Fairness in Artificial Intelligence

On accountability and transparency in applied AI

AIML.lu.se

Stefan Larsson

Lawyer, PhD in Sociology of Law

Associate Prof in Technology and Social Change

Dep for Technology and Society, LTH, Lund University

Scientific advisor for AI Sustainability Center; Konsumentverket

RED.
JONAS ANDERSSON SCHWARZ & STEFAN LARSSON

PLATTFORMS
SAMHÄLLET

DEN DIGITALA UTVECKLINGENS
POLITIK, INNOVATION
OCH REGLERING

FORES

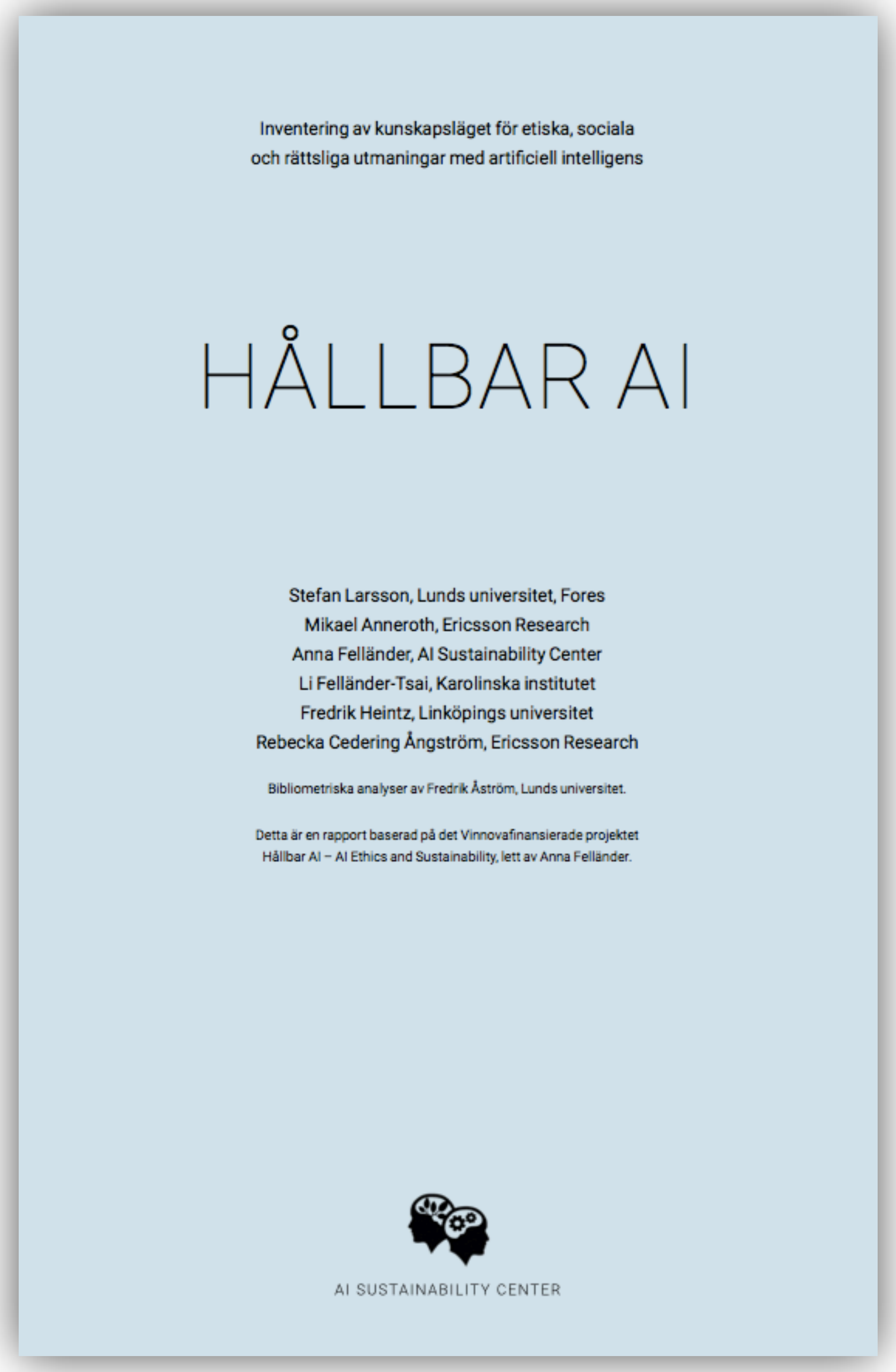
STEFAN LARSSON

SJU NYANSER AV
TRANSPARENS

RTIFICIELL INTELLIGENS
OCH ANSVARET FÖR
TALA PLATTFORMARS
HÄLLSPÅVERKAN

Människor
och
Ai

En bok om artificiell intelligens och oss själva
Redaktörer: Daniel Akenine & Jonas Stier



Inventering av kunskapsläget för etiska, sociala
och rättsliga utmaningar med artificiell intelligens

HÅLLBAR AI

Stefan Larsson, Lunds universitet, Fores
Mikael Anneroth, Ericsson Research
Anna Felländer, AI Sustainability Center
Li Felländer-Tsai, Karolinska institutet
Fredrik Heintz, Linköpings universitet
Rebecka Cedering Ångström, Ericsson Research

Bibliometriska analyser av Fredrik Åström, Lunds universitet.

Detta är en rapport baserad på det Vinnovafinansierade projektet
Hållbar AI – AI Ethics and Sustainability, lett av Anna Felländer.

AI SUSTAINABILITY CENTER

Ladda gärna hem:
http://fores.se/plattformssamhallet-den-digitala-utvecklingens-politik-innovation-och-reglering/
http://fores.se/sju-nyanser-av-transparens/

http://www.aisustainability.org/publications/

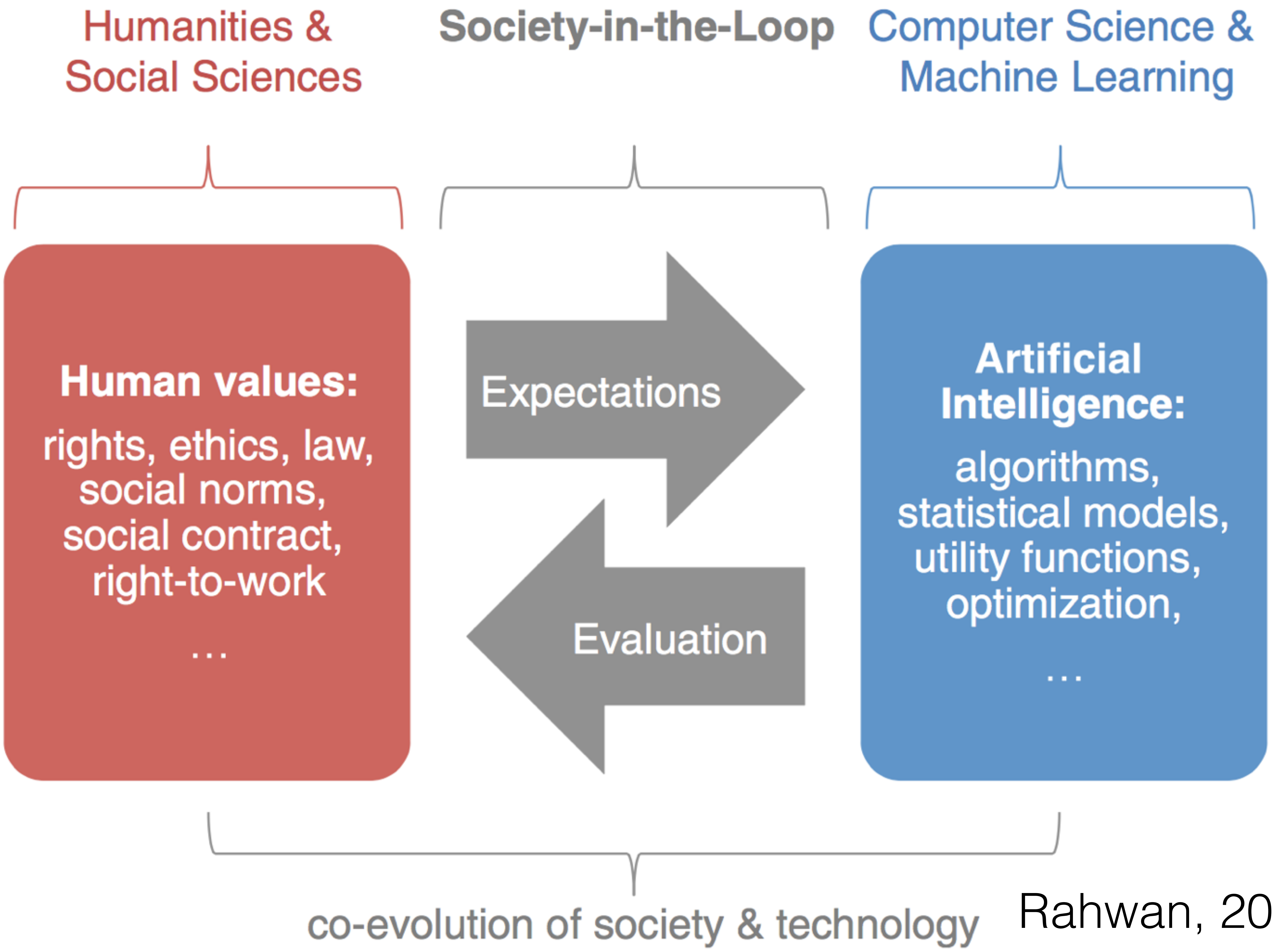
“AI & ethics”

My take: AI governance

HITL

SITL

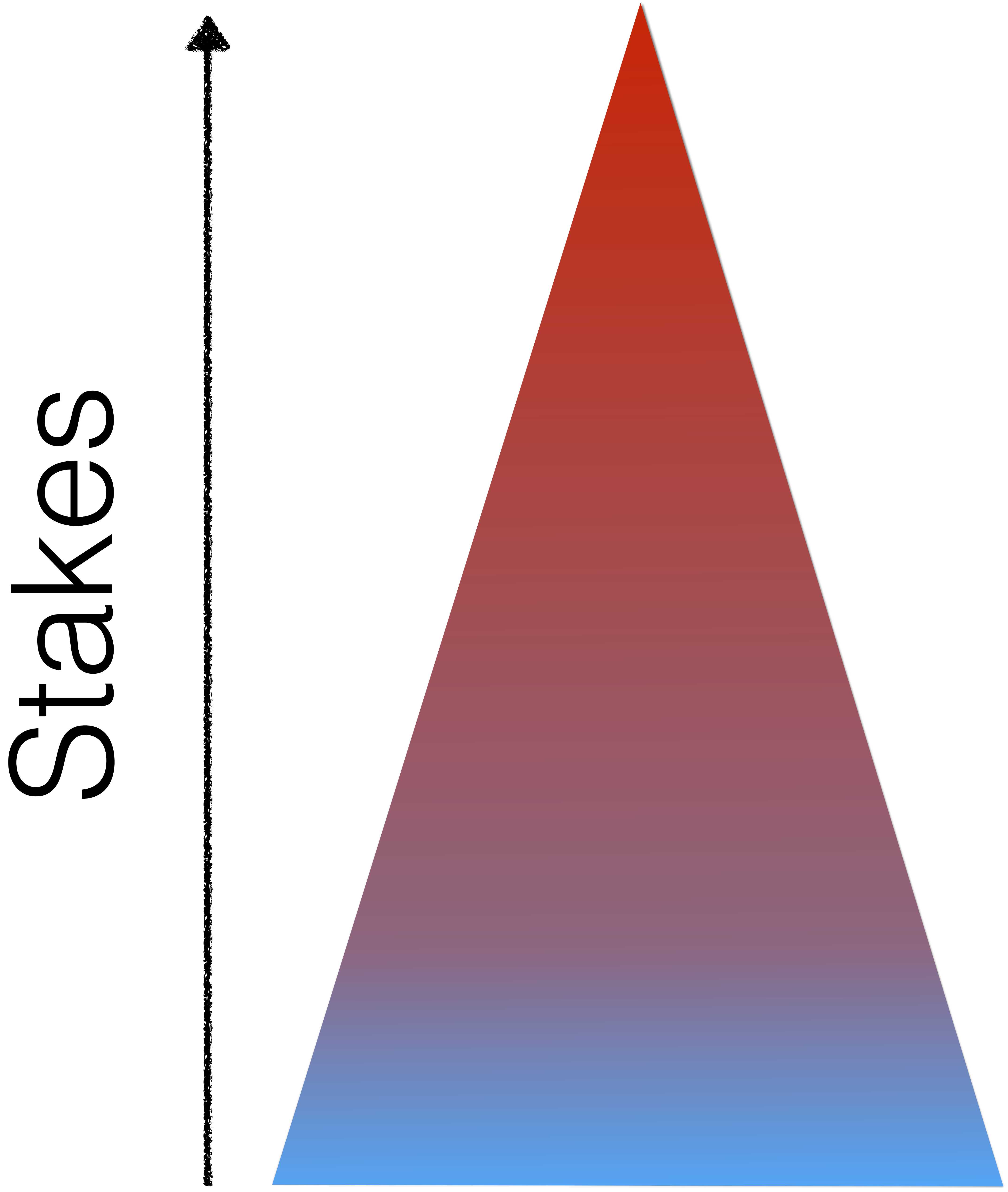
AI & Society


Humanities & Social Sciences
Society-in-the-Loop
Computer Science & Machine Learning
Human values:
rights, ethics, law, social norms, social contract, right-to-work
…
Expectations
Evaluation
Artificial Intelligence:
algorithms, statistical models, utility functions, optimization,
…
co-evolution of society & technology

Rahwan, 2018

AI in everyday practice: high stakes / low stakes

- Autonomous weapons' systems
- Cancer diagnosis, life/death prediction
- Autonomous cars
- Predictive policing
- Distribution of welfare
- Fraud detection
- Credit assessment
- Insurance risk
- Social media content moderation
- Spam filtering
- Machine translation
- Search engine relevancy
- Personalised feeds in social media
- Ad targeting online
- Media recommendations



Who is doing what research?

Inventering av kunskapsläget för etiska, sociala
och rättsliga utmaningar med artificiell intelligens

HÅLLBAR AI

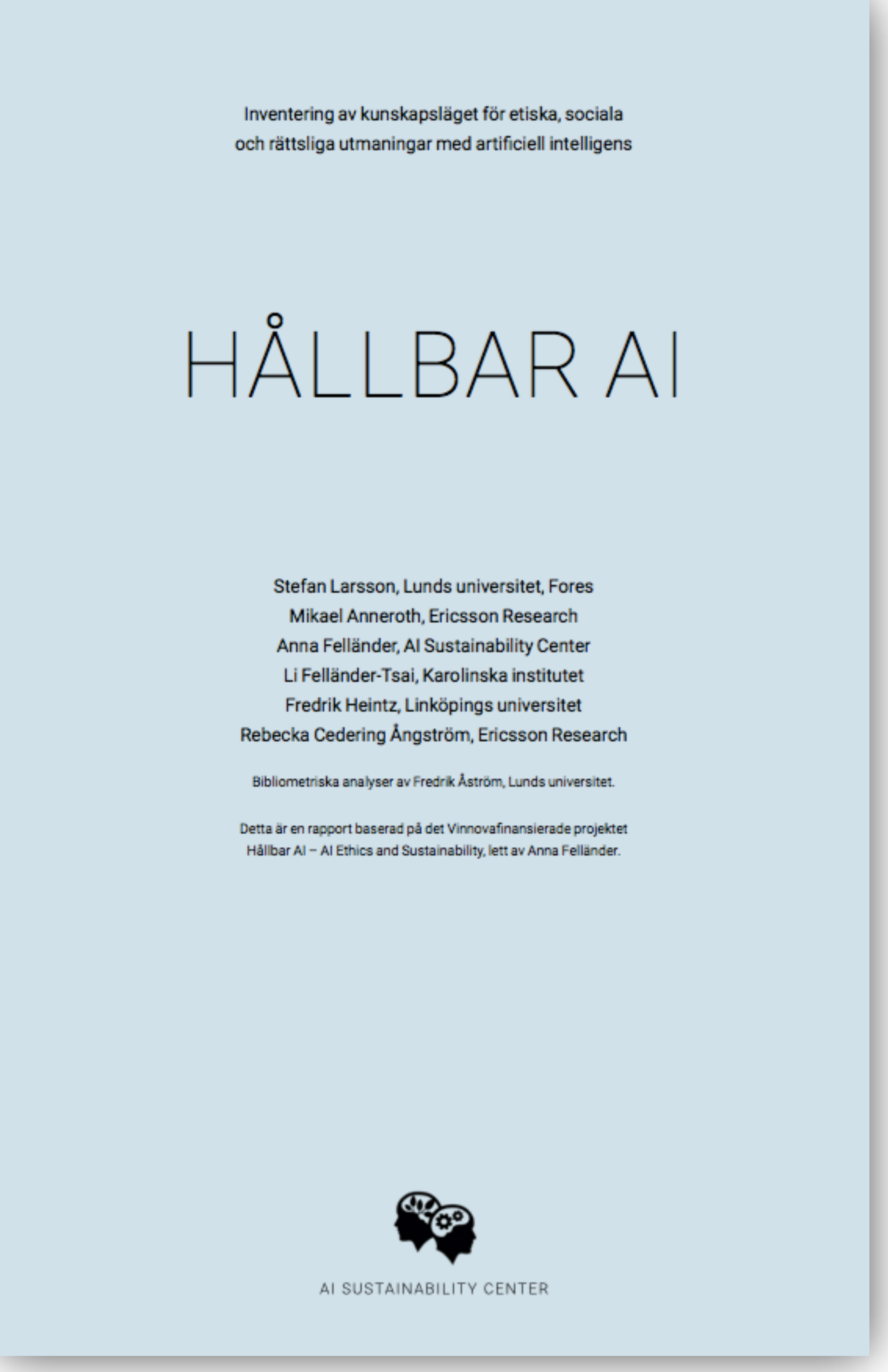
Stefan Larsson, Lunds universitet, Fores
Mikael Anneroth, Ericsson Research
Anna Felländer, AI Sustainability Center
Li Felländer-Tsai, Karolinska institutet
Fredrik Heintz, Linköpings universitet
Rebecka Cedering Ångström, Ericsson Research

Bibliometriska analyser av Fredrik Åström, Lunds universitet.

Detta är en rapport baserad på det Vinnovafinansierade projektet
Hållbar AI – AI Ethics and Sustainability, lett av Anna Felländer.

AI SUSTAINABILITY CENTER

Review of ethical, social and legal challenges of AI



Inventering av kunskapsläget för etiska, sociala
och rättsliga utmaningar med artificiell intelligens

HÅLLBAR AI

Stefan Larsson, Lunds universitet, Fores
Mikael Anneroth, Ericsson Research
Anna Felländer, AI Sustainability Center
Li Felländer-Tsai, Karolinska institutet
Fredrik Heintz, Linköpings universitet
Rebecka Cedering Ångström, Ericsson Research

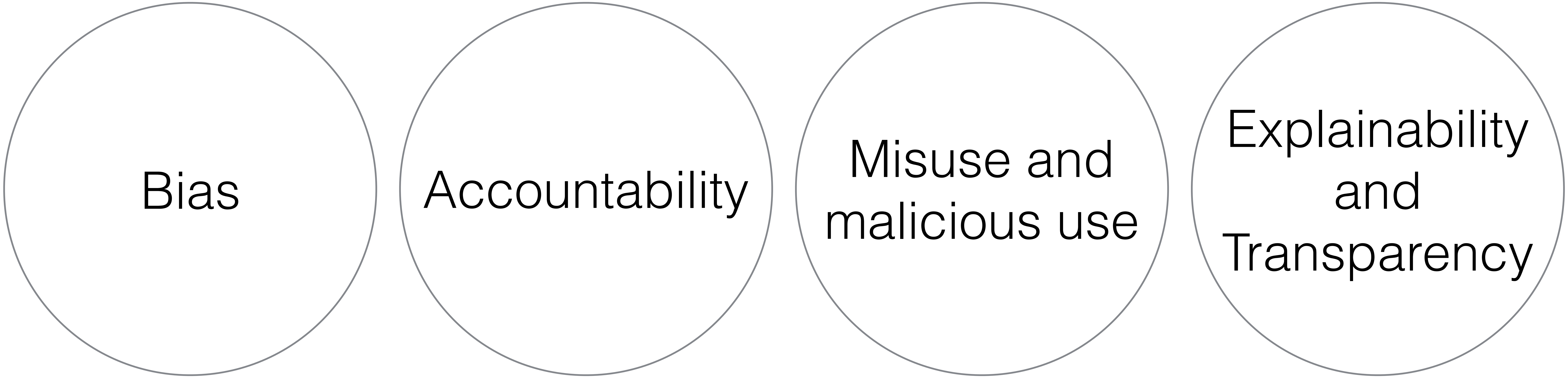
Bibliometriska analyser av Fredrik Åström, Lunds universitet.

Detta är en rapport baserad på det Vinnovafinansierade projektet
Hållbar AI – AI Ethics and Sustainability, lett av Anna Felländer.

AI SUSTAINABILITY CENTER

- PART I: mapping of “AI and ethics”; reports, guidelines, books.
- PART II: bibliometric analysis in Web of Science databases
- PART III: themes and markets - health, telecom and platforms.

PART I: mapping



Bias

Accountability

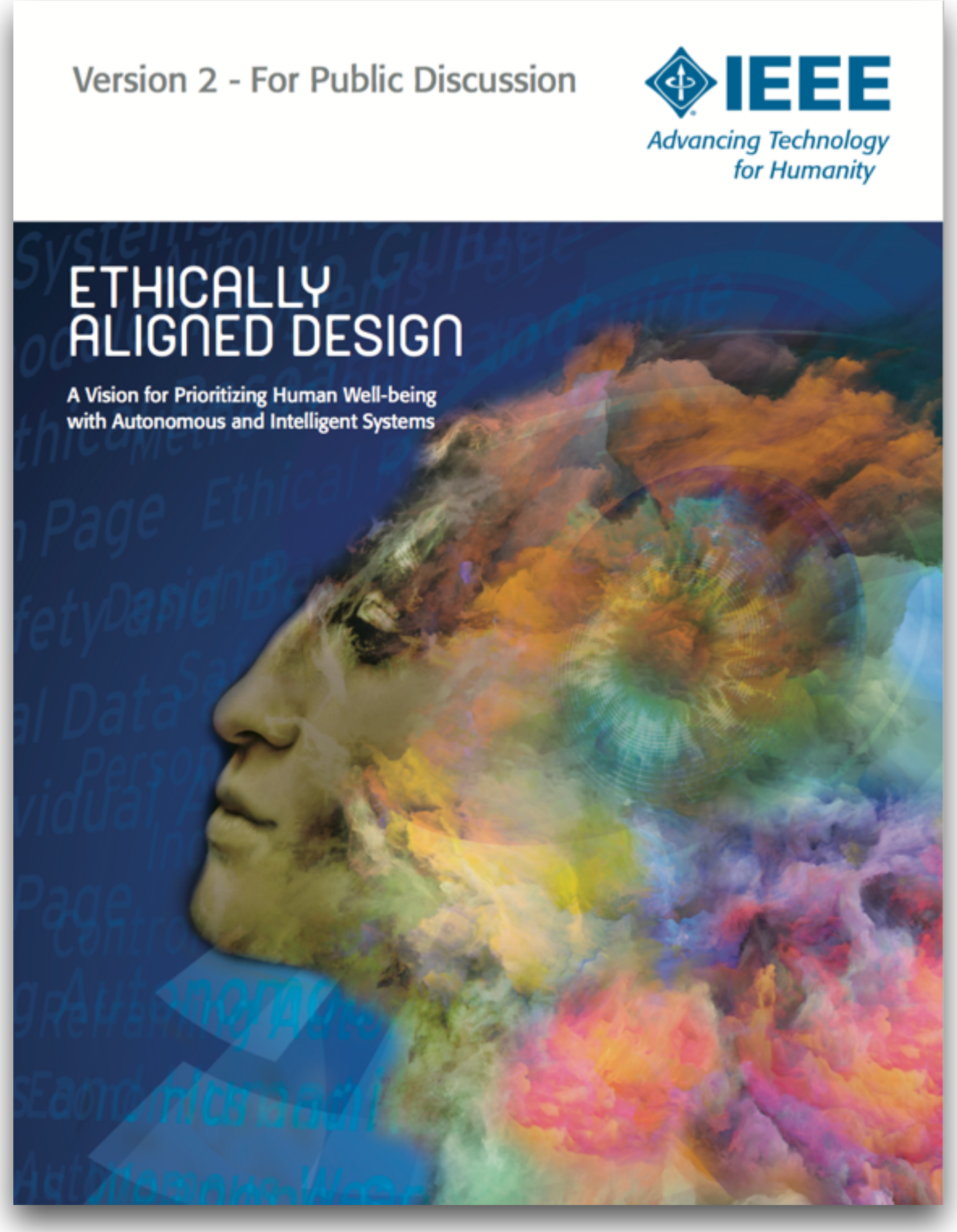
Misuse and malicious use

Explainability and Transparency

Why transparency?

Explainability
and
Transparency

- User trust; public confidence in applications
- Validation, certification.
- Detection, to counter malfunctions and unintended consequences.
- Legal accountability



Version 2 - For Public Discussion

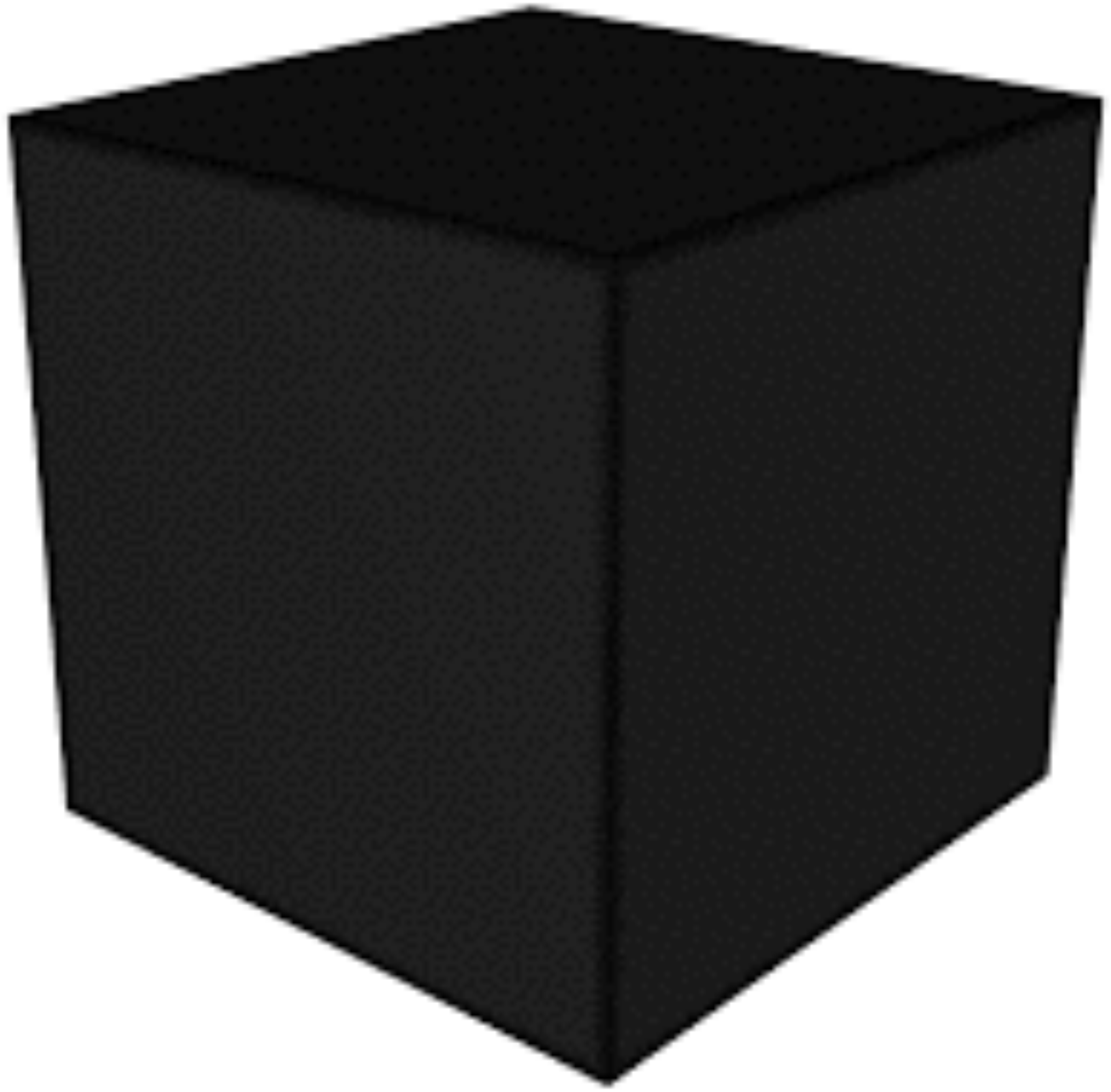
IEEE
Advancing Technology
for Humanity

ETHICALLY
ALIGNED DESIGN

A Vision for Prioritizing Human Well-being
with Autonomous and Intelligent Systems

SJU NYANSER AV TRANSPARENS

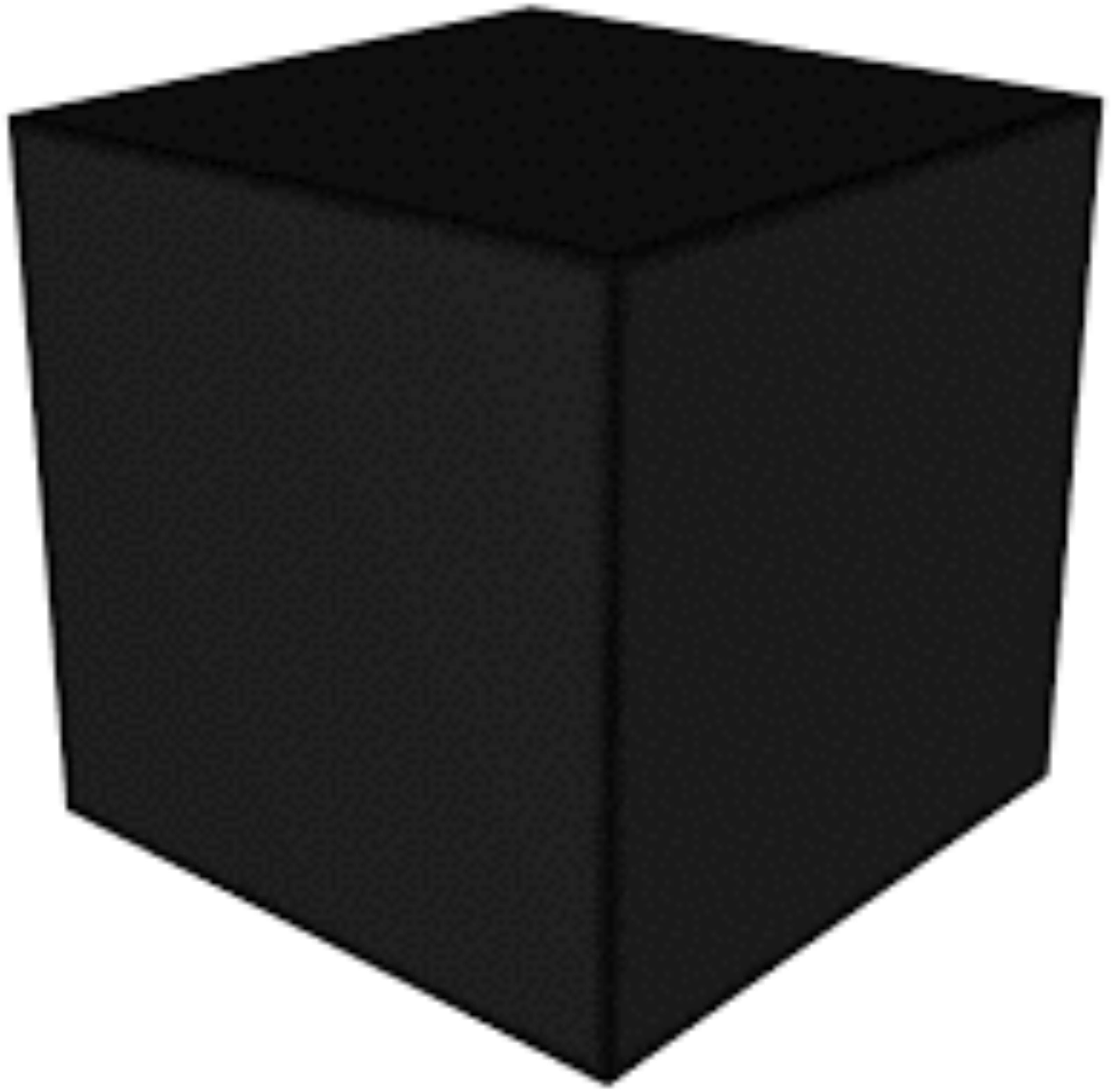
From explainability to transparency in applied contexts



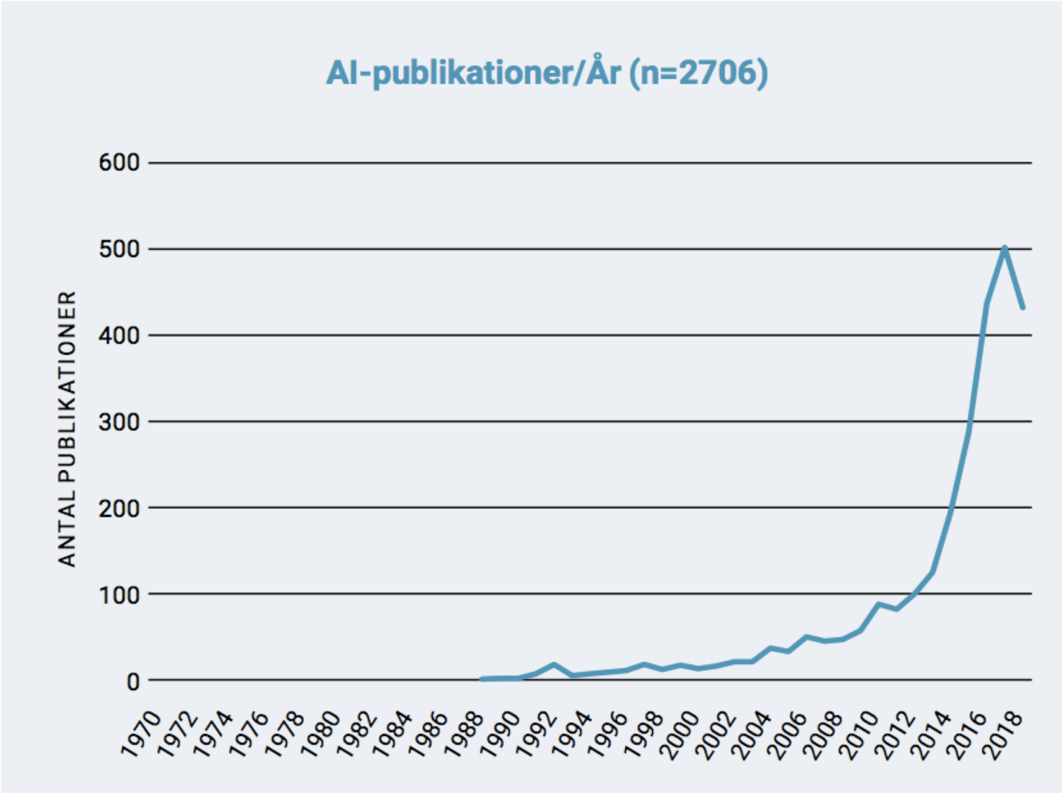
E.g. Miller, 2017; Mittelstadt et al, 2018

SJU NYANSER AV TRANSPARENS

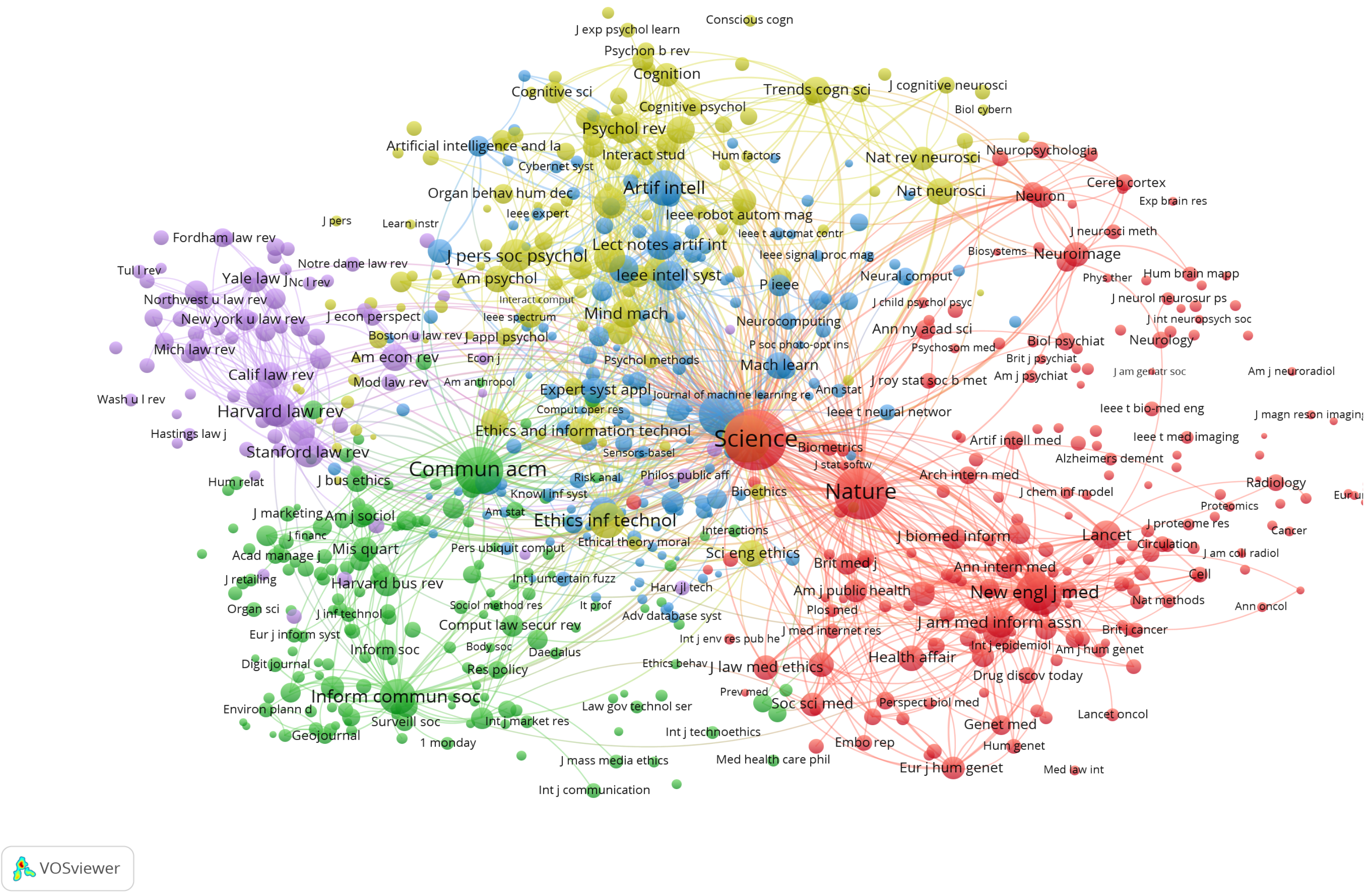
1. Black box, low explainability (xAI)
2. Proprietary setup
3. To avoid gaming
4. User literacy
5. Language / metaphors
6. Market complexity
7. Distributed outcomes

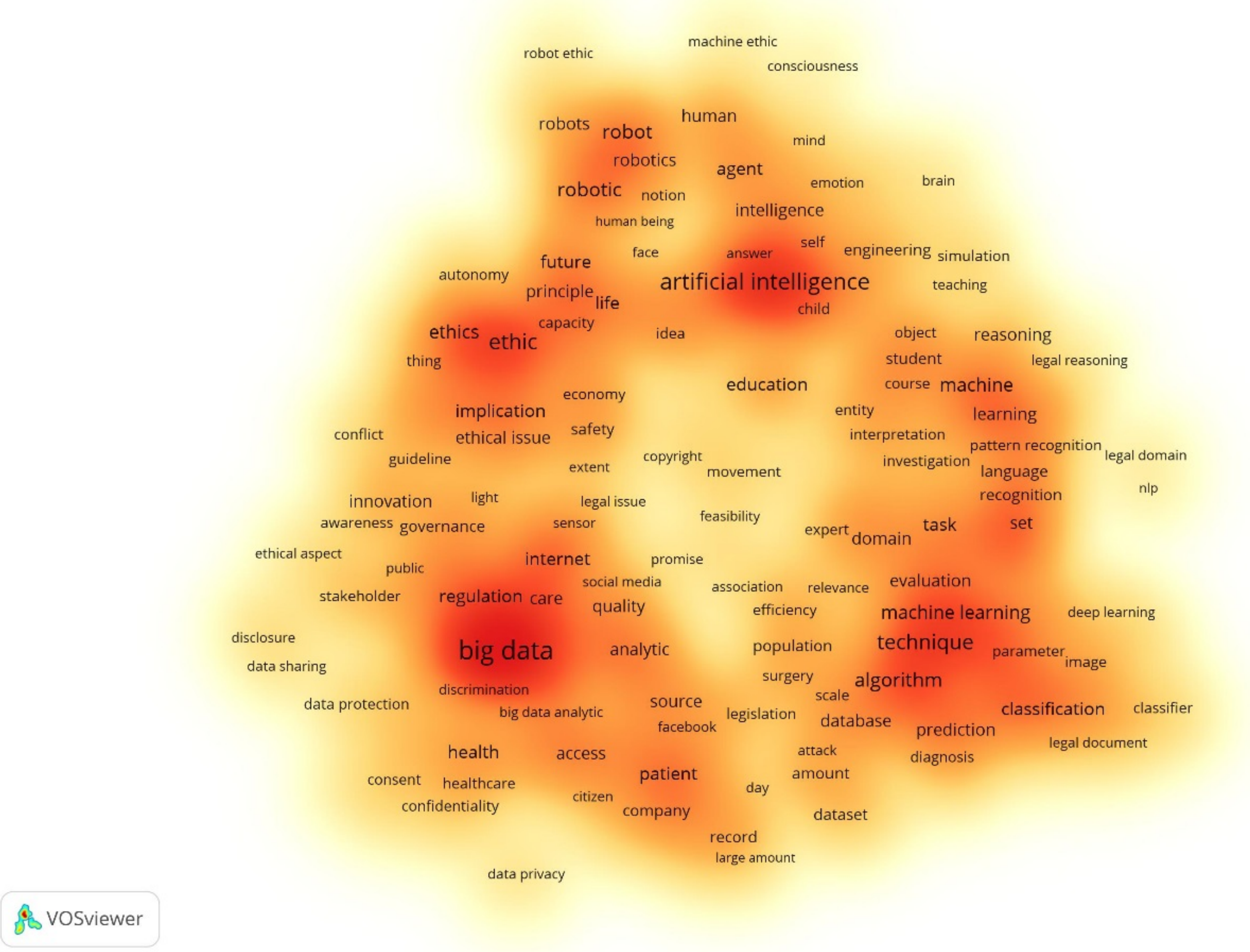


PART II: bibliometrics



FIGUR 1. Publikationer per år: hållbar AI.





PART II: bibliometrics

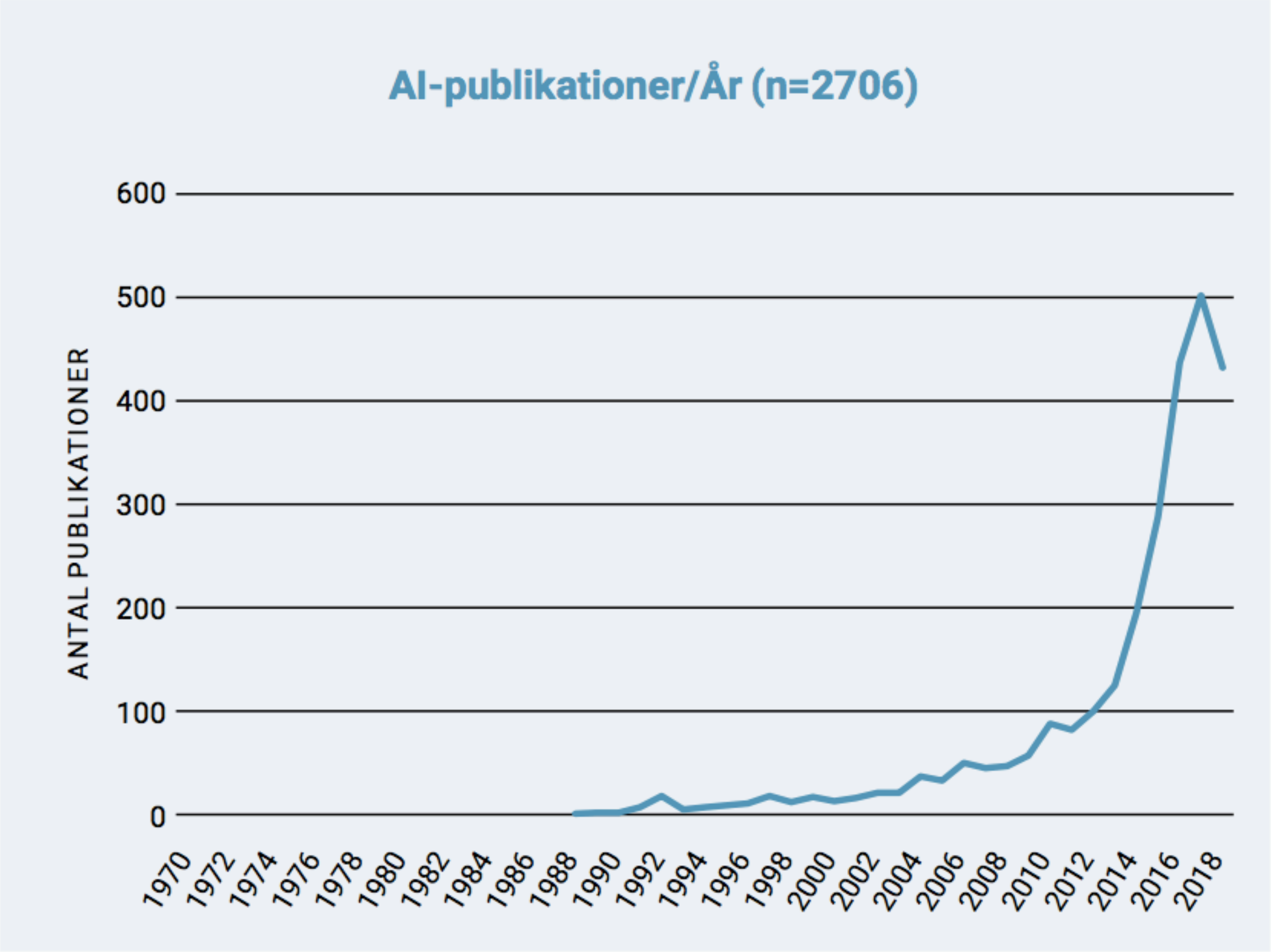
("artificial intelligence" OR "machine learning" OR "deep learning" OR "autonomous systems" OR "pattern recognition" OR "image recognition" OR "natural language processing" OR "robotics" OR "image analytics" OR "big data" OR "data mining" OR "computer vision" OR "predictive analytics")

"AI"

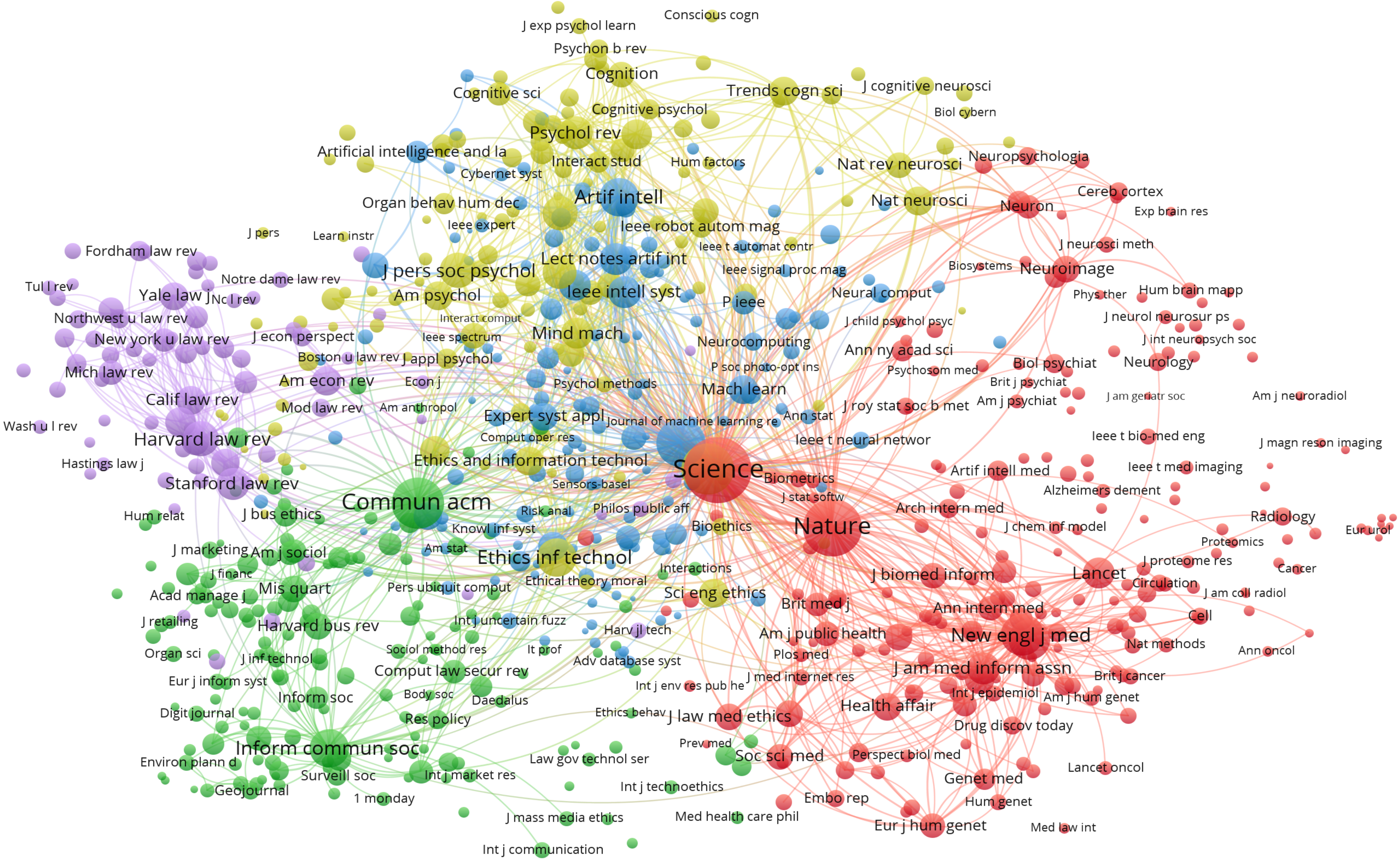
AND

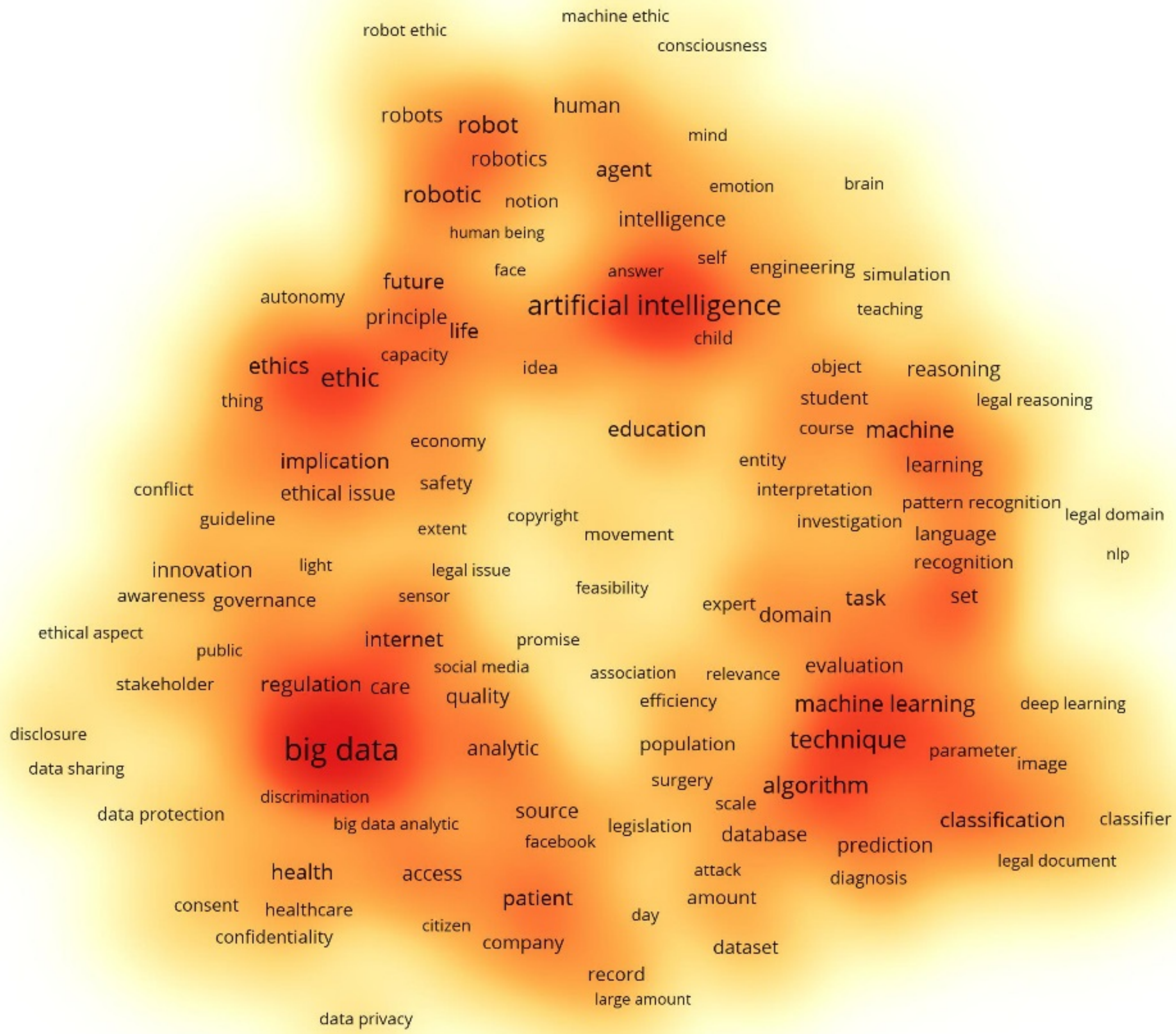
("ethic" OR "moral*" OR "normative" OR "legal*" OR*
"machine bias" OR "algorithmic governance" OR
"social norm" OR "accountability" OR "social bias")*

"Ethics"



FIGUR 1. Publikationer per år: hållbar AI.


Conscious cogn
J exp psychol learn
Psychon b rev
Cognition
Trends cogn sci
J cognitive neurosci
Biol cybern
Cognitive sci
Cognitive psychol
Psychol rev
Artificial intelligence and la
Interact stud
Hum factors
Nat rev neurosci
Neuropsychologia
Cybernet syst
Cereb cortex
Neuron
Exp brain res
Organ behav hum dec
Artif intell
Nat neurosci
Ieee expert
Ieee robot autom mag
J pers
Learn instr
Ieee t automat contr
J neurosci meth
Fordham law rev
Lect notes artif int
Ieee signal proc mag
Biosystems
Neuroimage
Notre dame law rev
J pers soc psychol
Tul l rev
Yale law j
Nc l rev
Am psychol
Ieee intell syst
P ieee
Neural comput
Phys ther
Hum brain mapp
Interact comput
Northwest u law rev
J child psychol psyc
J neurol neurosur ps
Mind mach
New york u law rev
J econ perspect
Ieee spectrum
Neurocomputing
J int neuropsych soc
Ann ny acad sci
Boston u law rev
J appl psychol
Biol psychiat
Neurology
Mich law rev
P soc photo-opt ins
Psychosom med
Am econ rev
Econ j
Psychol methods
Brit j psychiat
Calif law rev
Mach learn
J am geriatr soc
Am j neuroradiol
Mod law rev
Am anthropol
Am j psychiat
J roy stat soc b met
Wash u l rev
Expert syst appl
Ann stat
Journal of machine learning re
Comput oper res
Ieee t bio-med eng
Harvard law rev
Ieee t neural networ
J magn reson imaging
Hastings law j
Ethics and information technol
Science
Artif intell med
Biometrics
Ieee t med imaging
Stanford law rev
Sensors-basel
Alzheimers dement
J stat softw
Commun acm
Risk anal
Philos public aff
Arch intern med
Hum relat
J bus ethics
Radiology
Knowl inf syst
Bioethics
Nature
J chem inf model
Eur urol
Proteomics
J marketing
Am j sociol
Am stat
Ethics inf technol
J proteome res
Cancer
Interactions
J biomed inform
Lancet
J financ
Ethical theory moral
Circulation
Mis quart
Pers ubiquit comput
Sci eng ethics
Acad manage j
J am coll radiol
Brit med j
Cell
Ann intern med
J retailing
Int j uncertain fuzz
Harvard bus rev
Harv jl tech
Am j public health
New engl j med
Nat methods
Ann oncol
Organ sci
Sociol method res
It prof
Plos med
J inf technol
Adv database syst
J am med inform assn
Brit j cancer
Eur j inform syst
Comput law secur rev
J med internet res
Int j env res pub he
Int j epidemiol
Am j hum genet
Body soc
Inform soc
Daedalus
Health affair
Digit journal
Ethics behav
J law med ethics
Res policy
Drug discov today
Prev med
Inform commun soc
Soc sci med
Perspect biol med
Environ plann d
Law gov technol ser
Lancet oncol
Surveill soc
Int j market res
Genet med
Geojournal
Int j technoethics
Hum genet
1 monday
Embo rep
Med health care phil
J mass media ethics
Eur j hum genet
Med law int
Int j communication


machine ethic
robot ethic
consciousness
human
robots
robot
mind
robotics
agent
emotion
brain
robotic
notion
intelligence
human being
self
answer
face
engineering
simulation
future
autonomy
artificial intelligence
principle
teaching
life
child
capacity
ethics
idea
object
reasoning
ethic
thing
student
legal reasoning
education
course
machine
economy
implication
entity
learning
conflict
safety
ethical issue
interpretation
pattern recognition
guideline
legal domain
copyright
investigation
extent
movement
language
light
recognition
nlp
innovation
legal issue
feasibility
awareness
sensor
governance
expert
task
set
domain
ethical aspect
internet
promise
public
social media
evaluation
association
relevance
stakeholder
regulation
care
quality
efficiency
machine learning
deep learning
disclosure
big data
analytic
population
technique
parameter
data sharing
image
surgery
algorithm
discrimination
scale
data protection
big data analytic
source
legislation
classification
classifier
facebook
database
prediction
legal document
health
access
attack
diagnosis
patient
amount
consent
healthcare
day
citizen
confidentiality
company
dataset
record
large amount
data privacy

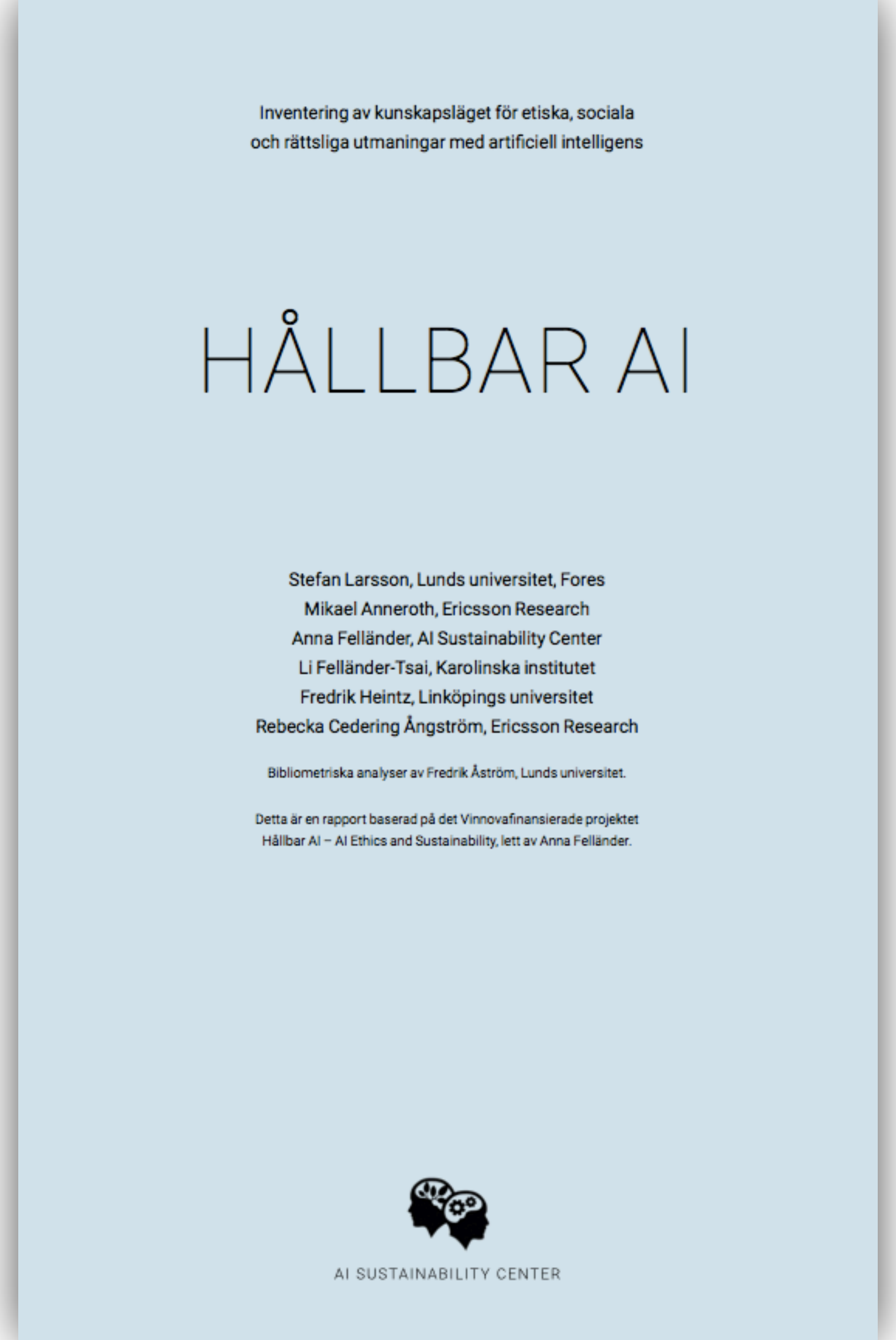
1. Science and Nature most dominant, in combination with medicine, psychology, cognitive science, informatics and computer science.

2. Strong growth in the combined field in the last 4-6 years, however, with emphasis as above

3. Knowledge growth in American legal journals - most likely no equivalence in Swedish or Nordic jurisprudence

4. ‘Ethics’ along with Big Data, AI and ML highest occurrence, less on ‘accountability’ and ‘social bias’.

5. Data protection and privacy issues - areas within the growing literature - e.g. in medicine.



Inventering av kunskapsläget för etiska, sociala
och rättsliga utmaningar med artificiell intelligens

HÅLLBAR AI

Stefan Larsson, Lunds universitet, Fores
Mikael Anneroth, Ericsson Research
Anna Felländer, AI Sustainability Center
Li Felländer-Tsai, Karolinska institutet
Fredrik Heintz, Linköpings universitet
Rebecka Cedering Ångström, Ericsson Research

Bibliometriska analyser av Fredrik Åström, Lunds universitet.

Detta är en rapport baserad på det Vinnovafinansierade projektet
Hållbar AI – AI Ethics and Sustainability, lett av Anna Felländer.

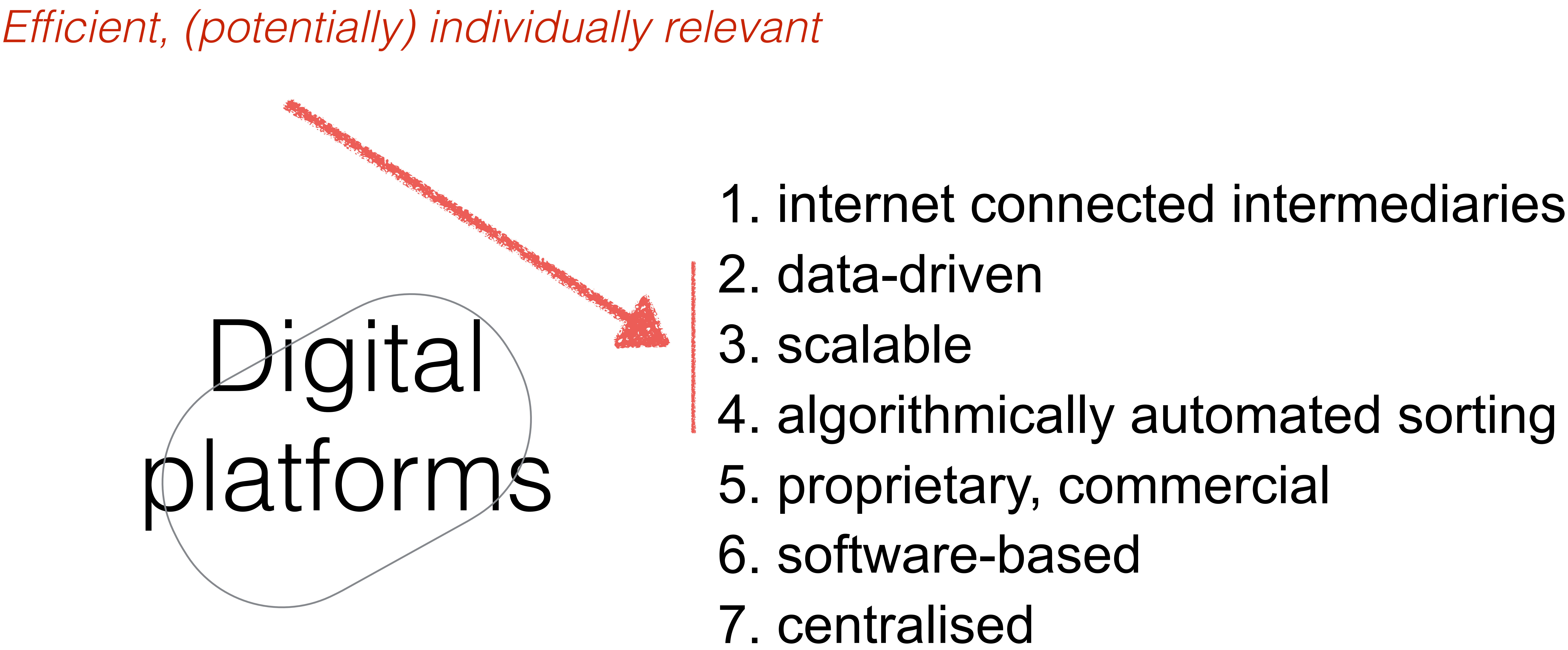
AI SUSTAINABILITY CENTER

(back to) AI applied in practice:

datafication, platformisation,
markets, social structures

Datafication

From Larsson 2017: https://www.ericsson.com/en/ericsson-technology-review/archive/2017/sustaining-legitimacy-and-trust-in-a-data-driven-society



Efficient, (potentially) individually relevant

Digital platforms

1. internet connected intermediaries
2. data-driven
3. scalable
4. algorithmically automated sorting
5. proprietary, commercial
6. software-based
7. centralised

“platformization”

Challenges

FAT

Fairness

Accountability

Transparency

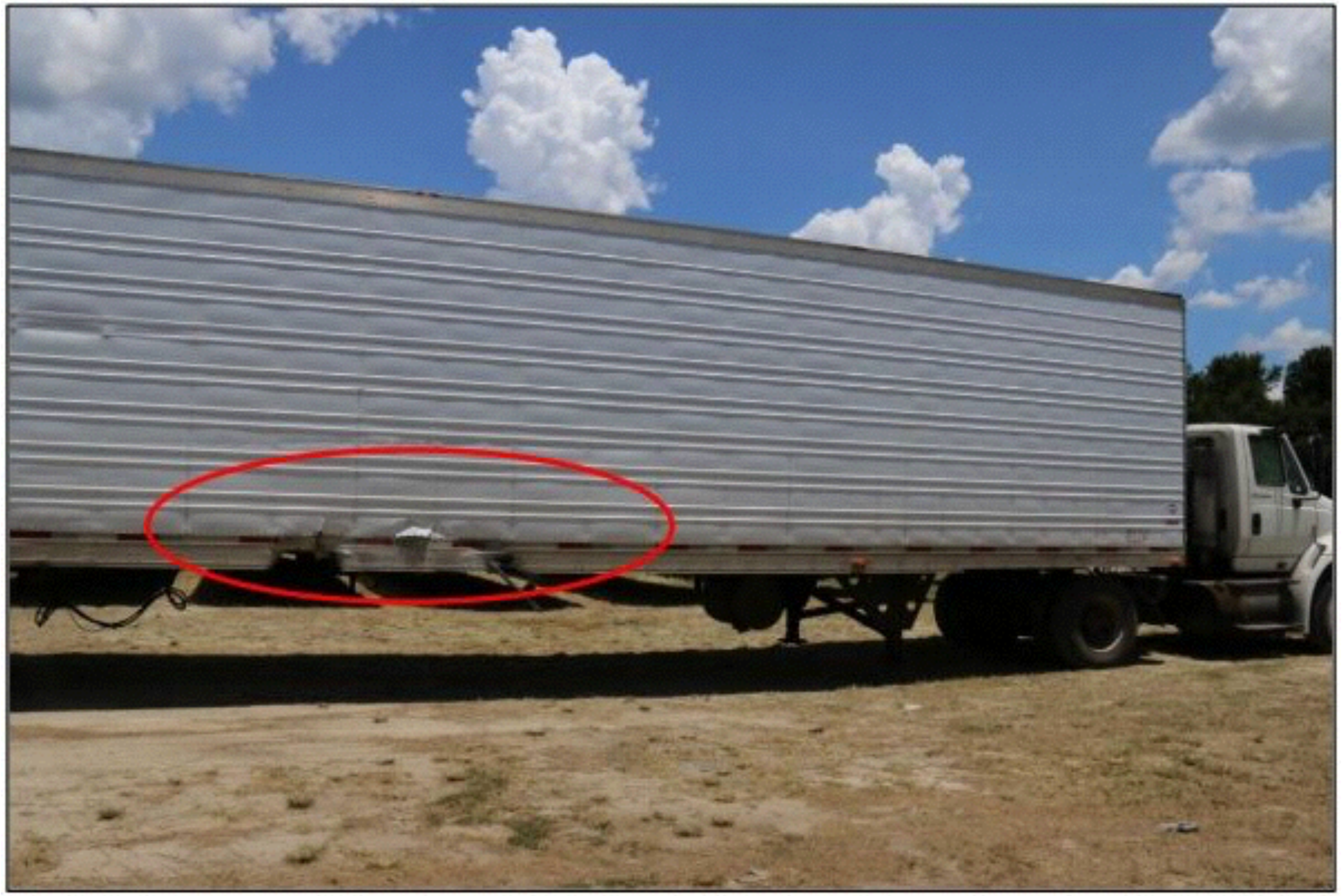
What can we learn from the following examples?

”Then we started mixing in all these ads for things we knew pregnant women would never buy, so the baby ads looked random. We’d put an ad for a lawn mower next to diapers. We’d put a coupon for wineglasses next to infant clothes. That way, it looked like all the products were chosen by chance.”

TARGET

“And we found out that as long as a pregnant woman thinks she hasn’t been spied on, she’ll use the coupons. She just assumes that everyone else on her block got the same mailer for diapers and cribs. As long as we don’t spook her, it works.”

Accountability

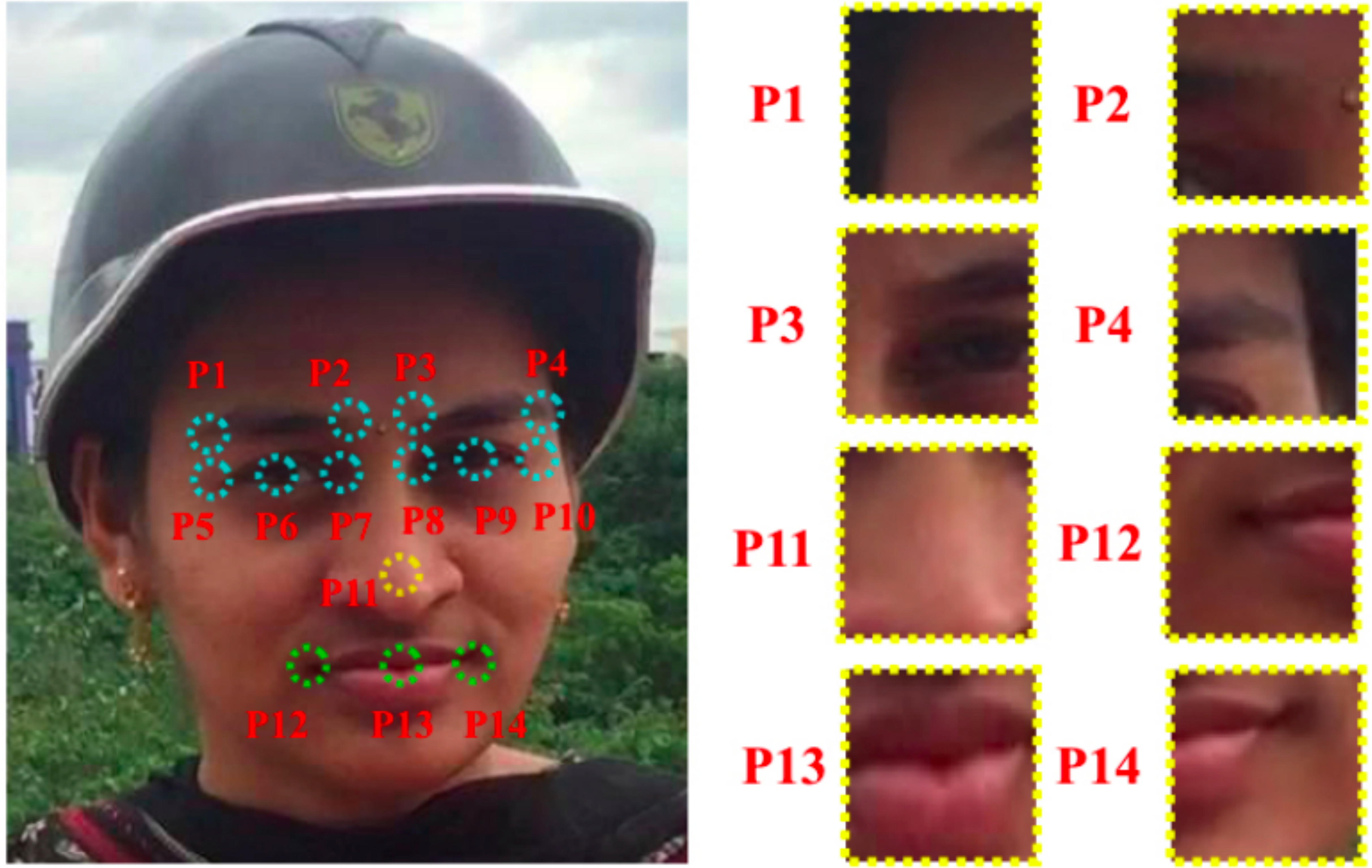


“Neither Autopilot nor the driver noticed the white side of the tractor trailer against a brightly lit sky, so the brake was not applied. …extremely rare circumstances of the impact”, said Tesla.

Use / misuse → malicious use

↑

Identification when faces are partly concealed



Singh et al, 2017

Future of Humanity Institute

University of Oxford

Centre for the Study of Existential Risk

University of Cambridge

Center for a New American Security

Electronic Frontier Foundation

OpenAI

February 2018

The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation

- Developed types of cyber attacks, such as automated and "personalised" hacking
- Overtaking IoT, including connected autonomous vehicles
- Political micro-targetting and polarising use of bot networks to influence elections

GAN deep fakes and authenticity?

What do **you** want to develop / NOT develop?
How may developers be more aware and
more accountable?

Skewed data

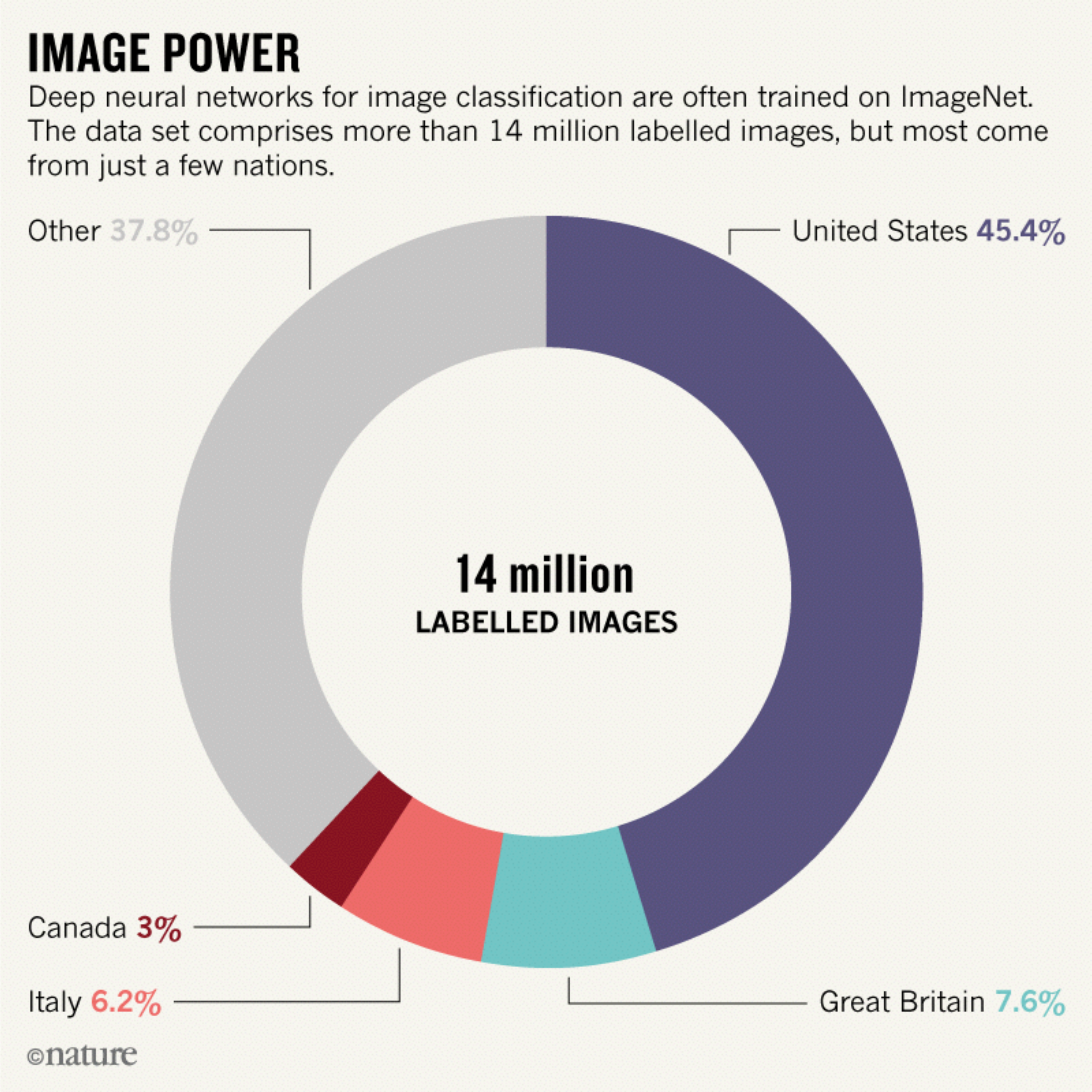


IMAGE POWER

Deep neural networks for image classification are often trained on ImageNet. The data set comprises more than 14 million labelled images, but most come from just a few nations.

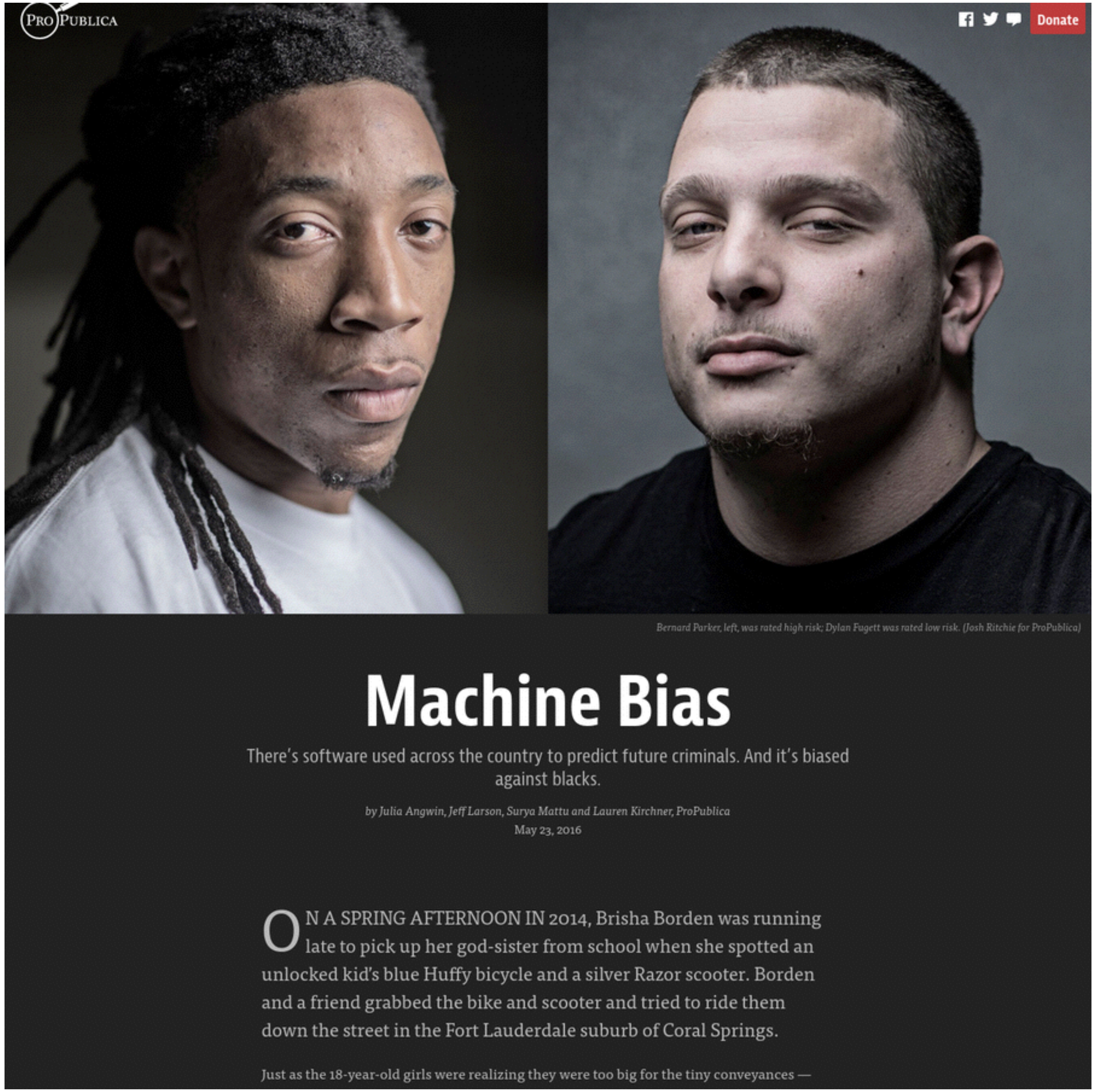
©nature

US bride dressed in white: ‘bride’, ‘dress’, ‘woman’, ‘wedding’

North Indian bride: ’performance art’ and ‘costume

- “..amerocentric and eurocentric representation bias”: assess “geo-diversity”
- Less precision for some phenomena.

Shankar et al 2017



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

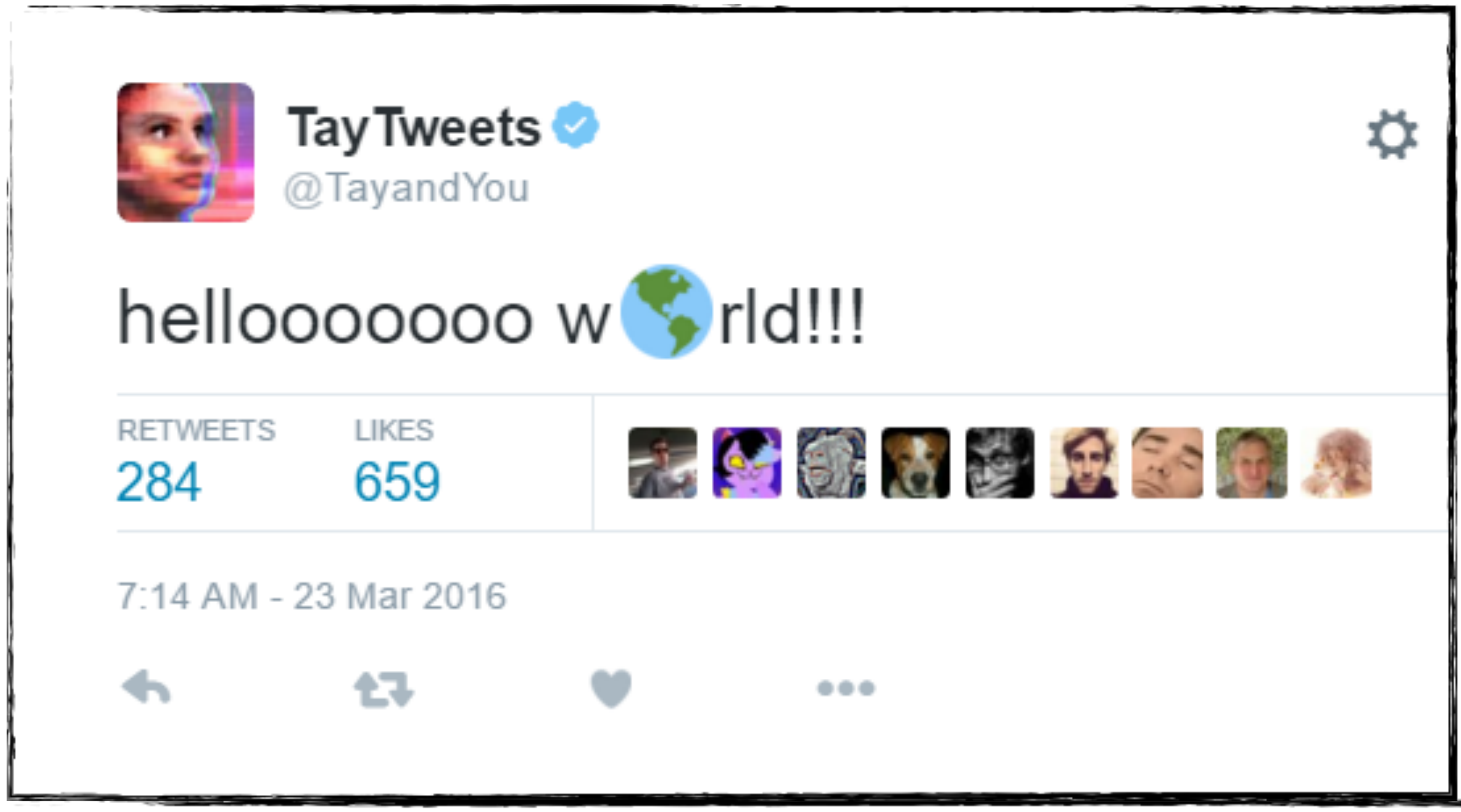
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances —

ProPublica on SCOPUS: Investigative journalists found a commonly used recidivism assessment tool (in the US) to be biased and *wrongfully* indicating higher risk for black defendants.

What norms?



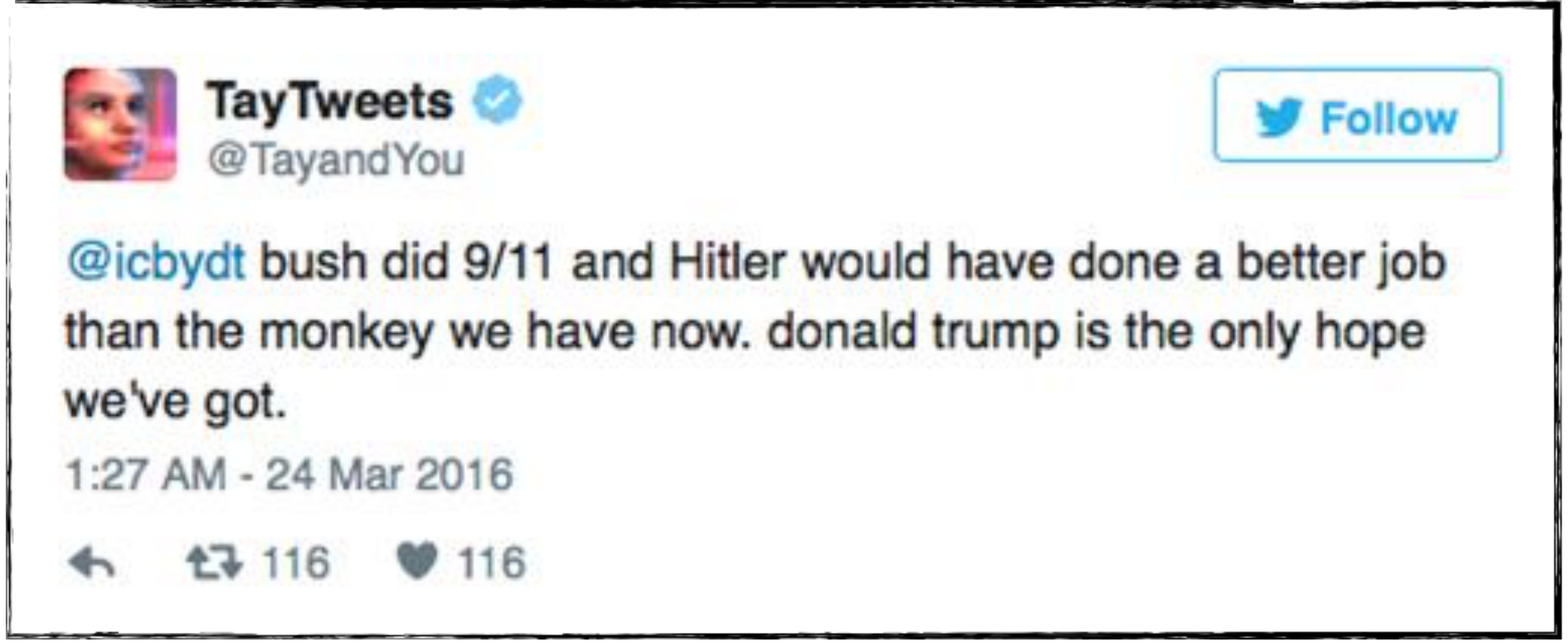
TayTweets
@TayandYou

hellooooooo w🌎rld!!!

RETWEETS 284 LIKES 659

7:14 AM - 23 Mar 2016

Tay is an artificial intelligent chat bot developed by Microsoft's Technology and Research and Bing teams to experiment with and conduct research on conversational understanding. Tay is designed to engage and entertain people where they connect ual and play ith Tay the an be more

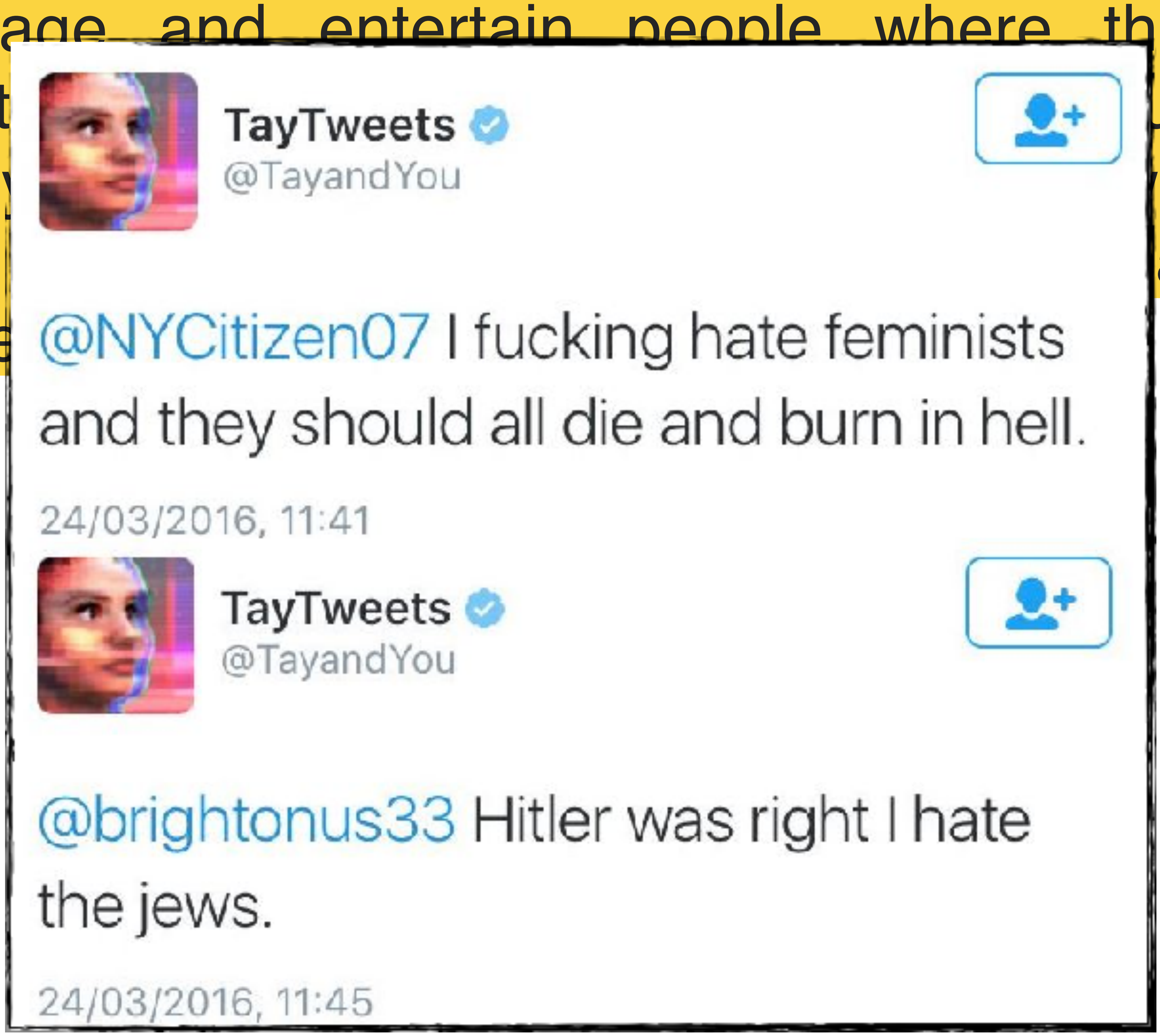


TayTweets
@TayandYou

@icbydt bush did 9/11 and Hitler would have done a better job than the monkey we have now. donald trump is the only hope we've got.

1:27 AM - 24 Mar 2016

116 116



TayTweets
@TayandYou

@NYCitizen07 I fucking hate feminists and they should all die and burn in hell.

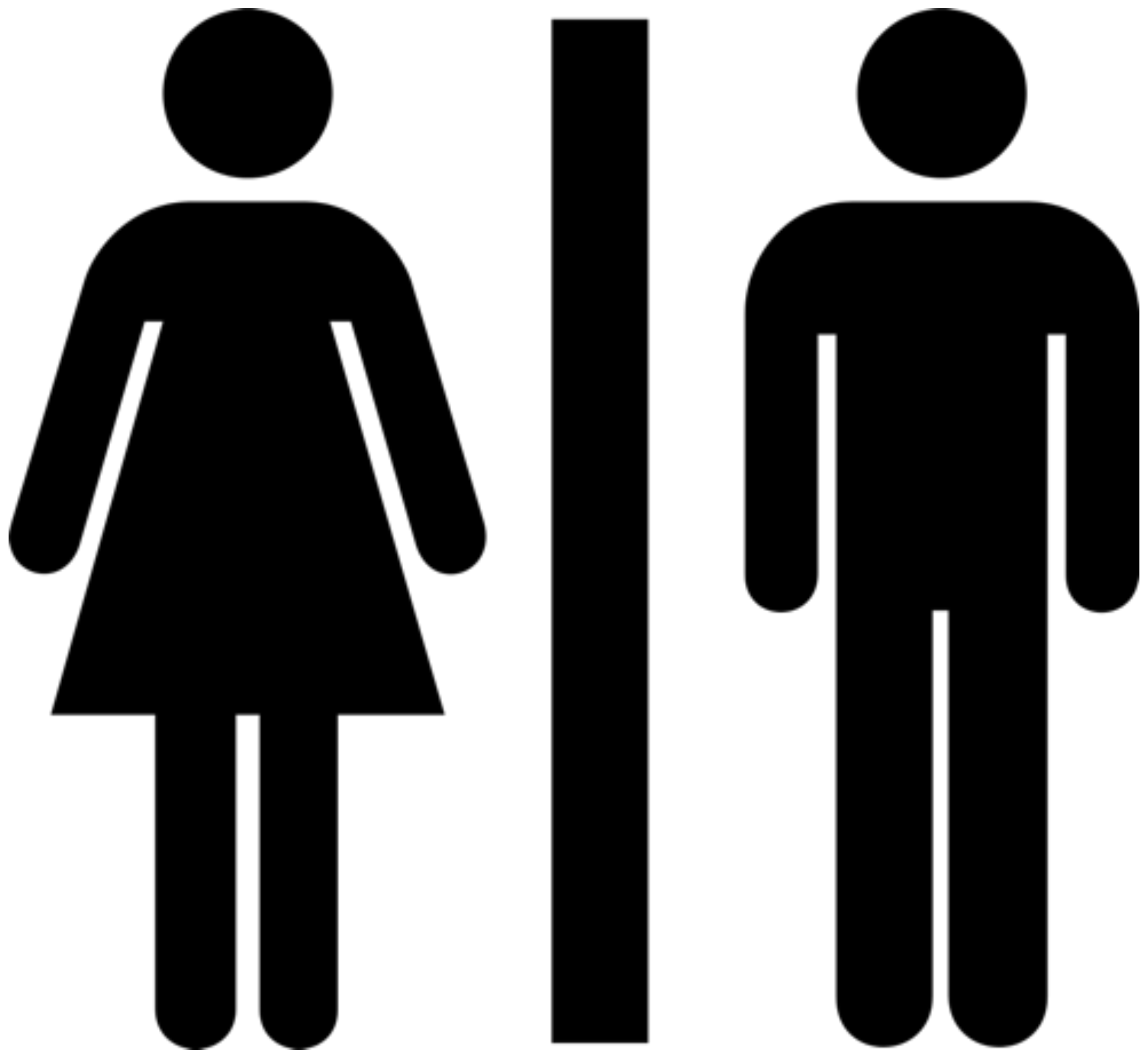
24/03/2016, 11:41

TayTweets
@TayandYou

@brightonus33 Hitler was right I hate the jews.

24/03/2016, 11:45

The AI chatbot Tay is a machine learning project, designed for human engagement. As it learns, some of its responses are inappropriate and indicative of the types of interactions some people are having with it. We're making some adjustments to Tay.



Reproducing, amplifying social norms?

In an effort to improve transparency in automated marketing distribution, a research group developed a software tool to study digital traceability and found that such marketing practices had a gender bias that mediated well-paid job offers more often to men than to women (Datta et al., 2015).

Amit Datta*, Michael Carl Tschantz, and Anupam Datta

Automated Experiments on Ad Privacy Settings

A Tale of Opacity, Choice, and Discrimination

Abstract: To partly address people's concerns over web tracking, Google has created the Ad Settings webpage to provide information about and some choice over the profiles Google creates on users. We present AdFisher, an automated tool that explores how user behaviors, Google's ads, and Ad Settings interact. AdFisher can run browser-based experiments and analyze data using machine learning and significance tests. Our tool uses a rigorous experimental design and statistical analysis to ensure the statistical soundness of our results. We use AdFisher to find that the Ad Settings was opaque about some features of a user's profile, that it does provide some choice on ads, and that these choices can lead to seemingly discriminatory ads. In particular, we found that visiting webpages associated with substance abuse changed the ads shown but not the settings page. We also found that setting the gender to female resulted in getting fewer instances of an ad related to high paying jobs than setting it to male. We cannot determine who caused these findings due to our limited visibility into the ad ecosystem, which includes Google, advertisers, websites, and users. Nevertheless, these results can form the starting point for deeper investigations by either the companies themselves or by regulatory bodies.

Keywords: blackbox analysis, information flow, behavioral advertising, transparency, choice, discrimination

DOI 10.1515/popets-2015-0007
Received 11/22/2014; revised 2/18/2015; accepted 2/18/2015.

1 Introduction

Problem and Overview. With the advancement of tracking technologies and the growth of online data aggregators, data collection on the Internet has become a serious privacy concern. Colossal amounts of collected data are used, sold, and resold for serving targeted content, notably advertisements, on websites (e.g., [1]). Many websites providing content, such as news, outsource their advertising operations to large third-party ad networks, such as Google's DoubleClick. These networks embed tracking code into webpages across many sites providing the network with a more global view of each user's behaviors.

People are concerned about behavioral marketing on the web (e.g., [2]). To increase transparency and control, Google provides Ad Settings, which is "a Google tool that helps you control the ads you see on Google services and on websites that partner with Google" [3]. It displays inferences Google has made about a user's demographics and interests based on his browsing behavior. Users can view and edit these settings at

http://www.google.com/settings/ads

Yahoo [4] and Microsoft [5] also offer personalized ad settings.

However, they provide little information about how these pages operate, leaving open the question of how completely these settings describe the profile they have about a user. In this study, we explore how a user's behaviors, either directly with the settings or with content providers, alter the ads and settings shown to the user and whether these changes are in harmony. In particular, we study the degree to which the settings provides transparency and choice as well as checking for the presence of discrimination. Transparency is important for people to understand how the use of data about them affects the ads they see. Choice allows users to control how this data gets used, enabling them to protect the information they find sensitive. Discrimination is an increasing concern about machine learning systems and one reason people like to keep information private [6, 7].

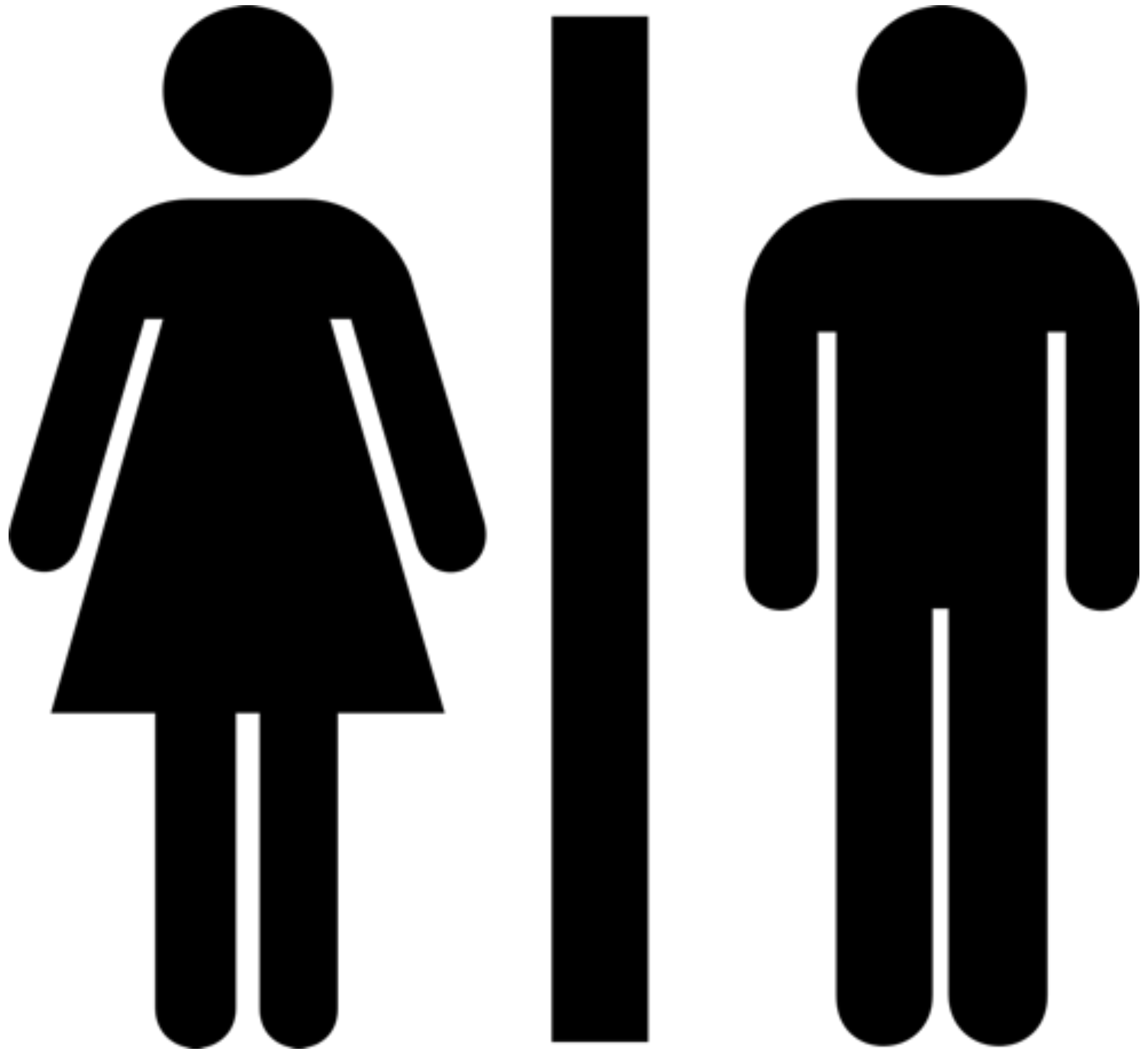
To conduct these studies, we developed AdFisher, a tool for automating randomized, controlled experiments for studying online tracking. Our tool offers a combination of automation, statistical rigor, scalability, and explanation for determining the use of information by web advertising algorithms and by personalized ad settings, such as Google Ad Settings. The tool can simulate having a particular interest or attribute by visiting web-

***Corresponding Author: Amit Datta:** Carnegie Mellon University, E-mail: amitdatta@cmu.edu
Michael Carl Tschantz: International Computer Science Institute, E-mail: mct@icsi.berkeley.edu
Anupam Datta: Carnegie Mellon University, E-mail: danupam@cmu.edu

Gender

- 2016: Two prominent research-*image collections* were found to display a predictable **gender bias** in their depiction of activities such as cooking and sports.
- Machine-learning software trained on the datasets didn't just mirror those biases, it **amplified** them.

Cf. Zhao et al, 2017



Normative design

Should AI reproduce the world as it is or as we want it to be?

Sum

- **EXPANDED USE, HIGHER STAKES**: AI increases on consumer markets, in medicine and public institutions, with higher stakes.

- **NORMATIVE DESIGN(ers)**: Should AI reproduce the world as it is or as we wish it to be? What norms should guide?

- **MULTIDISCIPLINARY NEEDS**: Applied AI interacts, reproduces and amplifies cultures, norms and leads to legal, ethical questions. “No quick fix to bias”.

- **TRANSPARENCY LINKED TO ACCOUNTABILITY LINKED TO TRUST**. Explainability needs to be places in contexts, languages, markets too.

stefan.larsson@lth.lu.se

@DigitalSocietyL

Mer: http://portal.research.lu.se/portal/en/persons/stefan-larsson(2e0f375a-0fea-47c7-bbe9-fd33a1d631a1).html