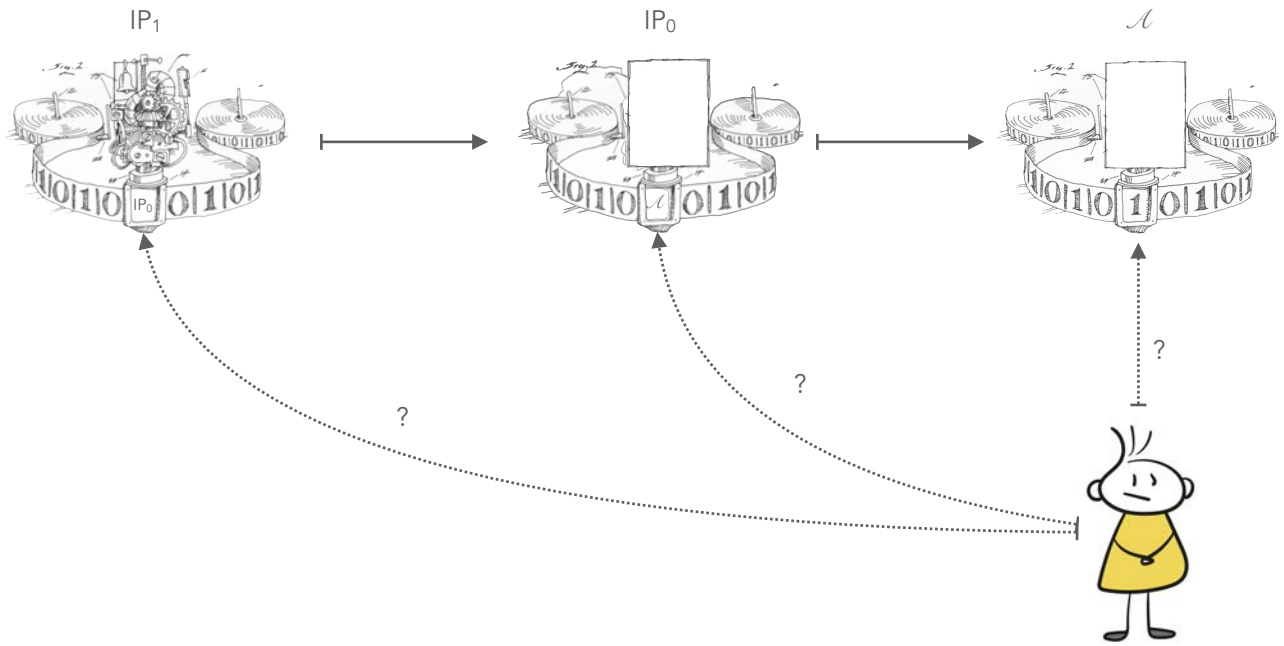
Transparency regress



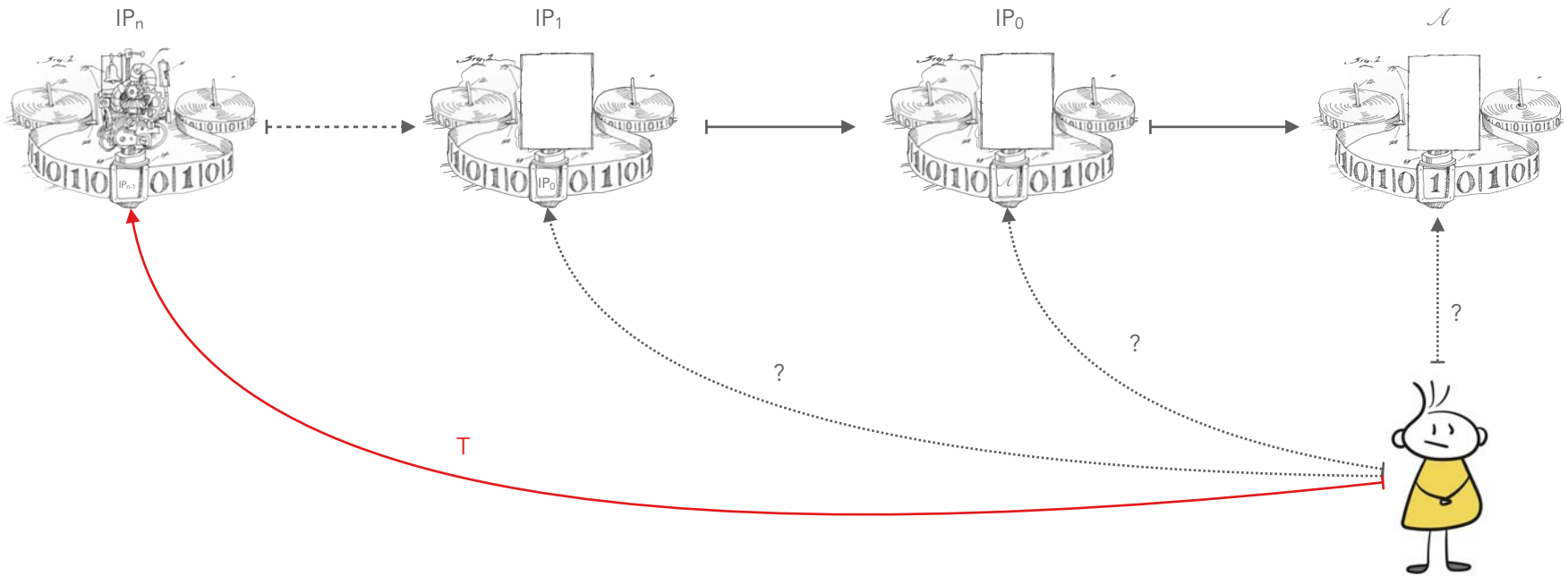
Transparency regress



Transparency regress

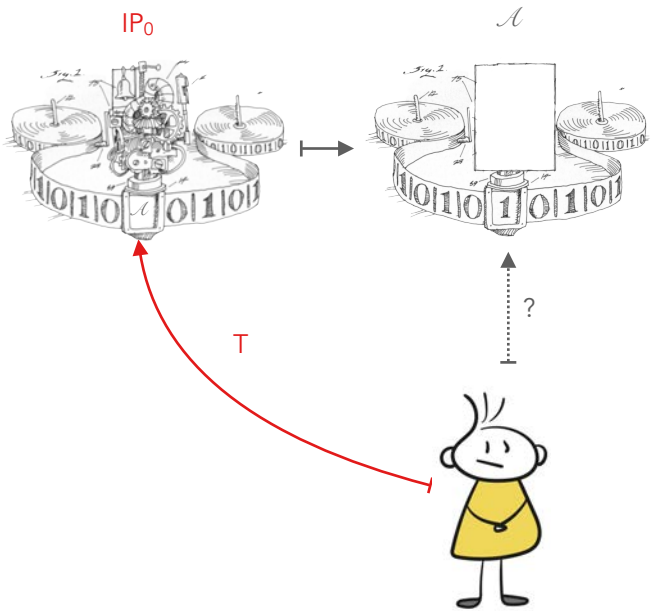


Transparency regress



Solution: Level-0 Transparency

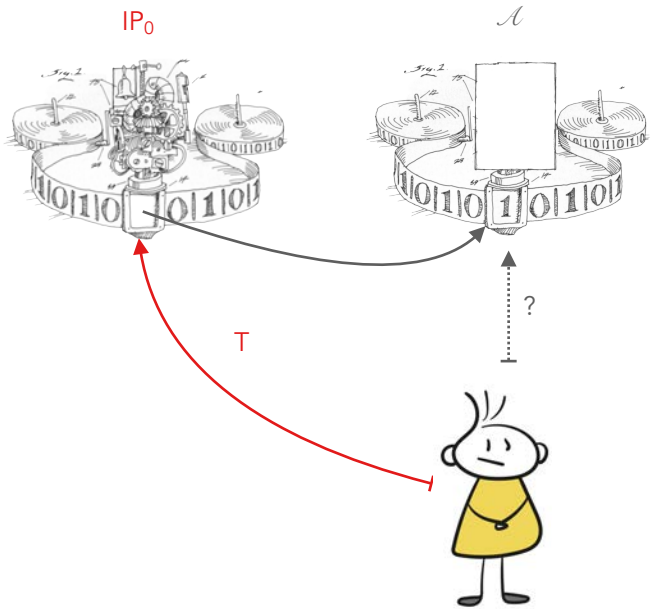
Level-0 Transparency: IP_0



IP_0, external to $\mathcal{A}$, transparent, which provides supporting evidence to believe $\mathcal{A}$, which in turn provides reasons to believe that *O*

Guidotti et al (2018)

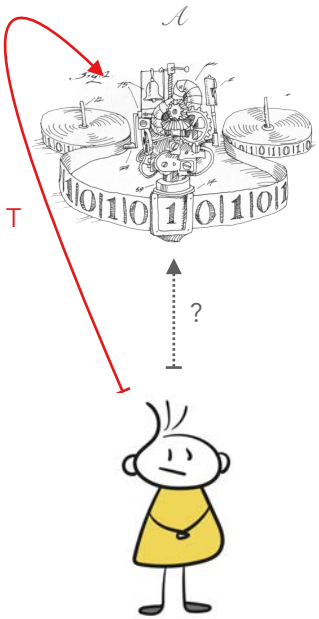
Level-0 Transparency: Agnostic models



IP_0, external to $\mathcal{A}$, transparent, provides supporting evidence to believe that *O*

Model agnostics such as LIME

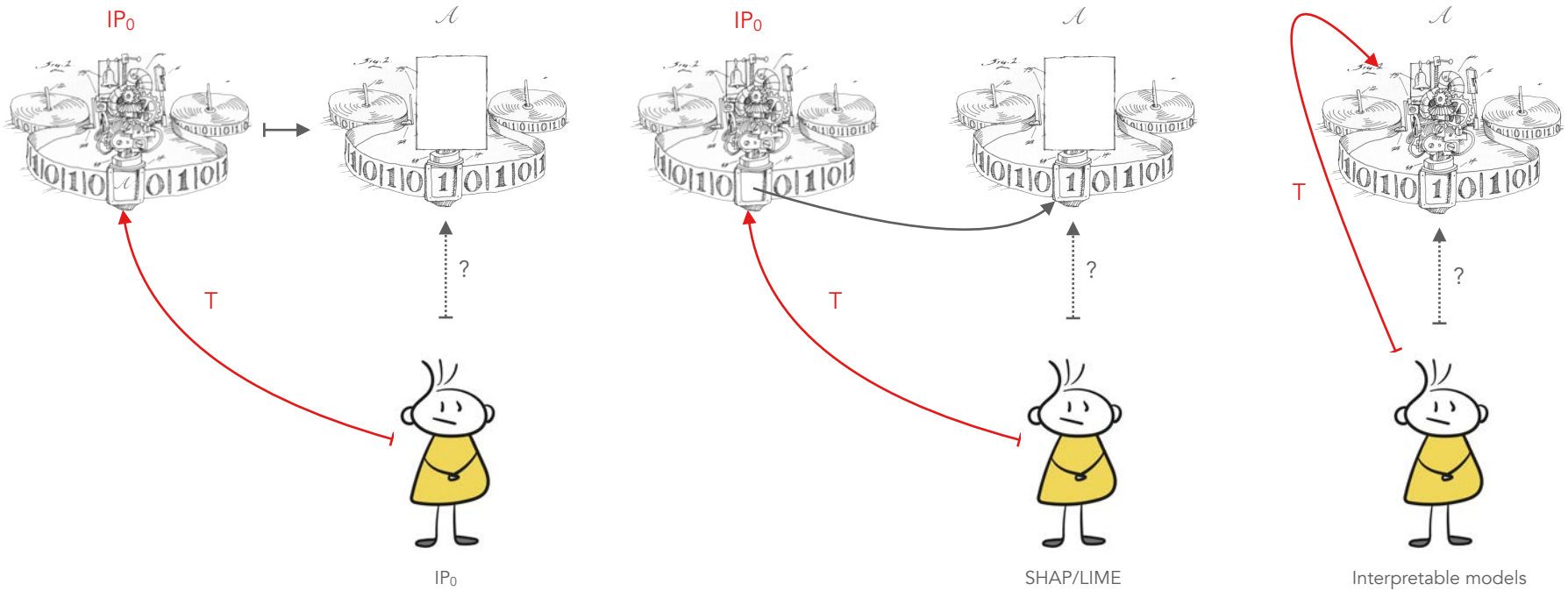
Level-0 Transparency: Interpretability



$\mathcal{A}$ itself transparent gives supporting evidence to believe that O

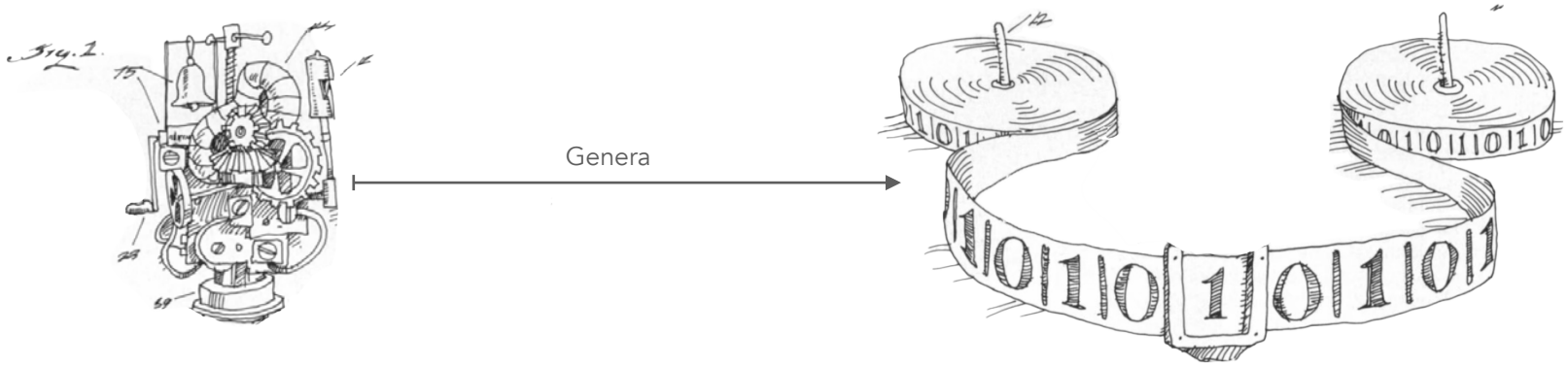
Interpretable models — Rudin (2019)

Level-0 Transparency

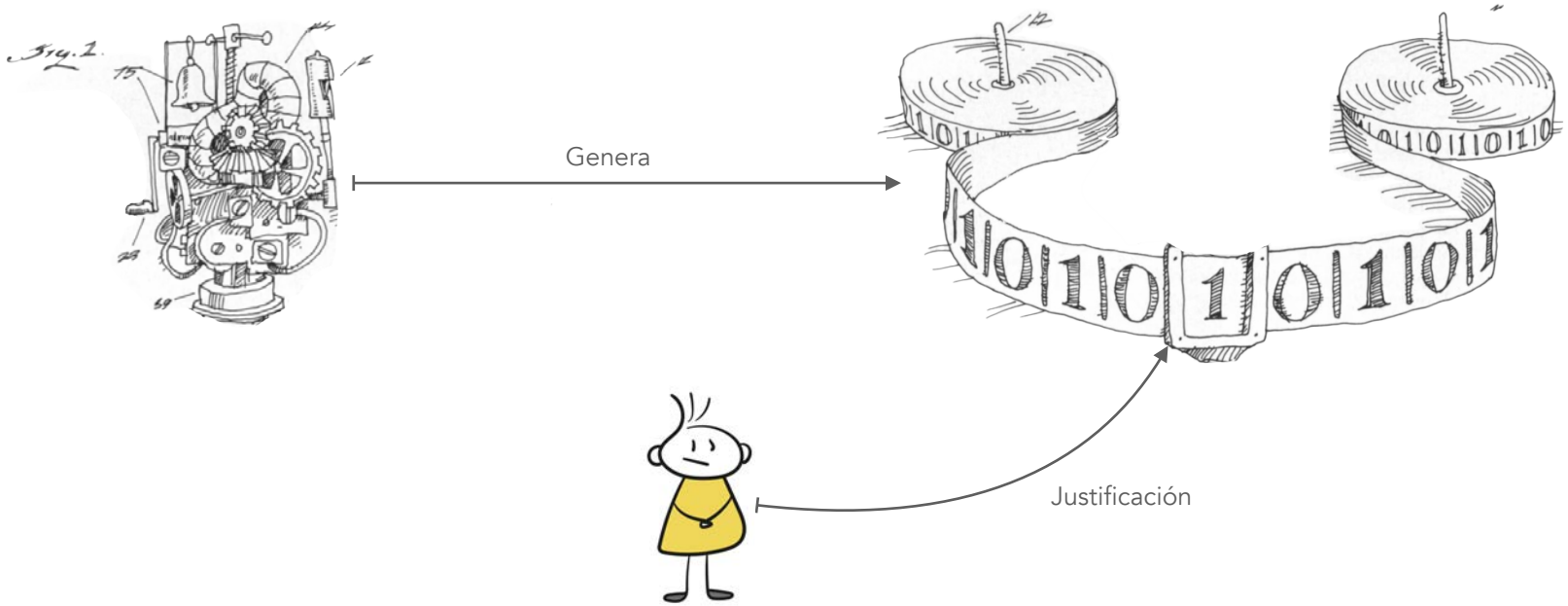


§2 Self-justification

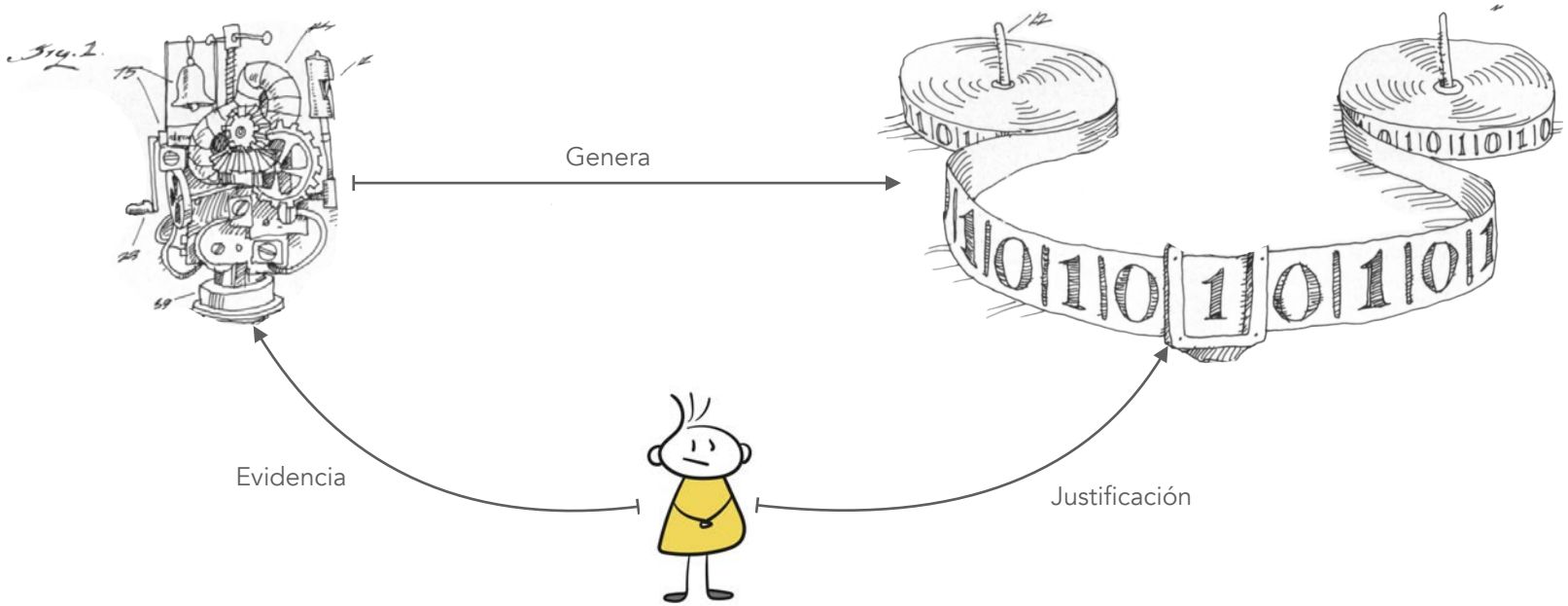
Self-justification



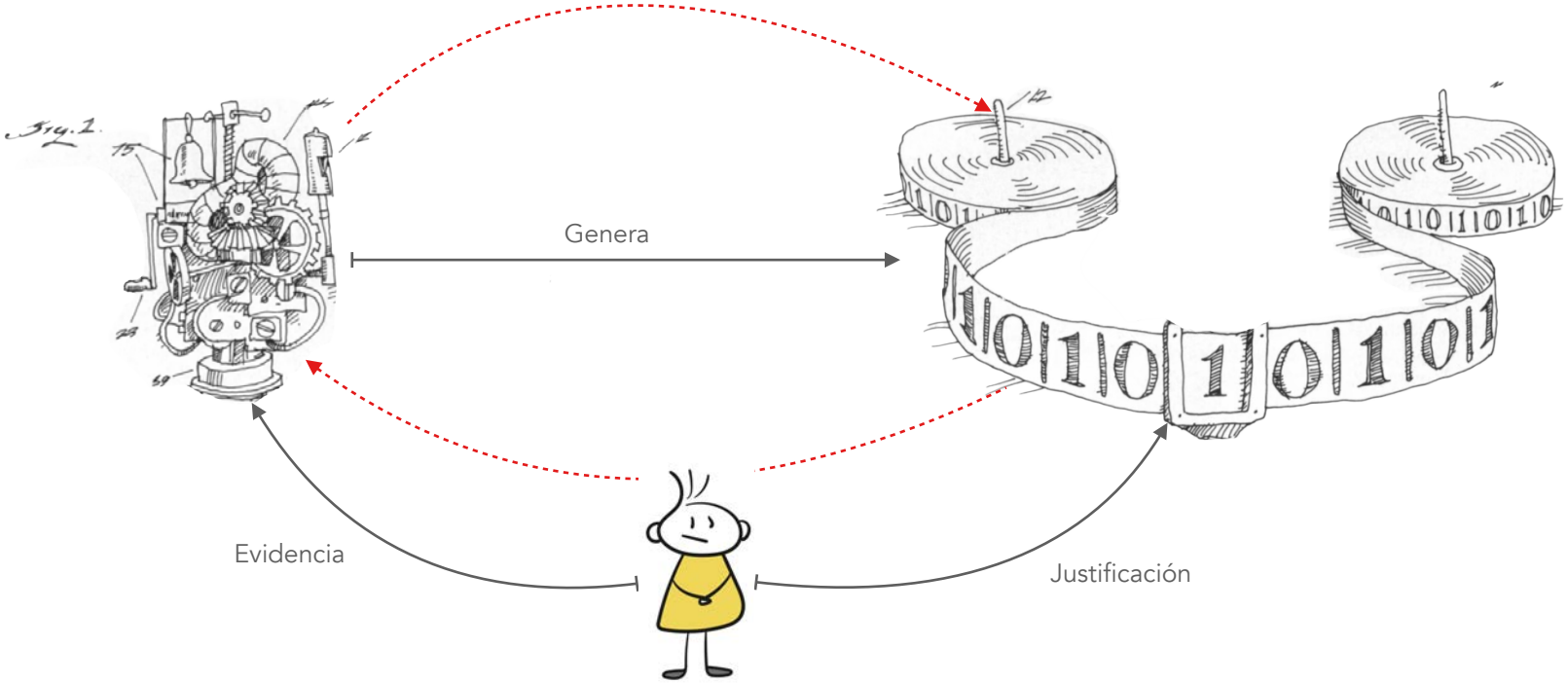
Self-justification



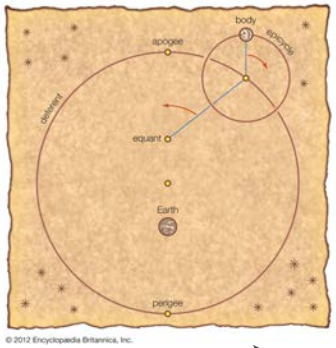
Self-justification



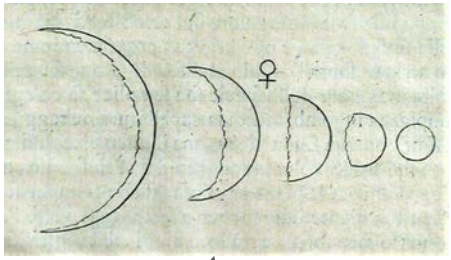
Self-justification



Self-justification: an analogy



Genera



Evidencia

Saliency maps

Fig. 2 | Saliency does not explain anything except where the network is looking. We have no idea why this image is labelled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes. Credit: Chaofen Chen, Duke University

(Rudin, 2019, 209)

Este ejemplo muestra cómo hay una circularidad en la justificación. La evidencia que apoya la tesis que es un Husky no tiene prioridad epistémica sobre la evidencia que apoya la tesis que es una flauta traversa. Transparencia no provee razones suficiente sobre los méritos de un conjunto de funciones sobre la otra.

Justificación y Transparencia

- T-Regress lleva a una parálisis epistémica: si una creencia require una serie infinita de justificaciones, desembarcamos en un escepticismo puesto que ninguna creencia puede establecerse

- Self-Justification lleva a un aislamiento epistémico: el sistema de creencias está completamente aislado de crítica externa pues la justifcación es auto-referencial.

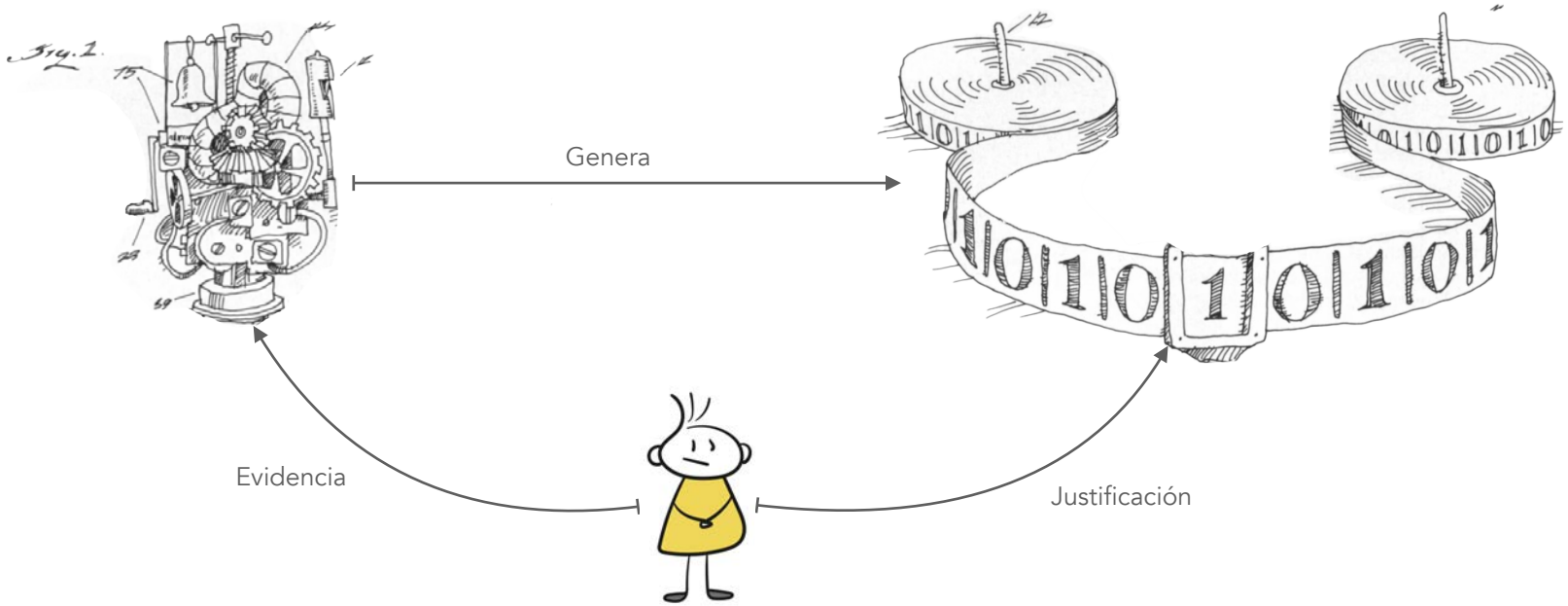
Relajando la definición

Transparencia contextual

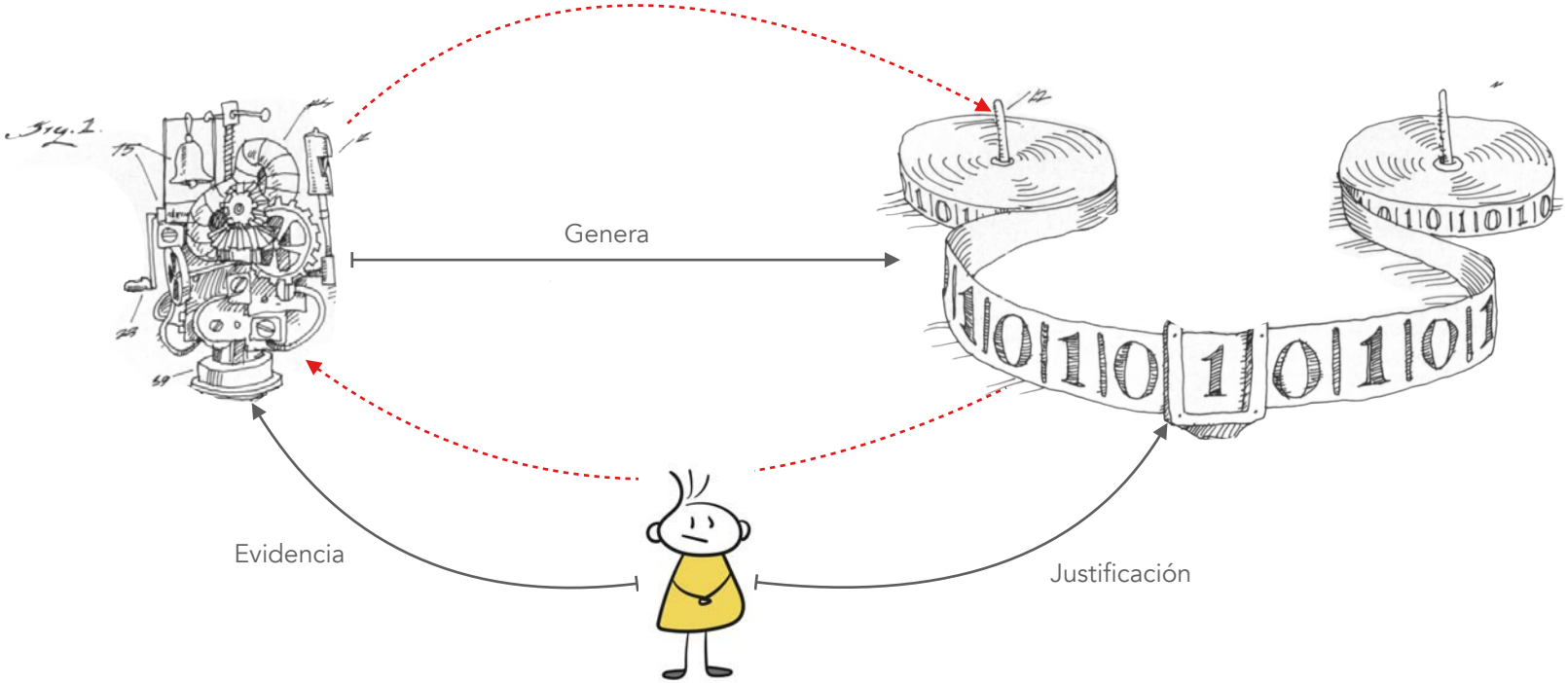
Zednik: ``rendering an opaque system transparent [...] **require[s] knowledge of the environmental patterns and regularities that are being tracked** and of the abstract representational structures that are tracking them"

Burrell: "In this emerging critique of 'algorithms' carried out by scholars in law and in the social sciences, few have considered in much depth their mathematical design. Many of these critics instead take a broad socio-technical approach looking at 'algorithms in the wild.' The algorithms in question are studied for the way they are situated within a corporation, under the pressure of profit and shareholder value, and as they are applied to particular real-world user populations (and the data these populations produce). **Thus something more than the algorithmic logic is being examined**."

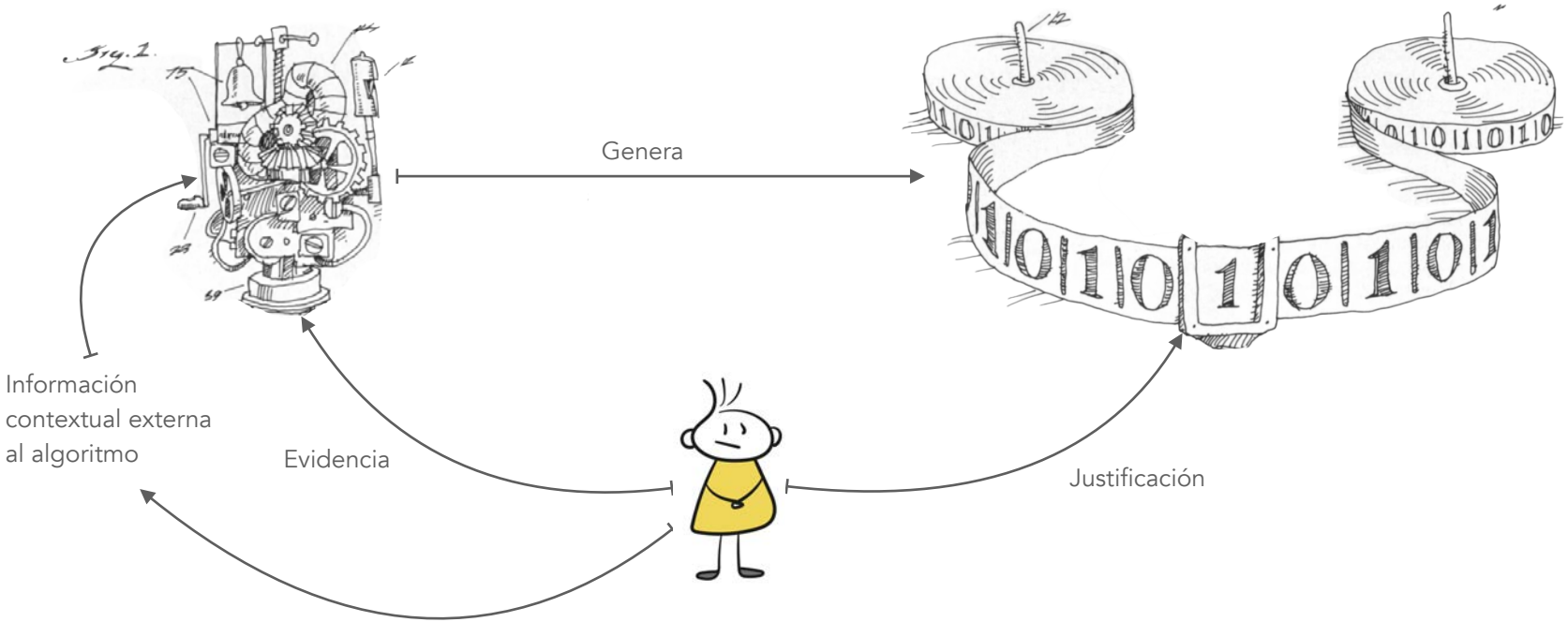
Self-justification



Self-justification



Transparencia contextual



Transparencia contextual

- El amige de la transparencia todavía necesita hacer lo siguiente:

1. Dado que transparencia contextual no da respuesta a T-Regress y/o Self-justification, entonces se provee de un argumento que tenga a la transparencia como epistemología fundacional
2. Demostrar cómo información contextual justifica SIN convertirse en la base primaria de justificación (ya que nuestro amigue quiere mantener una epistología internalista al algoritmo)—NB: 1. Todavía espera una solución, pues es un problema para cualquier epistemología internalista
3. Admitir que información contextual tiene más peso justificatorio que la lógica interna del algoritmo. Esto implica reconocer que una epistemogía internalista es defectiva e inadecuada y que es necesario cambiarla por una epistemología externalista.
    1. El segundo paso es mostrar cómo información externa (más información internal) justifican

¿Y ahora?

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)

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)

Abandon transparency?

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

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 connection with computer simulations rely on the assumption that results of simulations are trustworthy. Examples of these include the debate on the experimental role of computer simulations (Parker in Synthese 169(3):483–496, 2009; Morrison in Philos Stud 143(1):33–57, 2009), the nature of computer data (Barberousse and Vorms, in: Durán, Arnold (eds) Computer simulations and the changing face of scientific experimentation, Cambridge Scholars Publishing, Barcelona, 2013; Humphreys, in: Durán, Arnold (eds) Computer simulations and the changing face of scientific experimentation, Cambridge Scholars Publishing, Barcelona, 2013), and the explanatory power of computer simulations (Krohs in Int Stud Philos Sci 22(3):277–292, 2008; Durán in Int Stud Philos Sci 31(1):27–45, 2017). The aim of this article is to show that these authors are right in assuming that results of computer simulations are to be trusted when computer simulations are reliable processes. After a short reconstruction of the problem of *epistemic opacity*, the article elaborates extensively on *computational reliabilism*, a specified form of process reliabilism with computer simulations located at the center. The article ends with a discussion of four sources for computational reliabilism, namely, verification and validation, robustness analysis for computer simulations, a history of (un)successful implementations, and the role of expert knowledge in simulations.

Keywords Computer simulations · Reliabilism · Epistemic opacity · Verification and validation · Robustness analysis · History of success · Expert knowledge

Beyond transparency: computational reliabil as an externalist epistemology of algorithms

forthcoming in

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

Synthese Library

Juan M. Durán

Abstract This chapter examines the epistemology of algorithms, framing cussion as a question of epistemic justification. Current approaches empha gorithmic transparency, which involves elucidating internal mechanisms— functions and variables—and demonstrating how (or that) these compute Thus, the mode of justification through transparency is contingent on w be shown about the algorithm and, in this sense, is *internal* to the algori contrast, I propose an *externalist* epistemology of algorithms called *compu reliabilism* (CR). While I have previously developed CR in the context of cc simulations ([60, 74, 12]), this chapter extends the framework to a broader r algorithms used across scientific disciplines, particularly in machine learn deep neural networks. At its core, CR posits that an algorithm's output is j if it is generated by a reliable algorithm, where reliability is determined b bility indicators. These indicators arise from formal methods, algorithmic r expert competencies, research cultures, and other scientific practices. The cl primary objectives are to delineate the foundations of CR, explain its operationa mechanisms, and outline its potential as an externalist epistemology of algorithms.

1 Introduction

The use of algorithms for scientific purposes is delivering remarkable results. A couple of examples will suffice to illustrate this. In molecular biology, AlphaFold can

Juan M. Durán
Department of Values, Technology and Innovation

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

PAPER IN GENERAL PHILOSOPHY OF SCIENCE

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 seeks to defend reliabilist epistemologies of algorithms against the challenge posed by SIS-Errors. To this end, I draw upon c
framework and articulate epistemological
and thus preserve algorithmic reliability.

Keywords Reliabilist epistemologies of a
SIS-Errors · Paul Humphreys

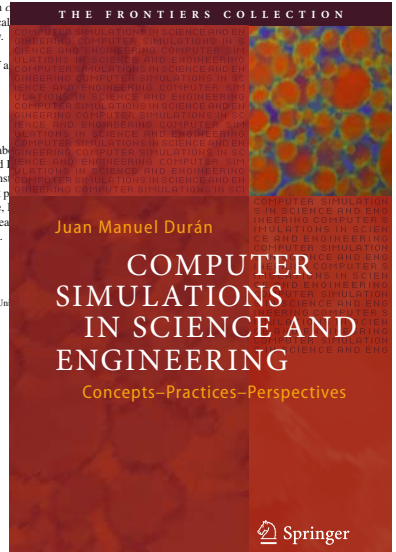
1 Introduction

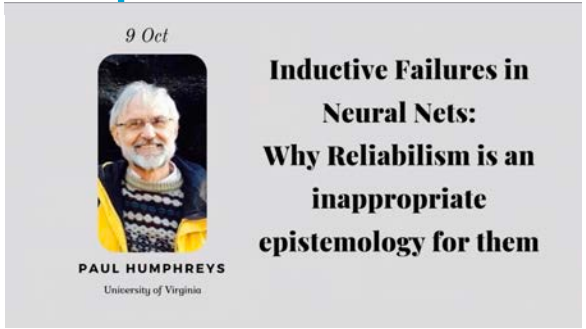
In 2009, there was a short-lived debate ab
tions in the scientific domain. Frigg and
novel, computer simulations did not const
lutionary departure from everything that p
(Frigg & Reiss, 2009, 601). In response,
in the context of computer science that bea
One of these issues is *epistemic opacity*.

✉ Juan M. Durán
j.m.duran@tudelft.nl

[1] Technology, Policy and Management, Delft Un
The Netherlands

Published online: 11 June 2025





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.

A number of commentators have noted that these definitions apply to processes that are not computational, such as the workings of sophisticated scientific instruments. That is correct and to appreciate the differences between epistemic opacity and what I shall call *epistemic inaccessibility* it is important to recall the context in which epistemic opacity was

Synthese (2022) 200:491
https://doi.org/10.1007/s11229-022-03975-6

ORIGINAL RESEARCH

Instruments, agents, and artificial intelligence: novel epistemic categories of reliability

Eamon Duede[1,2,3,4]

Received: 17 October 2021 / Accepted: 7 November 2022

Abstract
Deep learning (DL) has become increasingly central to science, primarily due to its capacity to quickly, efficiently, and accurately predict and classify phenomena of scientific interest. This paper seeks to understand the principles that underwrite scientists' epistemic entitlement to rely on DL in the first place and argues that these principles are philosophically novel. The question of this paper is not whether scientists can be justified in trusting in the reliability of DL. While today's artificial intelligence exhibits characteristics common to both scientific instruments and scientific experts, this paper argues that the familiar epistemic categories that justify belief in the reliability of instruments and experts are distinct, and that belief in the reliability of DL cannot be reduced to either. Understanding what can justify belief in AI reliability represents an occasion and opportunity for exciting, new philosophy of science.

Keywords Deep learning · Scientific knowledge · Models · Reliability · Trust and Justification

Received: 13 December 2023 | Revised: 12 March 2024 | Accepted: 16 April 2024
DOI: 10.1111/phc3.12974

ARTICLE

WILEY

Reliability in Machine Learning

Thomas Grote[1] | Konstantin Genin[1] | Emily Sullivan[2]

[1]Cluster of Excellence: Machine Learning: New Perspectives for Science, University of Tübingen, Tübingen, Germany
[2]Department of Theoretical Philosophy, Utrecht University, Utrecht, The Netherlands

Correspondence
Thomas Grote, Cluster of Excellence: Machine Learning: New Perspectives for Science, University of Tübingen, Maria von Linden Str. 6, D-72076 Tübingen, Germany.
Email: thomas.grote@uni-tuebingen.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: BE5601/4-1; Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Number: VI.Veni.201F.051; Carl-Zeiss-Stiftung

Abstract
Issues of reliability are claiming center-stage in the epistemology of machine learning. This paper unifies different branches in the literature and points to promising research directions, whilst also providing an accessible introduction to key concepts in statistics and machine learning – as far as they are concerned with reliability.

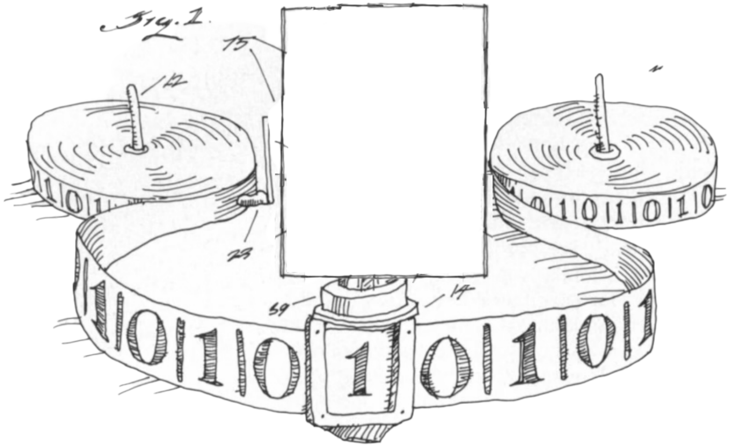
1 | INTRODUCTION

Machine learning models often achieve impressive accuracy under training conditions, but fail in spectacular or unexpected ways when they are deployed in real-world settings. Is there some way to guarantee that predictive accuracy in training carries over to the settings in which models are actually deployed? In other words: can we be justified in relying on machine learning models on the basis of their performance in training?

The underlying challenge is that there are various threats to reliability, arising at the time of (i) model development, (ii) model deployment and (iii) adapting the socio-technical environment to accommodate the model. Concerning (i), unlike traditional statistical models, there is no widely accepted mathematical theory that explains why and when state-of-the-art models such as deep neural networks generalize well. As for (ii), machine learning models are commonly used in unstable environments, or even induce changes to the environment itself, and they can be fooled by humanly imperceptible manipulations to the data. Regarding (iii), one basic issue is to align the model output with the existing epistemic norms in a given domain, which often involves aggregating the model output with other kinds of evidence. In addition, there is another overarching problem: machine learning models are

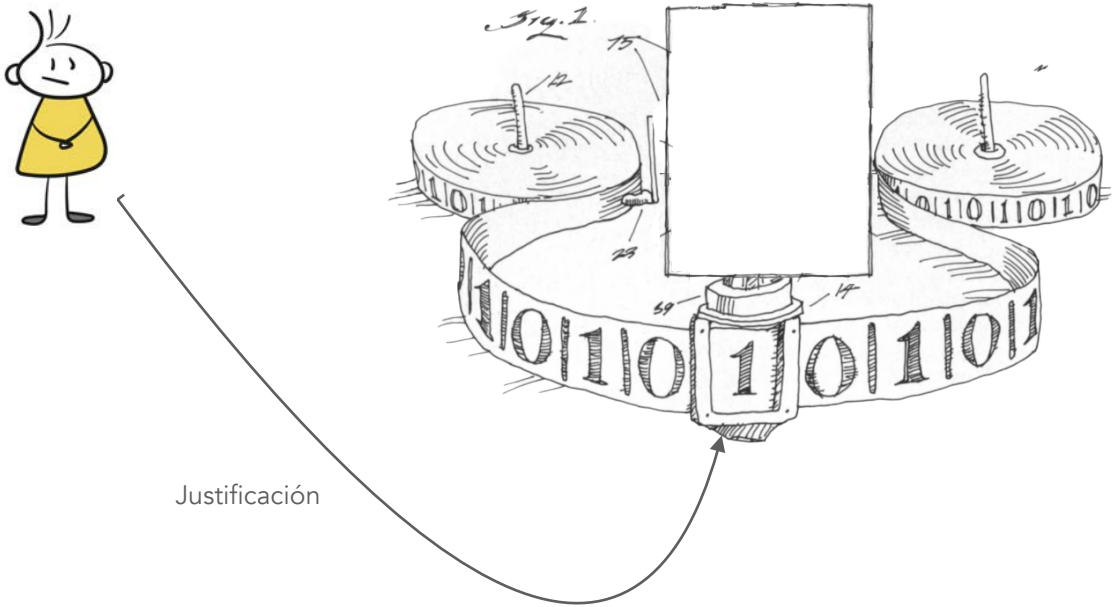
¿Cómo se justifica via CR?

Justificación via CR (a sketch)



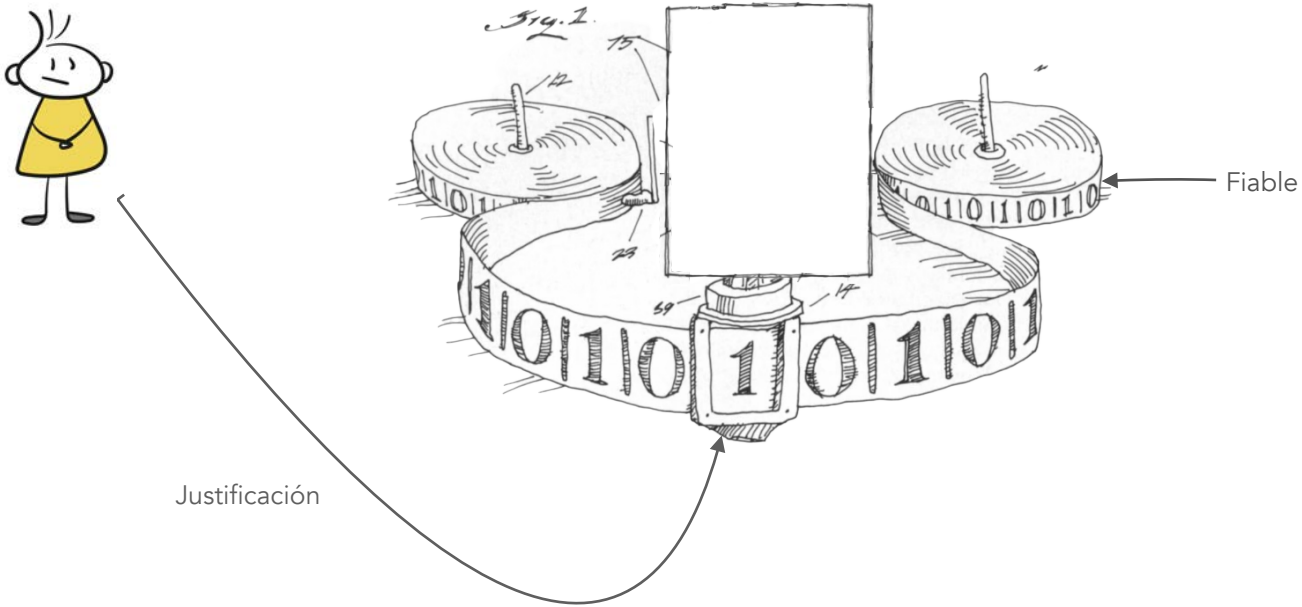
http://bridgethevoid.blogspot.com/

Justificación via CR (a sketch)



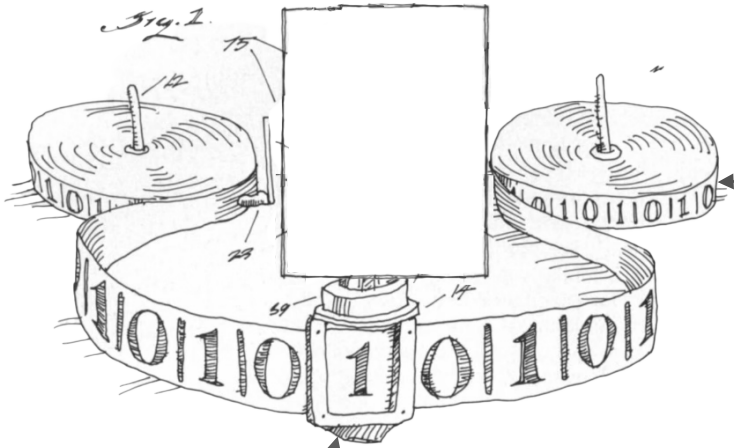
http://bridgethevoid.blogspot.com/

Justificación via CR (a sketch)



http://bridgethevoid.blogspot.com/

Justificación via CR (a sketch)



Fiable

Indicadores de fiabilidad

Justificación

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

Fiabilismo computacional

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

Bibliography

- Boge, F. (2020) "Two Dimensions of Opacity and the Deep Learning Predicament." *Minds and Machines*, 30, 2020: 187–208
- Chen, Lin, Rudin, Shaposhnik, Wang, Wang (2018) "An interpretable model with globally consistent explanations for credit risk" *NIPS 2018 Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy*, Montréal, Canada.
- Citron, D. K., & Pasquale, F. (2014). "The scored society: Due process for automated predictions." *Washington Law Review.* 89(1): 1–33
- Creel, K (2020) ``Transparency in complex computational systems'' *Philosophy of Science*, 87: 568–589.
- Duede, E. (2022). "Instruments, agents, and artificial intelligence: Novel epistemic categories of reliability" *Synthese*, 200(6):1-20.
- Durán, JM, Formanek, N. (2018) "Grounds for trust: Essential Epistemic Opacity and Computational Reliabilism" *Minds and Machines*. 28:645–666
- Durán, JM (forthcoming) "Beyond transparency: computational reliabilism as an externalist epistemology of algorithms" in *Philosophy of science for machine learning: core issues, new perspectives*. Juan M Durán and Giorgia Pozzi (eds.)
- Grote, T., Genin, K., and Sullivan, E. (2024). "Reliability in machine learning" *Philosophy Compass*, e12974.
- Guidotti, R, Monreale, A, Ruggieri, S, Turini, F, Giannotti, F, and Pedreschi, D. (2018) "A survey of methods for explaining black box models." *ACM Computing Surveys*, 51(5): 1-24, article 93.

Bibliography

- Humphreys, P (2009) "The philosophical novelty of computer simulation methods". *Synthese*. 169(3): 615-626
- Humphreys, P (online) "Epistemic Opacity and Epistemic Inaccessibility"
- Kroll, J. A., Huey, J, Barocas, S, Felten, E, Reidenberg, J, Robinson, D, and Yu, H. (2017) *"Accountable algorithms." University of Pennsylvania Law Review*, 165(3) 633–705.
- Lipton, Z. (2016) "The mythos of model interpretability." *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*. 96–100
- Mitchell, M, Wu, S, Zaldivar, A, Barnes, P, Vasserman, L, Hutchinson, B, Spitzer, E, Raji, I and Gebru, T. (2019) "Model cards for model reporting" *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*. 220–229
- O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing Group.
- Ribeiro, M T, Singh, S and Guestrin, C. (2016) "Why should I trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.
- Rudin, C. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." Nature Machine Intelligence, vol. 1, 2019, pp. 206–215. doi:10.1038/s42256-019-0048-x.
- Vallor, S. (2016). "Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting."
- Zerilli, J. "Explaining Machine Learning Decisions'' Philosophy of Science (2022), 89, 1–19

Thank you!

j.m.duran@tudelft.nl

Johannes Vermeer
Ansicht von Delft

https://www.mauritshuis.nl/