THE NEW 'METHODENSTREIT' IN ML

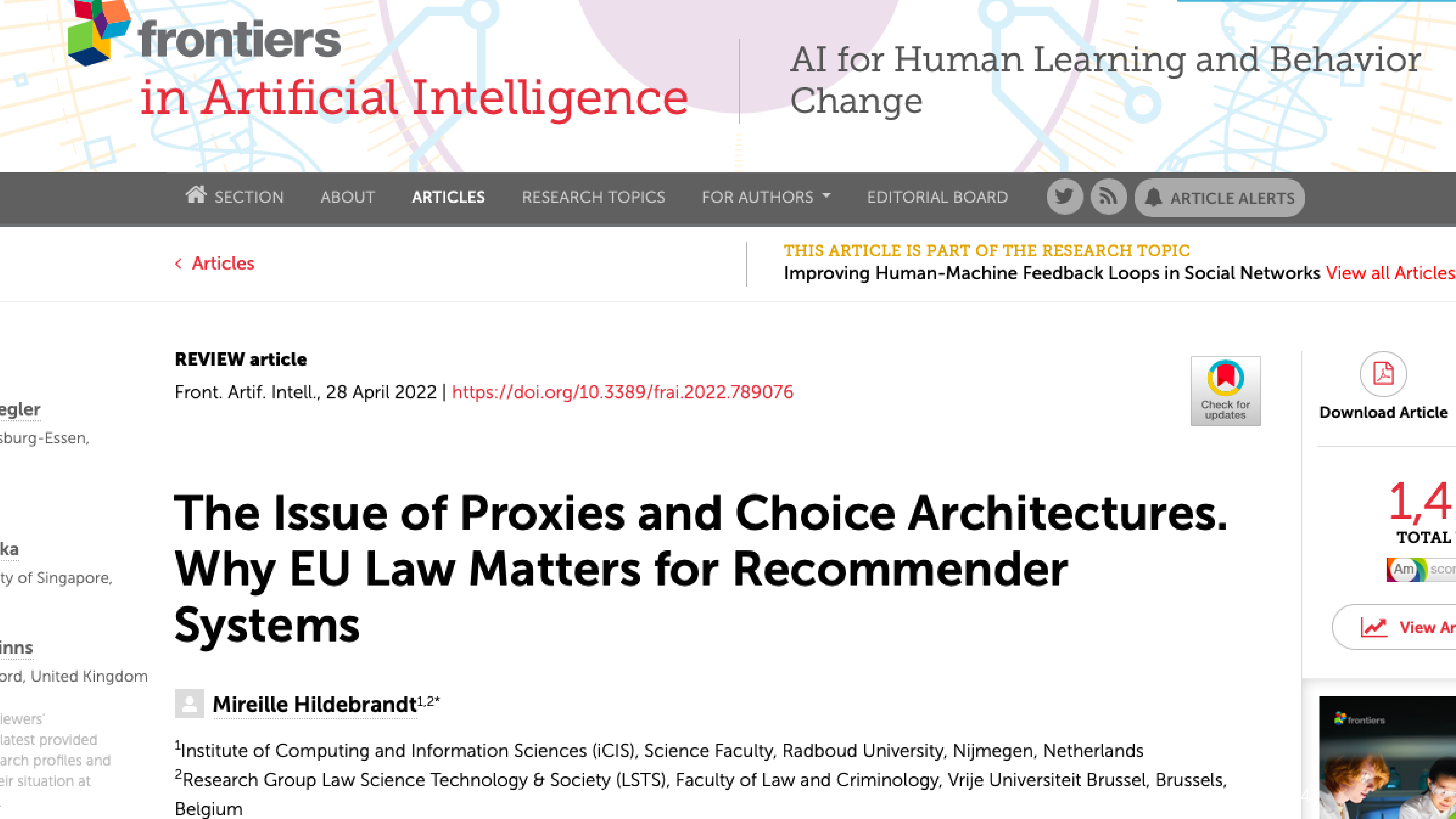
Mireille Hildebrandt

PROXIES AS BRIDGING UNDERSTANDING AND EXPLANATION

Mireille Hildebrandt

- My interest:
  - *Detect upstream design decisions that make a difference (FRIA)*

- We need:
  - *An internal critique of machine learning (explaining/understanding)*



THIS ARTICLE IS PART OF THE RESEARCH TOPIC

Improving Human-Machine Feedback Loops in Social Networks View all Articles

REVIEW article

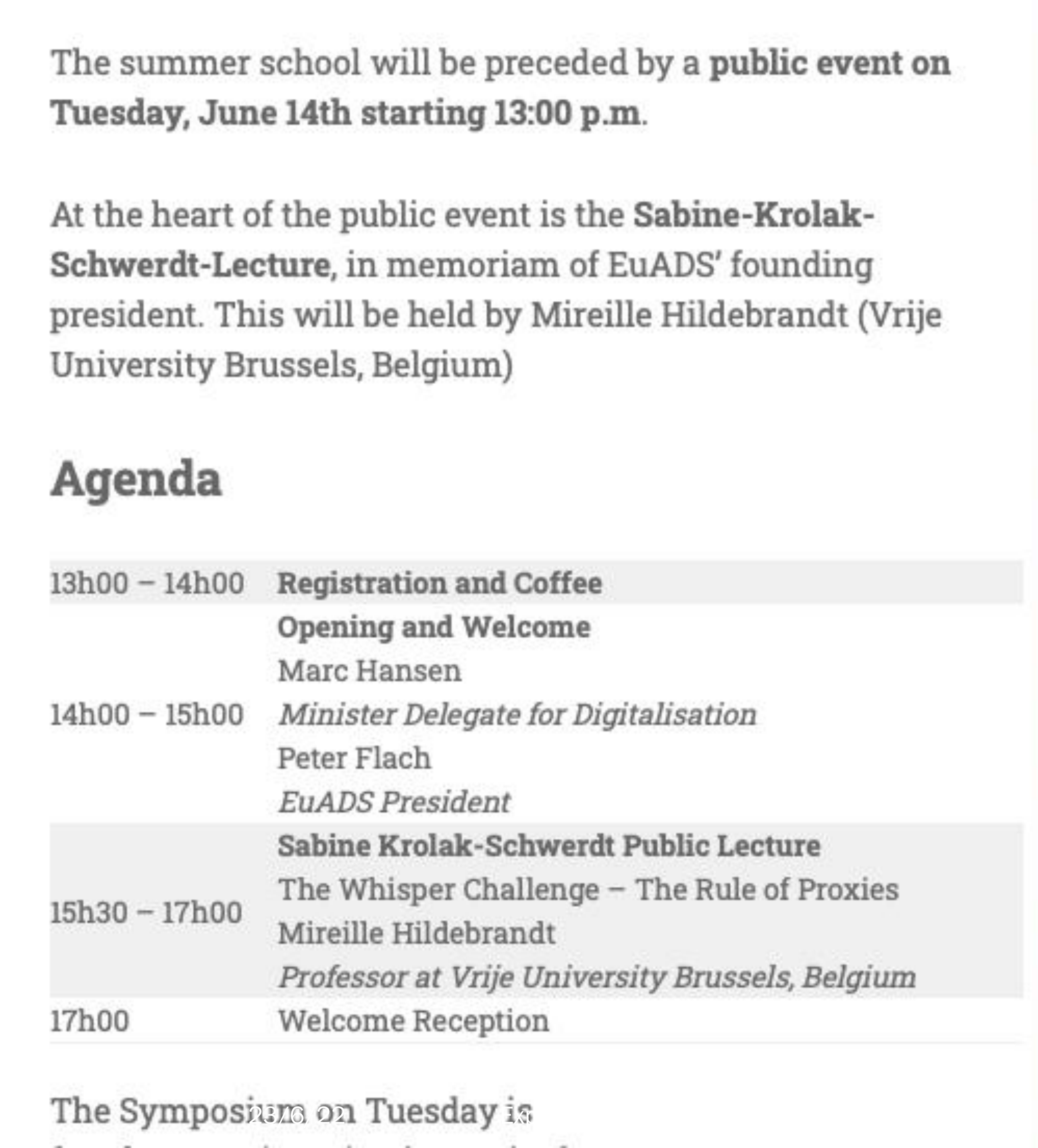
Front. Artif. Intell., 28 April 2022 | https://doi.org/10.3389/frai.2022.789076

The Issue of Proxies and Choice Architectures. Why EU Law Matters for Recommender Systems

Mireille Hildebrandt[1,2*]

[1]Institute of Computing and Information Sciences (iCIS), Science Faculty, Radboud University, Nijmegen, Netherlands

[2]Research Group Law Science Technology & Society (LSTS), Faculty of Law and Criminology, Vrije Universiteit Brussel, Brussels, Belgium



The summer school will be preceded by a **public event on Tuesday, June 14th starting 13:00 p.m**.

At the heart of the public event is the **Sabine-Krolak-Schwerdt-Lecture**, in memoriam of EuADS' founding president. This will be held by Mireille Hildebrandt (Vrije University Brussels, Belgium)

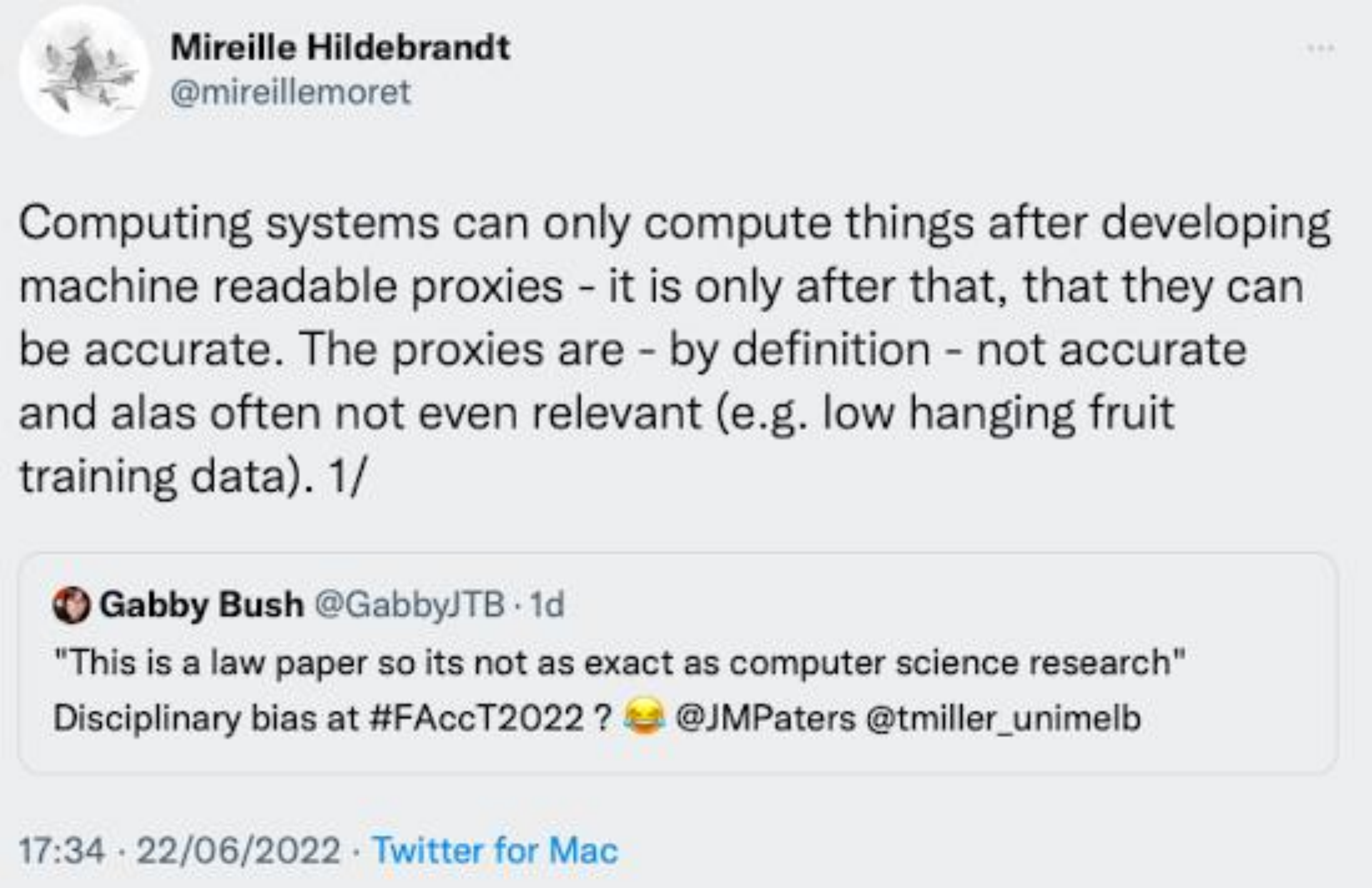
Agenda

13h00 – 14h00	**Registration and Coffee**
14h00 – 15h00	**Opening and Welcome** Marc Hansen *Minister Delegate for Digitalisation* Peter Flach *EuADS President*
15h30 – 17h00	**Sabine Krolak-Schwerdt Public Lecture** The Whisper Challenge – The Rule of Proxies Mireille Hildebrandt *Professor at Vrije University Brussels, Belgium*
17h00	Welcome Reception

The Symposium on Tuesday is

The Whisper Challenge - The Rule of Proxies in Data Science

Mireille Hildebrandt (Vrije Universiteit Brussel)

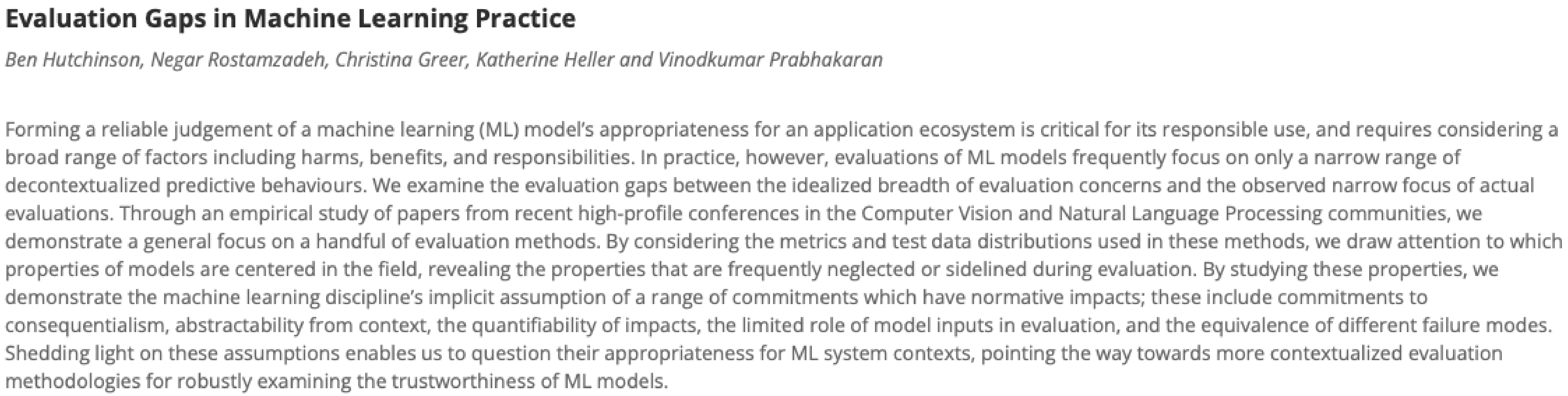


Mireille Hildebrandt
@mireillemoret

Computing systems can only compute things after developing machine readable proxies - it is only after that, that they can be accurate. The proxies are - by definition - not accurate and alas often not even relevant (e.g. low hanging fruit training data). 1/

> **Gabby Bush** @GabbyJTB · 1d
> "This is a law paper so its not as exact as computer science research" Disciplinary bias at #FAccT2022 ? 😂 @JMPaters @tmiller_unimelb

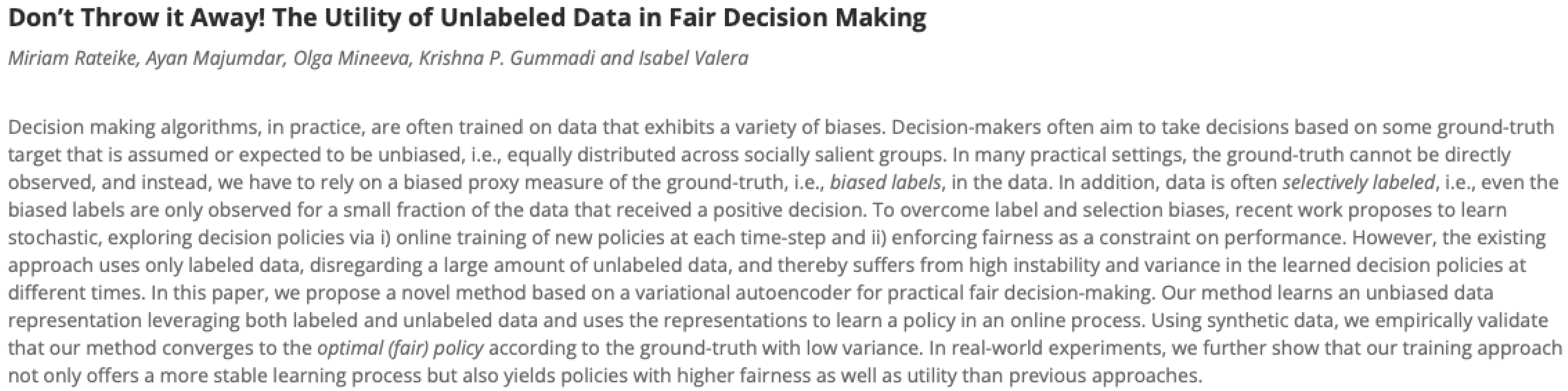
17:34 · 22/06/2022 · Twitter for Mac



Evaluation Gaps in Machine Learning Practice

Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller and Vinodkumar Prabhakaran

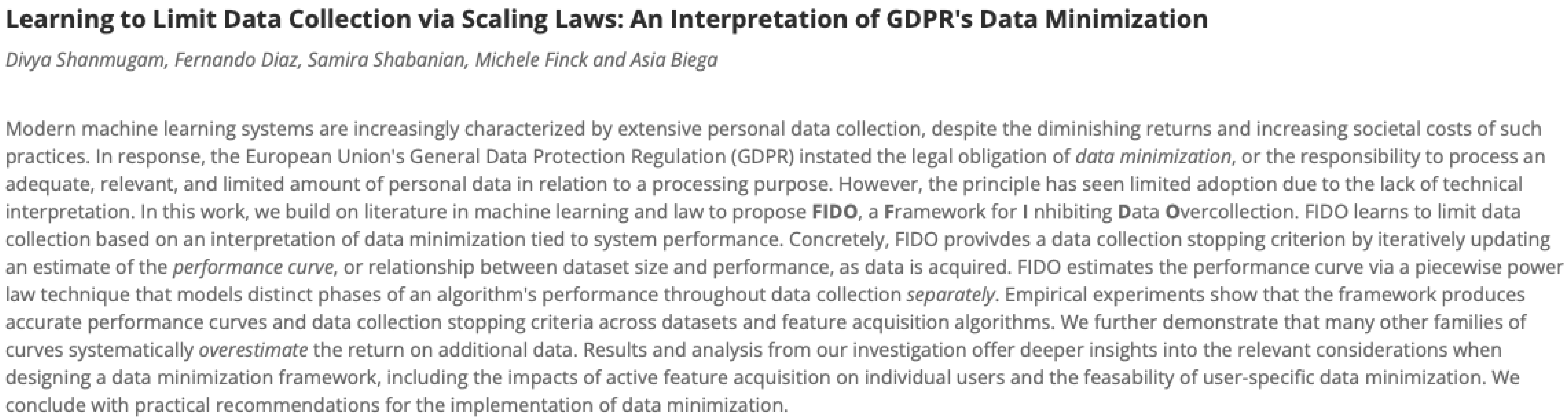
Forming a reliable judgement of a machine learning (ML) model's appropriateness for an application ecosystem is critical for its responsible use, and requires considering a broad range of factors including harms, benefits, and responsibilities. In practice, however, evaluations of ML models frequently focus on only a narrow range of decontextualized predictive behaviours. We examine the evaluation gaps between the idealized breadth of evaluation concerns and the observed narrow focus of actual evaluations. Through an empirical study of papers from recent high-profile conferences in the Computer Vision and Natural Language Processing communities, we demonstrate a general focus on a handful of evaluation methods. By considering the metrics and test data distributions used in these methods, we draw attention to which properties of models are centered in the field, revealing the properties that are frequently neglected or sidelined during evaluation. By studying these properties, we demonstrate the machine learning discipline's implicit assumption of a range of commitments which have normative impacts; these include commitments to consequentialism, abstractability from context, the quantifiability of impacts, the limited role of model inputs in evaluation, and the equivalence of different failure modes. Shedding light on these assumptions enables us to question their appropriateness for ML system contexts, pointing the way towards more contextualized evaluation methodologies for robustly examining the trustworthiness of ML models.



Don’t Throw it Away! The Utility of Unlabeled Data in Fair Decision Making

Miriam Rateike, Ayan Majumdar, Olga Mineeva, Krishna P. Gummadi and Isabel Valera

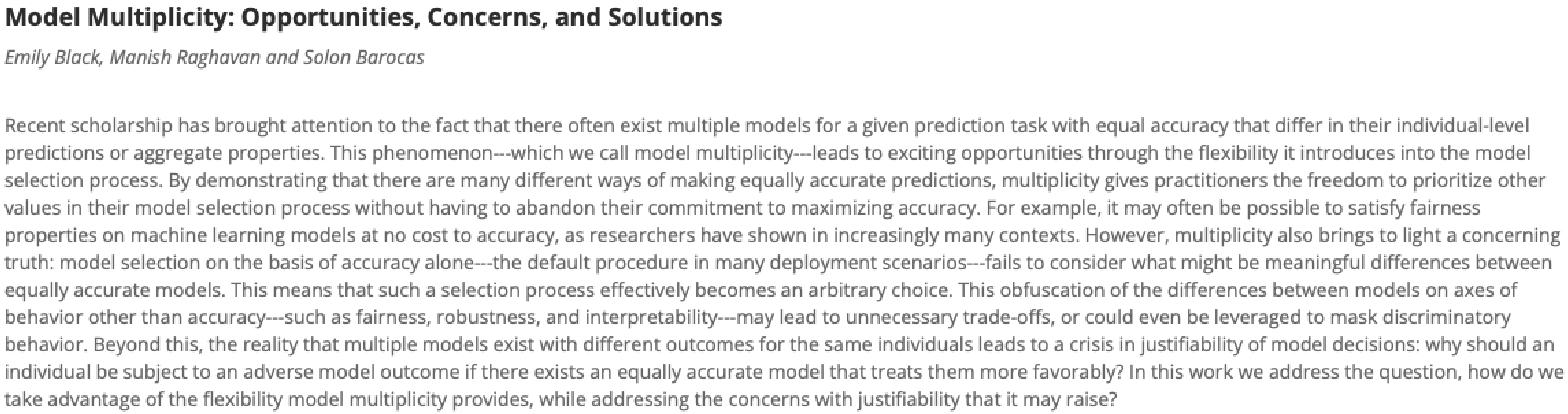
Decision making algorithms, in practice, are often trained on data that exhibits a variety of biases. Decision-makers often aim to take decisions based on some ground-truth target that is assumed or expected to be unbiased, i.e., equally distributed across socially salient groups. In many practical settings, the ground-truth cannot be directly observed, and instead, we have to rely on a biased proxy measure of the ground-truth, i.e., *biased labels*, in the data. In addition, data is often *selectively labeled*, i.e., even the biased labels are only observed for a small fraction of the data that received a positive decision. To overcome label and selection biases, recent work proposes to learn stochastic, exploring decision policies via i) online training of new policies at each time-step and ii) enforcing fairness as a constraint on performance. However, the existing approach uses only labeled data, disregarding a large amount of unlabeled data, and thereby suffers from high instability and variance in the learned decision policies at different times. In this paper, we propose a novel method based on a variational autoencoder for practical fair decision-making. Our method learns an unbiased data representation leveraging both labeled and unlabeled data and uses the representations to learn a policy in an online process. Using synthetic data, we empirically validate that our method converges to the *optimal (fair) policy* according to the ground-truth with low variance. In real-world experiments, we further show that our training approach not only offers a more stable learning process but also yields policies with higher fairness as well as utility than previous approaches.



Learning to Limit Data Collection via Scaling Laws: An Interpretation of GDPR's Data Minimization

Divya Shanmugam, Fernando Diaz, Samira Shabanian, Michele Finck and Asia Biega

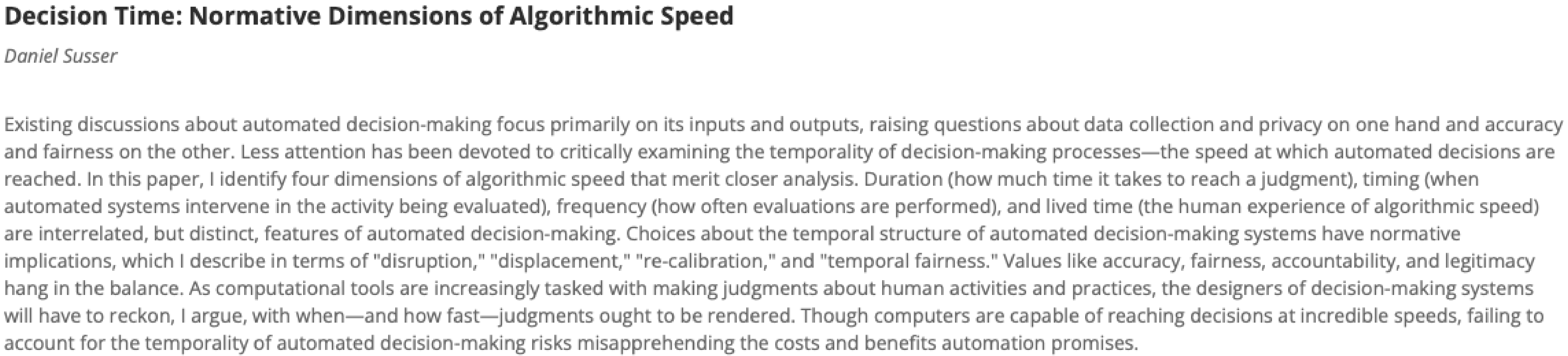
Modern machine learning systems are increasingly characterized by extensive personal data collection, despite the diminishing returns and increasing societal costs of such practices. In response, the European Union's General Data Protection Regulation (GDPR) instated the legal obligation of *data minimization*, or the responsibility to process an adequate, relevant, and limited amount of personal data in relation to a processing purpose. However, the principle has seen limited adoption due to the lack of technical interpretation. In this work, we build on literature in machine learning and law to propose **FIDO**, a **F**ramework for **I** nhibiting **D**ata **O**vercollection. FIDO learns to limit data collection based on an interpretation of data minimization tied to system performance. Concretely, FIDO provivdes a data collection stopping criterion by iteratively updating an estimate of the *performance curve*, or relationship between dataset size and performance, as data is acquired. FIDO estimates the performance curve via a piecewise power law technique that models distinct phases of an algorithm's performance throughout data collection *separately*. Empirical experiments show that the framework produces accurate performance curves and data collection stopping criteria across datasets and feature acquisition algorithms. We further demonstrate that many other families of curves systematically *overestimate* the return on additional data. Results and analysis from our investigation offer deeper insights into the relevant considerations when designing a data minimization framework, including the impacts of active feature acquisition on individual users and the feasability of user-specific data minimization. We conclude with practical recommendations for the implementation of data minimization.



Model Multiplicity: Opportunities, Concerns, and Solutions

Emily Black, Manish Raghavan and Solon Barocas

Recent scholarship has brought attention to the fact that there often exist multiple models for a given prediction task with equal accuracy that differ in their individual-level predictions or aggregate properties. This phenomenon---which we call model multiplicity---leads to exciting opportunities through the flexibility it introduces into the model selection process. By demonstrating that there are many different ways of making equally accurate predictions, multiplicity gives practitioners the freedom to prioritize other values in their model selection process without having to abandon their commitment to maximizing accuracy. For example, it may often be possible to satisfy fairness properties on machine learning models at no cost to accuracy, as researchers have shown in increasingly many contexts. However, multiplicity also brings to light a concerning truth: model selection on the basis of accuracy alone---the default procedure in many deployment scenarios---fails to consider what might be meaningful differences between equally accurate models. This means that such a selection process effectively becomes an arbitrary choice. This obfuscation of the differences between models on axes of behavior other than accuracy---such as fairness, robustness, and interpretability---may lead to unnecessary trade-offs, or could even be leveraged to mask discriminatory behavior. Beyond this, the reality that multiple models exist with different outcomes for the same individuals leads to a crisis in justifiability of model decisions: why should an individual be subject to an adverse model outcome if there exists an equally accurate model that treats them more favorably? In this work we address the question, how do we take advantage of the flexibility model multiplicity provides, while addressing the concerns with justifiability that it may raise?

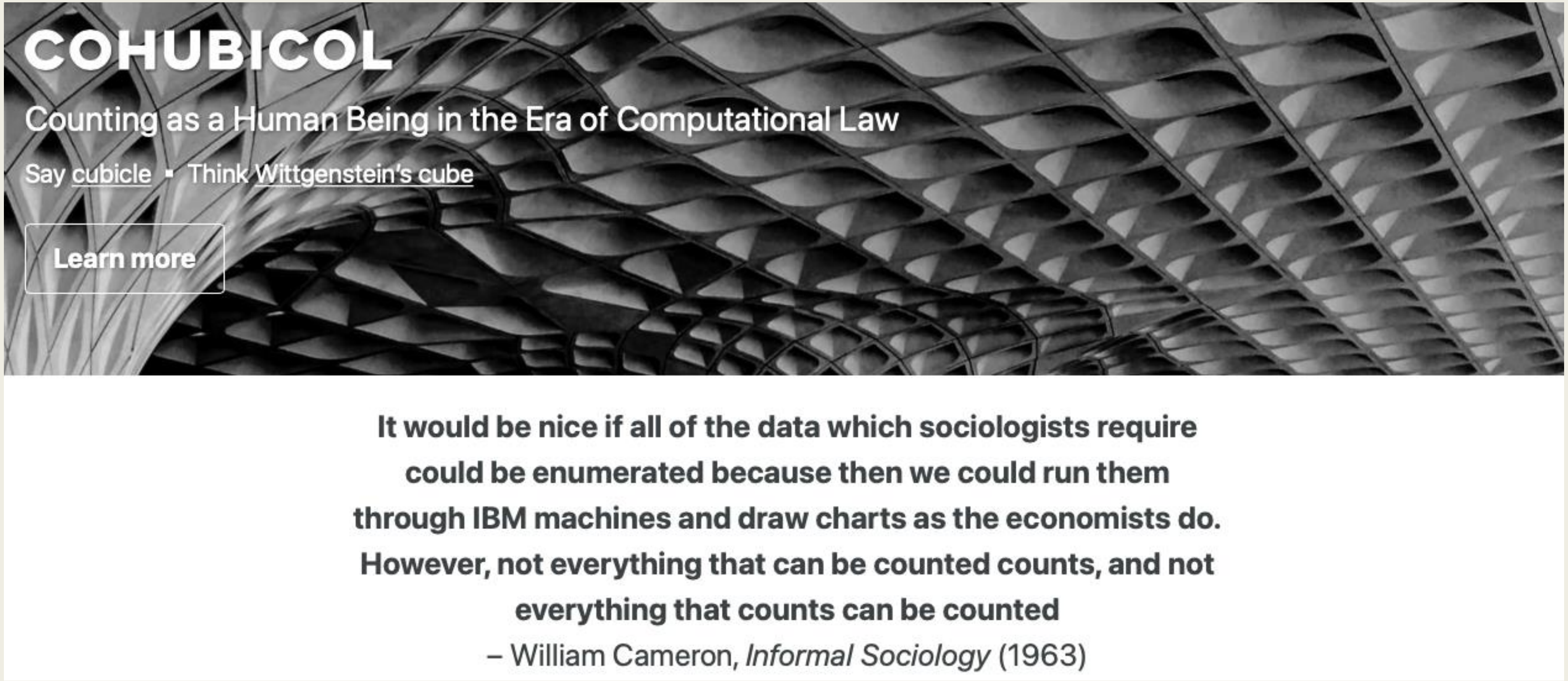


Decision Time: Normative Dimensions of Algorithmic Speed

Daniel Susser

Existing discussions about automated decision-making focus primarily on its inputs and outputs, raising questions about data collection and privacy on one hand and accuracy and fairness on the other. Less attention has been devoted to critically examining the temporality of decision-making processes—the speed at which automated decisions are reached. In this paper, I identify four dimensions of algorithmic speed that merit closer analysis. Duration (how much time it takes to reach a judgment), timing (when automated systems intervene in the activity being evaluated), frequency (how often evaluations are performed), and lived time (the human experience of algorithmic speed) are interrelated, but distinct, features of automated decision-making. Choices about the temporal structure of automated decision-making systems have normative implications, which I describe in terms of "disruption," "displacement," "re-calibration," and "temporal fairness." Values like accuracy, fairness, accountability, and legitimacy hang in the balance. As computational tools are increasingly tasked with making judgments about human activities and practices, the designers of decision-making systems will have to reckon, I argue, with when—and how fast—judgments ought to be rendered. Though computers are capable of reaching decisions at incredible speeds, failing to account for the temporality of automated decision-making risks misapprehending the costs and benefits automation promises.

- Datum = something given
  - *Raw data is an oxymoron (Gitelman), data is a construction*
  - *Data are proxies (traces, representations, imprints)*
- Factum = something made
  - *Les faits sont faits*
  - *Facts are made and tested, that’s when they are real*
  - *Dewey: an artificial lake is not an imaginary lake*



COHUBICOL

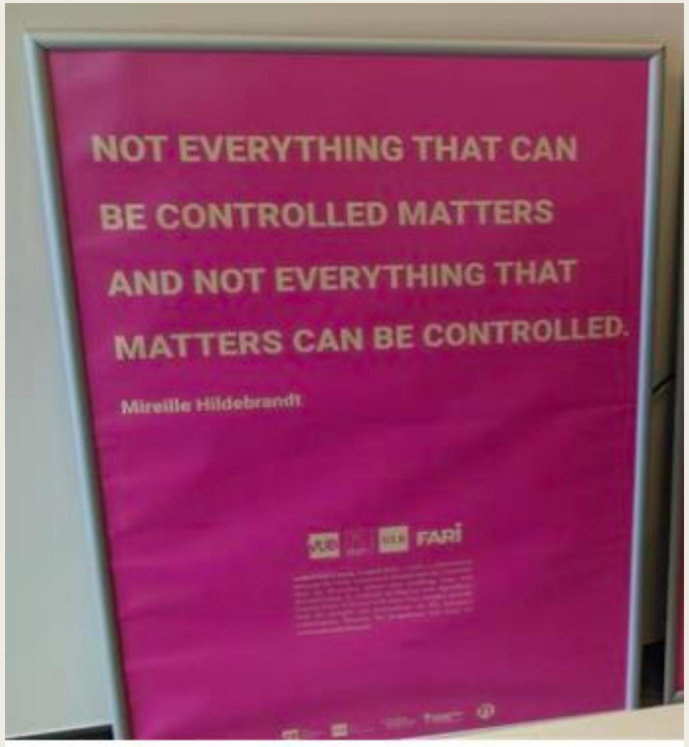
Counting as a Human Being in the Era of Computational Law

Say cubicle ▪ Think Wittgenstein's cube

It would be nice if all of the data which sociologists require could be enumerated because then we could run them through IBM machines and draw charts as the economists do. However, not everything that can be counted counts, and not everything that counts can be counted

– William Cameron, *Informal Sociology* (1963)

1. What matters is incomputable
2. It can nevertheless be made computable
3. In different ways – and that difference matters



NOT EVERYTHING THAT CAN BE CONTROLLED MATTERS AND NOT EVERYTHING THAT MATTERS CAN BE CONTROLLED.

Mireille Hildebrandt

Hartmut Rosa

Unverfügbarkeit

Unablässig versucht der moderne Mensch,
die Welt in Reichweite zu bringen:
Dabei droht sie uns jedoch stumm und
fremd zu werden: Lebendigkeit entsteht nur
aus der Akzeptanz des Unverfügbaren.

RESIDENZ VERLAG

CYBERNETICS

What's next?

- Why do we speak about ‘explanations’?
- Causes and reasons: explanation and justification
- Explanation and understanding
- Proxies in machine learning

What's next?

- Why do we speak about 'explanations'?
- Causes and reasons: explanation and justification
- Explanation and understanding
- Proxies in machine learning

GDPR

- the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject (art. 15.1(h)).

- in a concise, transparent, intelligible and easily accessible form, using clear and plain language (art. 12.1).

GDPR

- Art. 15 concerns post-hoc (local) explanations (a right of the data subject)
- Art. 13-14 concern ex ante explanations (obligations for controllers)

- in a concise, transparent, intelligible and easily accessible form, using clear and plain language (art. 12.1).

GDPR

- such processing should be subject to suitable safeguards, which should include specific information to the data subject and 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. (recital 71)

- the controller should use appropriate mathematical or statistical procedures for the profiling,
- implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised,
- secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and
- that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect.

- (recital 71)

AI Act art. 14.4
Human Oversight

- understand the capacities and limitations of the high-risk AI system
- remain aware of automation bias
- be able to correctly interpret the output
- be able not to use it, overrule or stop the system

What's next?

- Why do we speak about 'explanations'?
- Causes and reasons: explanation and justification
- Explanation and understanding
- Proxies in machine learning

What's next?

- Why do we speak about ‘explanations’?
- Causes and reasons: explanation and justification
- Explanation and understanding
- Proxies in machine learning

‘Methodenstreit’

- Social sciences wanting to be ‘like’ the natural sciences
- Causes: the rise of behaviourism (observable, measurable primitives)
- Reasons: motivation rather than motive, normative (institutional facts)
  - *Logical, deontological, defeasible reasoning*
- Reasons and legal justifications: reasons that constrain decision space of courts
  - *Law attributes legal effect based on stipulated set of conditions*
  - *Them being articulated in natural language they are contestable*
- The explanation of the ADM required by GDPR does not justify the decision
  - *That depends on other domains of law*
  - *But it contributes to contesting the decisions on grounds of fact*

Inversification of proxy real world relations

Behaviourism (Pavlov, Skinner, Watson) underpinning behavioural economics:

- The primitive (principal) is an observable behaviour
- The proxy is a natural language concept (vague, imprecise, ambiguous)
- Cognitive bias distracts from the primitives, need to be removed

Machine learning

- Fairness or justice are impossible concepts: vague, imprecise, ambiguous
- The proxy is a machine readable distribution deemed to be fair or just
- Or fairness/justice are just proxies for a fair distribution in the data?

Inversification of proxy real world relations

Rational choice theory (Coase, Elstar) underpinning neoclassical (neoliberal) economics:

- The primitive (principal) is individual rational choice in the context of game theory
- The proxy is a natural language concept (vague, imprecise, ambiguous)
- Concepts with open texture distract from the primitives, need to disambiguate and discretize

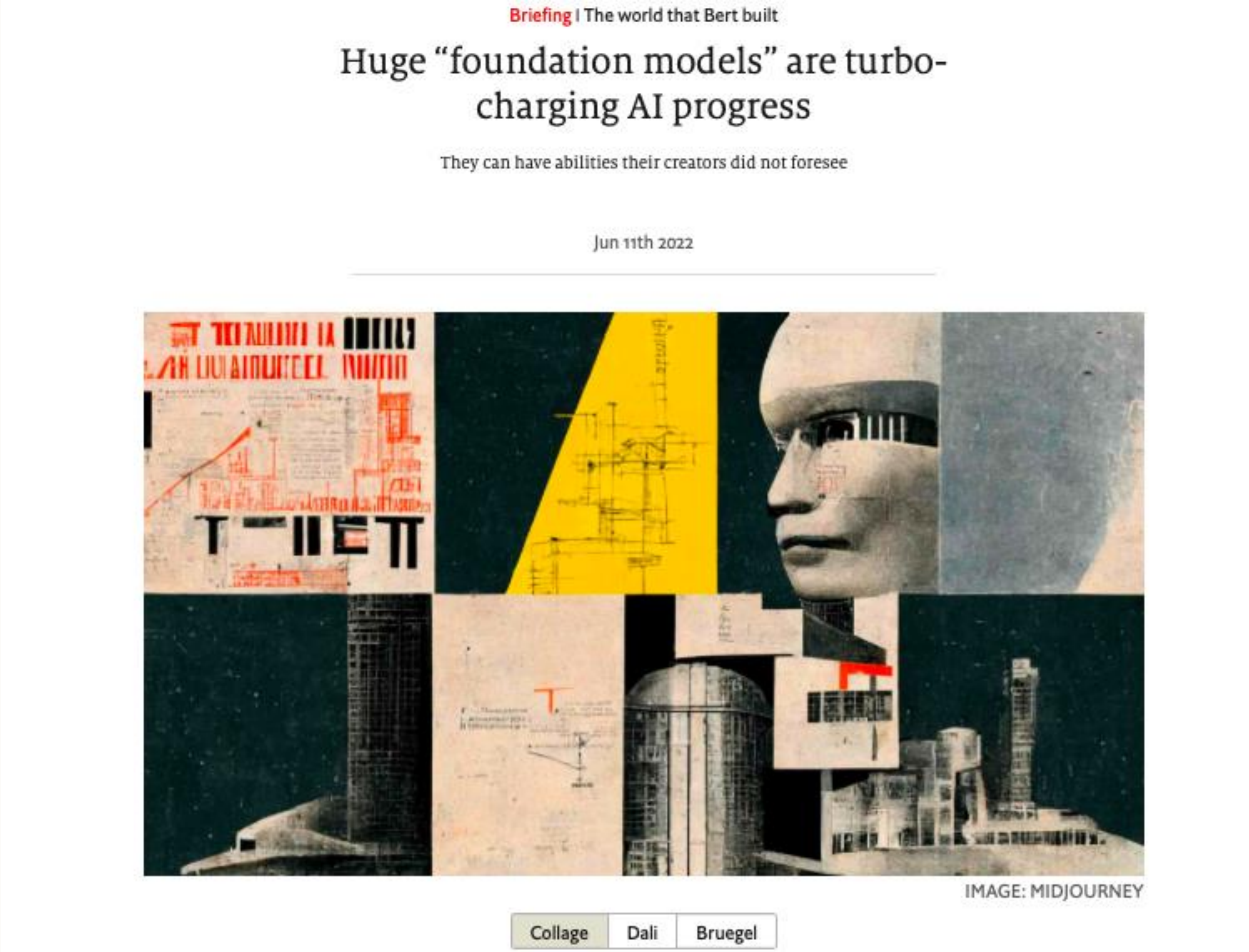
Machine learning

- Fairness or justice are impossible concepts: vague, imprecise, ambiguous
- The proxy is e.g. a multi agent system based on game theoretical assumptions
- Or fairness/justice are just proxies for the outcome of the MAS?

What's next?

- Why do we speak about 'explanations'?
- Causes and reasons: explanation and justification
- Explanation and understanding
- Proxies in machine learning

- Communication is a successful misunderstanding (ZIZEK)



Briefing | The world that Bert built

Huge “foundation models” are turbo-charging AI progress

They can have abilities their creators did not foresee

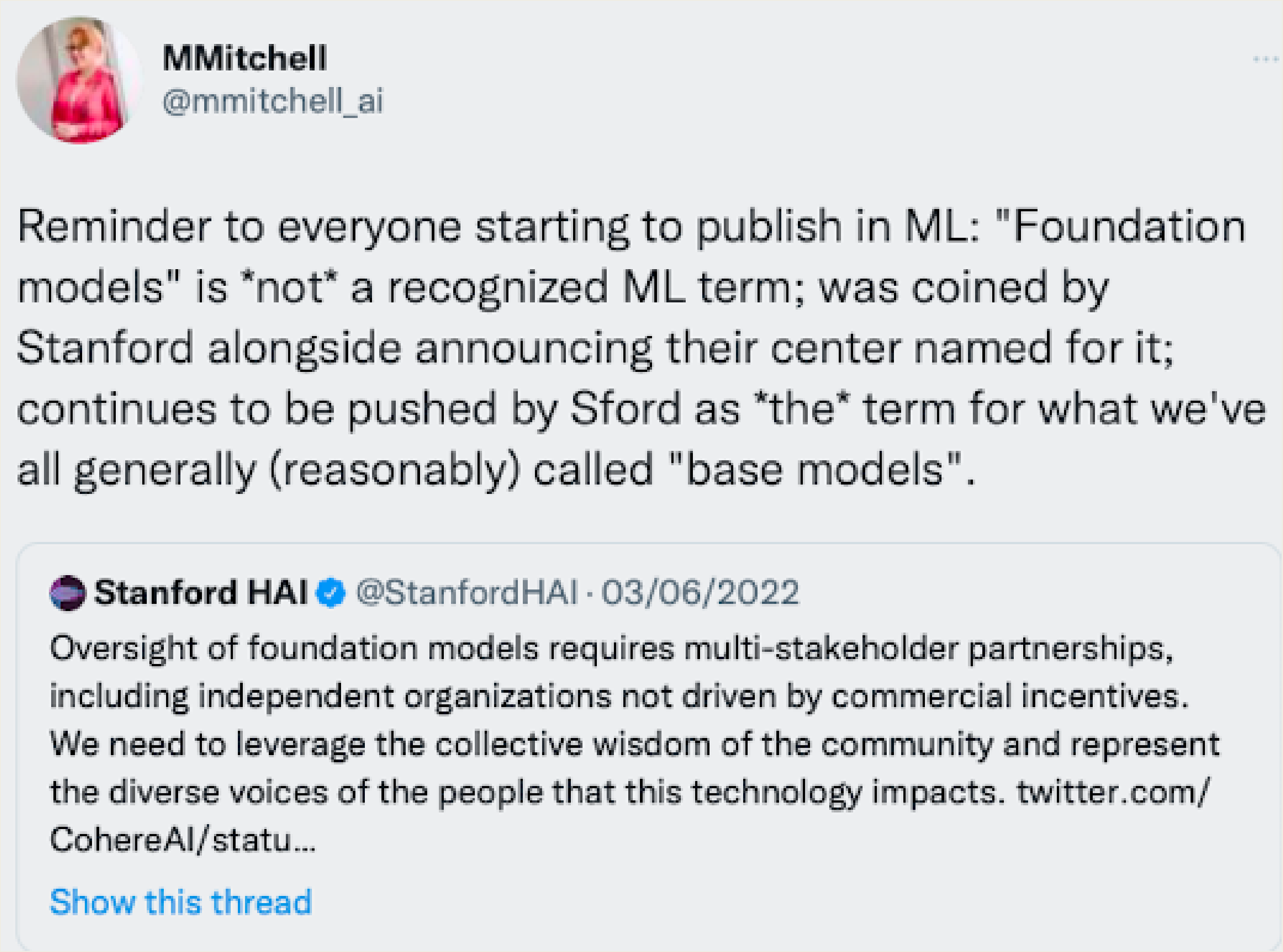
Jun 11th 2022

IMAGE: MIDJOURNEY

Why speak of

foundation model

instead of base model?



MMitchell
@mmitchell_ai

Reminder to everyone starting to publish in ML: "Foundation models" is *not* a recognized ML term; was coined by Stanford alongside announcing their center named for it; continues to be pushed by Sford as *the* term for what we've all generally (reasonably) called "base models".

Stanford HAI @StanfordHAI · 03/06/2022

Oversight of foundation models requires multi-stakeholder partnerships, including independent organizations not driven by commercial incentives. We need to leverage the collective wisdom of the community and represent the diverse voices of the people that this technology impacts. twitter.com/CohereAI/statu...

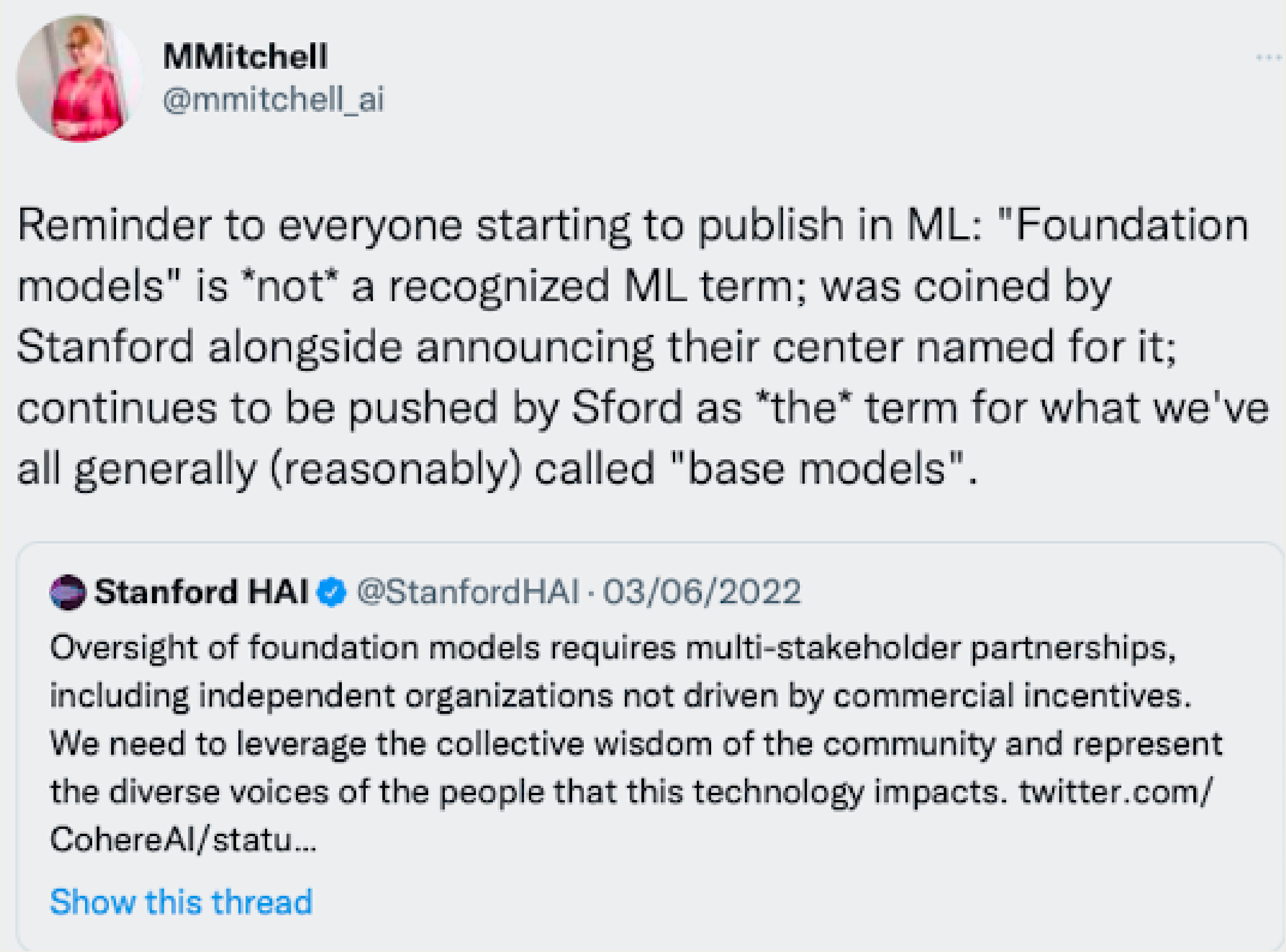
It’s the base

of other models, yes -

but the foundation is

the real world.

Or is it?

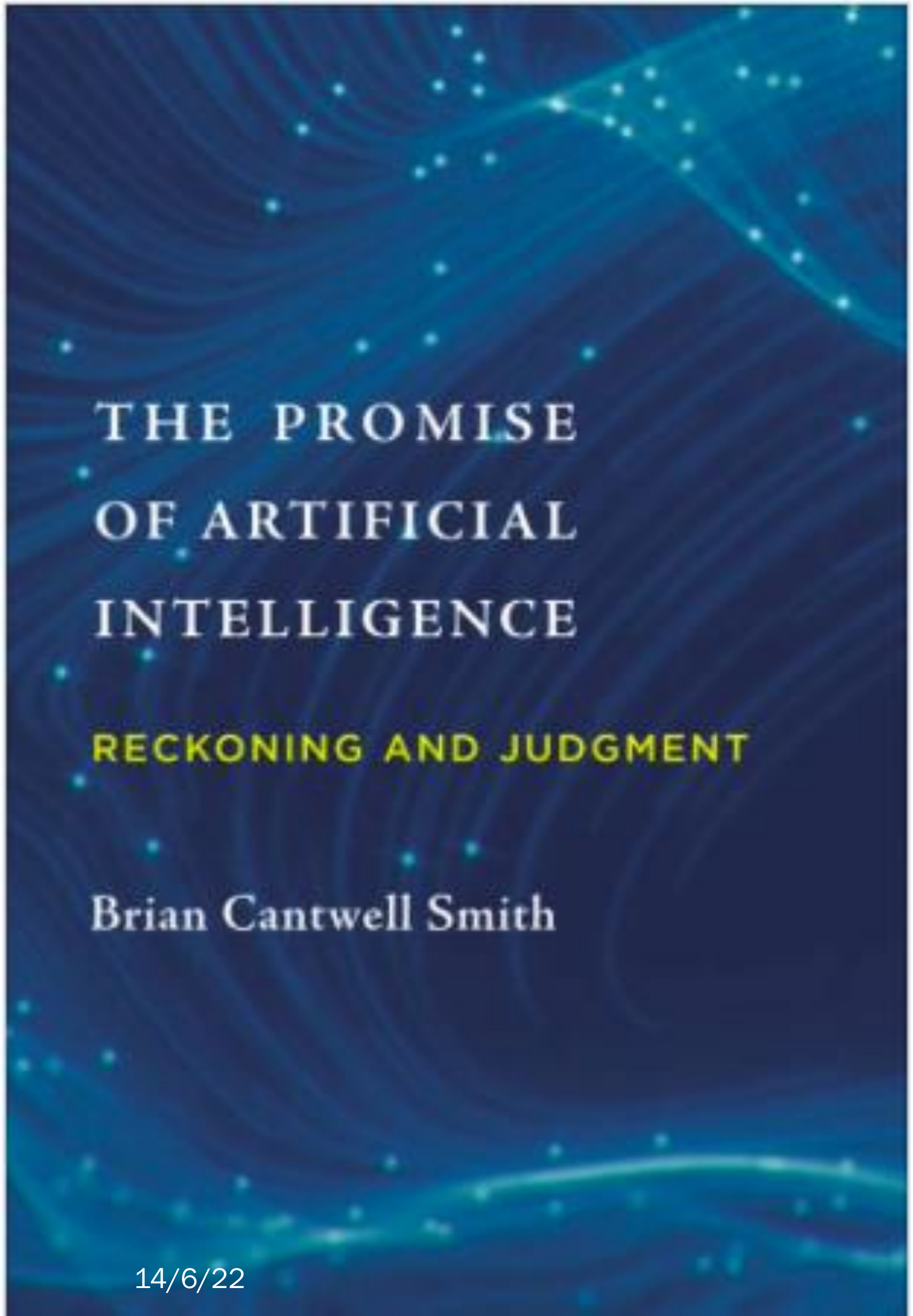


MMitchell
@mmitchell_ai

Reminder to everyone starting to publish in ML: "Foundation models" is *not* a recognized ML term; was coined by Stanford alongside announcing their center named for it; continues to be pushed by Sford as *the* term for what we've all generally (reasonably) called "base models".

Stanford HAI @StanfordHAI · 03/06/2022

Oversight of foundation models requires multi-stakeholder partnerships, including independent organizations not driven by commercial incentives. We need to leverage the collective wisdom of the community and represent the diverse voices of the people that this technology impacts. twitter.com/CohereAI/statu…



THE PROMISE OF ARTIFICIAL INTELLIGENCE

RECKONING AND JUDGMENT

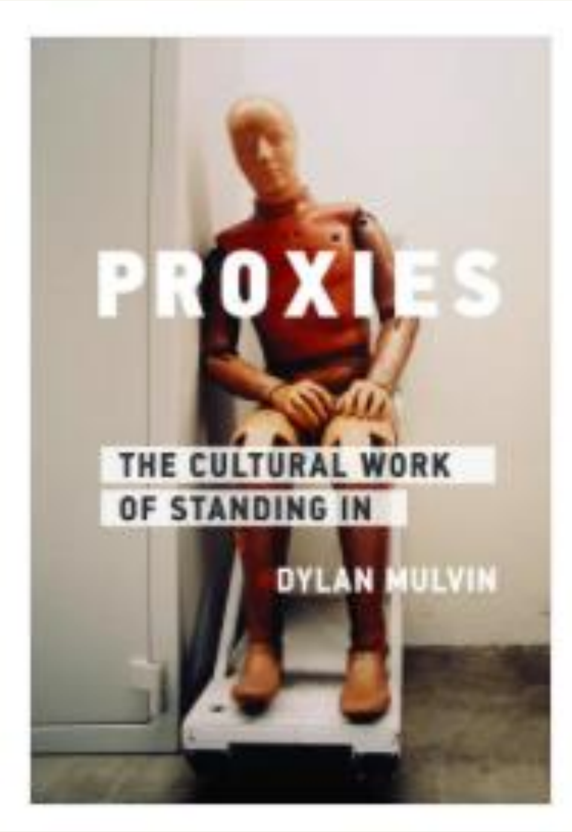
Brian Cantwell Smith

THE DIFFERENCE THAT MAKES A DIFFERENCE

BATESON (1972)

Proxies in data science:

- Data
  - *legal text corpora as a proxy for positive law when training for legal search*
  - *labelled X-rays as a proxy for correct diagnoses when training for medical diagnostics*
  - *www-data as a proxy for 'language acquisition(?)' when training foundation models*
- Variables
  - *income as a proxy for well being or wealth when training for equality*
  - *negative or positive labels as a proxy for emotional engagement when training for sentiment analysis*
- Parameters
  - *weights in an ANN as a proxy for correlations between variables when using backpropagation*
- Tasks
  - *legal text classification as a proxy for ordering relevant documents when training for legal search*
- Mathematical patterns
  - *base models eg BERT, GPT3, DALL-E as proxies for 'language acquisition (?)' when used for further training*



THE STAND IN

Proxies and principal

Meaning (via google, based on https://languages.oup.com/google-dictionary-en/):

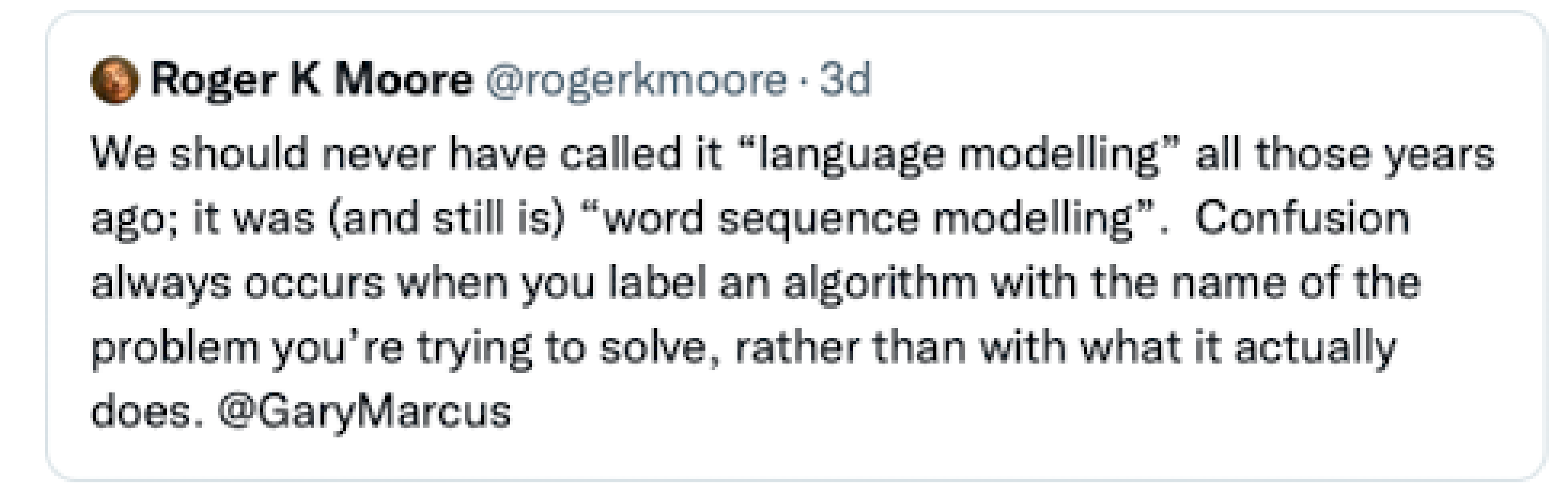
- the authority to represent someone else, especially in voting
- **a figure that can be used to represent the value of something in a calculation**.

Ethymology (https://www.etymonline.com/word/proxy)

- Procuratio (caring for, management, administration)

The issue of proxies

- One thing ’standing in’ for another:
  - *in mathematics numbers don’t necessarily ‘stand in for’ something else*
  - *E.g. –6 – 3 = – 9 (what, apples?), or square root of 2*
  - *In statistics and applied math (social science, computer science):*
    - A variable (x, y, z) stands for a feature/category/type with dedicated values:
      - a symbol (usually a letter) standing in for an unknown numerical value in an equation (https://www.britannica.com/topic/variable-mathematics-and-logic)
      - algebra (functions, equations)
      - imagine how this enabled abstraction



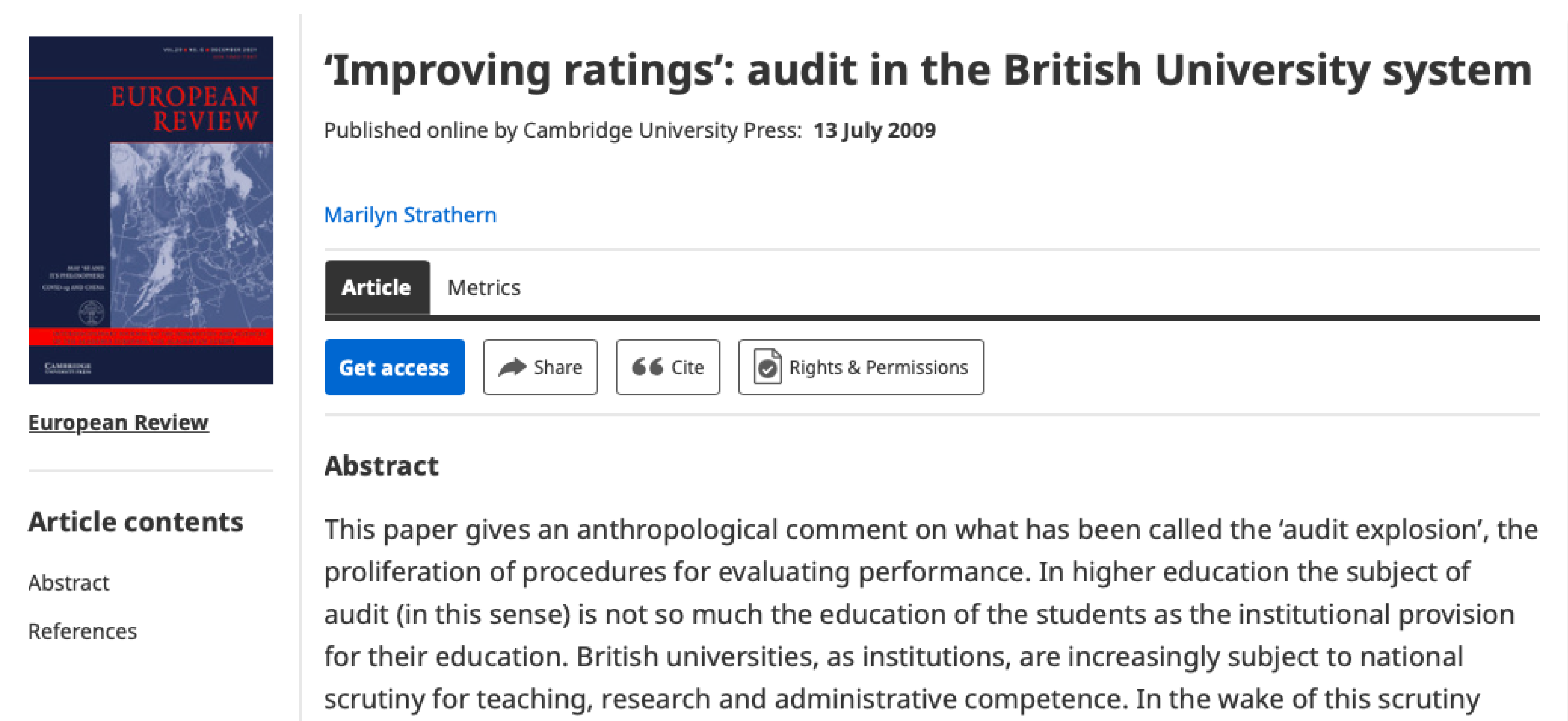
Roger K Moore @rogerkmoore · 3d

We should never have called it “language modelling” all those years ago; it was (and still is) “word sequence modelling”. Confusion always occurs when you label an algorithm with the name of the problem you’re trying to solve, rather than with what it actually does. @GaryMarcus

The issue of proxies

One thing ’standing in’ for another:

- a proxy in algebra and ML serves as the tertium comparationis
- E.g. a variable brings together different things under the same ‘heading’
- quantification is contingent upon prior qualification
- language as word sequencing (on LLMs)
- justice as fairness
  - *Fairness as a specific type of distribution in a dataset (outcome oriented)*
  - *Fairness as being heard and taken into account (process oriented)*
- quality of academic research
  - *Volume of publications in double blind peer reviewed international journals*
  - *Citation score (impact factor)*



Article contents

Abstract

References

'Improving ratings': audit in the British University system

Published online by Cambridge University Press: **13 July 2009**

Marilyn Strathern

Abstract

This paper gives an anthropological comment on what has been called the 'audit explosion', the proliferation of procedures for evaluating performance. In higher education the subject of audit (in this sense) is not so much the education of the students as the institutional provision for their education. British universities, as institutions, are increasingly subject to national scrutiny for teaching, research and administrative competence. In the wake of this scrutiny

- The relevance of the proxy depends on the purpose
- This is even more important in the case of ‘general purpose’ systems
- What is relevant for the myriad downstream purposes?
  - *acknowledging that N=ALL is a hoax*
  - *even all data on the web is not equivalent with ’real life’ or ‘real world’*

End