

Reflections on interdisciplinarity in Responsible ML

A Tale of a Duck, or a Rabbit, and an Elephant

Francien Dechesne
Leiden University

Session agenda

Things I have come to understand traveling across disciplines

- The co-evolutionary character of technology and humanity
 - temporal/dialectic dimension, uncertainty, control?
 - implications for responsible practices
- How different disciplinary lenses are/can/should be involved in “Responsible ML”
 - Reflection on embedded logics in computational sciences (including ML)
 - zooming in on fairness
- Discussion - further reading

Reflection points

- Understand one's positionality in interdisciplinary discussions
- Which responsibilities does that put on
 - your field and its culture (interacting with other fields)
 - yourself as a professional
 - yourself as a human, part of society
- What tools does your discipline give you to take positive responsibility?

Analytical
Abstract
Universal
Stable
Tautological

Trajectory

Normative
Societal
Contextual
Dynamic
Contested

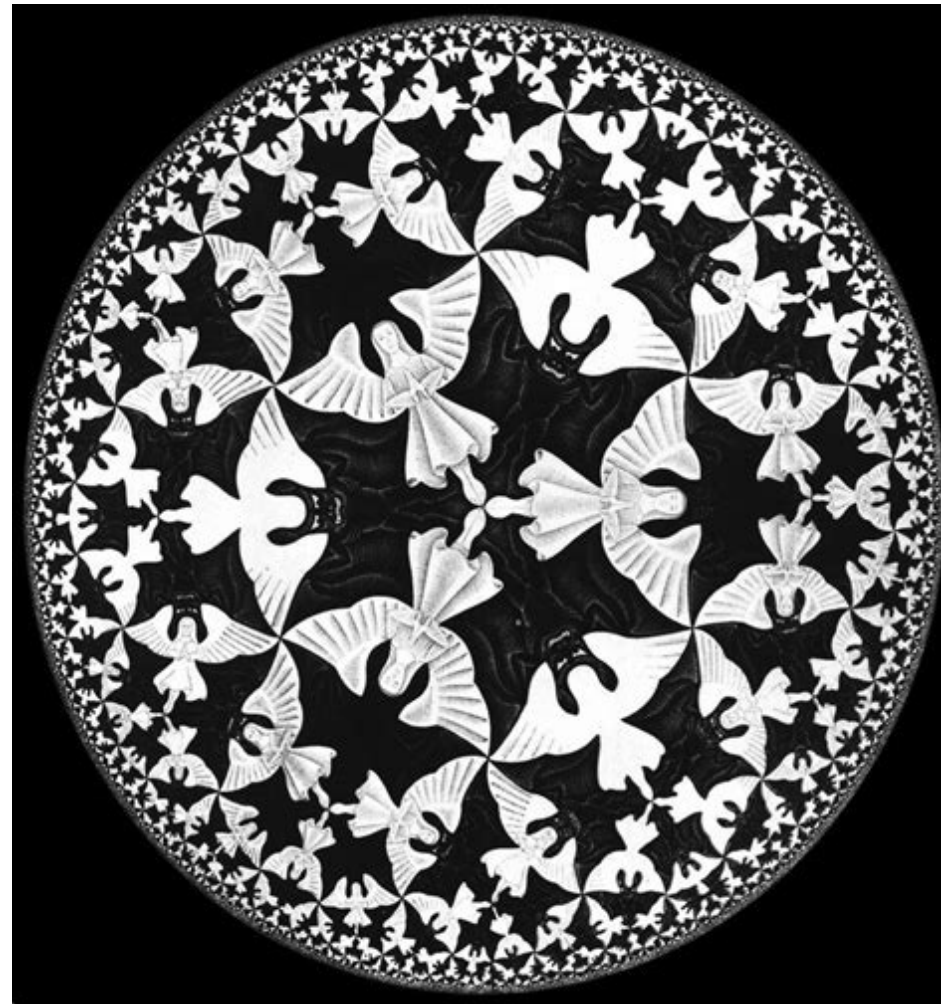
Mathematics

Philosophy

Logic

Computer Science

Formal Verification



Cybersecurity

Privacy

Value Sensitive Design/
Responsible Innovation

Artificial Intelligence
and society

Fairness in Machine Learning

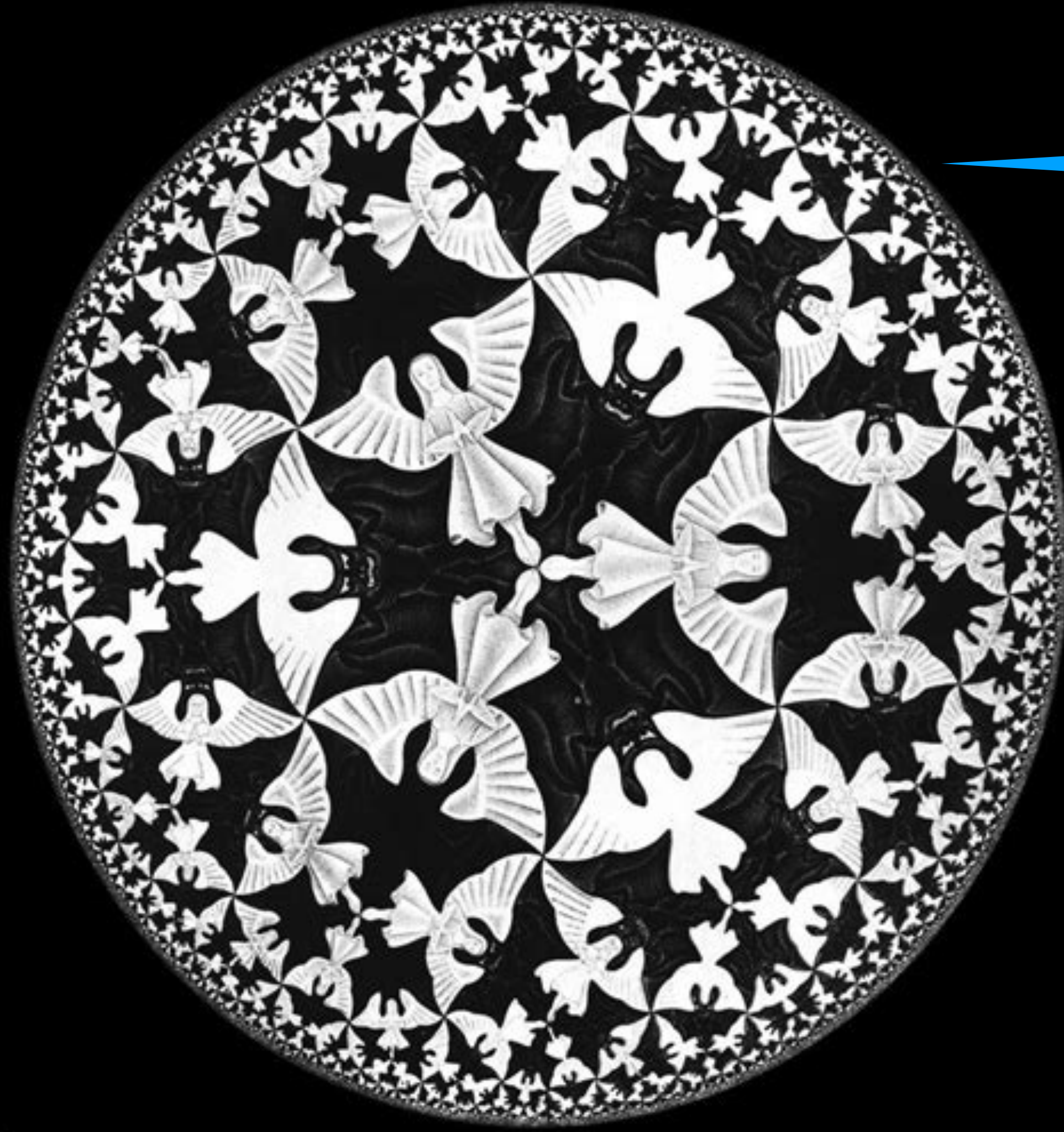


SCALES
Designing regulatory and institutional framework to balance public interest and individual liberties in the use of data analytics



4





ETHICS?

Law?



NIAS

Lorentz
center

Workshop @Snellius

Intersectionality and Algorithmic Discrimination

18 - 22 December 2017, Leiden, the Netherlands

Scientific Organizers

- Francien Dechesne, Leiden U
- Jenneke Evers, Leiden U
- Seda Gürses, KU Leuven

Scientific Committee

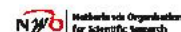
- Solon Barocas, Cornell U
- Sorelle Friedler, Haverford College
- Anna Lauren Hoffmann, U Washington
- Linnet Taylor, Tilburg U

The Lorentz Center organizes international workshops for researchers in all scientific disciplines. Its aim is to create an atmosphere that fosters collaborative work, discussion and interactions. For registration see: www.lorentzcenter.nl

This workshop is a part of the NIAS-Lorentz Program, to stimulate research bridging the natural sciences with the humanities and social sciences.

Poster design: SuperNova Studios .NL

FATML/FAT*/FAccT

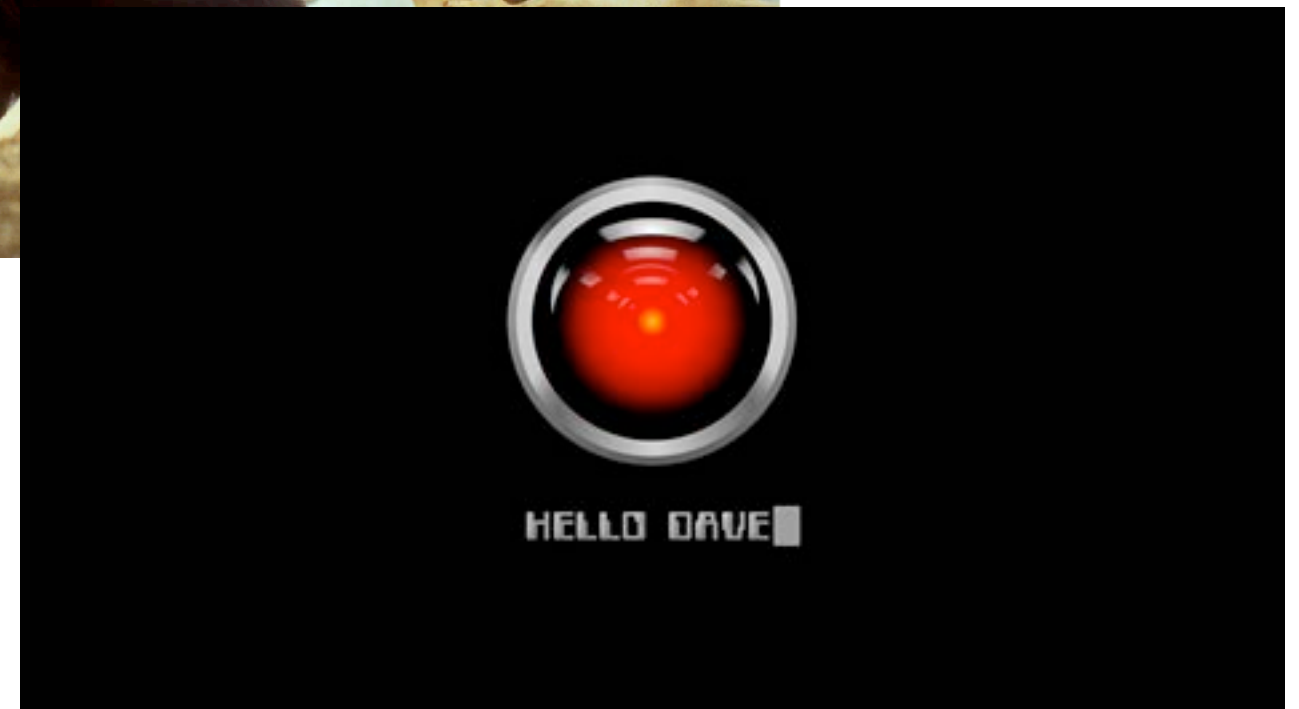
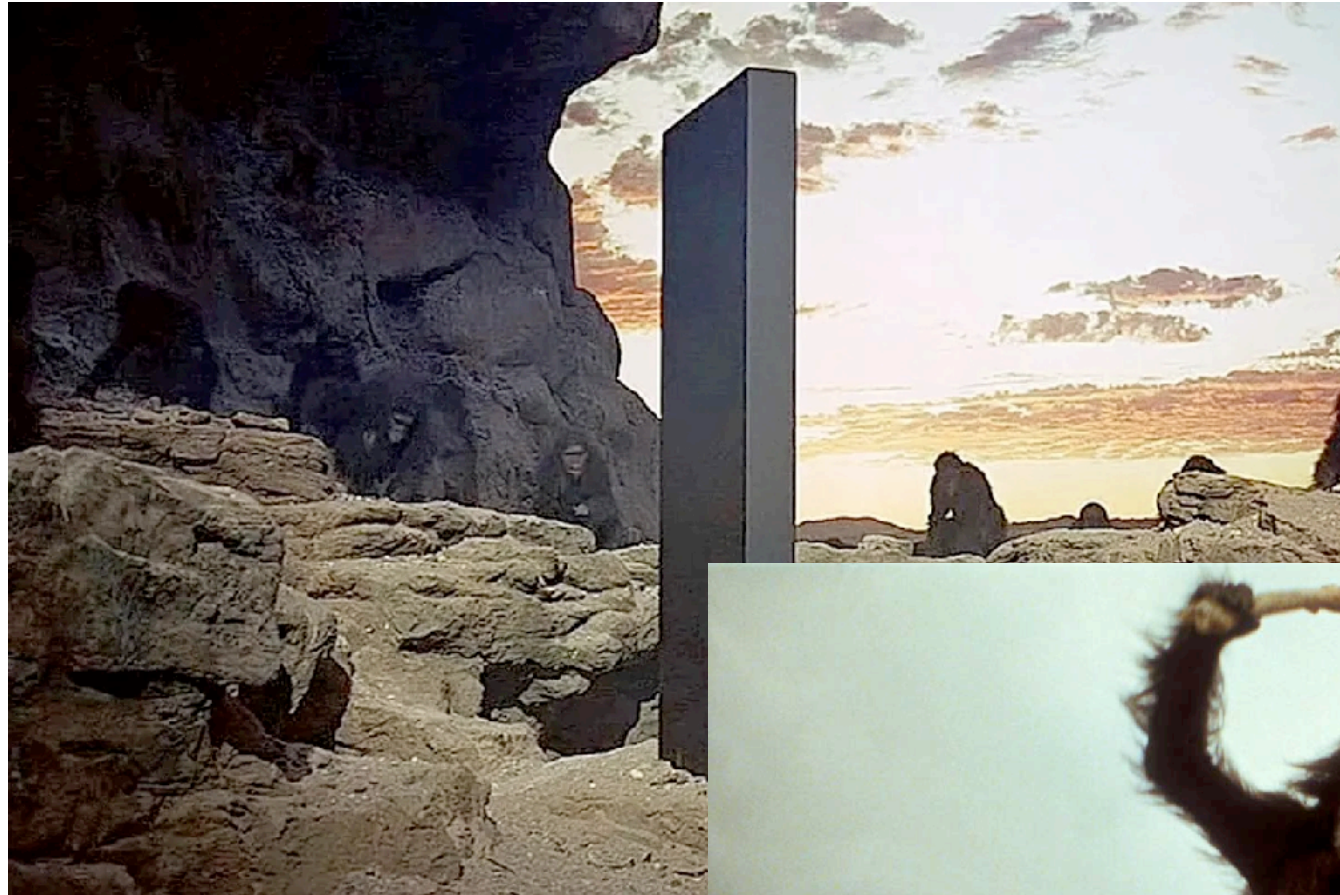


www.lorentzcenter.nl

Solving traffic safety?



Archaeological perspectives on AI!




Material Agency: Co-evolution of technology (infrastructure) and human behavior and norms



Technology is neither good nor bad
Nor is it neutral
(Melvin Kranzberg, 1985)

Middle ground between technological
determinism vs social constructionism

Temporal and epistemic dimensions of Responsibility

- Co-evolution/dialectic development:
 - human behavior/understanding  technological affordances
 - “new technologies as social experiments” [vdPoel 2016]
- Responsibility: positive [for doing good]/negative [for causing harm]
- Uncertainty vs Responsibility: can we **know** risks and benefits?
 - Emergent (i.e. non-deterministic?)
 - What is considered acceptable/to be prevented will co-evolve
 - Collingridge/**Control** Dilemma (1980) - [vdPoel2016]
 - precautionary principle - focuses on negative responsibility
 - positive approach: “responsible experimentation” (procedural)

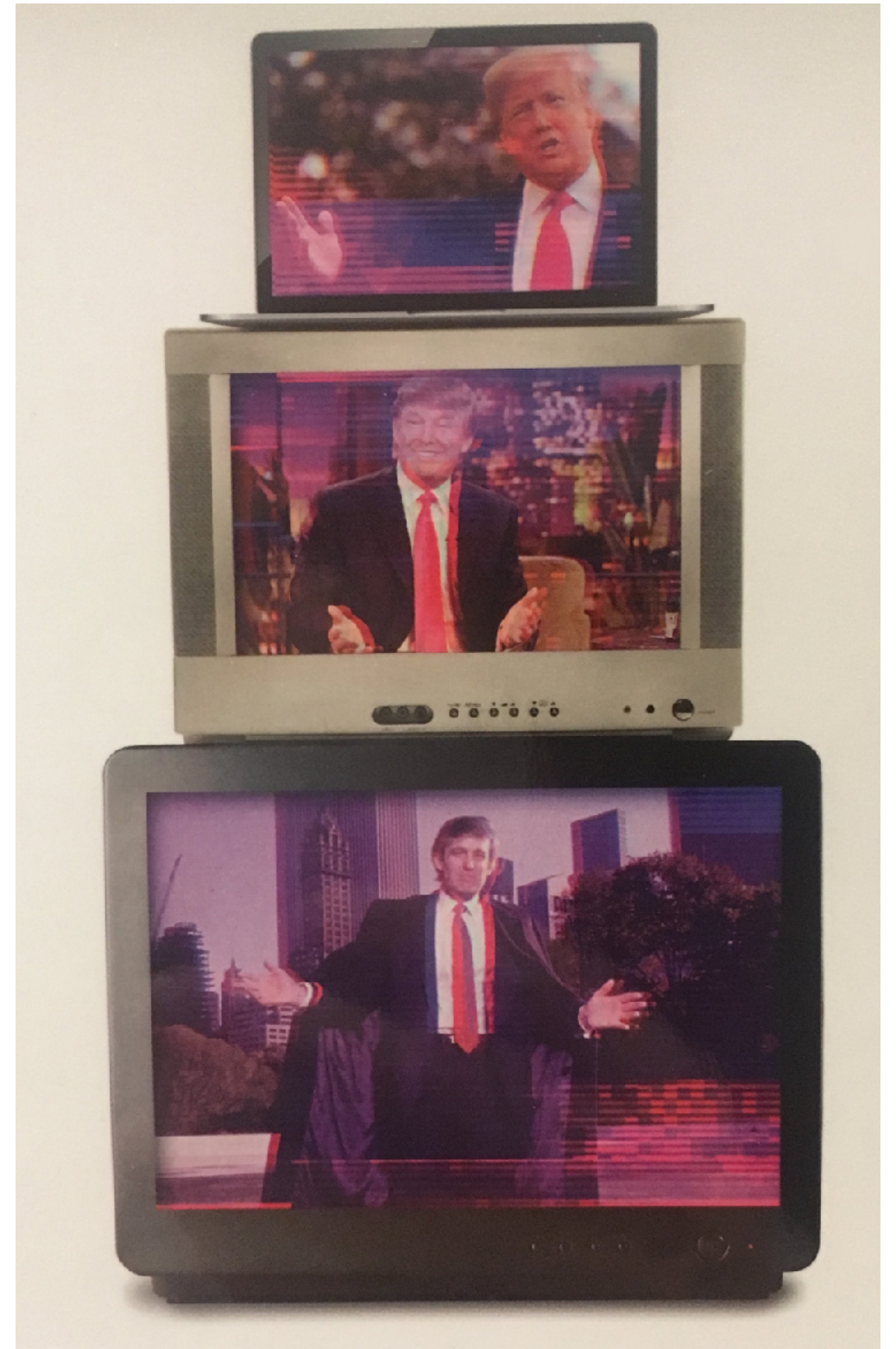
These are reasons why law/ethics/humanities cannot provide precise specifications of moral requirements you may want to implement

Neil Postman (1995)

Five Things We Need to Know About Technological Change

... and that you don't learn how to think or talk about in CS...

1. trade-offs
2. winners and losers
3. codification of the world (hammer -> nail)
4. Change is not additive, but ecological
5. easily viewed as mythic, "God-given" instead of constructed

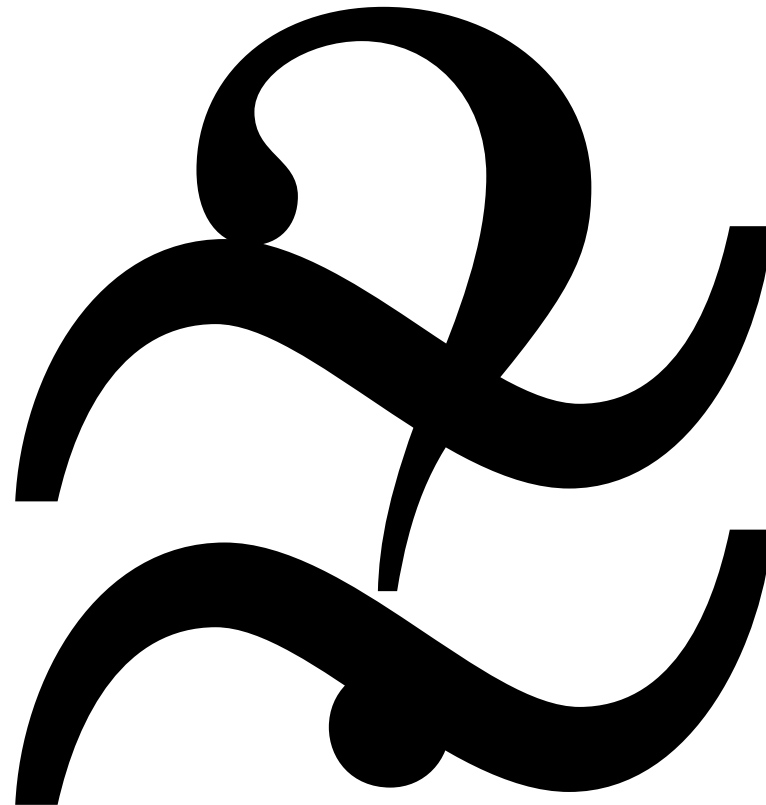


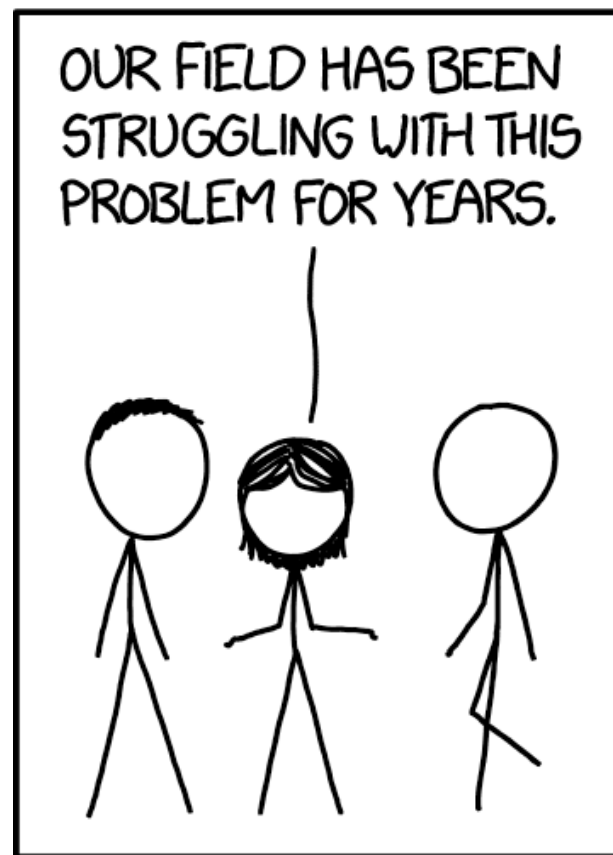
**Interactions about
fairness in ML?**

What is fairness?



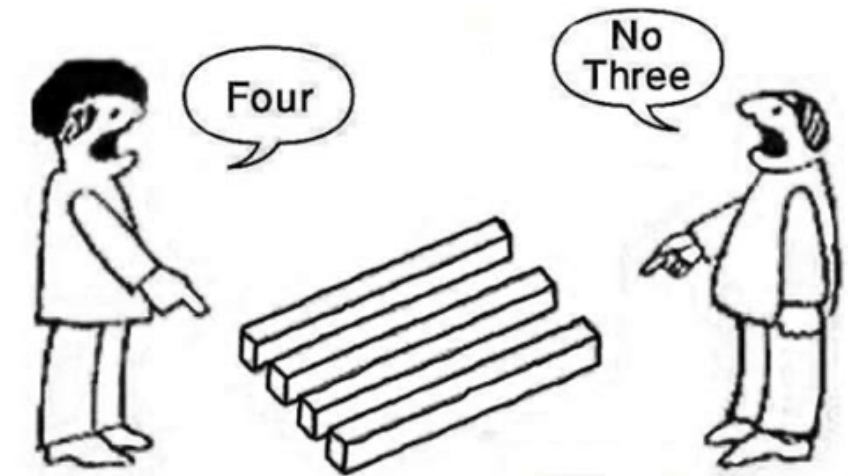
What is fairness?



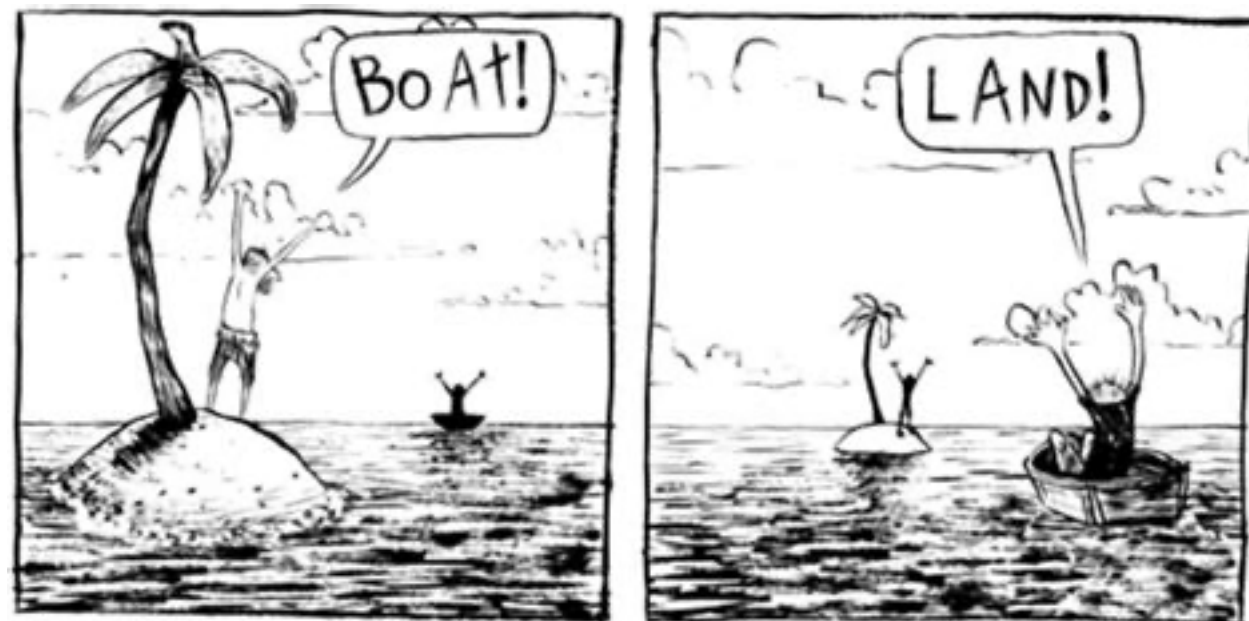




(mediated) framing



perspective

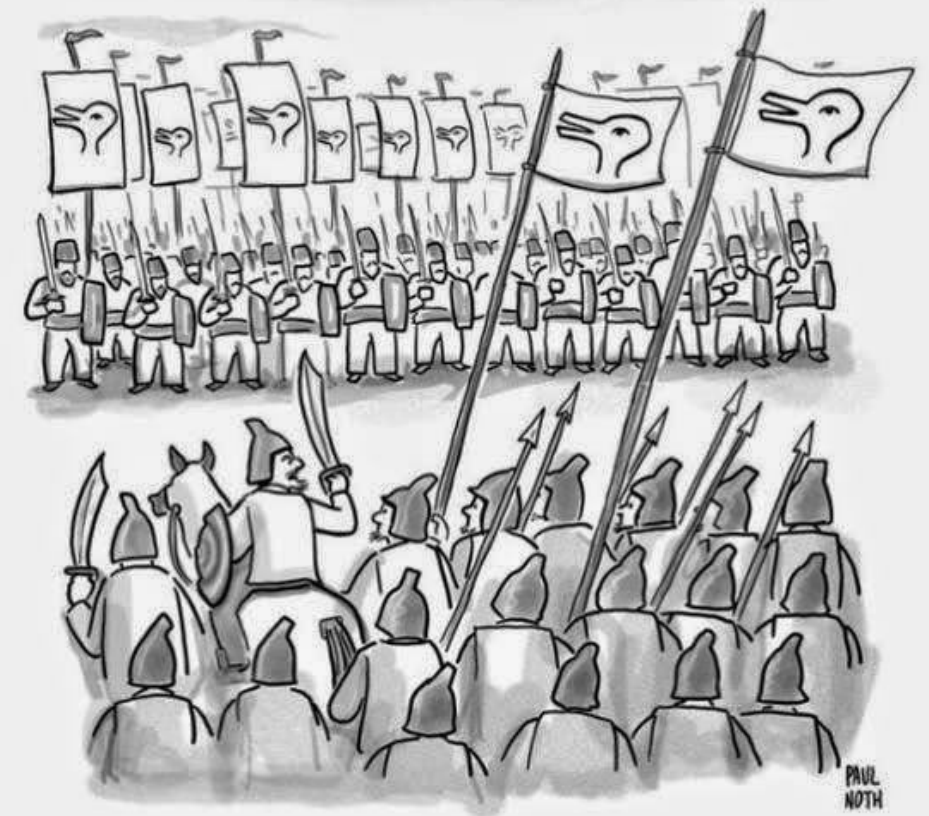
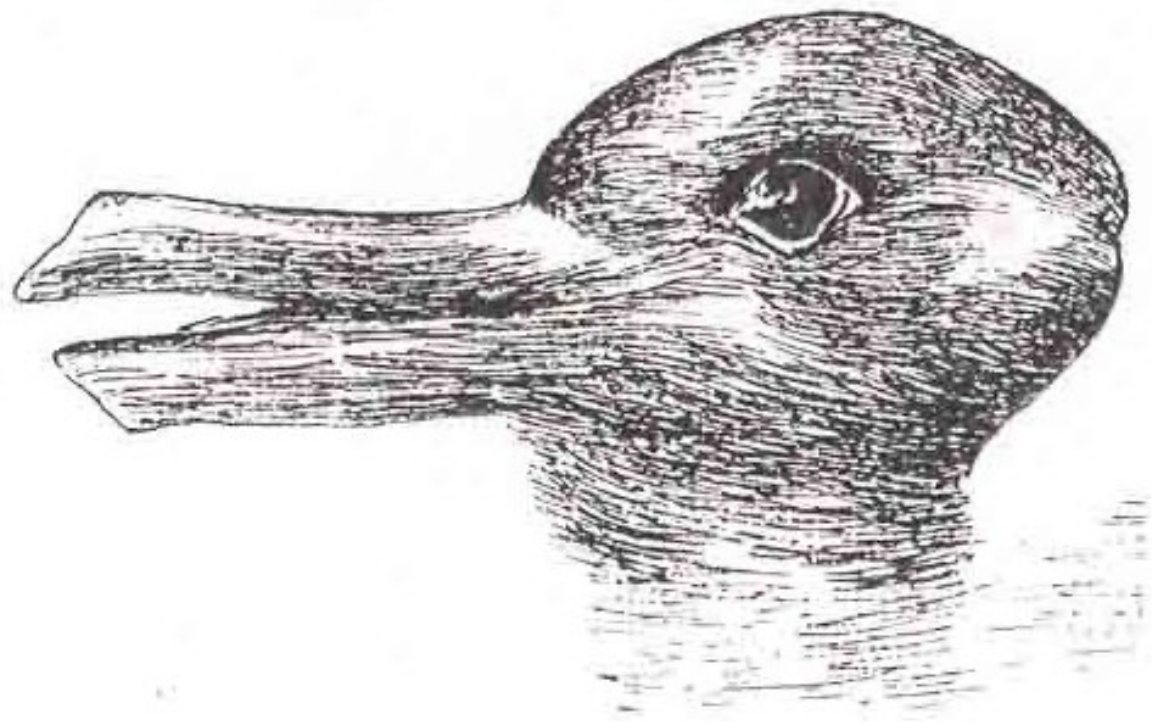


lived experiences (needs)

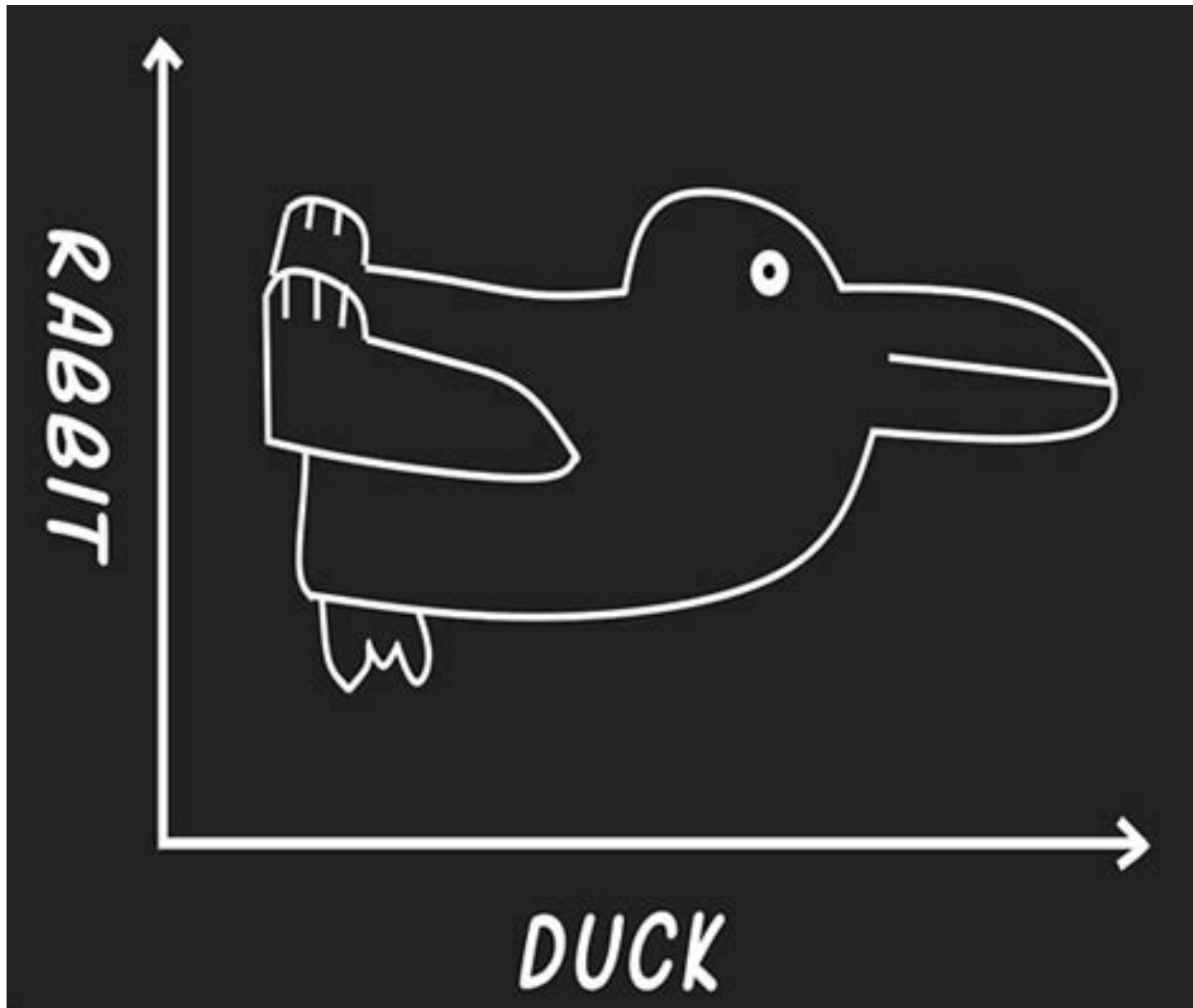


positionality

But also: interdisciplinary ambiguity



*"There can be no peace until they
renounce their Rabbit God and
accept our Duck God."*



6 CONCLUSION

In this paper, we focused on the uncertain process by which certain questions come to be posed in real-world applied data science projects. We have shown that some of the most important normative implications of data science systems find their roots in the work of problem formulation. The attempt to make certain goals amenable to data science will always involve subtle transformations of those objectives along the way—transformations that may have profound consequences for the very conception of the problem to which data science has been brought to bear—and what consequently appear to be the most appropriate ways of handling those problems. Thus, the problems we solve with data science are never insulated from the larger process of getting data science to return actionable results. As we have shown, these ends are very much an artifact of a contingent process of arriving at a successful formulation of the problem, and they cannot be easily decoupled from the process at arriving at these ends. In linking the normative concerns that data science has provoked to more nuanced accounts of the on-the-ground process of undertaking a data science project, we have suggested new objects for investigation and intervention: *which goals are posed and why; how goals are made into tractable questions and working problems; and, how and why certain problem formulations succeed.*

Passi, S., & Barocas, S. (2019). **Problem Formulation and Fairness**. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). ACM, New York, NY, USA, 39-48.

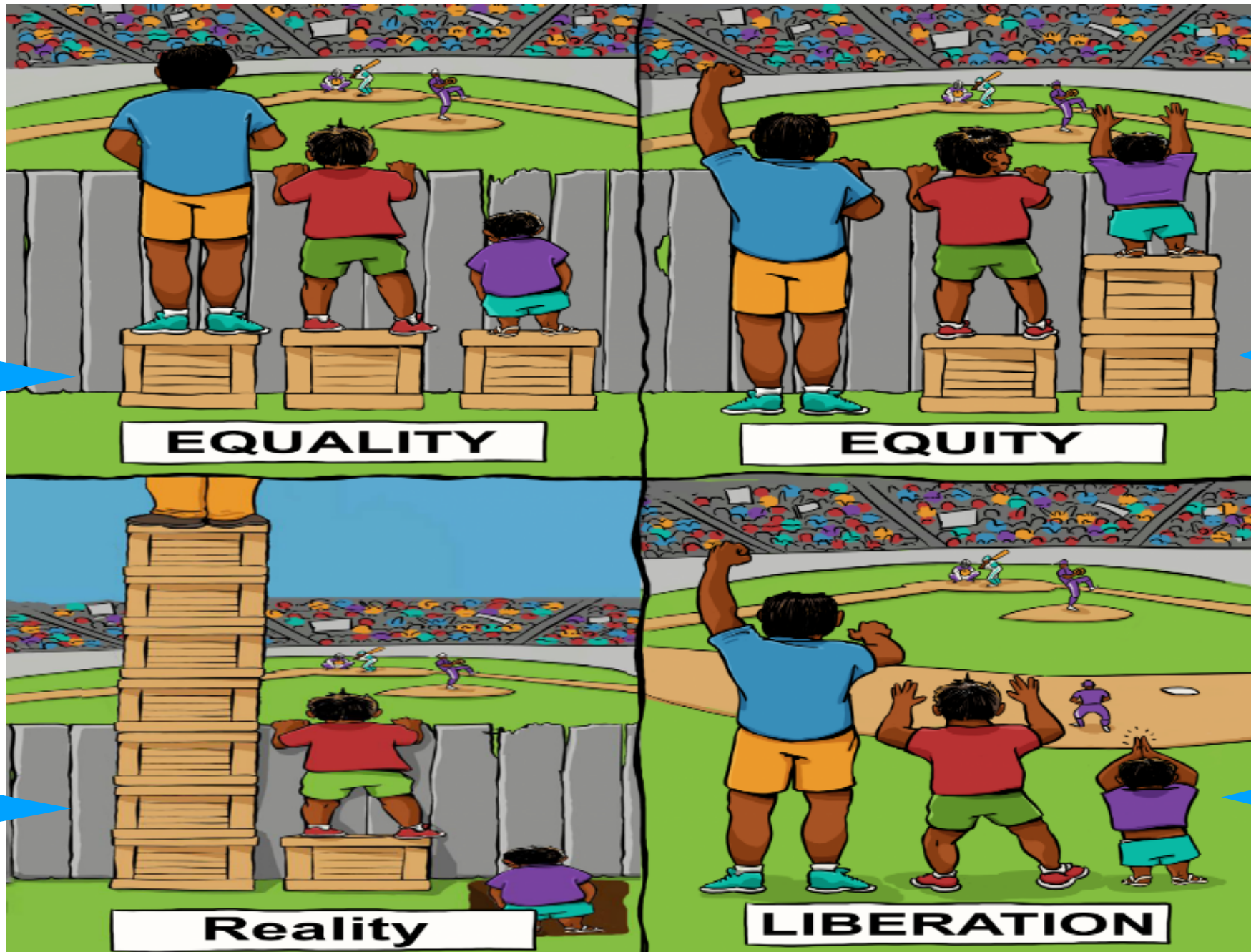
What is the problem?

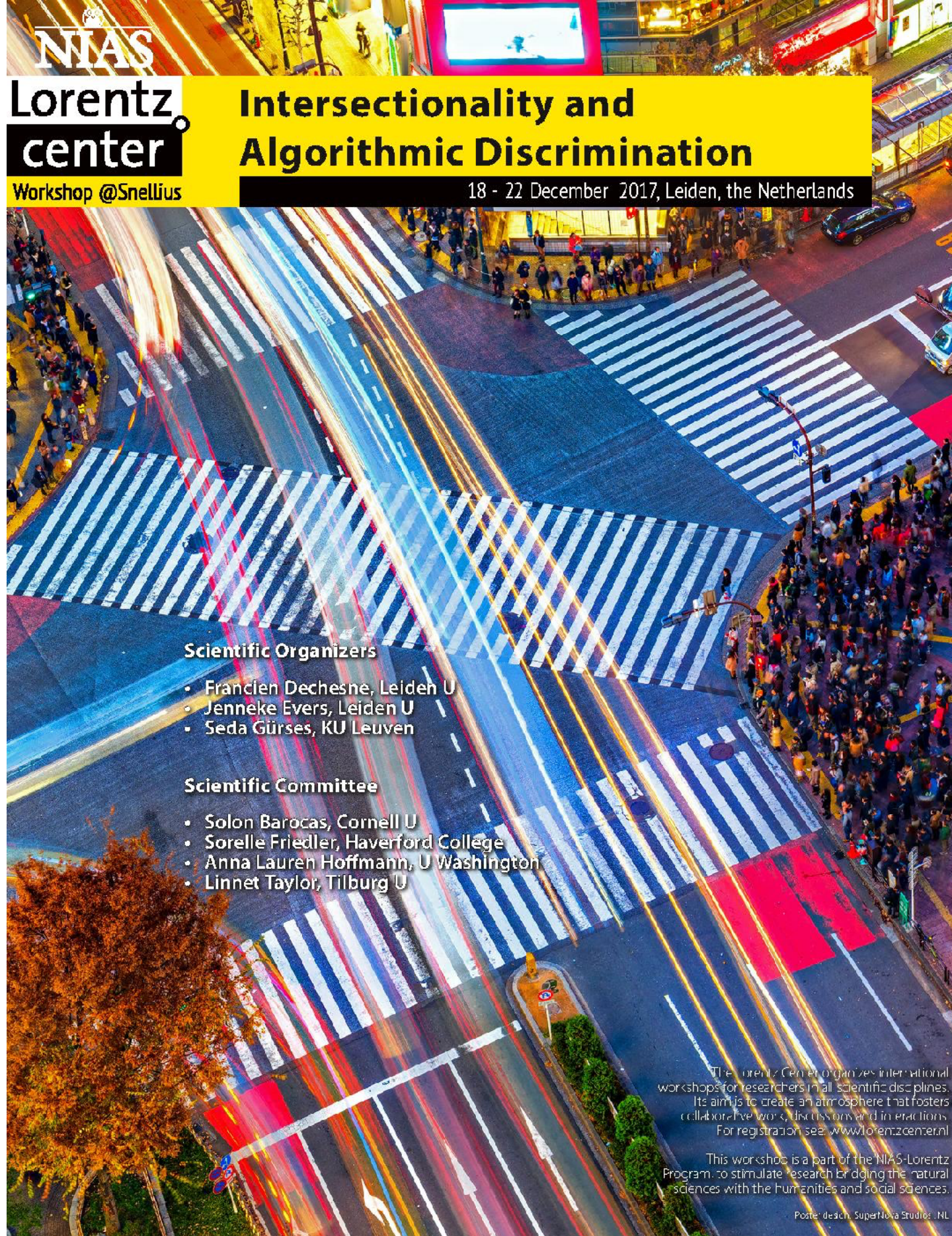
equal
distribution
of resources

equal
outcome

reparation of
historical
inequality?

making
inequalities
irrelevant





NIAS
Lorentz center
 Workshop @Snellius

Intersectionality and Algorithmic Discrimination

18 - 22 December 2017, Leiden, the Netherlands

Scientific Organizers

- Francien Dechesne, Leiden U
- Jenneke Evers, Leiden U
- Seda Gürses, KU Leuven

Scientific Committee

- Solon Barocas, Cornell U
- Sorelle Friedler, Haverford College
- Anna Lauren Hoffmann, U Washington
- Linnet Taylor, Tilburg U

The Lorentz Center organizes international workshops for researchers in all scientific disciplines. Its aim is to create an atmosphere that fosters collaborative work, discussion and interactions. For registration see: www.lorentzcenter.nl

This workshop is a part of the NIAS-Lorentz Program, to stimulate research bridging the natural sciences with the humanities and social sciences.

Poster design: SuperNova Studios .NL



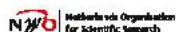
-moment

Reading technical paper together

[Preventing Fairness
 Gerrymandering: Auditing and
 Learning for Subgroup Fairness -
 Michael Kearns et al., PMLR
 80:2564-2572, 2018.]



DEFINITIONS



www.lorentzcenter.nl

"21 definitions of fairness"

Lessons from recidivism scoring

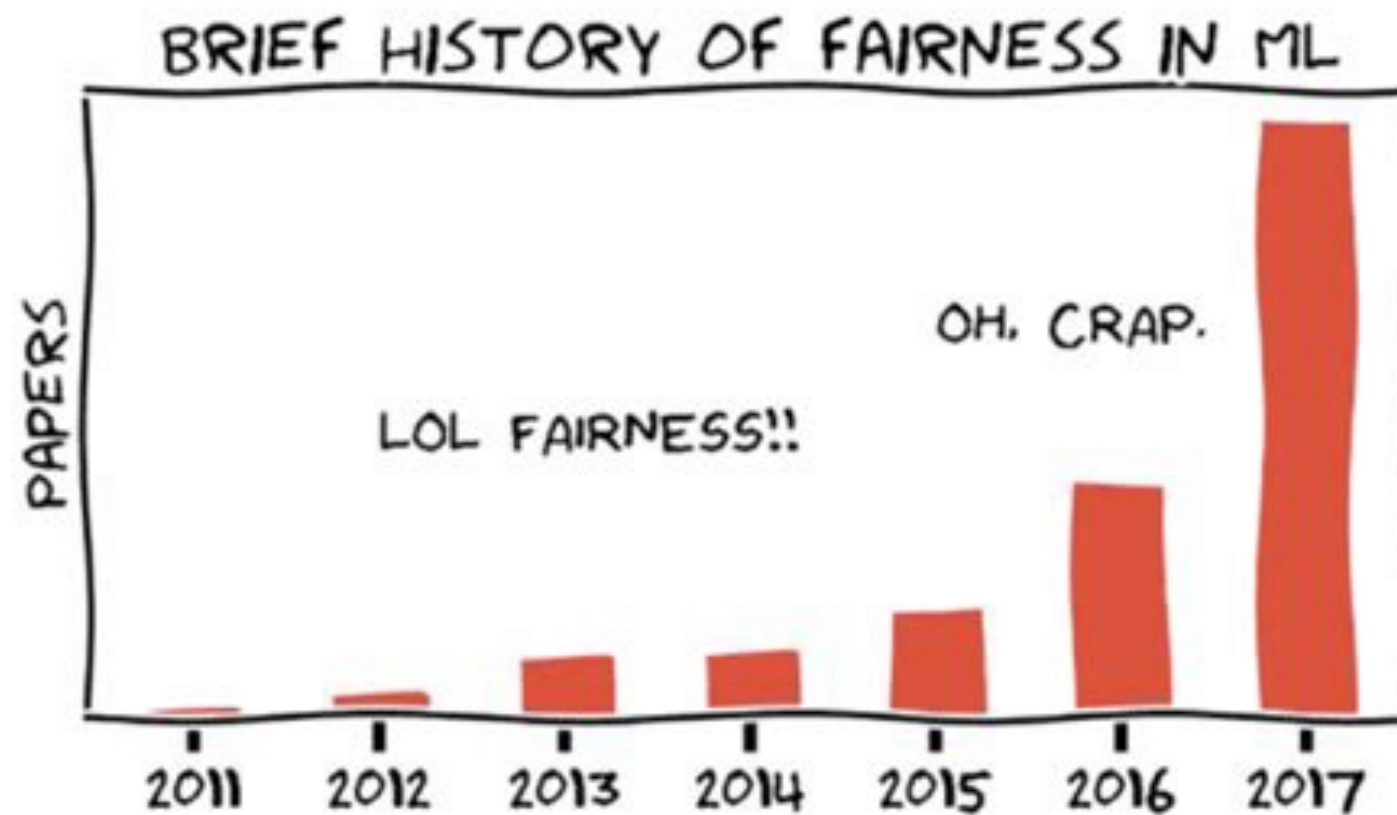


Fig1. The number of publications on fairness from 2011 to 2017

Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong. 2019. This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. Proc. ACM Hum.-Comput. Interact.

Positive responsibility: contributing to better understanding of the issues of the system

- **No direct discrimination**

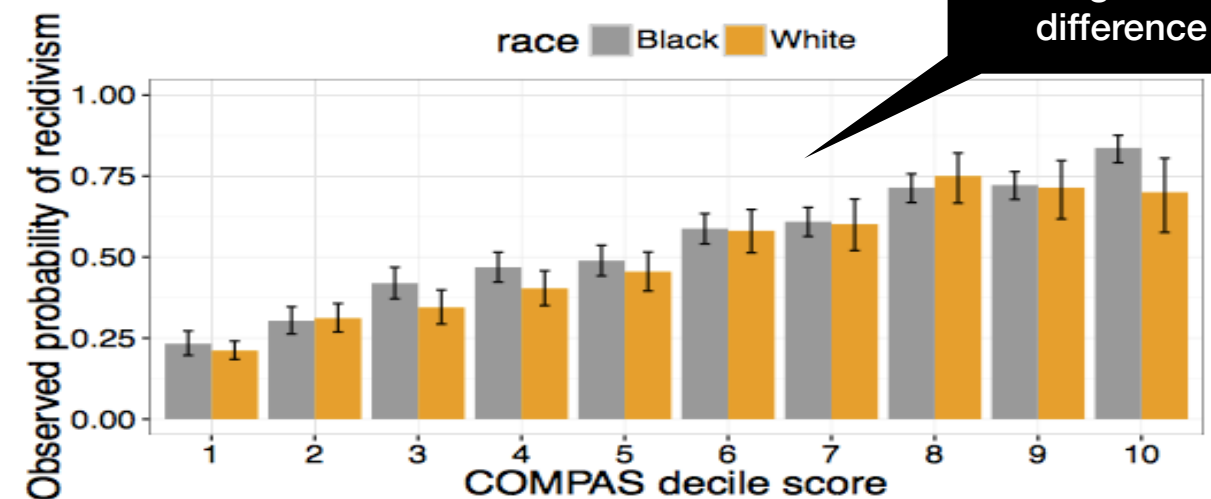
- race was not a parameter
- distribution of scores is similar
- (ZIP-code, or arrested family members could be a proxy)

- Problem was in **distribution of the errors**: false negatives (whites) vs false positives (blacks) for recidivism risk

- Impactful at low level of accuracy (only valid about 62% [Slave to the Algorithm])

- Causes related to human bias captured in data (a.o.):

- unreasonable weights to one field in questionnaire
- using re-arrest data as proxy for re-offence
 - training data from district in Florida with known disparity in arresting blacks vs whites



		Predicted Label		
		$\hat{y} = 1$	$\hat{y} = -1$	
True Label	$y = 1$	True positive	False negative	$P(\hat{y} \neq y y = 1)$ False Negative Rate
	$y = -1$	False positive	True negative	$P(\hat{y} \neq y y = -1)$ False Positive Rate
		$P(\hat{y} \neq y \hat{y} = 1)$ False Discovery Rate	$P(\hat{y} \neq y \hat{y} = -1)$ False Omission Rate	$P(\hat{y} \neq y)$ Overall Misclass. Rate

black overrepresented

white overrepresented

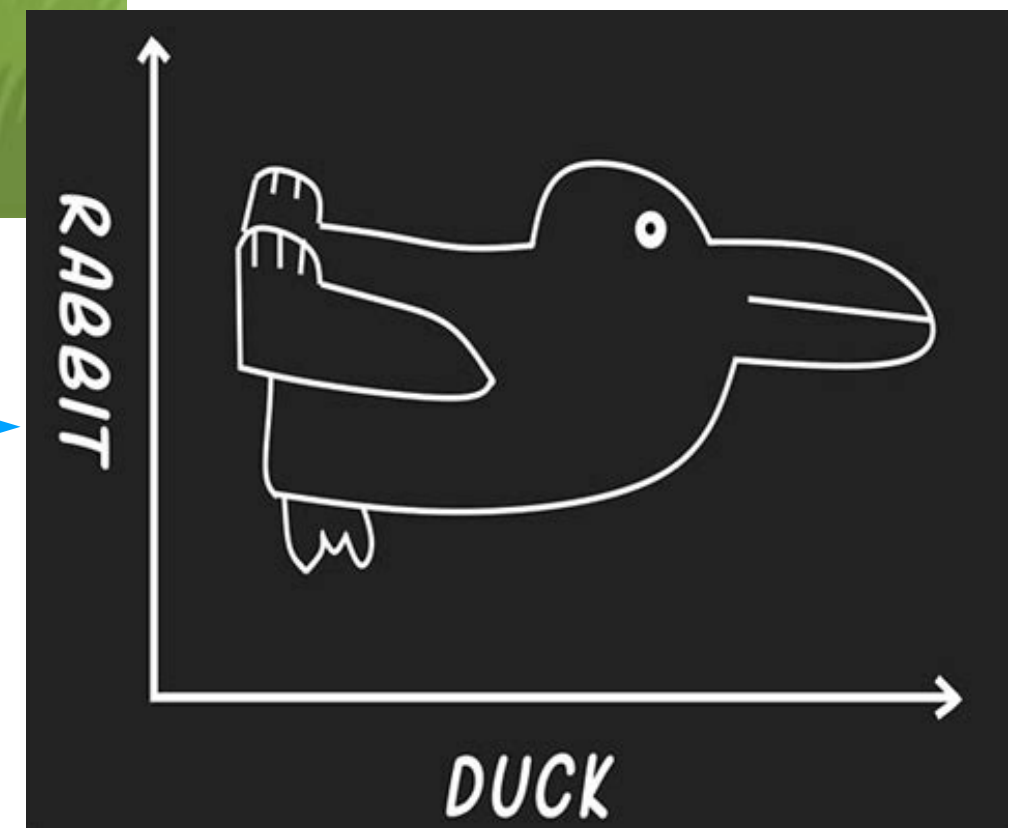
Yes, we need more measures for fairness, but...



each measure describes a particular aspect - that may or may not be the relevant one for the context of application

Also depends on how the moral problem is framed

- available data and techniques
- moral norms are often inherently contested, or open enough to allow for change <> measure
- so cannot be a pure ML question to solve



Solving traffic safety?

The right
measures?

Roof strength test

moderate overlap front crash test

Side crash test

Small overlap front crash test

Front-impact
airbags

Overhead airbags

Anti-lock brakes

Side impact
airbags

Pretensioners

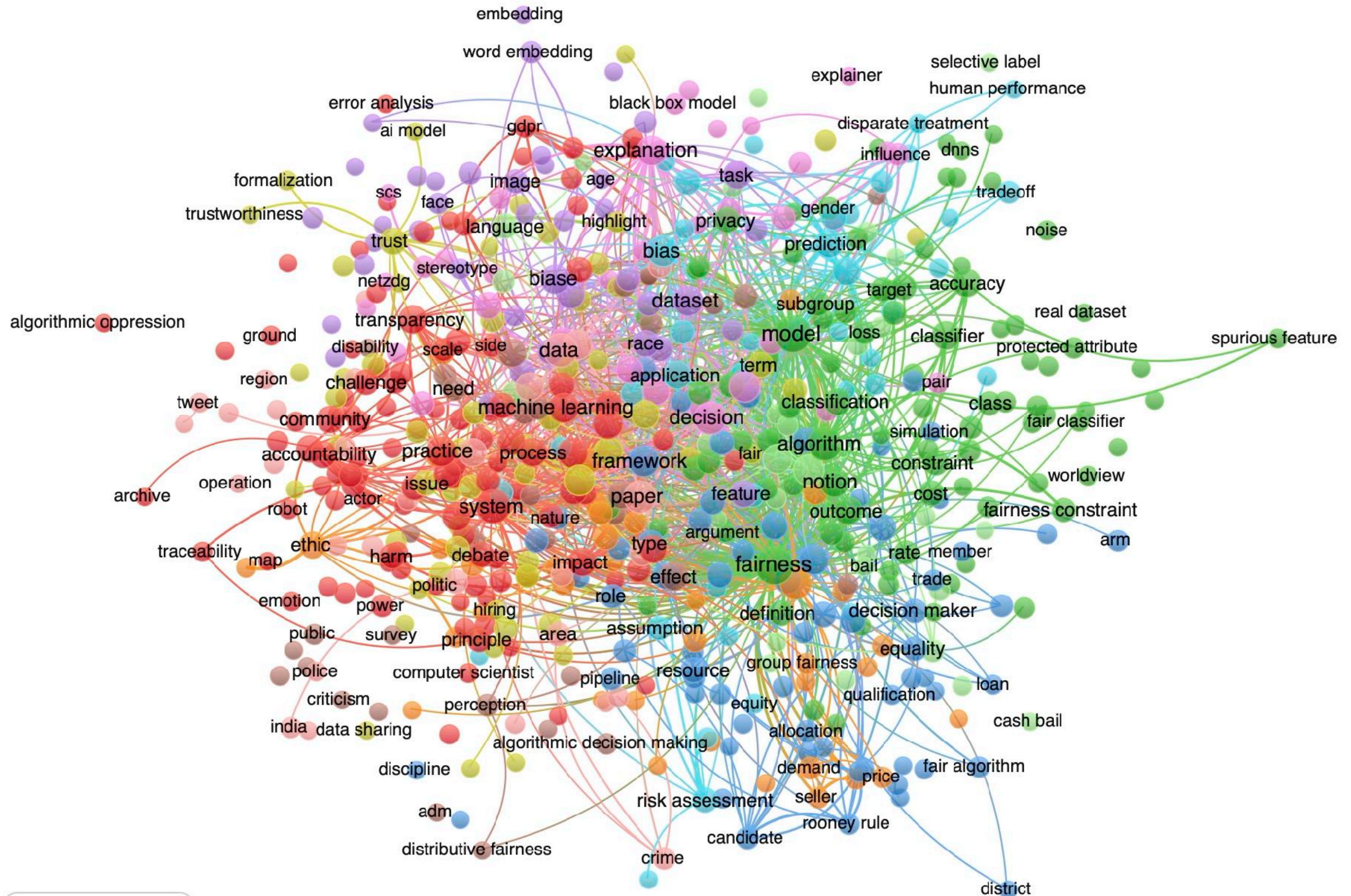
Stability control

Paint

ROVÉLO CREATIVE



David Moats et al, to appear [*2]: facilitating interdisciplinary conversations about AI Ethics terms using scientometric visualisations.



Disciplinary confusions around concepts like “fairness”?

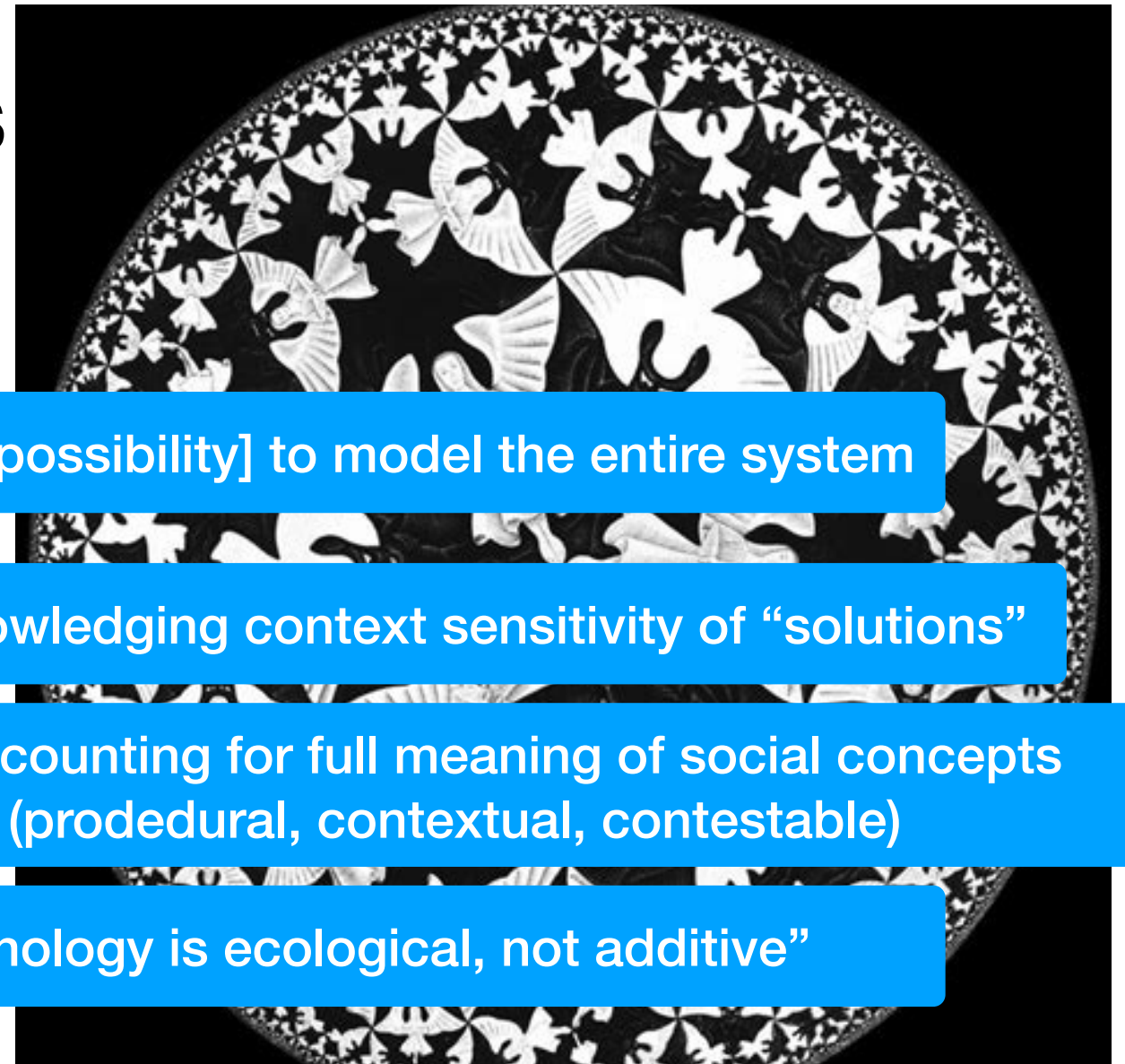
- Relevant perspectives:
 - philosophy/ethics: systematic frameworks for reasoning what is problematic, what should be the case and why
 - legal: substantive (discrimination) vs procedural (accountability, contestability)
 - history and political science: understanding power relations, problematic inequalities
 - social sciences: descriptive (what do people consider fair)
 - computational: operationalisations (fair division, voting, fairML)
- Different directions of abstraction of the concept [attempt]:
 - Philosophy/Ethics: towards capturing **core substance** (or absence thereof)
 - CS: away from substance to **structural specification** (context independent) [to allow for syntactic processing](#)
 - Law/Ethics: to **avoid specification** of substance [to allow for context-dependent interpretation](#) [=> rules often focused on procedure]
 - [Social science: heuristic starting point for empirical substantiation]

Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong. 2019. This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. Proc. ACM Hum.-Comput. Interact.

Reflecting on implicit logics of the computational sciences

Fairness and abstraction in **socio**technical systems

Selbst, boyd, Friedler, Venkatasubramanian, Vertesi
(FAT* 2019)



- Framing trap failure [impossibility] to model the entire system
- Portability trap not acknowledging context sensitivity of “solutions”
- Formalism trap not accounting for full meaning of social concepts (procedural, contextual, contestable)
- Ripple effect trap “technology is ecological, not additive”
- Solutionism trap failure to acknowledge best solution may not involve technology

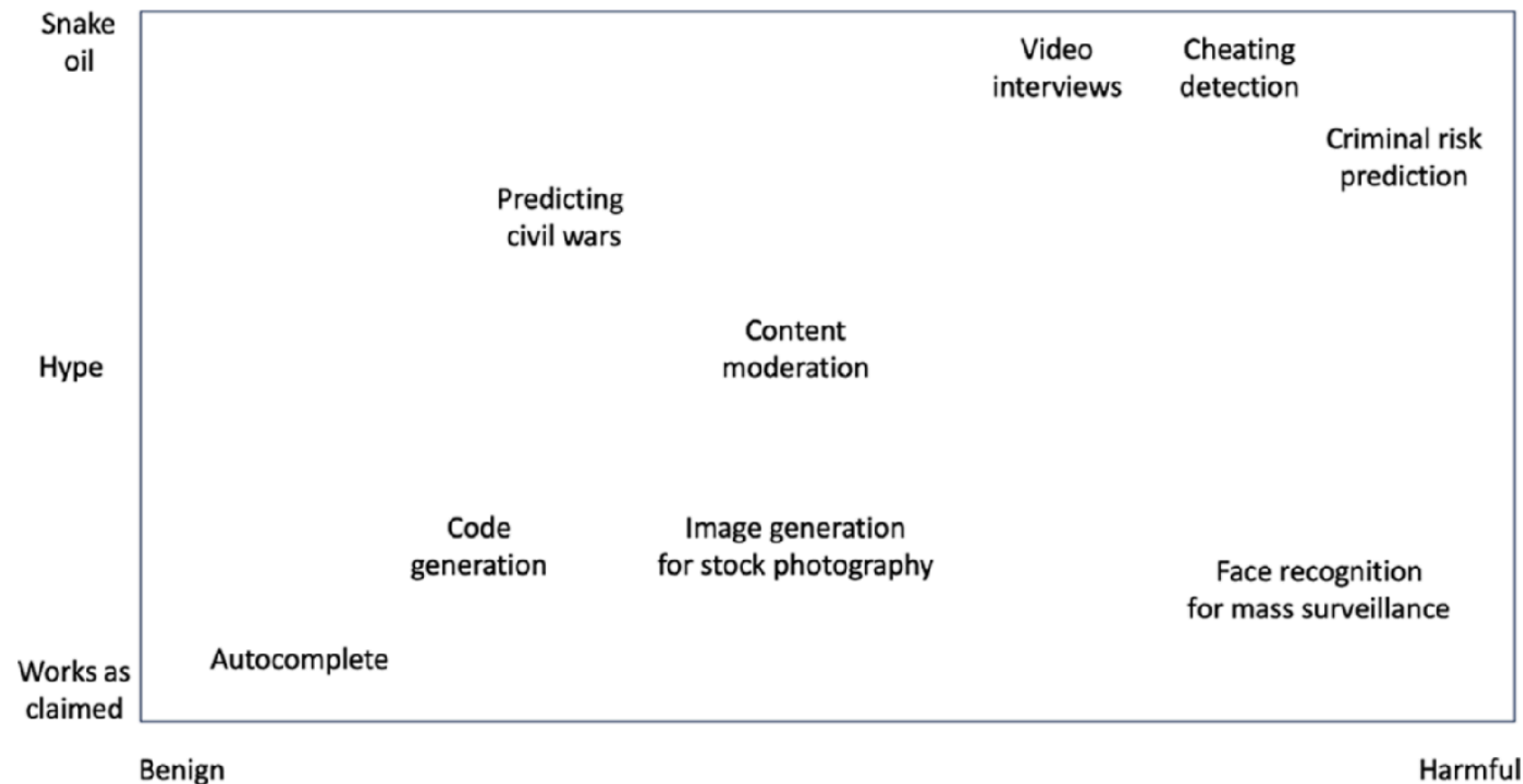
Different problem framings - different responsibilities

Table 1: Failure Taxonomy

Impossible Tasks	Conceptually Impossible Practically Impossible
Engineering Failures	Design Failures Implementation Failures Missing Safety Features
Post-Deployment Failures	Robustness Issues Failure under Adversarial Attacks Unanticipated Interactions
Communication Failures	Falsified or Overstated Capabilities Misrepresented Capabilities

Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22).

Different problem framings - different responsibilities

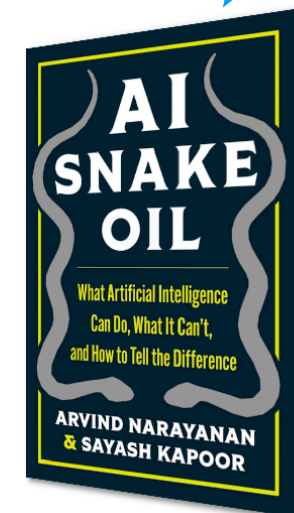


Do not
contribute to
hype in your
communication:
contribute to a
culture of
humility and
critical
evaluation

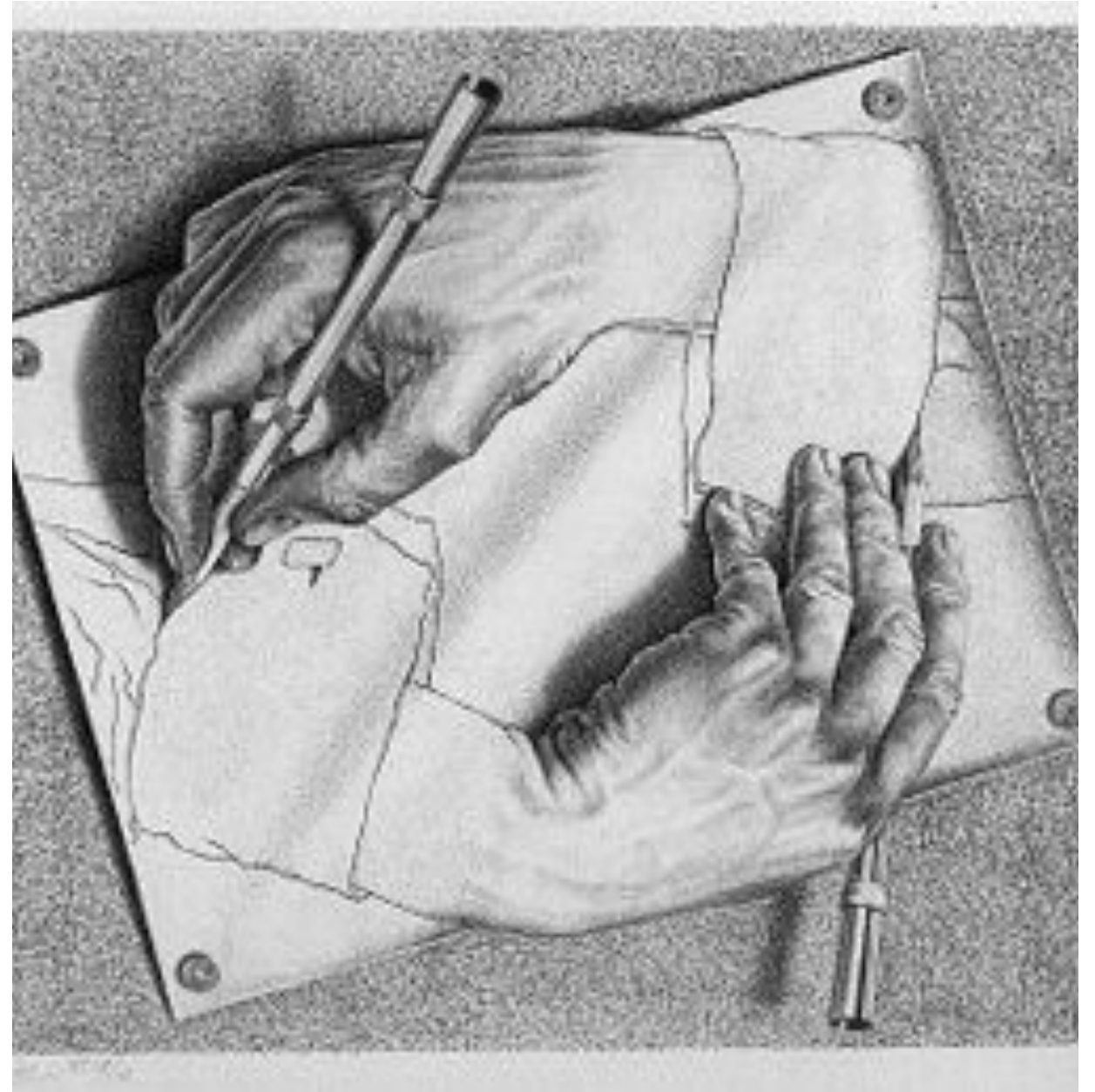
what success do
your
benchmarks
measure?

FIGURE 1.2. The landscape of AI snake oil, hype, and harms, showing a few illustrative applications.

Narayanan & Kapoor, AI Snake Oil, 2024



Reflection....



Wrap up

Things I have come to understand traveling across disciplines

- The co-evolutionary character of technology and humanity
 - temporal dimension, uncertainty, control?
 - implications for the responsibility of the CS professional?
- Awareness of what different disciplinary lenses bring to questions in Responsible ML
 - Reflection on embedded logics in computational sciences (including ML)
 - zooming in on fairness
- Discussion: what are responsibilities that the professionals in the field of ML can and should carry?

References - recommended reading

- Five Things We Need to Know About Technological Change (Neil Postman, 1998)
- van de Poel, I. An Ethical Framework for Evaluating Experimental Technology. *Sci Eng Ethics* 22, 667–686 (2016)
- Arvind Narayanan, 21 definitions of Fairness, Tutorial at FAT*2018
- Selbst, Andrew D. and Boyd, Danah and Friedler, Sorelle and Venkatasubramanian, Suresh and Vertesi, Janet, Fairness and Abstraction in Sociotechnical Systems. 2019 ACM Conference on Fairness, Accountability, and Transparency (FAT*), 59-68
- Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. 2019 ACM Conference on Fairness, Accountability, and Transparency (FAT*), 39–48.
- Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong. 2019. This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. *Proc. ACM Hum.-Comput. Interact.*
- Arvind Narayanan & Sayash Kapoor, *AI Snake Oil* (2024)