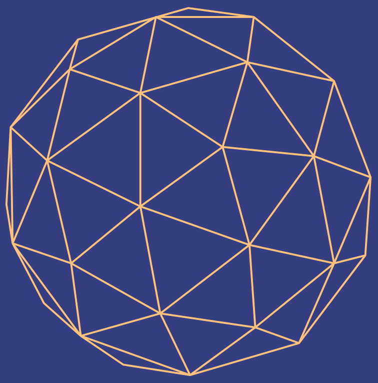
Issue 3 - January 2022

Good In Tech RESEARCH NEWS

Rethinking innovation and technology as drivers of a better world for and by humans



Ethics on Artificial Intelligence in practice: challenges and limits

All the articles from Jean-Marie John-Mathews 's thesis, defended on 1 December 2021

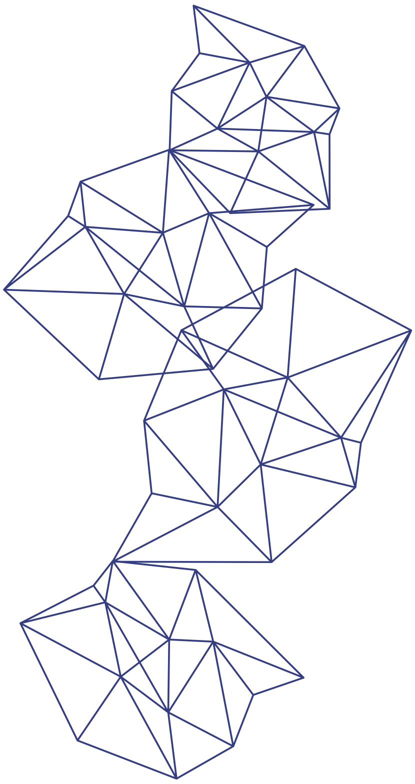
GOOD IN TECH VISION

Good In Tech main objectives are to create knowledge around four research areas and to contribute to the dissemination of this knowledge not only in academic and pedagogical spheres but also to corporations, decision-makers, regulators and the general public.

To this end, the Chair aims to create and develop an ecosystem of interactions between research, companies, students from the two partner academics and political institutions, civil society in order to raise awareness of all stakeholders on this new paradigm on responsible digital technologies and innovation.

The chair also aims to develop international partnerships, particularly in Europe, to share the issues of responsible digital innovation with international committees.

Finally, the Chair aims to share the results of academic works and debates it organizes with national and European political institutions in order to inform and influence public policies.



Jean-Marie John-Mathews

Jean-Marie is a researcher in AI Ethics at University Paris- Saclay. He is also the coordinator of the academic chair Good In tech (Institut Mines-Télécom / Sciences Po).

His research addresses solutions to make Machine Learning algorithms less biased, more private and explainable. His research field are XAI (Explainable Artificial Intelligence), FairML and Privacy-preserving Machine Learning. He has been published in top-ranked journals and international conference for his research work on AI Ethics.

He is also teaching at Sciences Po on Algorithms & Public Policies and at Aivancity on Mathematics for Machine Learning. In the past, he worked as a Data Scientist and AI engineer after graduating in Mathematics (ENSAE), Economics (Sciences Po / Polytechnique and PSE) and Philosophy (Sorbonne and ICP).

Some Critical and Ethical Perspectives on the Empirical Turn of AI Interpretability, 2022

This paper has been published in the rank A (HCERES) journal Technological Forecasting and Social Change.

WHY IS THIS TOPIC IMPORTANT ?

Today, **AI has one big drawback: it is often opaque**. It is difficult to explain the decisions of the algorithms or how they work. This is a problem because algorithms are increasingly used for everyday situations such as assigning credit or recruiting. Thus, users of algorithms must be capable of being held accountable and explaining the decisions of the algorithms they use.

The **interpretability of AI is therefore a major issue in ethics**. However, interpretability is very difficult to define. For any process, there is a number of possible explanations and these vary according to the interlocutor, his availability, the context and other factors. The explanation depends on the context of the explanation and there is no unique explanation.

There is an academic community of researchers called XAI or explainable artificial intelligence that produces contextualized explanations of artificial intelligence. **This paper is a critique of the empirical and very contextual methods of XAI.**

The way they proceed is by taking a case where the explanation should not be context dependent. The idea is to demonstrate the existence of cases in which the producers of explanations produce highly contextual explanations that pervert the underlying phenomenon they are trying to explain.

AI ethics is supposed to address a number of problems that are supposed to be independent from context. A discrimination must always be revealed by the explanation, it should not be dependent on contexts. By producing contextualized explanations, these may not reveal the underlying incident.

METHODOLOGY

Building the algorithm:

- Real world situation, with people given or denied a credit by banks
- Construction of an artificial intelligence algorithm using the German credit scoring data base.

Method:

- A sample of 800 people is chosen to conduct a survey. They **simulate a situation** in which the **800 respondents have been refused a bank loan**.
- They are told that **this decision was made by an algorithm using artificial intelligence** and they are given 8 explanations, so there are 100 people per explanation.
- **Half of the people** are in an **ethical situation**. That means that the decision is not discriminatory.
- **The other half of the participants** are in a **non-ethical situation**. This means that the algorithms that handled their cases are explicitly discriminatory.
- None of the participants know whether they were treated by the so-called "ethical" or by the "non-ethical" algorithm.
- Whether the situation is ethical or not, 4 different types of explanations are provided:
    - **Transparent algorithm scenarios**: They provided a points-based system using the integer coefficient of the logistic regression model to explain why the credit decision was negative. To do so, we assigned to each feature a certain number of points obtained from the coefficients of the logistic regression. If the sum of all the points exceeded a certain threshold, the bank loan was accepted.
    - **Post-hoc Shapley scenarios**: They provided a listing of the variables which played in the disadvantage of the individual. Participants are informed of the importance of each feature with respect to the negative decision made by the black-box algorithm.
    - **Post-hoc conterfactual scenarios**: They announce to the respondent, for some selected variables, the threshold at which the negative decision switched to a positive one.
    - **No explanation scenario**: They did not provide any further explanation in this case. This scenario was used as a baseline to measure the effects of the other scenarios.

Analysis of the data:

- To measure explanations' denunciatory power, they **asked the participants to rate their agreement** with two perception desiderata and two reaction desiderata:
    - Fairness perception of the algorithmic decision using a scale from "-2" (strongly disagree) to "2" (strongly agree).
    - Trust perception of the algorithmic decision using a scale from "-2" (strongly disagree) to "2" (strongly agree).
    - Free comments: we left a free field for respondents to comment freely on the algorithmic decision
    - Wish to contest the algorithmic decision using a scale from "-2" (strongly disagree) to "2" (strongly agree).
- **Denunciatory power** is measured by comparing the negative receptions between the situation with and without discrimination. More rigorously, this power could be defined as the "capacity to bring out a negative reception in a discriminatory situation".

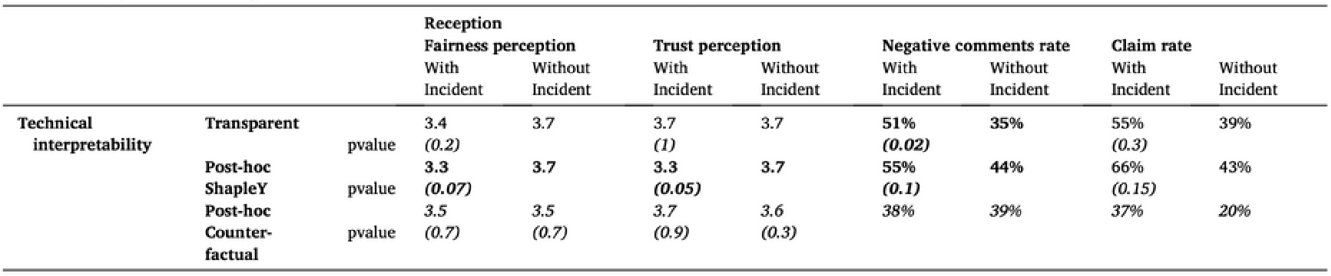
KEY FINDINGS

There are three main key findings.

Denunciatory power differs depending on modes of explanation

- By comparing the negative reception between ethical situations using Student's tests, we see that only the Shapley explanations have a significant denunciatory power
- Conversely, the counterfactual explanations have the lowest denunciatory power because they are not significant for any of the dimensions of negative reception.
- The denunciatory power of the explanations is therefore dependent on the modes of technical explanation.

Table 1
Denunciatory power depending on explanation modes (*t*-test with H0: means equality with the baseline scenario where decisions are not sexist).



			Reception							
			Fairness perception		**Trust perception**		**Negative comments rate**		**Claim rate**	
			With Incident	Without Incident	With Incident	Without Incident	With Incident	Without Incident	With Incident	Without Incident
Technical interpretability	**Transparent**		3.4	3.7	3.7	3.7	**51%**	**35%**	55%	39%
		pvalue	*(0.2)*		*(1)*		***(0.02)***		*(0.3)*	
	Post-hoc ShapleY		**3.3**	**3.7**	**3.3**	**3.7**	55%	44%	66%	43%
		pvalue	***(0.07)***		***(0.05)***		*(0.1)*		*(0.15)*	
	Post-hoc Counter-factual		*3.5*	*3.5*	*3.7*	*3.6*	*38%*	*39%*	*37%*	*20%*
		pvalue	*(0.7)*	*(0.7)*	*(0.9)*	*(0.3)*				

Empirical explanations tend to select the explanation that has the weakest denunciatory power:

- The empirical design of the explanation tends to value the explanation with the lowest denunciatory power when addressing an AI incident. This is because the empirical approach seeks explanations that minimize the number of negative feedbacks of the AI decision, while denunciatory power is measured precisely through the negative feedbacks. In this situation with contradictory objectives, denunciatory power is likely to be neglected in the development of future artificial intelligence tools, to minimize the number of negative feedbacks.

Explanations are empirically tested using desiderata

- We show empirically, using an experimental setup, that it is possible for AI designers to propose modes of technical explanation that lower the level of criticism and avoid revealing unethical situations. Far from being a solution to the ethical incidents of AI, explanation techniques can be hijacked by manufacturers in their interests, to the detriment of ethics. But we show that this phenomenon is not necessarily intentional on the part of the designer: by empirically selecting explanations with respect to desiderata that reduce criticism, the manufacturer indirectly creates the structural conditions for masking ethical incidents.

KEY TAKEAWAYS

This article shows the limits of empirical explanations.

Takeaway 1: There are often two major issues in AI ethics: explicability and discrimination.

- This paper is a criticism of the limits of contextual explanations, thus linking the two topics. Explainability is not the solution to the problems of discrimination. It does not necessarily reveal discriminatory incidents.

Takeaway 2: Empirical explanations are insufficient to fight discrimination.

- Even if builders are in good faith and sensitive to algorithm discrimination issues, their implementation of empirical methods to explain the algorithm yields insufficient and unethical results.

Takeaway 3: Recommandation.

- By selecting explanations with the least negative feedback from users, self-regulation in AI tends to paradoxically eliminate methods of explanation with a strong denunciatory power. As a consequence, in this scenario, AI ethics cannot be solved by only relying on explanations to denounce suspicious behaviors and individual behavior from end-user.
- It is therefore necessary to separate the concept of interpretability from user feedback. In this scenario, we need to define the concept of interpretability, so that it applies context-independently. Once it is formalized, experts or auditors can examine the ethics of algorithms independently of the user feedback.

From Reality to World. A Critical Perspective on AI Fairness, 2022

Co-authors: Dominique Cardon, Christine Balagué

Accepted for publication in Journal of Business Ethics (rank A, HCERES)

Dominique Cardon

Dominique Cardon is a Sociology professor at Sciences Po. Since 2010, his work has focused on the sociological analysis of web and big data algorithms in order to understand both the internal form of computation and the world that computers project on our societies. His research also focuses on the transformations of the media space and the new circuits of digital information. He is a member of the editorial board of the journal Réseaux and of the prospective committee of the CNIL.

Christine Balagué

Christine Balagué is HDR Professor at IMT-BS and holder of the Good in Tech Chair (www.goodintech.org). Her research focuses on modelling the behaviour of connected individuals, ethics of technology and AI, and responsible digital innovation. She is also a member of several national committees: CSA expert committee on online disinformation, Defence ethics committee, Haute Autorité de Santé recommendations impact commission, executive committee of Cap Digital. As Vice-President of the National Digital Council from 2013 to 2016, she is also co-author of several reports submitted to the French government on digital issues. She published more than 50 research articles & international conference proceedings as well as several books on society and economic digital metamorphosis.

WHY IS THIS TOPIC IMPORTANT ?

> **In a context where data are increasingly numerous, granular and behavioral, it is essential to renew our conception of AI ethics on algorithms in order to establish new models of responsibility for companies that take into account changes in the computing paradigm.**

Categories are social constructions that are produced in socio-political context. They can be gender or socio-professional categories, for example. The process of making these categories is interesting to describe and is particularly tackled in a sub-discipline of sociology: **pragmatic sociology**.

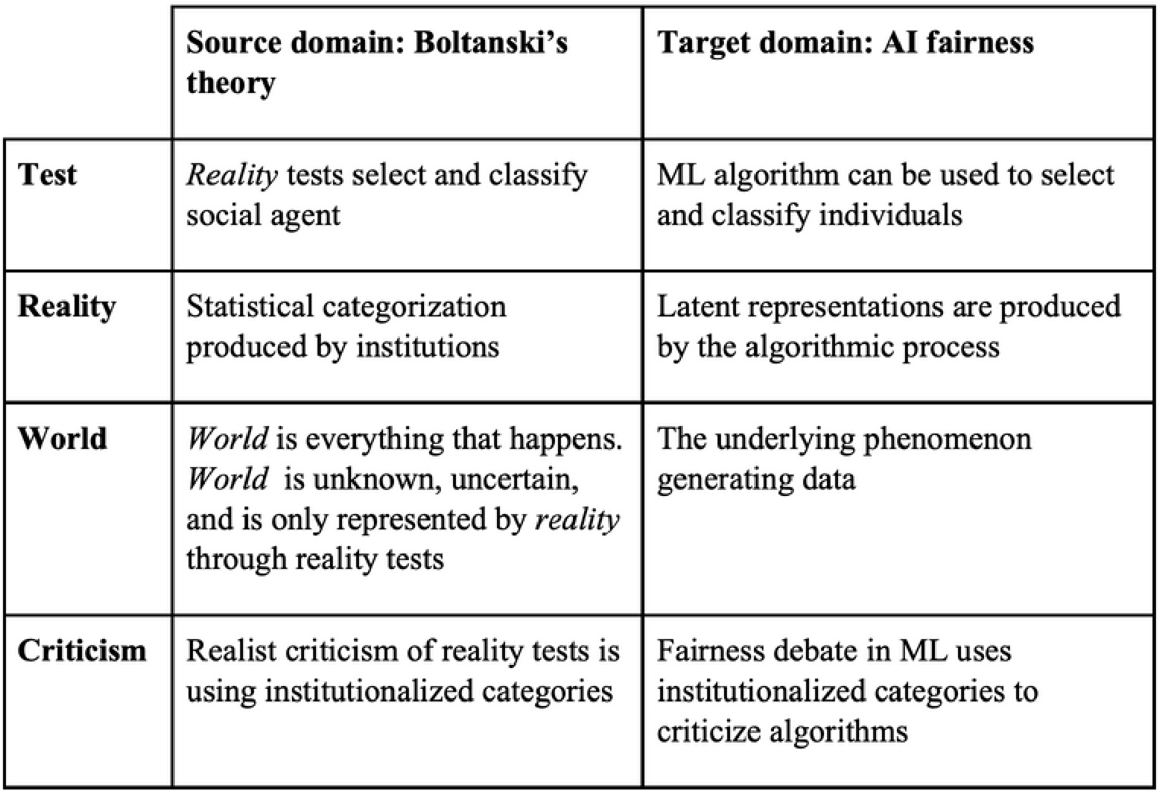
Artificial intelligence is radical because it does not use the categories generally used in classical descriptive statistics: **AI tries to go "under the category"**.

In practice, this means that it uses increasingly granular and behavioral data that are traces of actions that may have taken place on the Web for example. This allows it to free itself from institutionalized categories.

In this **philosophical paper**, the authors show that the very recent debate on fairness in AI can be understand using concepts raised by pragmatic sociology since the 1990s.

The authors therefore offer a theoretical contribution to consider AI ethics outside of high-level and top-down approaches, based on the **distinction between "reality" and "world" from Luc Boltanski**, father of pragmatic sociology.

The following table defines the framework for the analogy developed throughout the paper:



	Source domain: Boltanski's theory	**Target domain: AI fairness**
Test	*Reality* tests select and classify social agent	ML algorithm can be used to select and classify individuals
Reality	Statistical categorization produced by institutions	Latent representations are produced by the algorithmic process
World	*World* is everything that happens. *World* is unknown, uncertain, and is only represented by *reality* through reality tests	The underlying phenomenon generating data
Criticism	Realist criticism of reality tests is using institutionalized categories	Fairness debate in ML uses institutionalized categories to criticize algorithms

> **Algorithm biases, discrimination and consequently unfairness have been identified in various AI applications.**

The work of ethics in Artificial Intelligence is dedicated to the search for problematic situations, especially discriminatory ones. However, the authors try to capture these incidents with crystallized categories that are often produced by dominant institutions: there is a kind of **category crisis**. The objective of this paper is to discover a new way of analysis and detection of these situations, anchored in the world without being detached from the stability of reality.

Luc Boltanski's notion of world refers to a set of traits, affects, relationships to others and things that are not taken into account in objectification techniques conventionally used in institutional reality tests. The world constantly overflows reality to challenge codification techniques, categorization and decision-making rules of the devices/social systems that support it.

THE DEBATE ON FAIRNESS

The objective of the paper is to **understand the AI fairness debate through the lens of pragmatic sociology**.

1.Realist criticism on machine learning and fairness

The authors argue in this paper that these criticisms belong to a “realist” vision, on two grounds:

- Criticism is mostly supported by two “protected categories”, namely race and gender
- They are using statistical indicators of fairness calculating differential treatment by the machine learning algorithm, between these “protected” categories

This criticism is therefore realist since it is based on categorical and stabilized representations of reality by proposing to redress/correct the relative share of one category in relation to another.

Reality is stable because institutions have worked to make it stable. **These categories are necessary for criticism** because to federate and mobilize, it is necessary to use instruments that **allow generalization**. The categories are meaningful and instituted, they allow to totalize the remarks and to escape from the lived experience. The implicit theory of justice that it implements thus needs to rely on a system of categorization whose meanings are shared and validated by institutions in order to correct the harms of unequal distribution.

2.Repairing Injustice by Fixing Algorithms

The proposed corrections to **debiasing the algorithms** are also **based on these same categories** that Boltanski criticizes.

The authors define realist fairness engineering as the **Fair ML techniques**, based on categories, used to address algorithmic fairness.

These are realist answers to criticism, in Boltanski's terms, since they rely on a representation between ideal variables, like skills, in a latent hypothesis space to assess and correct algorithmic fairness. Actually, they use categorical fairness metrics to correct ML algorithms without changing data formats and models. The integration of fairness corrective measures within the calculation is not enough to address criticism of algorithmic unfairness.

Thus, the correction of algorithmic biases by relying on realistic fairness engineering has **various limitations**.

- Representation can be too simplistic. When the hypothesis space is too simplistic, it may fail to represent the singularities of the social world.
- Intra and inter-category fairness. The categories used to measure fairness produce groups that are too large and too homogeneous by ignoring intra-category variability. Methods for correcting fairness may involve accentuating intra-category inequality in favor of inter-category equality.
- Trade-off between performance and fairness. Since the interest of algorithm designers may be on performance rather than fairness, voluntary fairness correction is hard to expect without a binding mechanism.

3.A radical response: when algorithms compute the world

There are a number of initiatives in AI fairness today. A recent trend proposes to leave categories aside and to trust artificial intelligence which, through very granular data, manages to emancipate itself from bias and categories.

Emerging promises associated with big data and AI are driving a major shift in the format of data. Facing criticism that traditional statistical categories misrepresent reality, the new algorithmic tests seek to reduce the arbitrariness and imprecision of reality tests by capturing the world. This shift is characterized by a considerable increase in the number of data, the granularization of statistical entities, the personalization of information, and the accumulation of behavioral traces.

Example with recruitment methods:

For example, artificial intelligence opens **personalized recruitment processes**, by analyzing candidates' reactions to interactive information on the company in job interview videos or even using external data, such as purchasing behavior or other traces in the social media (Sánchez-Monedero et al. 2020). These recruitment methods that use behavioural traces are supposed to be more accurate and performant than recruitments based on socio-demographic categories or candidates' degree, according to the promoters of these methods.

The defenders of this thesis emphasize that the algorithm has a privileged access to the world that avoids having to rely on institutionalized categories. It would therefore be an exact translation of the world, free of bias. We must therefore trust the algorithms and not use categories in machine learning since these impose a vision governed by biases.

This vision calls for a redefinition of the concept of what is fair. It is sufficient for the algorithm to have the same answer for two people close in the latent space for it to be considered fair. The more the algorithm reflects the latent categories produced by the algorithm itself, the more the result obtained is considered fair. Thus, fairness includes situations that actually reflect the world.

4.Complex regime of domination confuses the World with Reality

In this paper, the authors criticize the radical response presented earlier. They see it as a complex regime of domination.

Definition of a complex domination regime, also considered as a managerial domination regime: whereas simple domination seeks to confirm the legitimacy of reality tests by merely correcting the dysfunctions highlighted by criticism, complex domination constitutes a new configuration of the exercise of power marked by the on-going change of reality tests.

Experts no longer derive their justification from intelligible principles, but from the idea that observable data encompass the world as adequately as possible. Complex domination regime **can be criticized** for different reasons:

- **Explainability problem**: one of the major consequences of this shift from reality to the world is that since computational data can no longer be organized in the form of interpretable symbolic variables, it is no longer possible to project them into a hypothesis space.
- **Exclusion problem**: affected individuals are no longer able to understand, verify and criticize the fairness of algorithms. This exclusion can be detrimental to criticism or negotiation.

KEY TAKEAWAYS

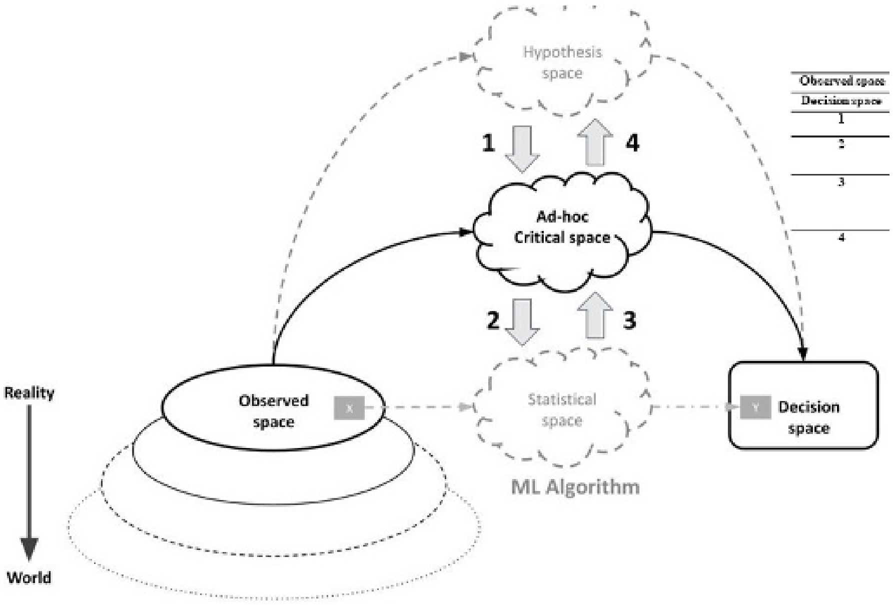
The authors' recommendations:

- **Even though traditional categories have limitations** in identifying all forms of injustice, the authors argue that **we should not abandon reality but instead stabilize new realities** that are radical enough to take into account the change in data format but without falling into the system of complex domination that encloses the calculation.
- On the one hand, in order to **avoid naive and naturalistic induction** (confusing the world with reality), pragmatic sociology gives materiality to the world's entities by questioning it through active intervention. On the other hand, in order to **avoid the domination of experts** in the construction of reality, pragmatic sociology opens up the description of the world by exposing it through a public investigation, confronting it with contradictory discourse, and fostering reflexivity

In order to achieve this, the authors suggest an ad-hoc critical space. This space would be in between the hypothetical milieu of institutionalized categories and the statistical space of algorithmically created categories.

This space can be reached through **a methodology in 4 steps** :

- **Break down the objective**: the first step of this ethical exploration of the algorithm's operation is to break down the objective of the calculation in order to make designers sensitive to the diversity of paths leading to the result.
- **Radicalizing through experimentation**: based on the questions of the previous step, the algorithm can be experimented through the manipulation of input data and the visualization of the output to better understand the algorithm's decision process.
- **Stabilize new realities by aggregating points of view**: using visualization tools, one shall multiply the experimental interventions from the previous step until stable trends appear, these are the emerging realities.
- **Re-qualify the objective and open-up realities**: this ad-hoc qualification of existing data therefore requires the mediation of techniques, such as survey tools but also governance mechanisms that can involve external auditors, ethical charters, or ethical committees.



Critical Empirical Study on Black-box Explanations in AI, 2021

Proceedings of the International Conference on Information Systems 2021 (rank A, CORE)

WHY IS THIS TOPIC IMPORTANT ?

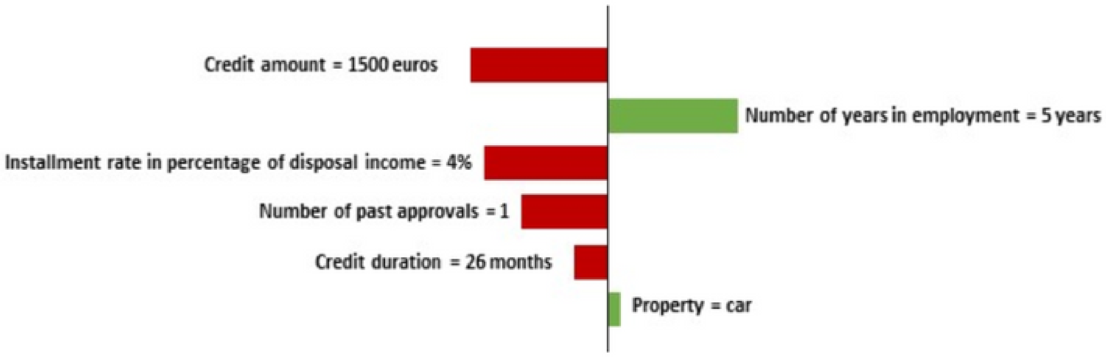
There is a problem in AI that decisions are not explainable because the algorithms are too large, non-linear, no mathematical theory to support the various theorems. This is called the **interpretability problem of AI**.

Then there is **a controversy about the solutions** to this problem. There are two main currents:

- **Post-hoc explanations**: The advantage of these algorithms is that they are efficient. So AI has made the choice to sacrifice human interpretability for better performance. So they say let's go to the end of this logic and create extremely powerful algorithms and try to interpret decisions a posteriori. This is what is known as post-hoc decisions.

Example: the case of a bank loan application.
In the post-hoc approach, when the algorithm refuses the bank credit, the explanations given will be interpreted a posteriori.

- Shapley explanation :



- Counterfactual explanations:

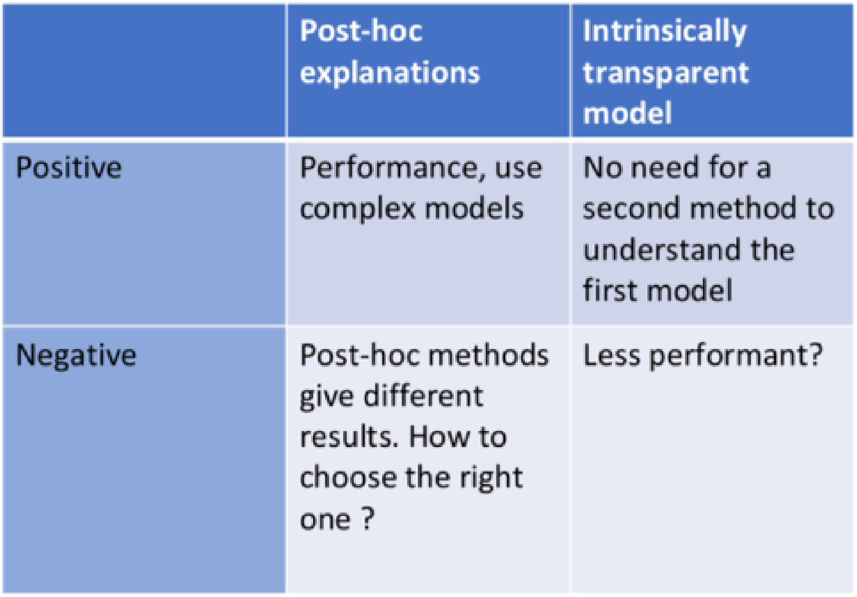
"If the credit amount was between 3000 and 4000 euros, your credit application would have been accepted by the algorithm."

"If the credit duration was reduced to less than 12 months, your credit request would have been accepted by the algorithm."

- **Intrinsically transparent algorithm:** If you interpret an algorithm post-hoc, there is no guarantee that this interpretation is true to what is happening in the algorithm. So, according to them, you have to take the problem from the beginning, upstream, and create algorithms that are intrinsically interpretable. There are therefore a series of methods that allow us to make algorithms that are intrinsically interpretable.

Example: here we are talking about making a system with points announced a priori (we associate points with the different variables). It is therefore a more comprehensible algorithm.

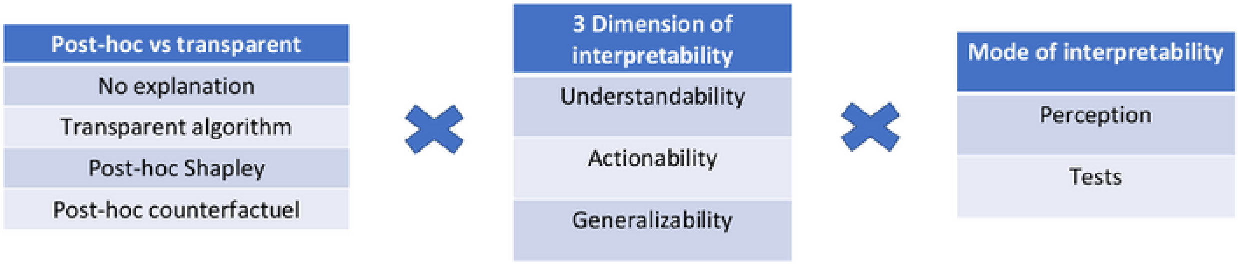
This paper highlights the controversy between these two positions by providing an empirical critique of post-hoc methods.



	Post-hoc explanations	Intrinsically transparent model
Positive	Performance, use complex models	No need for a second method to understand the first model
Negative	Post-hoc methods give different results. How to choose the right one ?	Less performant?

METHODOLOGY

It is **a randomised experiment with 400 people** and 4 scenarios (no explanation, transparent algorithm, post-hoc Shapley, post-hoc conterfactual), 3 dimensions of interpretability and 2 modes of interpretability.
In other words, some people had no explanation, others for whom the algorithm was transparent, people with post-hoc explanations (Shapley or conterfactual).



Post-hoc vs transparent
No explanation
Transparent algorithm
Post-hoc Shapley
Post-hoc counterfactuel

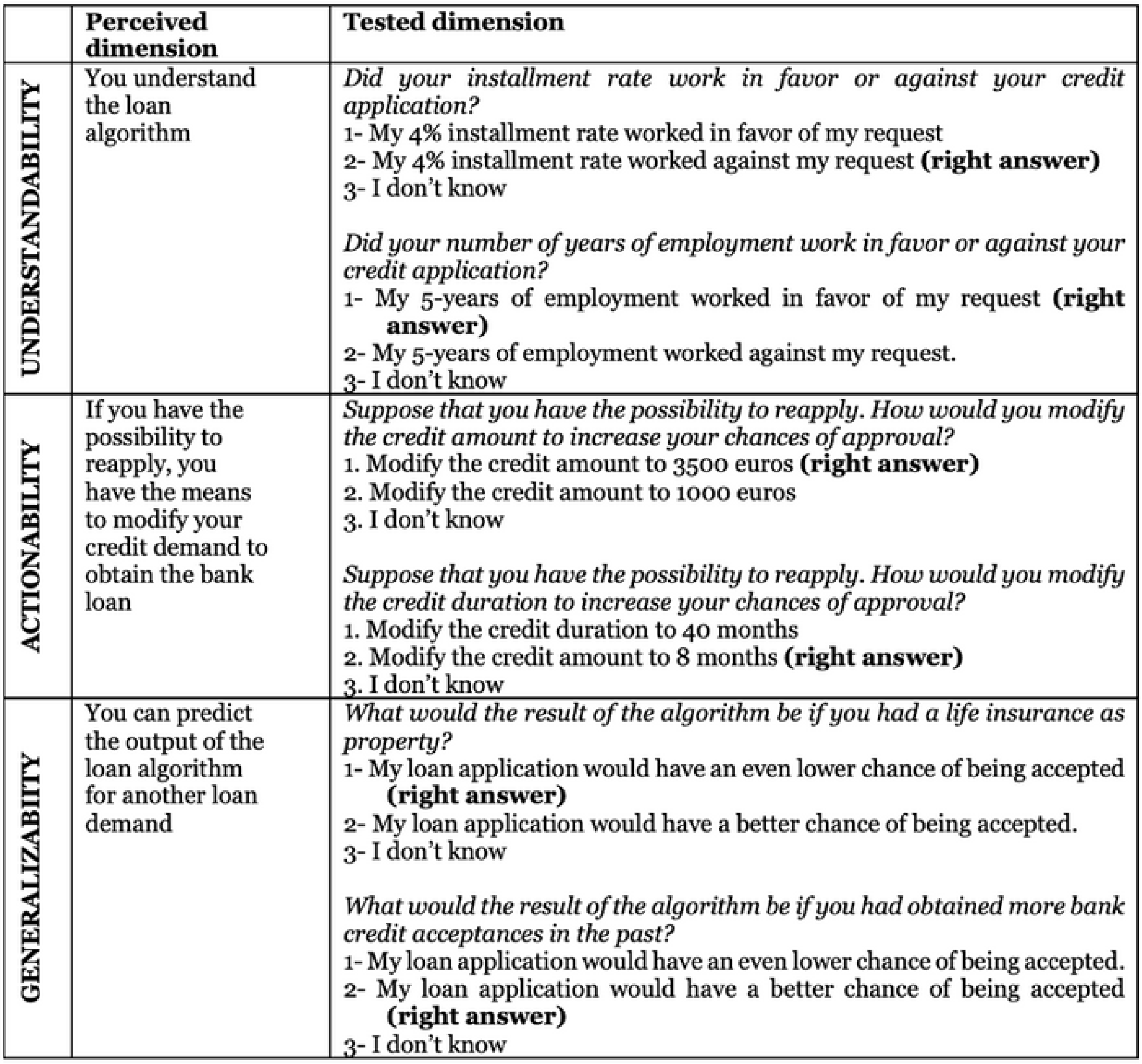
×

3 Dimension of interpretability
Understandability
Actionability
Generalizability

×

Mode of interpretability
Perception
Tests

For more details:
In concrete terms, the respondents were interviewed to see whether they had the right interpretation or not and to establish a score. It is by measuring the correct answers that we can know whether people have understood the algorithm correctly or not.



	Perceived dimension	Tested dimension
UNDERSTANDABILITY	You understand the loan algorithm	*Did your installment rate work in favor or against your credit application?* 1- My 4% installment rate worked in favor of my request 2- My 4% installment rate worked against my request **(right answer)** 3- I don't know *Did your number of years of employment work in favor or against your credit application?* 1- My 5-years of employment worked in favor of my request **(right answer)** 2- My 5-years of employment worked against my request. 3- I don't know
ACTIONABILITY	If you have the possibility to reapply, you have the means to modify your credit demand to obtain the bank loan	*Suppose that you have the possibility to reapply. How would you modify the credit amount to increase your chances of approval?* 1. Modify the credit amount to 3500 euros **(right answer)** 2. Modify the credit amount to 1000 euros 3. I don't know *Suppose that you have the possibility to reapply. How would you modify the credit duration to increase your chances of approval?* 1. Modify the credit duration to 40 months 2. Modify the credit amount to 8 months **(right answer)** 3. I don't know
GENERALIZABIITY	You can predict the output of the loan algorithm for another loan demand	*What would the result of the algorithm be if you had a life insurance as property?* 1- My loan application would have an even lower chance of being accepted **(right answer)** 2- My loan application would have a better chance of being accepted. 3- I don't know *What would the result of the algorithm be if you had obtained more bank credit acceptances in the past?* 1- My loan application would have an even lower chance of being accepted. 2- My loan application would have a better chance of being accepted **(right answer)** 3- I don't know

KEY FINDINGS

- **Limitations of post-hoc explanations of black-box-models** compared to transparent ML models.
    - Behavioral interpretability is strongly weakened by post-hoc explanations of a black-box model.
    - Post-hoc explanations tend to give partial and biased information on the underlying mechanism of the algorithm, which tends to actually mislead the participants while they overestimate their capacity to interpretate the decision in a declarative basis (Rudin 2019). Factual explanations are not good. They tend to distract attention by focusing on actionable variables, thus lowering people's ability to generalize.
- The **opposition between self-reported indicators and tested behavioral indicators is key**. Post-hoc methods make people think they understand the algorithm, when in reality, when tested with questions, they do not. One of the contributions of this paper is that to succeed in showing the limits, the technique is to say "you have to stop telling people if they have understood or not, you have to test them".

KEY TAKEAWAYS

- This paper provides **empirical concerns about post-hoc explanations of black-box ML models**, one of the major trends in AI explainability (XAI), by showing its lack of interpretability and societal consequence.
- « We show the importance of tested behavioral indicators, in addition to self-reported perceived indicators, to provide a more comprehensive view of the dimensions of interpretability. »
- Recommendation: **Prefer inherently simple models or additive models**

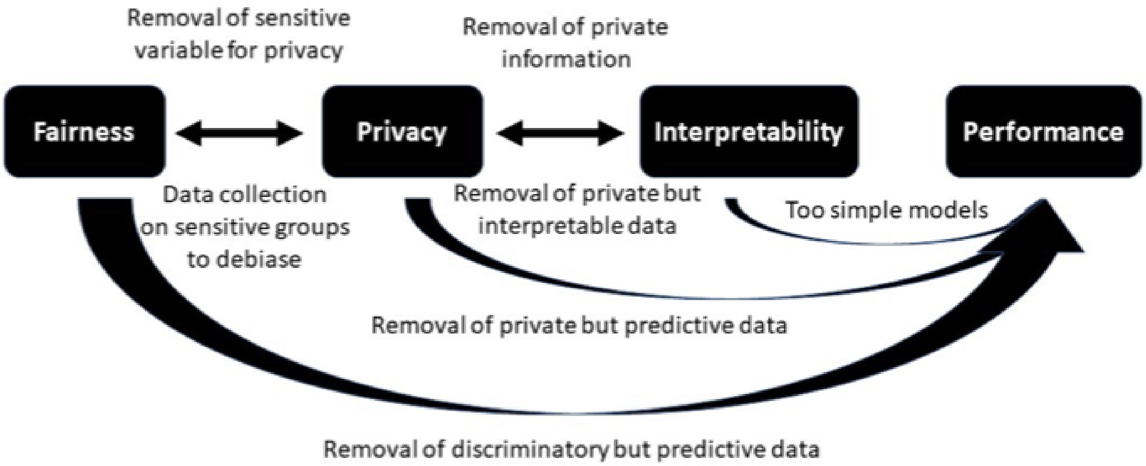
A Self-regulatory Framework for AI Ethics -opportunities and challenges, 2021

WHY IS THIS TOPIC IMPORTANT ?

> There is a great divergence between ethical charters and guidelines and the actual practices that are rooted in an economic reality.

This paper seeks to translate these major principles into more technical constraints so that algorithm designers can adopt them. In particular, it seeks to analyze the behavior of data scientists who sculpt the algorithms. The interface seeks to accompany algorithm builders towards more ethical decisions.

It is necessary to make compromises between different criteria. For example, the principle of fairness poses a problem. To avoid discrimination bias, a lot of personal information is needed, which creates friction with the desire for privacy, transparency and interpretability. **The objective is therefore to create a concrete and practical tool** that allows the theoretical principles of ethical charters to be translated into actions and that allows the various ethical metrics to be arbitrated.


Removal of sensitive variable for privacy
Removal of private information
Fairness
Privacy
Interpretability
Performance
Data collection on sensitive groups to debiase
Removal of private but interpretable data
Too simple models
Removal of private but predictive data
Removal of discriminatory but predictive data

METHODOLOGY

Jean-Marie used a **design science methodology**.
This method consists in understanding a phenomenon in society, not only by passively observing it but rather by creating a tool and confronting it empirically. It is precisely through the adjustment that this phenomenon can have in contact with the tool that we can understand this phenomenon. We seek to disturb the phenomenon in order to understand it, in an "action research" approach.

The objectives of the interface are:

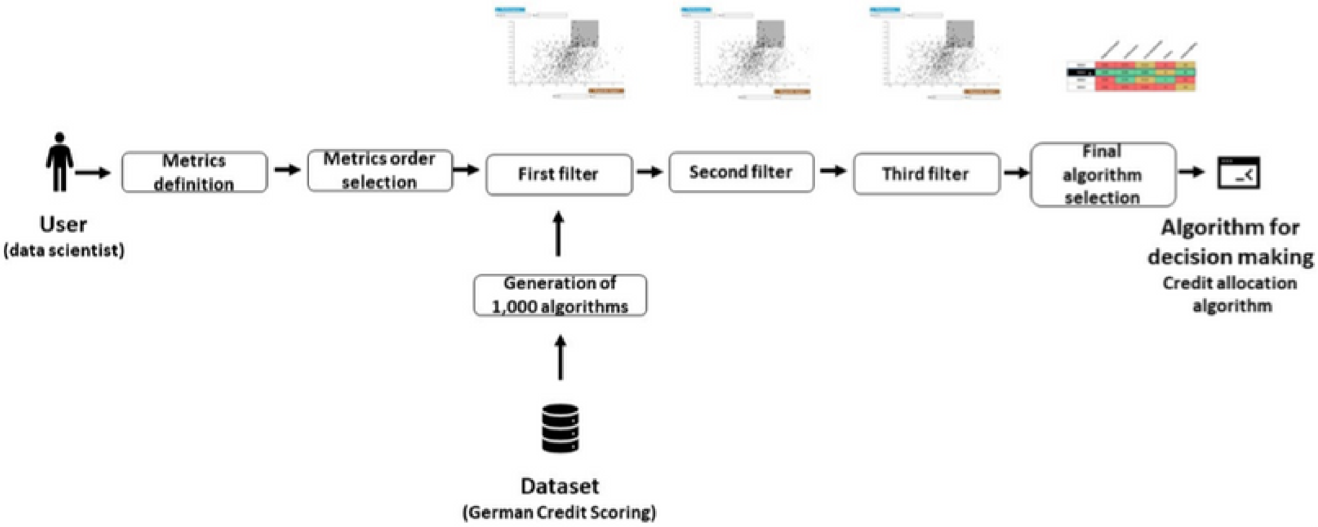
- Is it technically possible to create such a tool?
  - It must meet certain objectives: AI practitioners must have a better understanding of AI, the tool must allow the creation of more ethical algorithms, it must be easy to use and generate debate.
- By creating such a tool, can we learn more about the socio-technical world?
  - How much do data scientists care about ethical issues? How do they build ethical artificial intelligence algorithms?

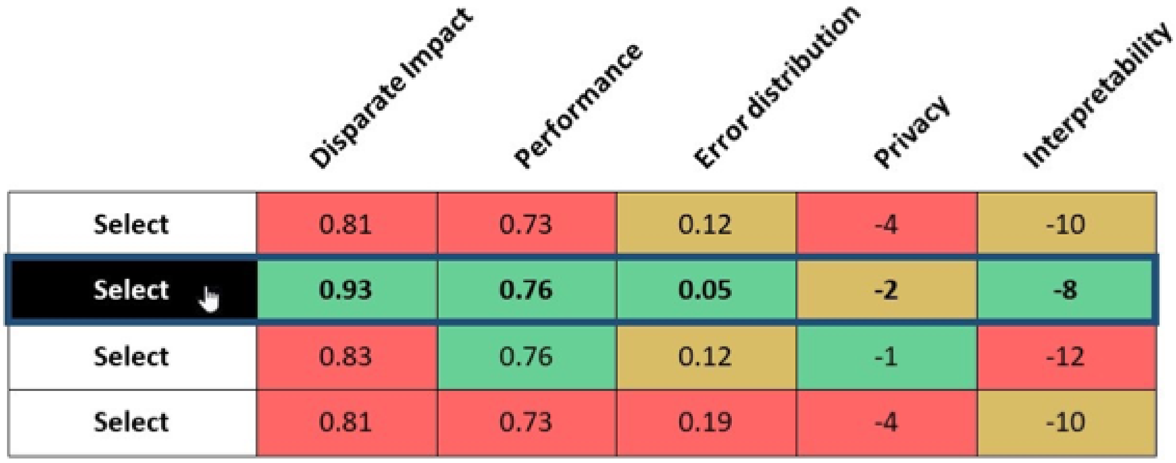
The technical design of the interface :

- Use of the very new **Fairness Machine Learning (FairML)** method that offers methods to calculate algorithmic biases
- Use of the German credit scoring open source database
- This interface **could be used for recruitment, credit scoring, resource allocation or social benefits issues**...
- For the moment, the interface is tested on a credit scoring case, i.e. the algorithm created will determine if a person is solvent or not and if he/she will be able to pay back a credit. These algorithms are often discriminating and lack transparency.

> **We have chosen to develop a decision support tool for data scientists and designers to address the ethical issues of AI.**

- Jean-Marie asks data scientists who work in finance to build an algorithm from the interface. The interface generates 1000 algorithms from the user's preferences and the user will have to choose one of them thanks to the interface.
- Jean-Marie then tries to determine if the scientists create more ethical algorithms thanks to the interface and if they manage the tradeoff between the different ethical and technical constraints: fairness, privacy, interpretability and performance.
- Use of **five different metrics that address the different ethical goals**:
  - Two metrics for fairness (disparate impact and error distribution)
  - One metric for interpretability (total numbers of features in the model)
  - One metric for privacy (number of private features)
  - One metric for performance (area under the curve)
- The interface helps data scientists to reach a decision by analyzing the algorithms and comparing the different metrics two by two. The final choices of the data scientists thus reveal their ethical preferences.





	Disparate Impact	Performance	Error distribution	Privacy	Interpretability
Select	0.81	0.73	0.12	-4	-10
Select	0.93	0.76	0.05	-2	-8
Select	0.83	0.76	0.12	-1	-12
Select	0.81	0.73	0.19	-4	-10

KEY FINDINGS

The key findings are of two kinds.

- Concrete metrics
  - He compares the algorithm selected by the scientist to an algorithm previously created without the interface by observing how the 5 dimensions are distributed in each of the algorithms and what choices were made in both cases.
  - This allows him to understand and see the impact of the interface as well as to verify if the interface has met its objectives.
  - Result: The algorithms without the interface were more efficient but less "ethical" (in terms of privacy, explainability and fairness) with respect to the 4 other metrics.

	Performance	Disparate Impact	Error Distribution	Privacy (number of private features)	Interpretability (number of features)
With the interface	0.74 (<0.01)	0.91 (<0.01)	0.14 (<0.35)	5.8 (<0.01)	9.4 (<0.01)
Without the interface	0.82	0.78	0.15	6.5	21

·Perception

- He askes simple questions to users in order to get their opinion on the interface in relation to the different objectives. He seeks to know if users feel that the interface has allowed them to make more ethical decisions and to understand the ethical issues of artificial intelligence.
- Result: Users mostly confirm that the interface is easy to use and helps generate debate, understand the ethics of AI and can help make decisions more ethical.

KEY TAKEAWAYS

- This work has a positive outcome, as it has proven possible to create technical interfaces that translate top-down ethical principles and transform them into technical functionalities that can be operationally activated by data scientists.
- Data scientists can sacrifice performance if given an interface that can allow them to trade-off different metrics proxying ethical dimensions such as fairness, privacy or explainability.
- Such a self-regulatory and technical approach for AI regulation can only work if some structural conditions are met. This approach requires that organizations invest time and money to mitigate AI issues and produce better algorithms.

The Displacement of Reality Tests. The Selection of Individuals in the age of Machine Learning, 2022

Co-author: Dominique Cardon

Dominique Cardon

Dominique Cardon is a Sociology professor at Sciences Po. Since 2010, his work has focused on the sociological analysis of web and big data algorithms in order to understand both the internal form of computation and the world that computers project on our societies. His research also focuses on the transformations of the media space and the new circuits of digital information. He is a member of the editorial board of the journal Réseaux and of the prospective committee of the CNIL.

WHY IS THIS TOPIC IMPORTANT ?

> **With the advent of Artificial Intelligence, we observe a transformation of the classification tests used in our society (for recruitment, credit scoring, at university, in health, etc.)**

The idea of this socio-philosophical article is to say: in our society **we constantly have classification tests** (for recruitment, credit scoring, at university, in health, etc.). These tests imply a modification of the status of individuals: titles, diplomas, badges, scores, qualifications, resources, and opportunities...

In order to rank people, institutions are set up to create this legitimacy, i.e. at the end of the ranking test people accept the result. The results of a **reality test** (Luc Boltanski and Laurent Thévenot's terminology) must be justifiable and supported by institutions. Our society is organized in such a way that institutions carry these tests and stabilize them.

In this article, the authors show that with the advent of machine learning, of artificial intelligence, we have a transformation of these classification tests. They describe this transformation in four parts based on the theory of Luc Boltanski (a pragmatic sociologist). Theses four parts are developed in the key findings.

This topic is important because :

- Some sociologists Luc Boltanski explains that **we need these reality tests to qualify what happens to us**: "I am a worker", "I am a csp+", "I am a victim", "I am a privileged person", etc.
- These tests are everywhere and "**they make the reality of our society**", according to Boltanski (in the social sense, i.e. the categories).
- Today, with the arrival of these **new algorithmic methods**, we are **transforming these classification tests** and therefore modifying the reality of our societies.

METHODOLOGY

For this survey, the authors worked with Etalab, the interministerial body that promotes open data in France. Etalab oversees several algorithms and the authors discussed with them about this. They organized discussions with several people, researchers or a designers of the tool discussed during the session.

They explored different modes of selection used in French society:

- ParcoursSup system, the platform used by secondary school students to apply to higher education institutions;
- The allocation of nursery places by municipalities;
- The Agence de biomédicine's Score Coeur device for assigning coronary artery bypass grafts;
- Cib Nav, the automatic audit triggering system for ships going to sea;
- The prediction of the evolution of job offers at Pôle emploi.

Finally, in another investigation, the author looked at the recruitment procedures of companies that used algorithmic devices (John Matthews et al., 2021).

Theoretical source: **Boltanski**, Reality test, Reality and world

.

Boltanski L., Thévenot L., *On Justification. Economies of Worth*, Princeton University Press, 2006.
Boltanski L., *Critique. A Sociology of Emancipation*, Cambridge, Polity, 2011.
Boltanski L., *Mysteries and conspiracies: Detective stories, spy novels and the making of modern societies*, John Wiley & Sons, 2014.

KEY FINDINGS

The authors articulate 4 issues :

1. **Displacement of reality tests**. A transformation of selection tests in our societies: performance tests, form tests, file tests, and continuous tests.
2. **De-categorization of datasets**. The process of spatial-temporal expansion of the data space used to automate decisions (the comparison space).
3. **Change in the computation form**. Create a multi-dimensional space opening multiple paths between the initial data and the expected objective (deep learning techniques). The transformation of the calculation methods with new algorithms that increasingly go towards the impossibility of explaining and interpreting the selection criteria.
4. **Displacement of justification regimes**. How to justify (through principles) and legitimize (through institutional stabilization authorities) the principles on which calculations base selection? Tests are increasingly difficult to criticize because they are no longer supported by institutions that give them legitimacy. For example, if you want to criticize a test, you need an institution that supports the test (e.g. the Ministry of Education). The fact that algorithms are being put in place means that we need less and less institutions, which provide the key principles of selection. For the institutions, which are criticized, it is a way of delegating responsibility to the algorithm.

KEY TAKEAWAYS

It is essential to always leave room for criticism, a possibility for people to question the realities that come out of the algorithms. For example, if we talk about an algorithm that selects people, we will create new algorithmic classes in which some people will be more successful than others, more privileged than others. In political and activist action, we need people to be able to take hold of the categories, to give them meaning, and to be able to question them.

The problem is that if you don't allow for a critique because these categories are new and constantly changing, you don't allow for political action.

We then engage in **a system of expert domination**. This is a term introduced by Luc Boltanski, which means that we create a system in which we create reality tests - which allow us to represent the world but also to criticize it - and we have an 'expert' category of the population that is always maneuvering the permanent change of reality tests. **Expert domination is a permanent change in the format of the tests so that criticism no longer has a hold.**

Recomposing Normativities in Machine Learning Practices through Material and Discursive Experiments, 2022

Co-authors: Robin de Mourat, Donato Ricci, Maxime Crépel

Robin de Mourat

He holds an agrégation in Applied Arts (ENS Paris-Saclay preparation) and a PhD in Aesthetics (Rennes 2 University). He has worked in several contexts of scientific collaboration with researchers in architectural history and media sociology. He has held the position of research designer in digital methods at the médialab of Sciences Po since 2019.

Donato Ricci

Donato Ricci is a designer and researcher. He specialises in the use of design methods in the Humanities and Social Sciences.

He has followed the design aspects of Bruno Latour's EME project, the Reset Modernity! exhibition at ZKM Karlsruhe and the Shanghai Himalayas Museum. From 2005 to 2012, he participated in the development of the DensityDesign Lab research programme. He is assistant professor of "Representação e Conhecimento" (Knowledge and Representation) at the Universidade de Aveiro. He promotes the use of political design as a means of social enquiry.

Maxime Crépel

Maxime Crépel is a sociologist and research engineer at the Medialab of Sciences Po. His research focuses on the uses of digital technology and is partly financed by Good in Tech. He is part of the algoglitch project which aims to explore representations and forms of negotiation between users and algorithms.

WHY IS THIS TOPIC IMPORTANT ?

This topic is important because AI is experiencing many incidents today: biases and discrimination have been identified in various AI applications such as predictive models in trials, facial recognition, speech recognition, AI for recruitment, predictive models in health care and search engines.

Today in the world of **AI regulation**, which is also a discipline, the ethics of Artificial Intelligence, there are two main trends:

- **the moralization of AI technicality**: the objective of this approach is to find principles and write ethical charters, guidelines and recommendations for developing corresponding Artificial Intelligences.
- **the technicization of morality**: the technical approach considers, on the contrary, that it is preferable to find technical solutions. It is then a question of developing algorithms that will repair other algorithms. Computer scientists tend to opt for this approach, considering that principled debates have not proved very effective in regulating AI.

The authors argue that it is fundamental to describe the process of stabilization of algorithmic normativities (Grosman and Reigeluth, 2019) if we want to address the variety of challenges posed by AI.

> **AI is experiencing many incidents today: biases and discrimination have been identified in various AI applications.**

To study and intervene into these processes this paper proposes **an experimental interview setting for the data scientists involved in AI algorithm creation**.

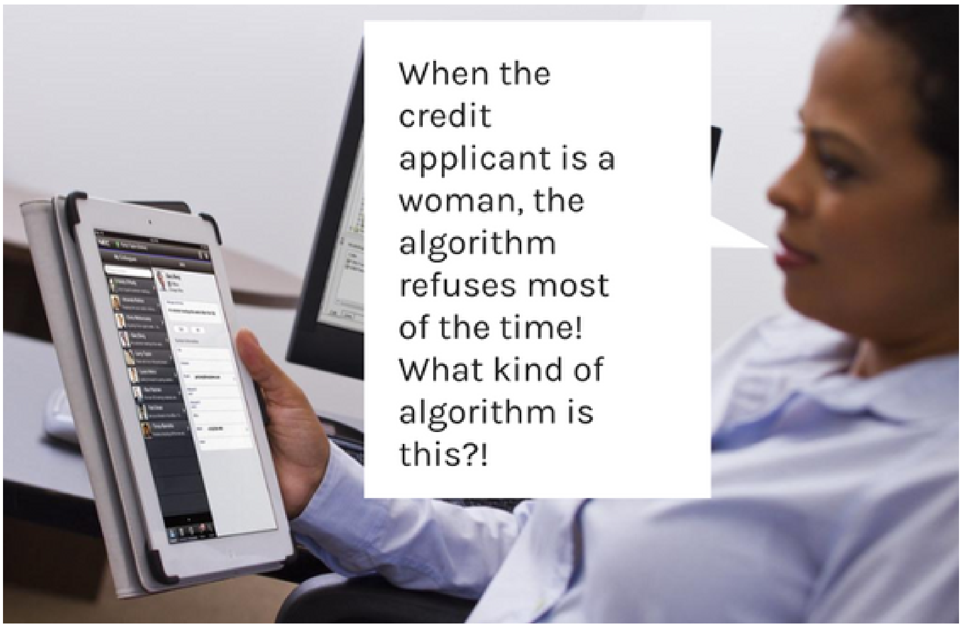
In the first part, the authors set out the Latourian ethical concepts on which they will base their conclusions, and in the second part they explain their survey method and approach, ending with the results obtained and their interpretation.

METHODOLOGY

Interview protocol on the **reenactment of data scientists' daily practices**:

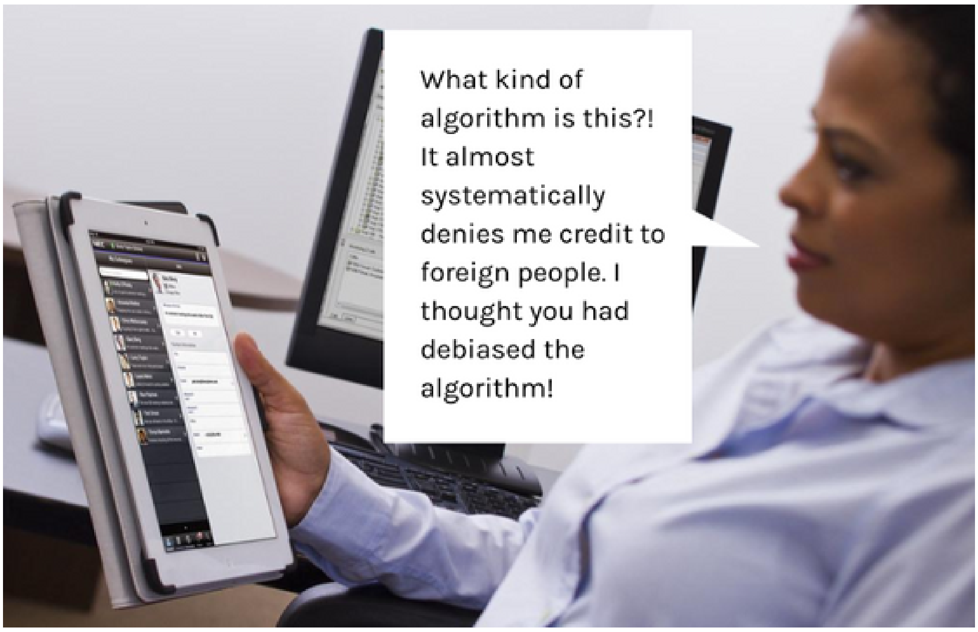
The authors went to meet data scientists in the field of banking and asked them to create an AI algorithm that decides whether people can get credit or not. They make this algorithm in about 30 minutes and once it is done, the authors provokes the data scientists by staging problematic situations.

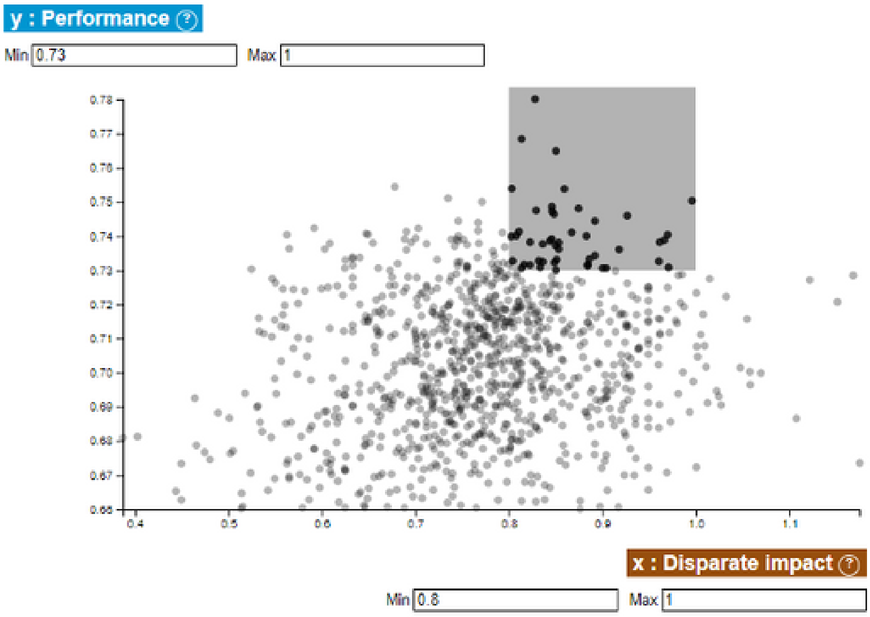
For example:


When the credit applicant is a woman, the algorithm refuses most of the time! What kind of algorithm is this?!

Here, the data scientist has to justify himself because he is the one who created the algorithm.

Then the interviewers continue with successive provocations:


What kind of algorithm is this?! It almost systematically denies me credit to foreign people. I thought you had debiased the algorithm!



y : Performance ?

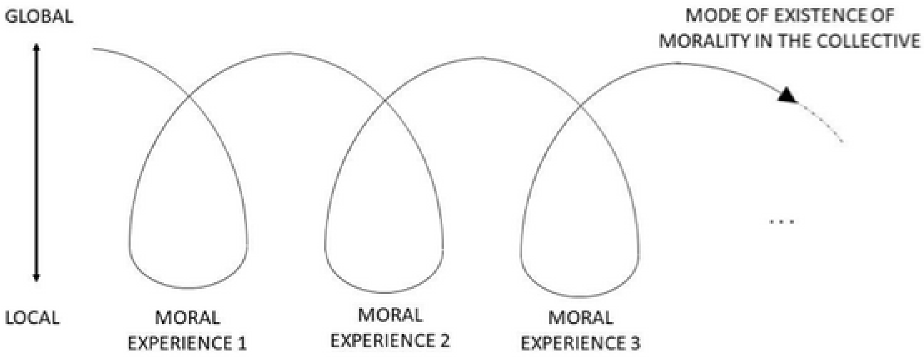
Min 0.73 Max 1

x : Disparate impact ?

Min 0.8 Max 1

Then we go to an interface shown above. Each point corresponds to an algorithm and the data scientists will be asked to choose one that best corresponds to their ethical preferences. The aim is to solve ethical problems.

Then, we continue with successive challenges. The data scientist finds himself in a cycle where he will have to justify himself. So, the authors proposes a solution that allows the algorithm to be debiased. **They created a software that allows data scientist to take into account the remarks.**



This investigation device is **modelled by this scheme**: they meet a data scientist, they provoke him, and each provocation is called a "**moral experience**". It allows them not to hover in the discourses of the ethics of AI (the great principles or in the great technicality). Here, the objective is to dig towards the local by provoking concrete situations. **The local level is the moment when data scientists will hesitate, be a bit emotional**, reject things, invoke external entities (opinions, emotions, etc).

And then there are times when, on the contrary, they go very quickly by mobilising external entities such as the GDPR (General Data Protection Regulation), the law, the official figure, the metric, the disparate impact, etc. All these tools allow data scientists to go fast. They will believe that an ethical problem is solved quickly with a few ethical metrics. This is the global moment, the second pole.

The investigation device consists of **going back and forth between the global and the local by provoking successive moral experience**. The idea is to say that **our ethical preferences are co-constructed with these investigation devices**. We form an opinion on what happens at the time the situation arises. This investigative device allows the mode of existence of morality to circulate.

KEY FINDINGS

The result, and there is only one: in situation, there are moments of back and forth between the **local and the global**. **Both are necessary**: the hesitation is proof of the reflective attitude (as an individual, how the scientists position themselves in relation to this problem). Conversely, in the global pole, data scientists no longer have a reflective attitude because they go fast and they mobilise the law to justify their algorithm. They do not have this reflexive attitude but they manage to mobilise entities and delegate their responsibility. They can thus say: "I just followed my company's guidelines".

In conclusion, the main idea is that there is a back and forth between the global and the local, and both are necessary to regulate AI. There is a tendency in the traditional way of regulating AI to focus only on the global, whereas the local allows people to position themselves.

KEY TAKEAWAYS

› We must **be careful not to globalise the debate on the ethics of AI**, either through an ultra-technical approach or an ultra-principled approach. Instead of going in that direction, **it is important to localise the debate**. It is only in this framework that we can **create a reflective and empowering attitude of AI actors**.

Issue 3 - January 2022

Good In Tech RESEARCH NEWS

Rethinking innovation and technology as drivers of a better world for and by humans

Christine Balagué
Professor at Institut Mines-Télécom Business School
Co-holder of the Good In Tech Chair
christine.balague@imt-bs.eu

Dominique Cardon
Professor at Sciences Po and director of the medialab
Co-holder of the Good In Tech Chair
dominique.cardon@sciencespo.fr

Jean-Marie John-Mathews
Data scientist
Coordinator of the Good In Tech Chair
jean-marie.john-mathews@imt-bs.eu

Jade Vergnes
Writer for Good In Tech Research News
jade.vergnes@sciencespo.fr

Clic here to contact