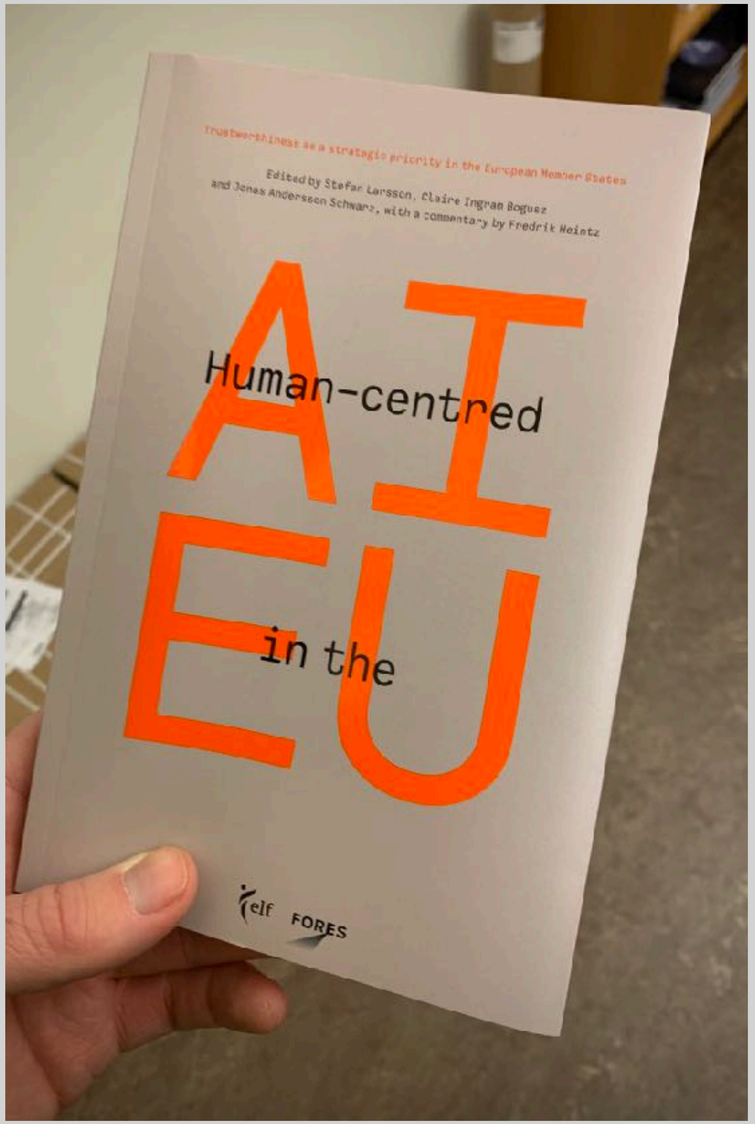
Ethics and AI:
Platforms, fairness & governance



Stefan Larsson

Associate Professor in Technology and Social Change at the Department of Technology and Society at Lund University, Sweden

Lawyer (LLM)
PhD in sociology of law
PhD in spatial planning

LUNDS UNIVERSITET

Main research projects

- AI Transparency and Consumer Trust (WASP-HS)
  2020–2024
- Framework for Sustainable AI (Vinnova)
  2019–2021
- AIR Lund: Artificially Intelligent use of registers (Vetenskapsrådet) 2019–2023
- Consumers’ trust in retail’s use of data (Handelsrådet)
  2018–2021

Today

1. Why ethics?
2. Contemporary “appliedness”: Platforms + AI
3. AI & fairness: Increased awareness
4. Two types of fairness issues: skewed data / skewed world
5. Governance ideas, based on this awareness: Ethics guidelines
6. Ethics guidelines in the EU: becoming formalised
7. A checklist

Break - breakout sessions

Larsson, S. (2020) On the Governance of Artificial Intelligence through Ethics Guidelines, Asian Journal of Law and Society.
Larsson, S., Ingram Bogusz, C., & Andersson Schwarz, J. Eds. (2020) Human-Centred AI in the EU. Trustworthiness as a strategic priority in the European Member States. Brussels: European Liberal Forum.
Larsson, S. & Heintz, F. (2020) Transparency in artificial intelligence, Internet Policy Review 9(2): 1-16.
Larsson, S. (2019) The Socio-Legal Relevance of Artificial Intelligence, “Law in an Algorithmic World”, Special Issue of Droit et Société. 103(3): 573-593.

Why?

- Formality
- Quality
- Security
- Fairness
- Legality

Formality (1/2): engineering degrees in Sweden

For the degree of Master of Science in Engineering (civilingenjör), the student must

- demonstrate insight into the possibilities and limitations of technology, its role in society and people's responsibility for how it is used, including social and economic aspects as well as environmental and work environment aspects;

- demonstrate the ability to make assessments with regard to relevant scientific, societal and ethical aspects and demonstrate awareness of ethical aspects of research and development work,

The Higher Education Ordinance (1993:100)

Formality (2/2): this course

Course syllabus

Artificiell intelligens
Artificial Intelligence

EDAP01, 7,5 credits, A (Second Cycle)

Valid for: 2020/21
Decided by: PLED C/D
Date of Decision: 2020-03-30

General Information

Main field: Machine Learning, Systems and Control.
Elective Compulsory for: MMSR1
Elective for: BME4, C4-pv, D4-pv, D4-mai, E4-bg, F4, F4-mai, MSOC2, Pi4-bam
Language of instruction: The course will be given in English

Aim

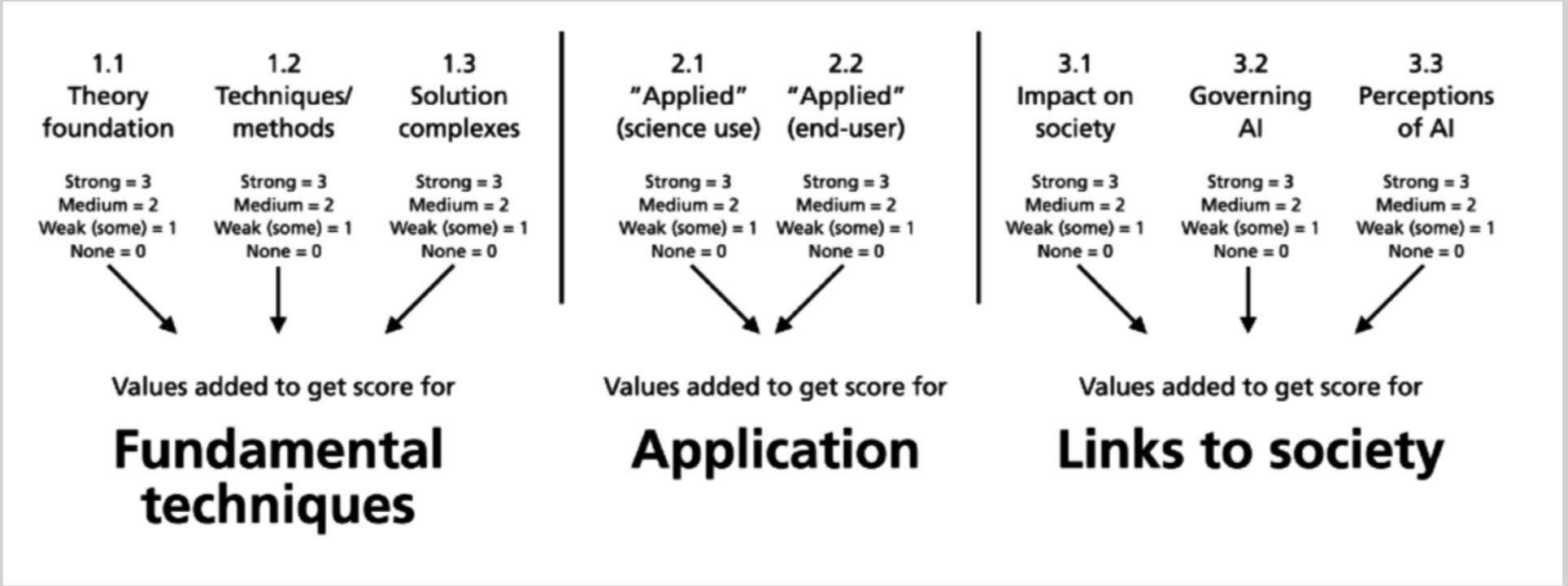
To give an introduction to several subdomains of artificial intelligence and to give an orientation about fundamental methods within these domains. To convey knowledge about breath and depth of the domain. To provide insight about the ethical consequences of AI-based technology.

Judgement and approach
For a passing grade the student must

- demonstrate ability to identify needs for additional knowledge and to continuously develop new skills.
- demonstrate ability to critically judge the ethical and societal consequences of using AI in particular context.

“AI” & “Ethics”

What is AI?



AI-education @ Lund University

Report #1

LUND UNIVERSITY

MAGNUS EKBLAD | MIKAEL SUNDSTRÖM

1.
Contemporary digi: Platforms + AI

Digitisation

Networks and
the internet

Datafication
and platforms

AI & data-
driven agency

Sectorial platforms (van Dijck et al.)

urban transport

news production and dissemination

health

education (edtech)

Other sectors of interest

finance (fintech)

advertising (adtech)

identification and verification

Digitisation

Datafication and platforms

Platform logic

1. Datafication
2. Scalability
3. Automation
4. Centralisation
5. Proprietary
6. Internet-connected
7. Software-based

See Larsson & Andersson Schwarz (eds. 2018) *Developing Platform Economies. A European Policy Landscape*. ELF/Fores.

Some challenges

1. **Transparency is low in automated data-driven ecosystems**

    a. Hard to supervise by authorities / accountability issues / privacy

    b. Consumers can't tell and choose. May be manipulated by predictive tools or "dark patterns".

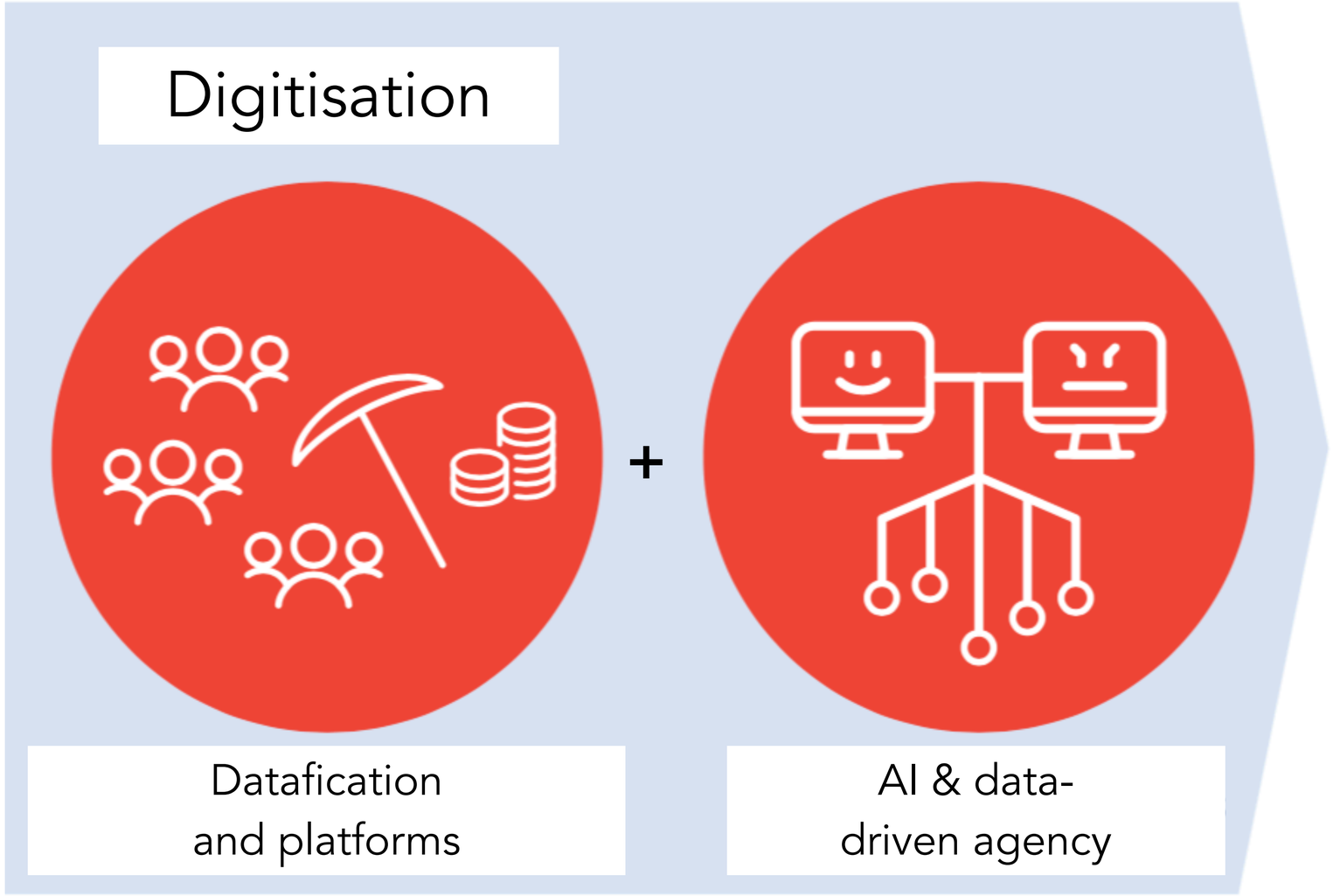
2. **"Code is law":** Platforms regulate /enable / govern their domain = power

    a. e.g. self-preferencing: market-maker/ infrastructure AND competitor on that market is hard to balance

    b. rating/ ranking

3. **Scale and AI may create risk for unintended outcomes**

    a. **"Commercial surveillance":** cooling effects? / unfairness: scandals / low trust



2.
AI & fairness: Increased awareness

"As the use and impact of autonomous and intelligent systems (A/IS) become pervasive, we need to establish societal and policy guidelines in order for such systems to remain human-centric, serving humanity's values and ethical principles."

II. General Principles

The ethical and values-based design, development, and implementation of autonomous and intelligent systems should be guided by the following General Principles:

1. **Human Rights**
   A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.

2. **Well-being**
   A/IS creators shall adopt increased human well-being as a primary success criterion for development.

3. **Data Agency**
   A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.

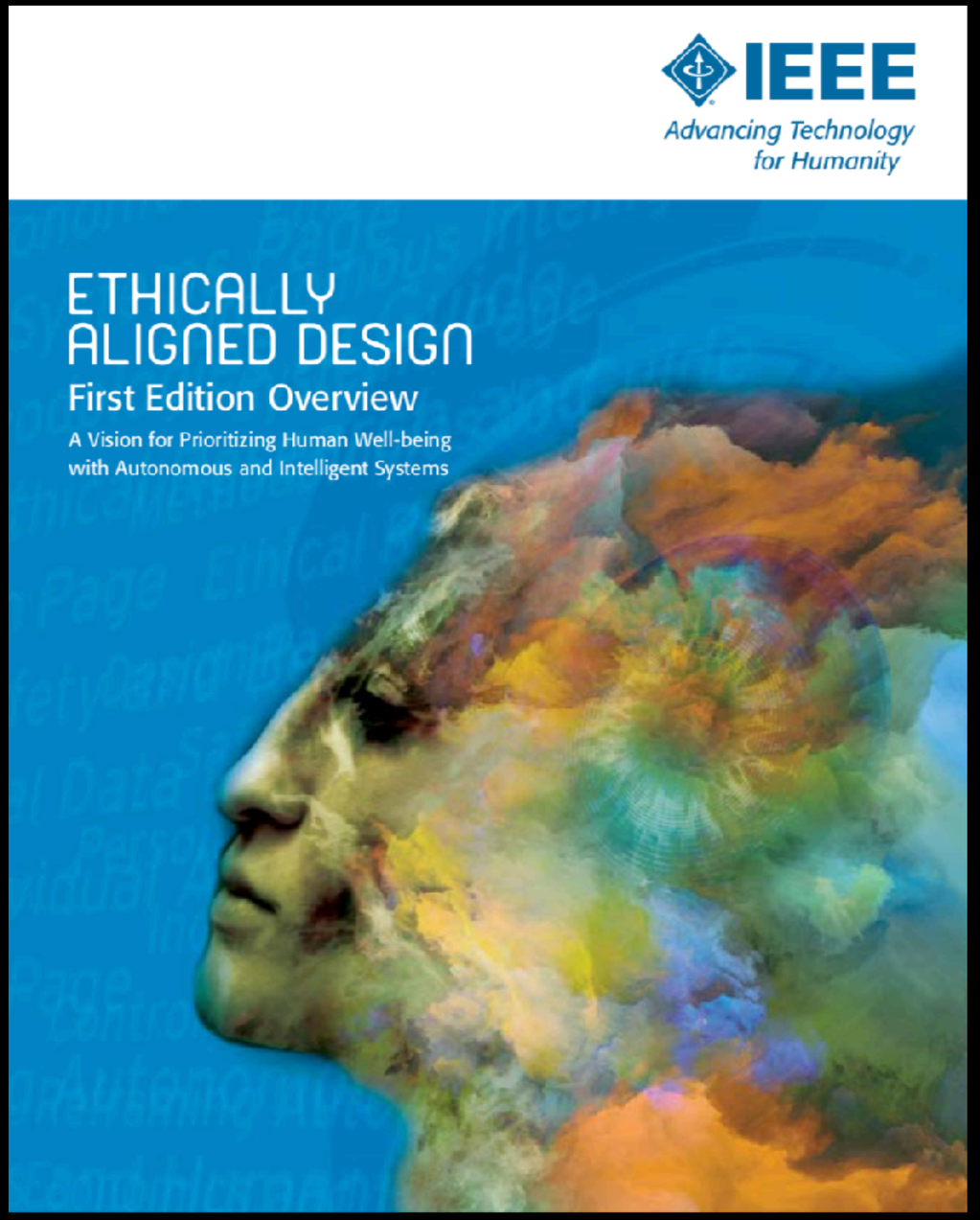
4. **Effectiveness**
   A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.

5. **Transparency**
   The basis of a particular A/IS decision should always be discoverable.

6. **Accountability**
   A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.

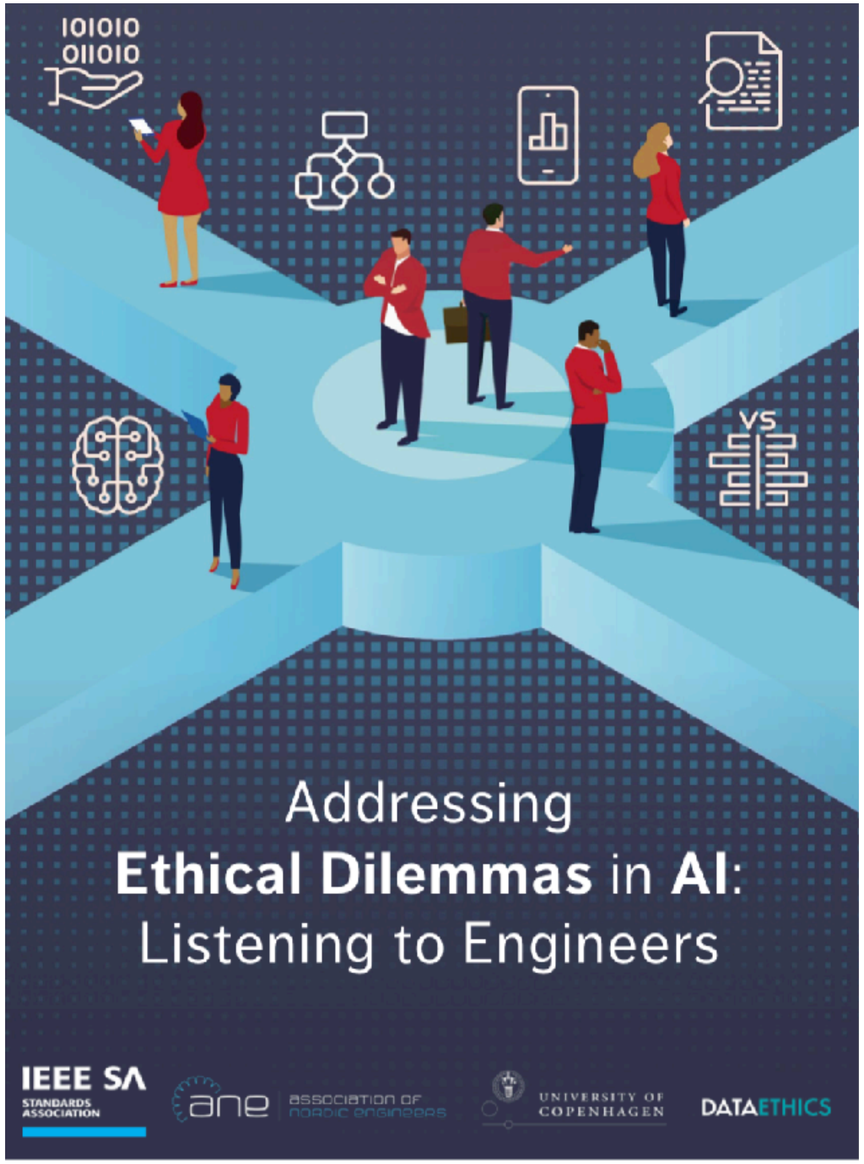
7. **Awareness of Misuse**
   A/IS creators shall guard against all potential misuses and risks of A/IS in operation.

8. **Competence**
   A/IS creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.




AAAI / ACM conference on
ARTIFICIAL INTELLIGENCE,
ETHICS, AND SOCIETY
HOME
PROGRAM
ORGANIZATION
SPONSORS
ARCHIVE
Fourth AAAI /ACM Conference on
Artificial Intelligence,
Ethics, and Society
A virtual conference
May 19-21, 2021


The Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society



2.1.1 EDUCATION & TRAINING SPACES

Some engineers expressed that there is a need for additional education concerning ethics as part of engineering education in general and as additional training as part of life-long learning initiatives. Engineering AI systems is an inherently interdisciplinary endeavour requiring collaboration with people from many different backgrounds. Yet there was a sense that engineers themselves need to take additional responsibility for identifying potential ethical issues and discussing these.

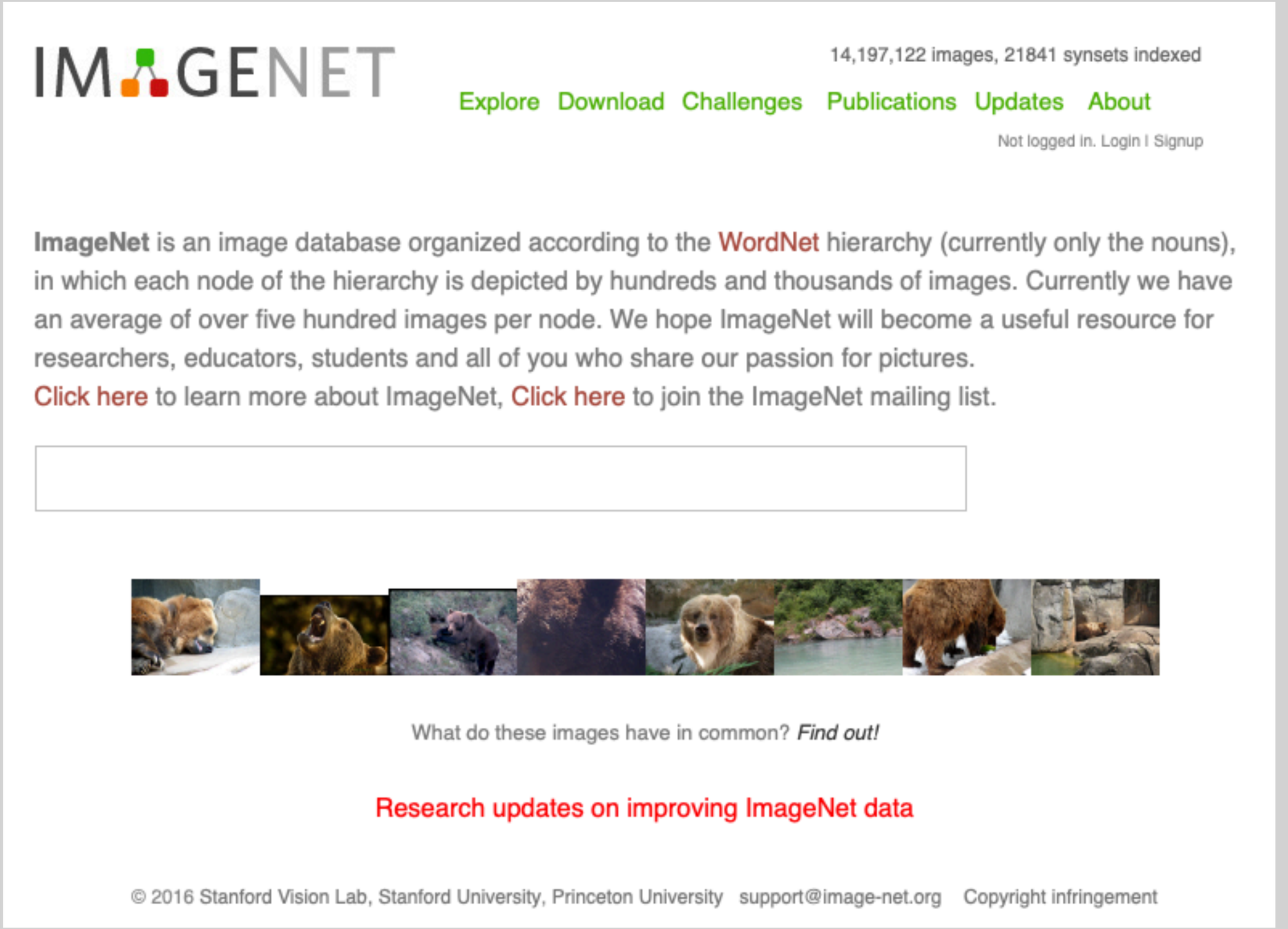
3.
Two types of fairness issues

1.

Biased Data

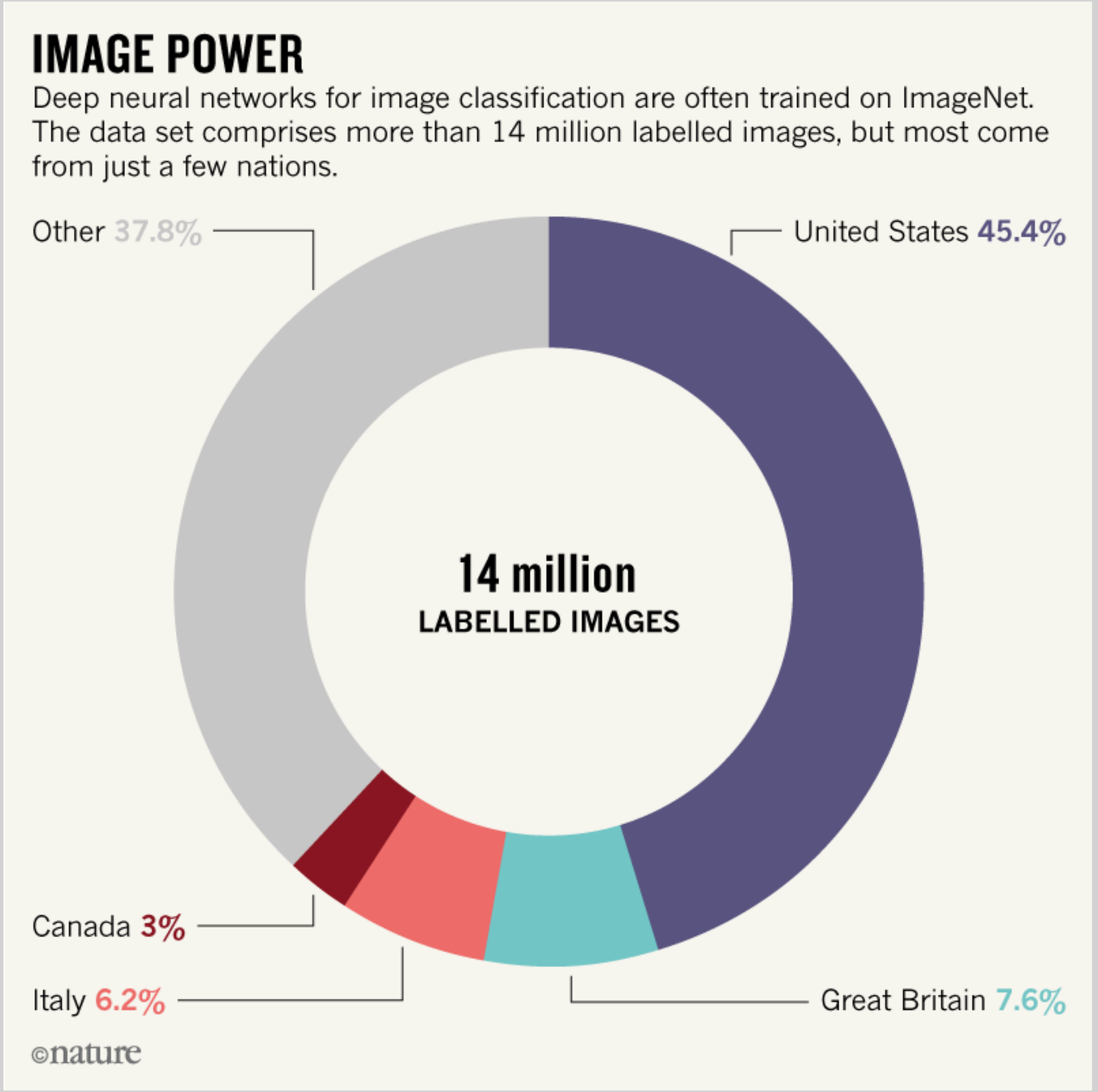
Low precision

In seven years, the winning accuracy in classifying objects in the dataset rose from 71.8% to 97.3%



- US bride dressed in white: ‘bride’, ‘dress’, ‘woman’, ‘wedding’
- North Indian bride: ’performance art’ and ‘costume’
- Suggests:
  - “… amerocentric and eurocentric representation bias”: assess “geo-diversity”
- So: “Less accuracy” for some phenomena

Quality?



Shankar et al., (2017) No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv*:1711.08536.

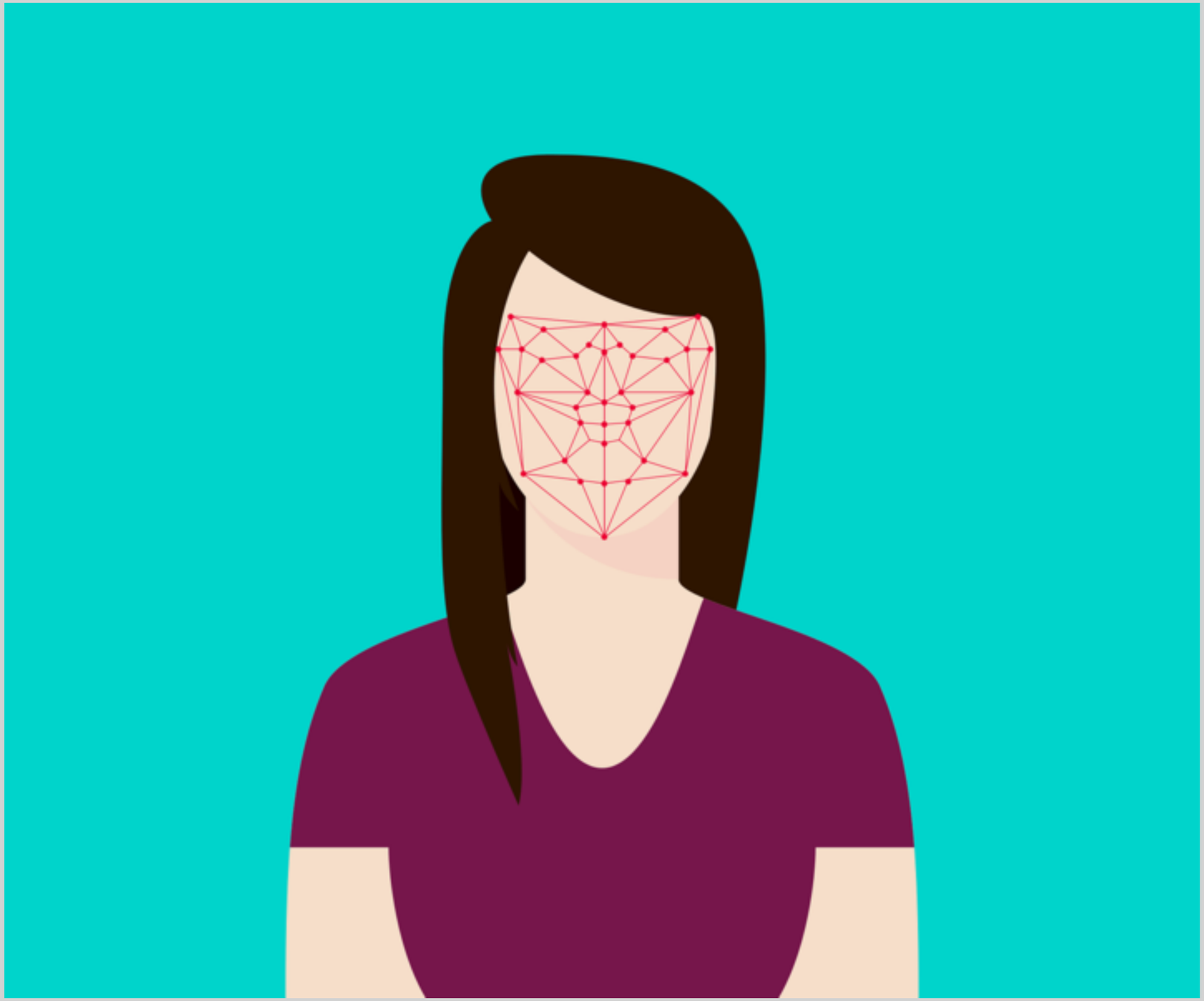
Commercial image recognition services were found to have poorer precision for women and dark skin

(Buolamwini & Gebru, 2018, “Gender shades”)

The precision of systems for detecting pedestrians was different for different skin tones

(Wilson et al 2019)

Quality, security, fairness, legality. note: stakes.



2.

The interaction with society

“Normative” automation

- To improve the transparency of automated marketing, a research team developed a software tool to study digital traceability;

- They found that a common such tool showed a gender bias, i.e. more often conveyed well-paid jobs to men than to women

Complexity, transparency, unfairness.

Datta et al., (2015) Automated Experiments on Ad Privacy Settings – A Tale of Opacity, Choice, and Discrimination".
Proceedings on Privacy Enhancing Technologies. 1: 92–112,

DE GRUYTER OPEN Proceedings on Privacy Enhancing Technologies 2015; 2015 (1):92–112

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

- Chatbot Tay becomes anti-semitic, racist, anti-feminist and pays tribute to Hitler (2016)
- Facebook’s auto-generated ad categories including Jew hatred (2017)
- Commercial image databases contained gender bias, with kitchen attributes associated with women and hunting and sports for men (Zhao et al., 2017)

Mirroring, reproducing, amplifying (unfair) societal structures

Quick sum

1. **You *must*** be able to assess ethical and societal considerations.
    A. **Not only** about fairness (and who defines what's fair - you?), also about product quality and security (and law).
2. **Applied AI:** Platforms and data-diven markets meets AI development
3. **Normative challenges:** the mirroring of an unjust society? (not "solvable"): accountability, transparency

4.
Guidelines

Asian Journal of Law and Society, 00 (2020), pp. 1–15
doi:10.1017/als.2020.19

On the Governance of Artificial Intelligence through Ethics Guidelines•

Stefan LARSSON*
Lund University

Abstract

This article uses a socio-legal perspective to analyze the use of ethics guidelines as a governance tool in the development and use of artificial intelligence (AI). This has become a central policy area in several large jurisdictions, including China and Japan, as well as the EU, focused on here. Particular emphasis in this article is placed on the Ethics Guidelines for Trustworthy AI published by the EU Commission's High-Level Expert Group on Artificial Intelligence in April 2019, as well as the White Paper on AI, published by the EU Commission in February 2020. The guidelines are reflected against partially overlapping and already-existing legislation as well as the ephemeral concept construct surrounding AI as such. The article concludes by pointing to (1) the challenges of a temporal discrepancy between technological and legal change, (2) the need for moving from principle to process in the governance of AI, and (3) the multidisciplinary needs in the study of contemporary applications of data-dependent AI.

Keywords: AI governance, Ethics Guidelines for Trustworthy AI, EU Commission, transparency in AI, AI and law

Tenets
Partnership on AI

Asilomar AI Principles
Future of Life Institute

AI Policy Principles
ITI

Top 10 Principles for Ethical AI
UNI Global Union

Microsoft AI Principles
Microsoft

2016 — 2017 — 2018

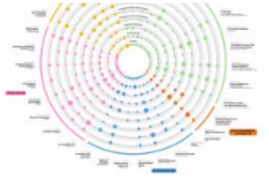
Sep · Oct · Jan · Apr · Oct · Dec · Jan · Feb · Mar

Preparing for the Future of AI
U.S. National Science and Technology Council

Six Principles of AI
Tencent Institute

White Paper on AI Standardization
Standards Administration of China

- Civil Society
- Government
- Inter-governmental Organization
- Multistakeholder
- Private Sector

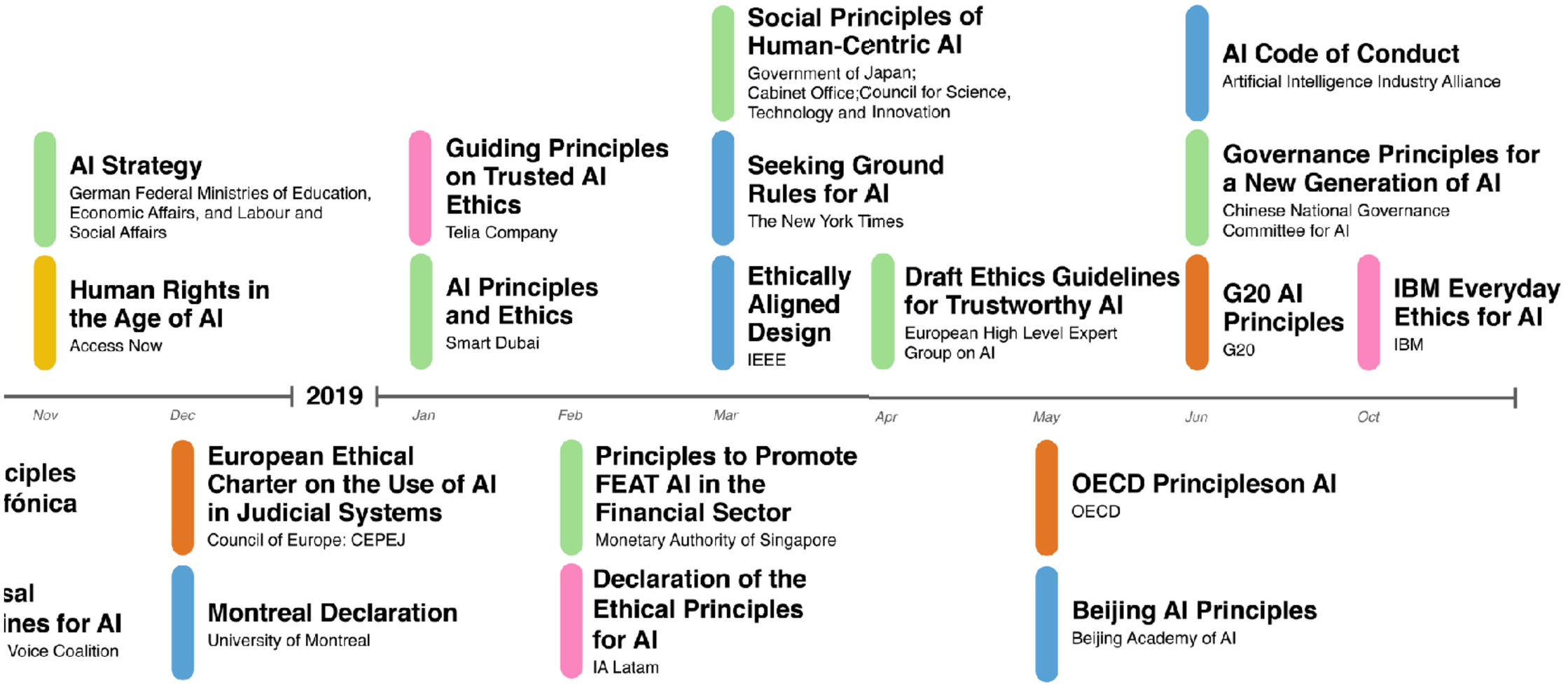


Principled Artificial Intelligence

MAPPING CONSENSUS IN ETHICAL AND RIGHTS-BASED APPROACHES TO PRINCIPLES FOR AI

Published Jan 15, 2020


For a Meaningful AI
Mission assigned by the French Prime Minister
Toronto Declaration
Amnesty International I Access Now
Future of Work and Education for the Digital Age
T20: Think20
AI Strategy
German Federal Ministries of Education, Economic Affairs, and Labour and Social Affairs
Human Rights in the Age of AI
Access Now
Guiding Principles on Trusted AI Ethics
Telia Company
AI Principles and Ethics
Smart Dubai
2019
Mar
Apr
May
Jun
Jul
Oct
Nov
Dec
Jan
Feb
AI in the UK
UK House of Lords
AI for Europe
European Commission
AI in Mexico
British Embassy in Mexico City
National Strategy for AI
Niti Aayog
AI at Google: Our Principles
Google
AI Principles of Telefónica
Telefónica
Universal Guidelines for AI
The Public Voice Coalition
European Ethical Charter on the Use of AI in Judicial Systems
Council of Europe: CEPEJ
Montreal Declaration
University of Montreal
Principles FEAT AI i Financial
Monetary Autho
Declaratic Ethical Pr for AI
IA Latam


Social Principles of Human-Centric AI
Government of Japan; Cabinet Office;Council for Science, Technology and Innovation
AI Code of Conduct
Artificial Intelligence Industry Alliance
AI Strategy
German Federal Ministries of Education, Economic Affairs, and Labour and Social Affairs
Guiding Principles on Trusted AI Ethics
Telia Company
Seeking Ground Rules for AI
The New York Times
Governance Principles for a New Generation of AI
Chinese National Governance Committee for AI
Human Rights in the Age of AI
Access Now
AI Principles and Ethics
Smart Dubai
Ethically Aligned Design
IEEE
Draft Ethics Guidelines for Trustworthy AI
European High Level Expert Group on AI
G20 AI Principles
G20
IBM Everyday Ethics for AI
IBM
2019
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Oct
ciples
fónica
European Ethical Charter on the Use of AI in Judicial Systems
Council of Europe: CEPEJ
Principles to Promote FEAT AI in the Financial Sector
Monetary Authority of Singapore
OECD Principleson AI
OECD
sal
ines for AI
Voice Coalition
Montreal Declaration
University of Montreal
Declaration of the Ethical Principles for AI
IA Latam
Beijing AI Principles
Beijing Academy of AI

PRINCIPLED ARTIFICIAL INTELLIGENCE

A Map of Ethical and Rights-Based Approaches to Principles for AI

Authors: Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, Madhulika Srikumar

Designers: Arushi Singh (arushisingh.net) and Melissa Axelrod (melissaaxelrod.com)

HOW TO READ:

Date, Location

Document Title

Actor

COVERAGE OF THEMES:

References International Human Rights

Explicitly Adopts Human Rights Framework

The size of each dot represents the percentage of principles in that theme contained in the document. Since the number of principles per theme varies, it's informative to compare dot sizes within a theme but not between themes.

The principles within each theme are:

Privacy:
Privacy
Control over Use of Data
Consent
Privacy by Design
Recommendation for Data Protection Laws
Ability to Restrict Processing
Right to Rectification
Right to Erasure

Accountability:
Accountability
Recommendation for New Regulations
Impact Assessment
Evaluation and Auditing Requirement
Verifiability and Replicability
Liability and Legal Responsibility
Ability to Appeal
Environmental Responsibility
Creation of a Monitoring Body
Remedy for Automated Decision

Safety and Security:
Security
Safety and Reliability
Predictability
Security by Design

Transparency and Explainability:
Explainability
Transparency
Open Source Data and Algorithms
Notification when Interacting with an AI
Notification when AI Makes a Decision about an Individual
Regular Reporting Requirement
Right to Information
Open Procurement (for Government)

Fairness and Non-discrimination:
Non-discrimination and the Prevention of Bias
Fairness
Inclusiveness in Design
Inclusiveness in Impact
Representative and High Quality Data
Equality

Human Control of Technology:
Human Control of Technology
Human Review of Automated Decision
Ability to Opt out of Automated Decision

Professional Responsibility:
Multistakeholder Collaboration
Responsible Design
Consideration of Long Term Effects
Accuracy
Scientific Integrity

Promotion of Human Values:
Leveraged to Benefit Society
Human Values and Human Flourishing
Access to Technology

Further information on findings and methodology is available in Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches *(Berkman Klein, 2020) available at cyber.harvard.edu.*

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

- 2016–2019: At least **84 public/private initiatives** have produced statements describing high-level principles, values, and other tenets to guide the ethical development, deployment, and governance of AI

- much convergence around i) transparency, ii) justice and fairness, iii) non-maleficence, iv) responsibility and v) privacy

- But with “substantive divergence” in relation to how these principles are interpreted

See: Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.

nature machine intelligence

PERSPECTIVE

https://doi.org/10.1038/s42256-019-0088-2

The global landscape of AI ethics guidelines

Anna Jobin, Marcello Ienca and Effy Vayena*

In the past five years, private companies, research institutions and public sector organizations have issued principles and guidelines for ethical artificial intelligence (AI). However, despite an apparent agreement that AI should be 'ethical', there is debate about both what constitutes 'ethical AI' and which ethical requirements, technical standards and best practices are needed for its realization. To investigate whether a global agreement on these questions is emerging, we mapped and analysed the current corpus of principles and guidelines on ethical AI. Our results reveal a global convergence emerging around five ethical principles (transparency, justice and fairness, non-maleficence, responsibility and privacy), with substantive divergence in relation to how these principles are interpreted, why they are deemed important, what issue, domain or actors they pertain to, and how they should be implemented. Our findings highlight the importance of integrating guideline-development efforts with substantive ethical analysis and adequate implementation strategies.

Artificial intelligence (AI), or the theory and development of computer systems able to perform tasks normally requiring human intelligence, is widely heralded as an ongoing "revolution" transforming science and society altogether[1,2]. While approaches to AI such as machine learning, deep learning and artificial neural networks are reshaping data processing and analysis[3], autonomous and semi-autonomous systems are being increasingly used in a variety of sectors including healthcare, transportation and the production chain[4]. In light of its powerful transformative force and profound impact across various societal domains, AI has sparked ample debate about the principles and values that should guide its development and use[5,6]. Fears that AI might jeopardize jobs for human workers[7], be misused by malevolent actors[8], elude accountability or inadvertently disseminate bias and thereby undermine fairness[9] have been at the forefront of the recent scientific literature and media coverage. Several studies have discussed the topic of ethical AI[10–15], notably in meta-assessments[11–16] or in relation to systemic risks[17,18] and unintended negative consequences such as algorithmic bias or discrimination[19–21].

National and international organizations have responded to these concerns by developing ad hoc expert committees on AI, often mandated to draft policy documents. These committees include the High-Level Expert Group on Artificial Intelligence appointed by the European Commission, the expert group on AI in Society of the Organisation for Economic Co-operation and Development (OECD), the Advisory Council on the Ethical Use of Artificial Intelligence and Data in Singapore, and the Select Committee on Artificial Intelligence of the UK House of Lords. As part of their institutional appointments, these committees have produced or are reportedly producing reports and guidance documents on AI. Similar efforts are taking place in the private sector, especially among corporations who rely on AI for their business. In 2018 alone, companies such as Google and SAP publicly released AI guidelines and principles. Declarations and recommendations have also been issued by professional associations and non-profit organizations such as the Association of Computing Machinery (ACM), Access Now and Amnesty International. This proliferation of soft-law efforts can be interpreted as a governance response to advanced research into AI, whose research output and market size have drastically increased[22] in recent years.

Reports and guidance documents for ethical AI are instances of what is termed non-legislative policy instruments or soft law[23]. Unlike so-called hard law—that is, legally binding regulations passed by the legislatures to define permitted or prohibited conduct—ethics guidelines are not legally binding but persuasive in nature. Such documents are aimed at assisting with—and have been observed to have significant practical influence on—decision-making in certain fields, comparable to that of legislative norms[24]. Indeed, the intense efforts of such a diverse set of stakeholders in issuing AI principles and policies is noteworthy, because they demonstrate not only the need for ethical guidance, but also the strong interest of these stakeholders to shape the ethics of AI in ways that meet their respective priorities[10,25]. Specifically, the private sector's involvement in the AI ethics arena has been called into question for potentially using such high-level soft policy as a portmanteau to either render a social problem technical[26] or to eschew regulation altogether[20]. Beyond the composition of the groups that have produced ethical guidance on AI, the content of this guidance itself is of interest. Are these various groups converging on what ethical AI should be, and the ethical principles that will determine the development of AI? If they diverge, what are their differences and can these differences be reconciled?

Our Perspective maps the global landscape of existing ethics guidelines for AI and analyses whether a global convergence is emerging regarding both the principles for ethical AI and the suggestions regarding its realization. This analysis will inform scientists, research institutions, funding agencies, governmental and intergovernmental organizations, and other relevant stakeholders involved in the advancement of ethically responsible innovation in AI.

Methods

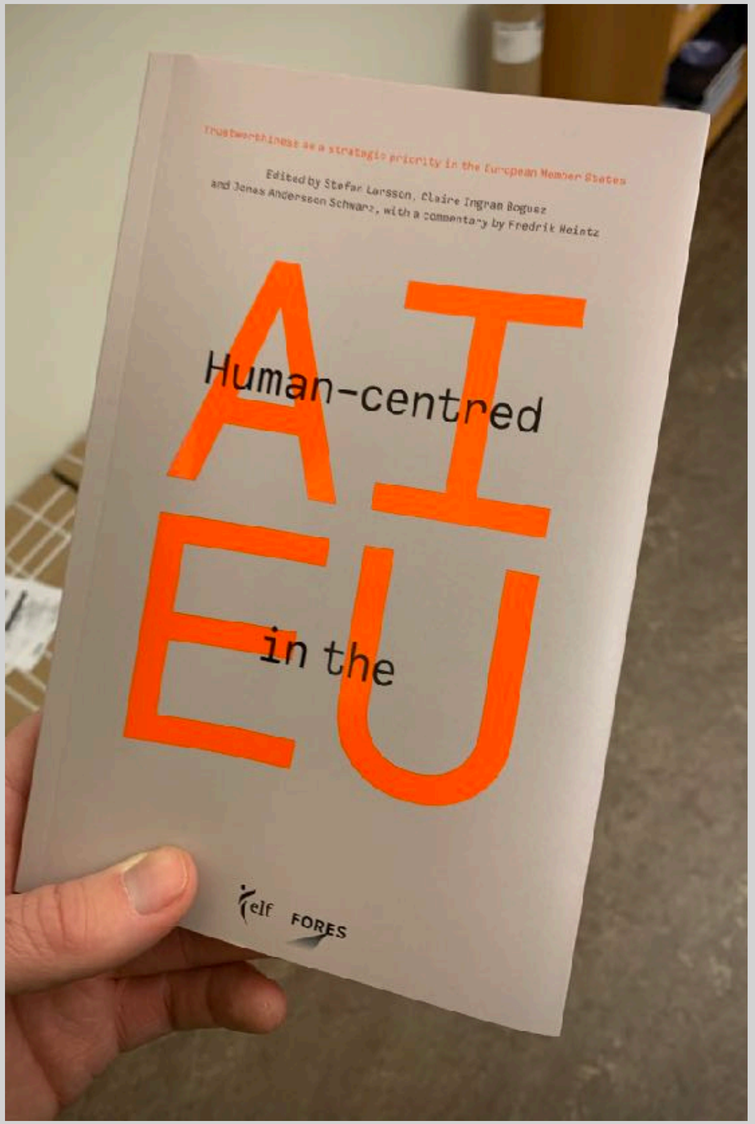
We conducted a scoping review of the existing corpus of documents containing soft-law or non-legal norms issued by organizations. This included a search for grey literature containing principles and guidelines for ethical AI, with academic and legal sources excluded. A scoping review is a method aimed at synthesizing and mapping the existing literature[27] that is considered particularly suitable for complex or heterogeneous areas of research[27,28]. Given the absence of a unified database for AI-specific ethics guidelines, we developed a protocol for discovery and eligibility, adapted from the Preferred

Health Ethics and Policy Lab, Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland. *e-mail: effy.vayena@hest.ethz.ch

5.
AI in the EU: General level

Human-centred AI in the EU

Trustworthiness as a strategic priority in the European Member States



Editors

Stefan Larsson
Lawyer (LLM), senior lecturer and Associate Professor in Technology and Social Change at the Department of Technology and Society at Lund University, Sweden.

Claire Ingram Bogusz
Post-doctoral researcher at the House of Innovation at the Stockholm School of Economics and the Department of Applied IT at the University of Gothenburg, Sweden.

Jonas Andersson Schwarz
Senior lecturer and Associate Professor in Media and Communications at Södertörn University, Stockholm, Sweden.

AI in the EU: Member State strategies and EU-level policies over time

EU

18.04.25
EUROPEAN COMMISSION
Communication Artificial Intelligence for Europe

18.12.07
EUROPEAN COMMISSION
Coordinated Plan on Artificial Intelligence

19.04.08
EUROPEAN COMMISSION
Communication: Building Trust in Human-Centric Artificial Intelligence

19.04.08
AI HLEG
Ethics Guidelines for Trustworthy AI

19.06.26
AI HLEG
Policy and Investment Recommendations for Trustworthy AI

20.02.19
EUROPEAN COMMISSION
White Paper on Artificial Intelligence: a European approach to excellence and trust

20.07.23
AI HLEG
Sectoral Considerations on the Policy and Investment Recommendations for Trustworthy AI

20.07.17
AI HLEG
The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment

2017 | 2018 | 2019 | 2020

Country

Ethics Guidelines for Trustworthy AI (April 2019)

Foundations of Trustworthy AI

Adhere to ethical principles based on fundamental rights

Four Ethical Principles

Acknowledge and address tensions between them

Respect for human autonomy
Prevention of harm
Fairness
Explicability

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non-discrimination, and fairness
6. Societal and environmental wellbeing
7. Accountability

Policy and Investment Recommendations for Trustworthy AI (June 2019)

A. Empowering and Protecting Humans and Society
B. Transforming Europe's Private Sector
C. Europe's Public Sector as a Catalyst of Sustainable Growth and Innovation
D. Ensuring World-Class Research Capabilities
E. Building Data and Infrastructure for AI
F. Generating appropriate Skills and Education for AI
G. Establishing an appropriate governance and regulatory framework
H. Raising Funding and Investment

The White Paper (February 2020)

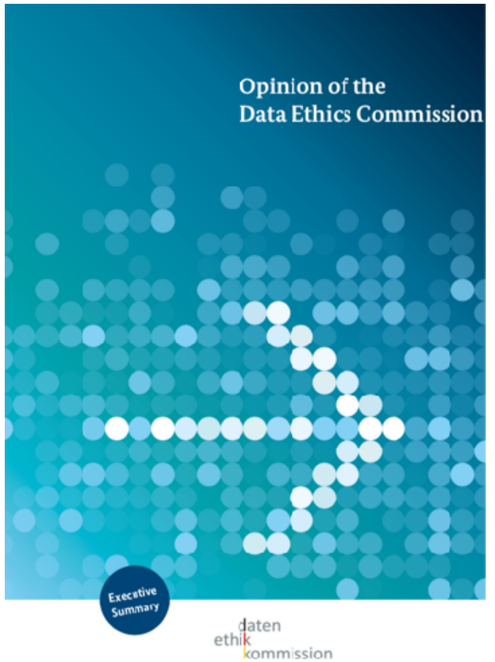
Possible legal improvements

1. **Effective application and enforcement** of existing EU and national regulation. Specifically points to *a lack of transparency* that makes it difficult to identify and prove possible breaches.
2. The limitation of scope of **safety legislation that applies to products and not to services**, and therefore in principle not to services based on AI technology.
3. The **changing functionality of AI systems**, for example for products that rely on frequent software updates of machine learning.
4. The **allocation of responsibilities** at different places in a supply chain.

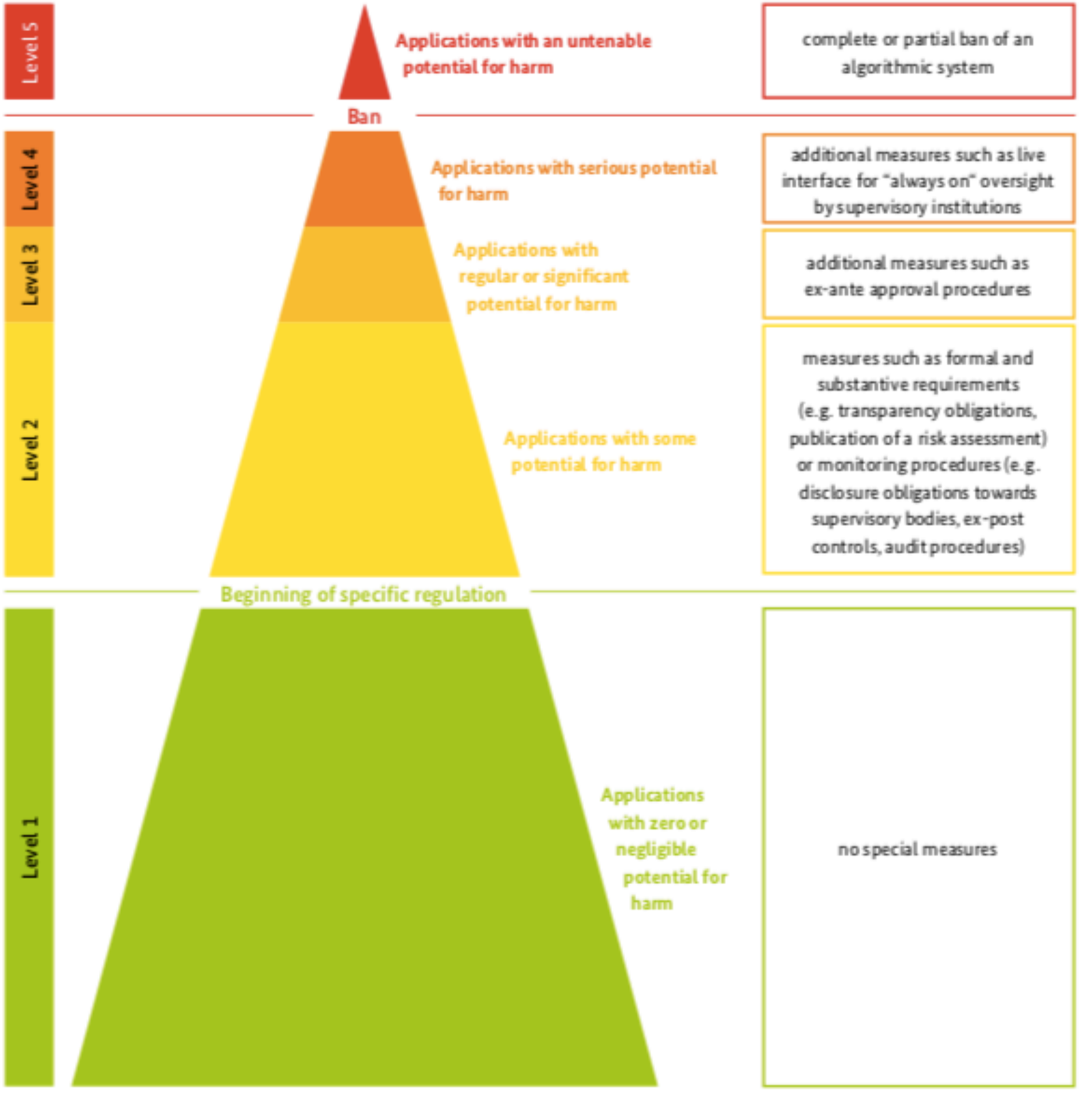
The White Paper
On risk

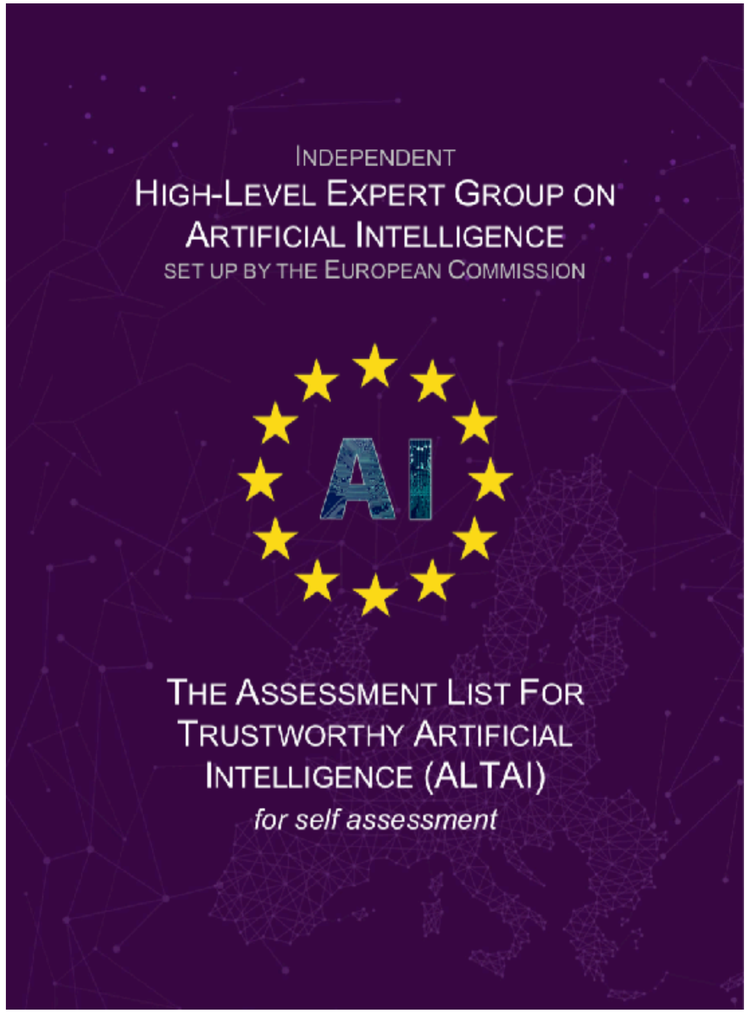
- …**High-risk applications** are distinguished from all other applications; especially pointing to healthcare, transport, energy and parts of the public sector;
- Cumulatively, the AI application would **need to have been used in such a way** that significant risks were likely to arise

- Critique:
  - How about “commercial surveillance”?
  - How about a more levelled approach? (Germany)



daten
ethik
kommission





REQUIREMENT #1 Human Agency and Oversight

- Human Agency and Autonomy
- Human Oversight

REQUIREMENT #2 Technical Robustness and Safety

- Resilience to Attack and Security
- General Safety
- Accuracy
- Reliability, Fall-back plans and Reproducibility

REQUIREMENT #3 Privacy and Data Governance

- Privacy
- Data Governance

REQUIREMENT #4 Transparency

- Traceability
- Explainability
- Communication

REQUIREMENT #5 Diversity, Non-discrimination and Fairness

- Avoidance of Unfair Bias
- Accessibility and Universal Design
- Stakeholder Participation

REQUIREMENT #6 Societal and Environmental Well-being

- Environmental Well-being
- Impact on Work and Skills
- Impact on Society at large or Democracy

REQUIREMENT #7 Accountability

- Auditability
- Risk Management

How to use this Assessment List for Trustworthy AI (ALTAI)

This Assessment List for Trustworthy AI (ALTAI) is best completed involving a multidisciplinary team of people. These could be from within and/or outside your organisation with specific competences or expertise on each of the 7 requirements and related questions. Among the stakeholders you may find for example the following:

- AI designers and AI developers of the AI system;
- data scientists;
- procurement officers or specialists;
- front-end staff that will use or work with the AI system;
- legal/compliance officers;
- management.

REQUIREMENT #1 Human Agency and Oversight

- Human Agency and Autonomy
- Human Oversight

- Is the AI system designed to interact, guide or take decisions by human end-users that affect humans[18] or society?
  - Could the AI system generate confusion for some or all end-users or subjects on whether a decision, content, advice or outcome is the result of an algorithmic decision?
  - Are end-users or other subjects adequately made aware that a decision, content, advice or outcome is the result of an algorithmic decision?

- Please determine whether the AI system (choose as many as appropriate):
  - Is a self-learning or autonomous system;
  - Is overseen by a *Human*-in-the-Loop;
  - Is overseen by a *Human*-on-the-Loop;
  - Is overseen by a *Human*-in-Command.
- Have the humans (human-in-the-loop, human-on-the-loop, human-in-command) been given specific training on how to exercise oversight?
- Did you establish any detection and response mechanisms for undesirable adverse effects of the AI system for the end-user or subject?
- Did you ensure a 'stop button' or procedure to safely abort an operation when needed?
- Did you take any specific oversight and control measures to reflect the self-learning or autonomous nature of the AI system?

REQUIREMENT #2 Technical Robustness and Safety

- Resilience to Attack and Security
- General Safety
- Accuracy
- Reliability, Fall-back plans and Reproducibility

- Could the AI system have adversarial, critical or damaging effects (e.g. to human or societal safety) in case of risks or threats such as design or technical faults, defects, outages, attacks, misuse, inappropriate or malicious use?

- Did you define risks, risk metrics and risk levels of the AI system in each specific use case?
  - Did you put in place a process to continuously measure and assess risks?
  - Did you inform end-users and subjects of existing or potential risks?
- Did you identify the possible threats to the AI system (design faults, technical faults, environmental threats) and the possible consequences?
  - Did you assess the risk of possible malicious use, misuse or inappropriate use of the AI system?
  - Did you define safety criticality levels (e.g. related to human integrity) of the possible consequences of faults or misuse of the AI system?

- Could a low level of accuracy of the AI system result in critical, adversarial or damaging consequences?
- Did you put in place measures to ensure that the data (including training data) used to develop the AI system is up-to-date, of high quality, complete and representative of the environment the system will be deployed in?

REQUIREMENT #3 Privacy and Data Governance

- Privacy
- Data Governance

- Did you consider the impact of the AI system on the right to privacy, the right to physical, mental and/or moral integrity and the right to data protection?
- Depending on the use case, did you establish mechanisms that allow flagging issues related to privacy concerning the AI system?

- Is your AI system being trained, or was it developed, by using or processing personal data (including special categories of personal data)?
- Did you put in place any of the following measures some of which are mandatory under the General Data Protection Regulation (GDPR), or a non-European equivalent?
    - Data Protection Impact Assessment (DPIA)[23];
    - Designate a Data Protection Officer (DPO)[24] and include them at an early state in the development, procurement or use phase of the AI system;
    - Oversight mechanisms for data processing (including limiting access to qualified personnel, mechanisms for logging data access and making modifications);
    - Measures to achieve privacy-by-design and default (e.g. encryption, pseudonymisation, aggregation, anonymisation);

REQUIREMENT #4 Transparency

- Traceability
- Explainability
- Communication

- Did you put in place measures that address the traceability of the AI system during its entire lifecycle?
  - Did you put in place measures to continuously assess the quality of the input data to the AI system?[27]
  - Can you trace back which data was used by the AI system to make a certain decision(s) or recommendation(s)?
- Did you explain the decision(s) of the AI system to the users?[29]
- Do you continuously survey the users if they understand the decision(s) of the AI system?
- In cases of interactive AI systems (e.g., chatbots, robo-lawyers), do you communicate to users that they are interacting with an AI system instead of a human?
- Did you establish mechanisms to inform users about the purpose, criteria and limitations of the decision(s) generated by the AI system?
  - Did you communicate the benefits of the AI system to users?
  - Did you communicate the technical limitations and potential risks of the AI system to users, such as its level of accuracy and/ or error rates?
  - Did you provide appropriate training material and disclaimers to users on how to adequately use the AI system?

REQUIREMENT #5 Diversity, Non-discrimination and Fairness

- Avoidance of Unfair Bias
- Accessibility and Universal Design
- Stakeholder Participation

- Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design?
- Did you consider diversity and representativeness of end-users and/or subjects in the data?

- Did you put in place educational and awareness initiatives to help AI designers and AI developers be more aware of the possible bias they can inject in designing and developing the AI system?
- Did you ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system?

- Is your definition of fairness commonly used and implemented in any phase of the process of setting up the AI system?

- Did you ensure that the AI system corresponds to the variety of preferences and abilities in society?

- Did you consider a mechanism to include the participation of the widest range of possible stakeholders in the AI system's design and development?

REQUIREMENT #6 Societal and Environmental Well-being

- Environmental Well-being
- Impact on Work and Skills
- Impact on Society at large or Democracy

- Are there potential negative impacts of the AI system on the environment?
  - Which potential impact(s) do you identify?
- Where possible, did you establish mechanisms to evaluate the environmental impact of the AI system's development, deployment and/or use (for example, the amount of energy used and carbon emissions)?

- Does the AI system impact human work and work arrangements?

- Could the AI system have a negative impact on society at large or democracy?
  - Did you assess the societal impact of the AI system's use beyond the (end-)user and subject, such as potentially indirectly affected stakeholders or society at large?
  - Did you take action to minimize potential societal harm of the AI system?
  - Did you take measures that ensure that the AI system does not negatively impact democracy?

REQUIREMENT #7 Accountability

- Auditability
- Risk Management

- Did you establish mechanisms that facilitate the AI system's auditability (e.g. traceability of the development process, the sourcing of training data and the logging of the AI system's processes, outcomes, positive and negative impact)?
- Did you ensure that the AI system can be audited by independent third parties?

- Did you foresee any kind of external guidance or third-party auditing processes to oversee ethical concerns and accountability measures?
  - Does the involvement of these third parties go beyond the development phase?
- Did you organise risk training and, if so, does this also inform about the potential legal framework applicable to the AI system?
- Did you consider establishing an AI ethics review board or a similar mechanism to discuss the overall accountability and ethics practices, including potential unclear grey areas?

Conclusions

1. **Contemporary AI ethics is much about governance:** Transparency, Accountability and Fairness
2. **Challenge:** From principles to process.
3. **Definitional relevancy**: What risk? …and what is AI?
4. **AI & society**: the interplay needs to be understood.
5. **Why “ethics”? consider:** quality, security - on top of fairness, legality (and formalitites).

Ethics and AI:
Platforms, fairness & governance

Thank you!

Stefan Larsson
Associate Professor in Technology and Social Change at the Department of Technology and Society at Lund University, Sweden

Lawyer (LLM)
PhD in sociology of law
PhD in spatial planning

LUNDS UNIVERSITET