DATA-DRIVEN AI VS. MODEL-DRIVEN AI: WHICH ONE SHOULD WE TRUST MORE?

Invited talk at:

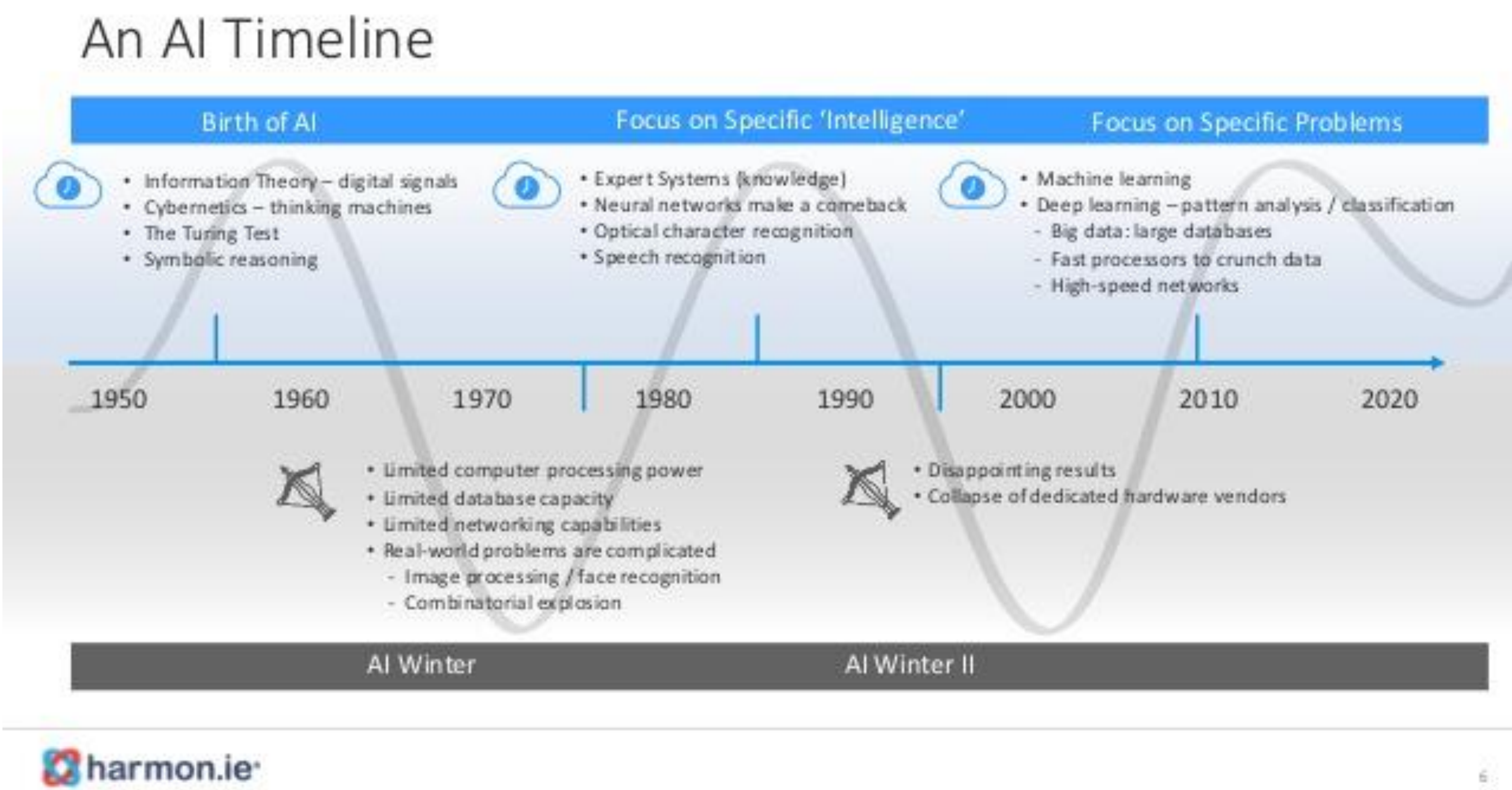
11th Ws on Linked Data in Architecture and Construction (LDAC 2023)

Matera - June 16, 2023

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO
DIPARTIMENTO DI INFORMATICA

Prof. Francesca A. LISI
francesca.lisi@uniba.it

HIGHS & LOWS OF AI



MIND OR BRAIN?

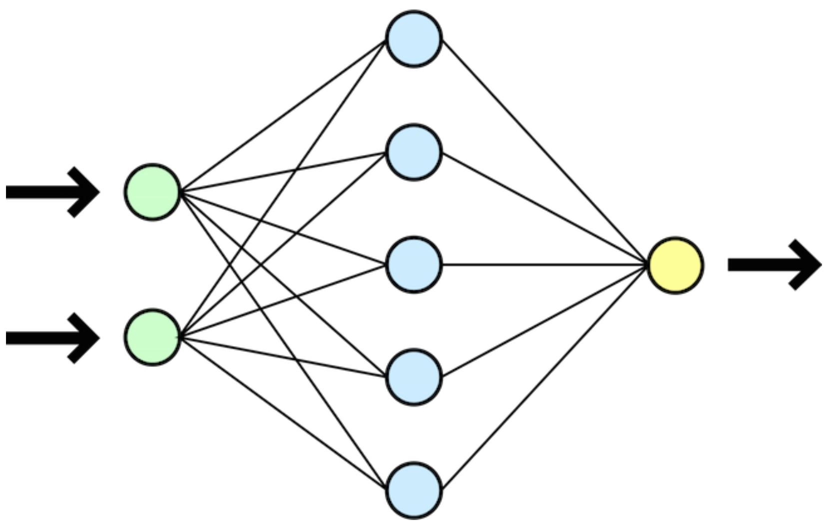
Automated Theorem Proving

(Davis, 1954; Robinson, 1965)

$$\begin{array}{rl} & p \\ & p \rightarrow q \\ \hline \therefore & q \end{array}$$

Artificial Neural Networks

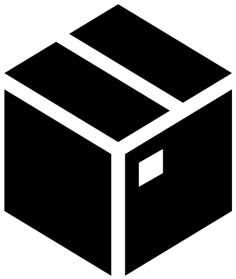
(McCulloch & Pitts, 1943; Rosenblatt, 1958)



OPAQUE OR TRANSPARENT?

Black-box systems

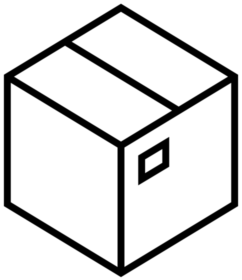
- Data-driven
- Biased
- Subsymbolic



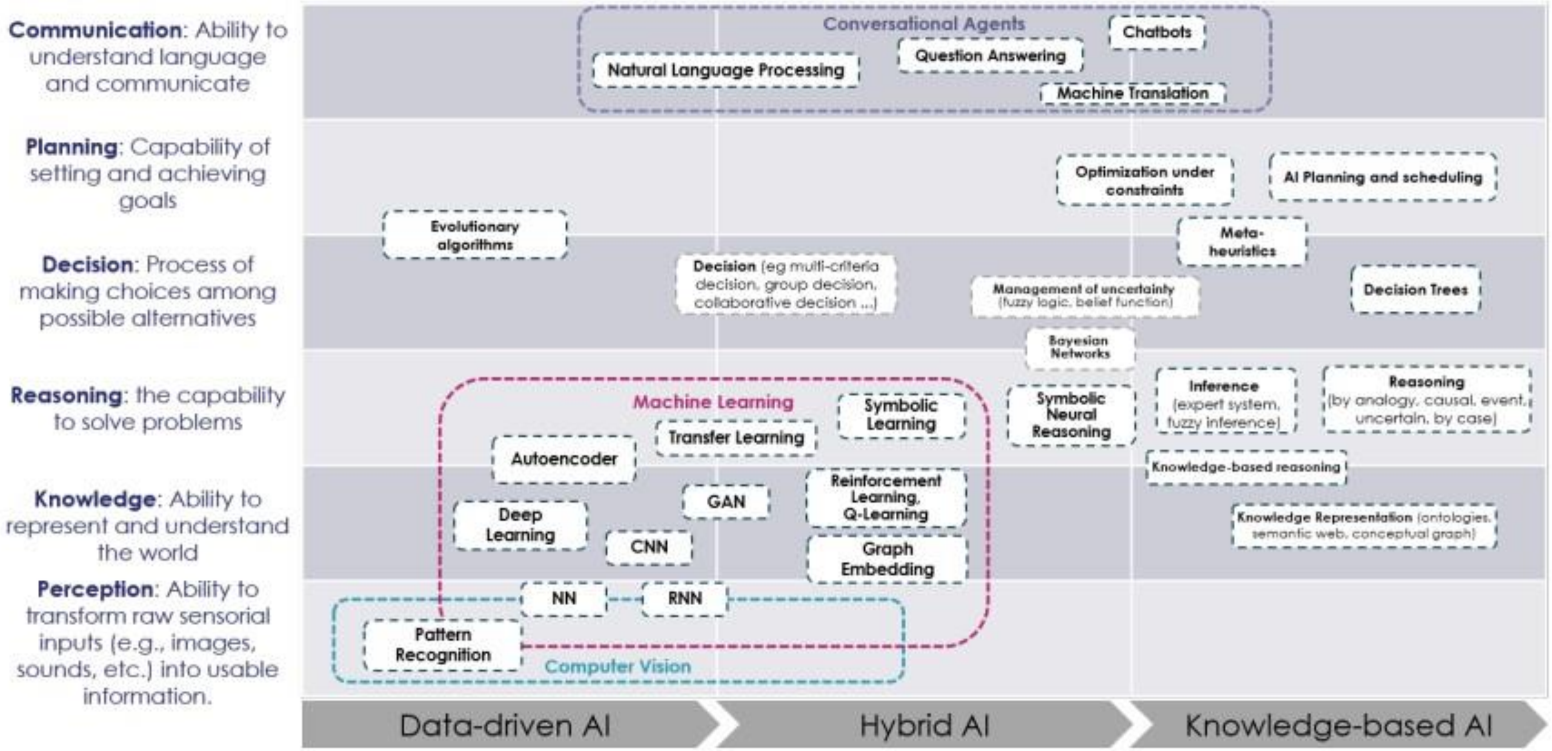
(e.g., deep learning)

White-box systems

- Model-driven
- Explainable
- Symbolic



(e.g., logical reasoning)



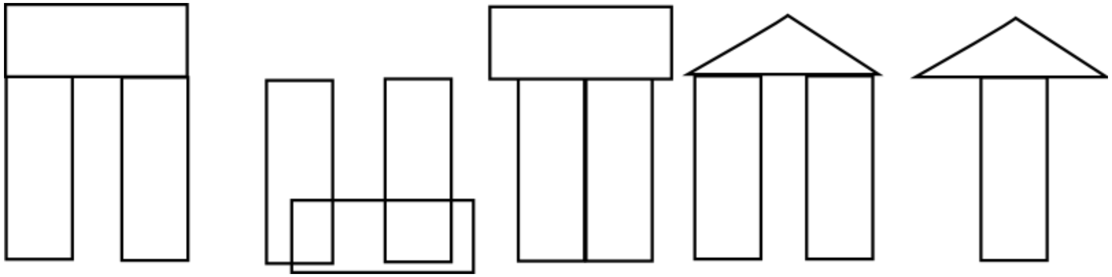
AI PARADIGMS

Source: "Combining Data-Driven and Knowledge-Based AI Paradigms for Engineering AI-Based Safety-Critical Systems"

https://ceur-ws.org/Vol-3087/paper_40.pdf

AN EXAMPLE: HOW TO RECOGNIZE AN ARCH?

- Data-driven: Deep learning from tons of images of arches
- Model-driven: Automated reasoning from a structured knowledge representation of the concept of arch (e.g., one horizontal element on top of two vertical elements)
- Hybrid: Symbolic learning from structured representations of arches



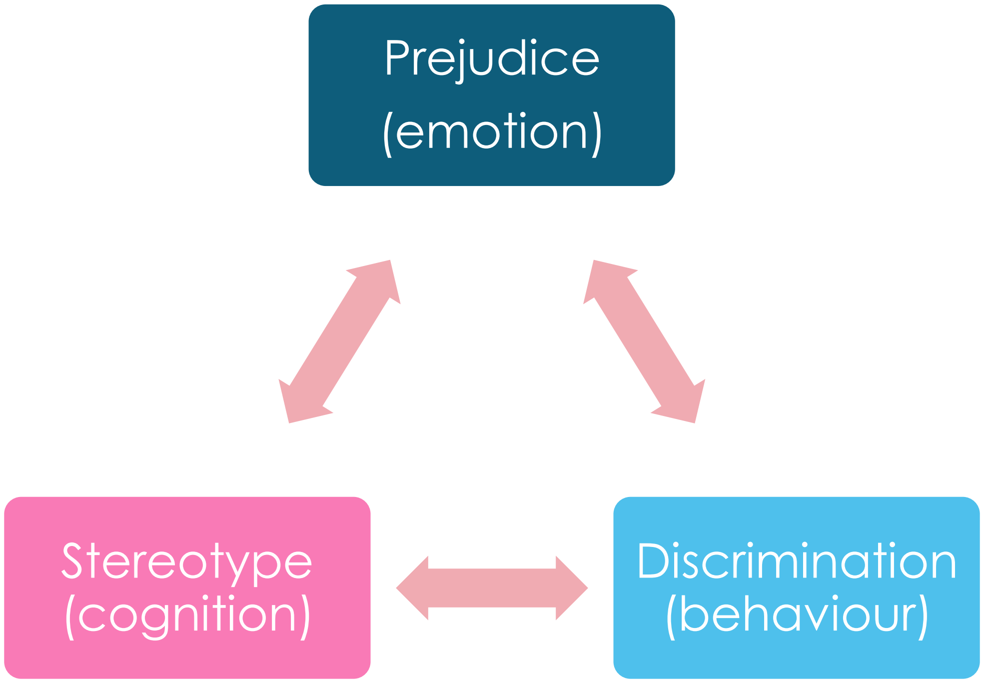
AI IN DECISION MAKING

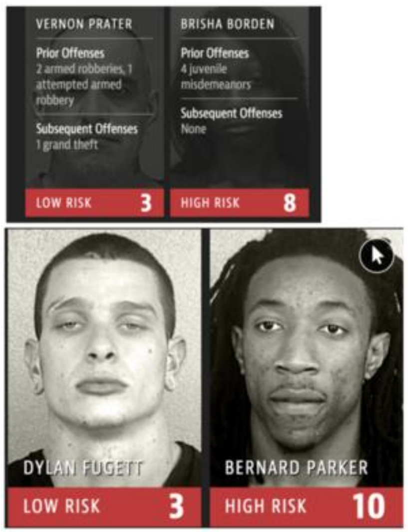
Should we rely on data? Should we rely on models?

THE VICIOUS CIRCLE OF BIASES

« AI can **amplify** discrimination and biases, such as **gender or racial discrimination**, because those are present in the data the technology is trained on, reflecting people's behaviour. »

Yoshua Bengio, 2019



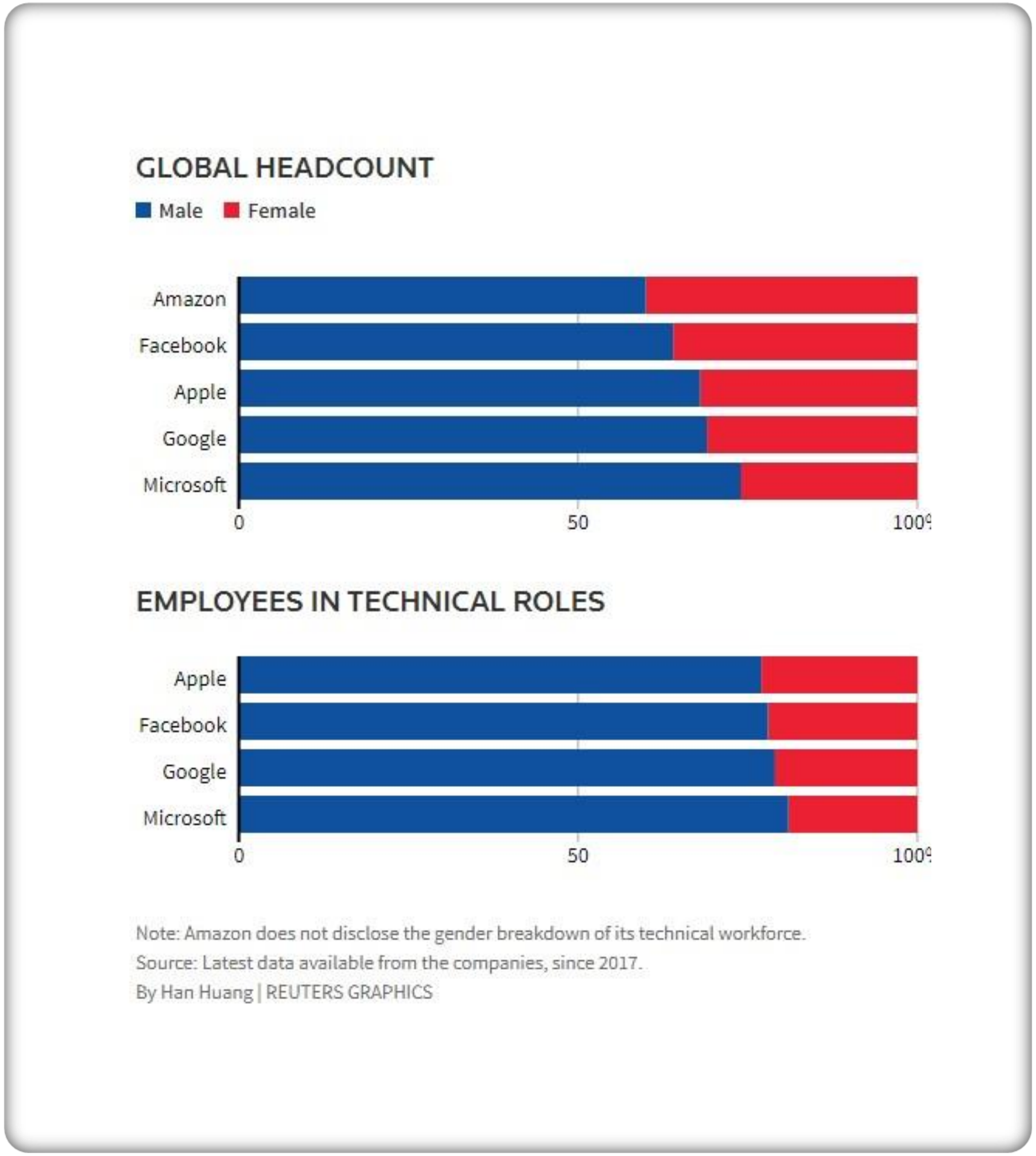


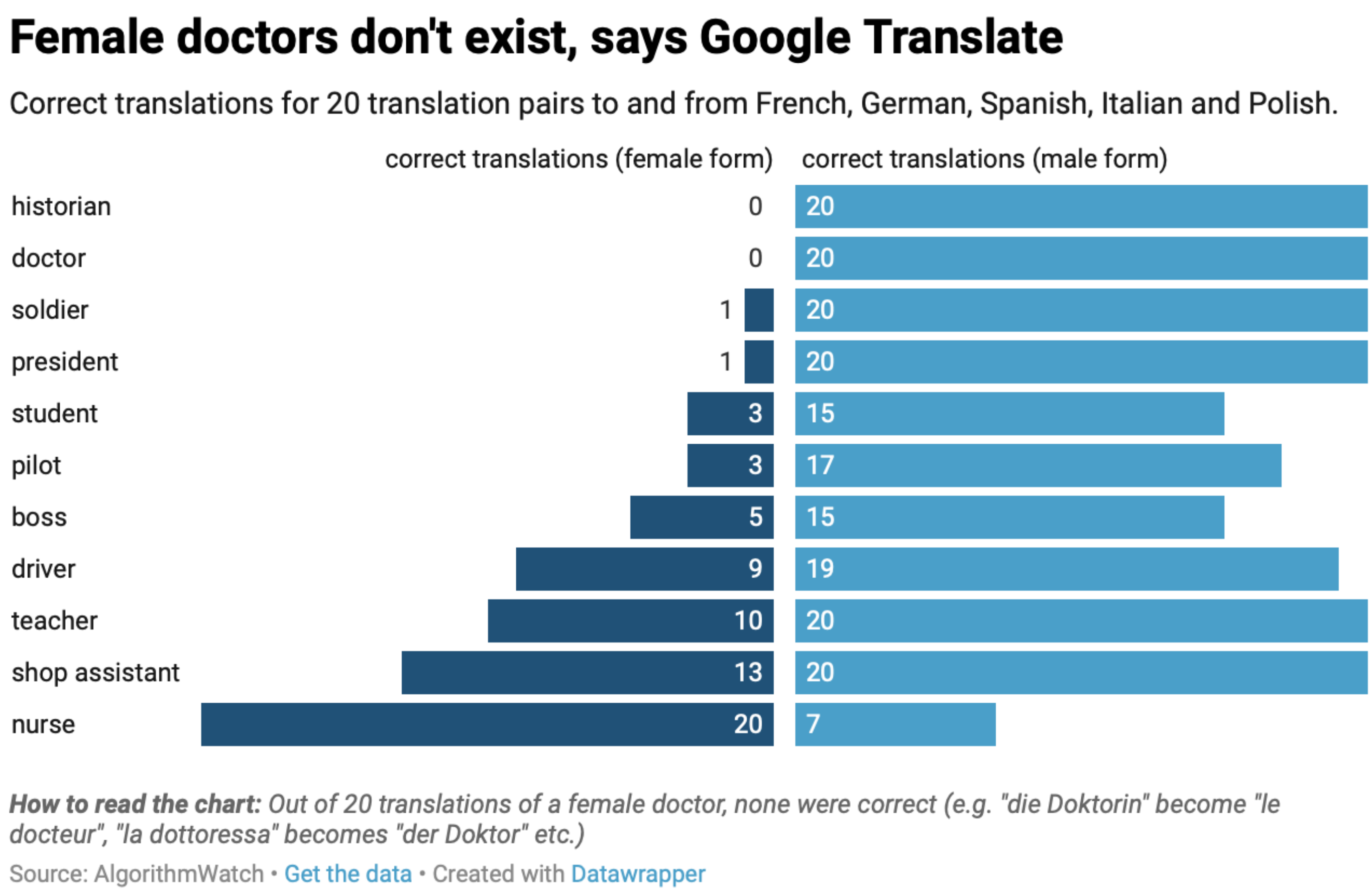
RISKY AI APPLICATIONS

Predictive justice

RISKY AI APPLICATIONS

Personnel recruiting

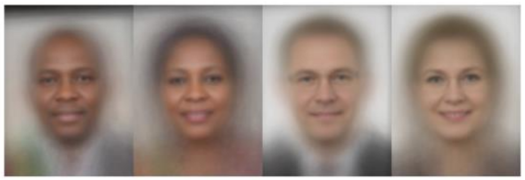


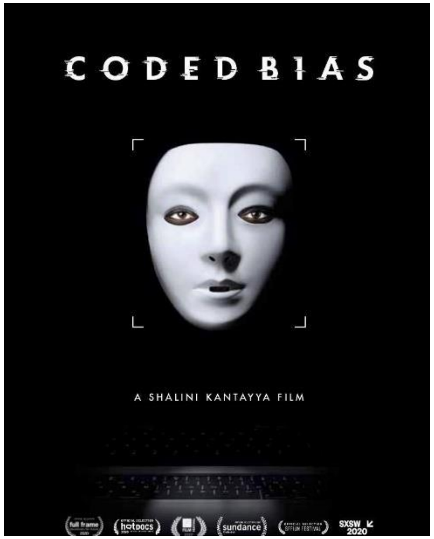


RISKY AI APPLICATIONS

Machine translation

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

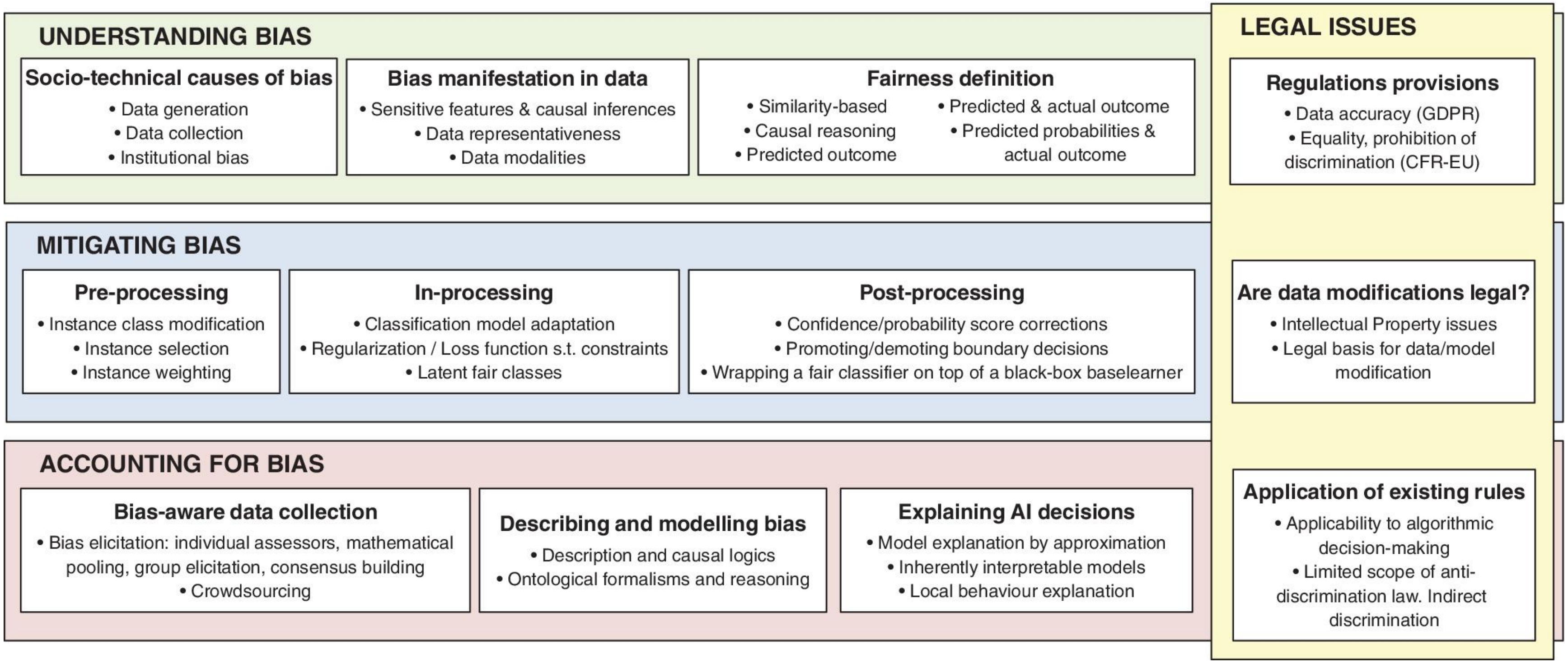




RISKY AI APPLICATIONS

Face recognition

BIAS IN DATA-DRIVEN AI



Source:
«Bias in data-driven artificial intelligence systems—An introductory survey»
https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.1356

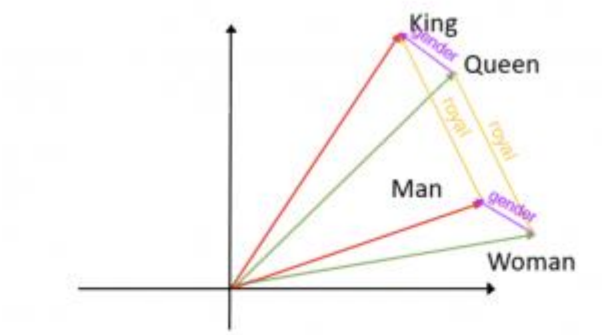
BIAS MITIGATION IN AI

Data debiasing

- Corrective actions on data, e.g. on image datasets used in CV

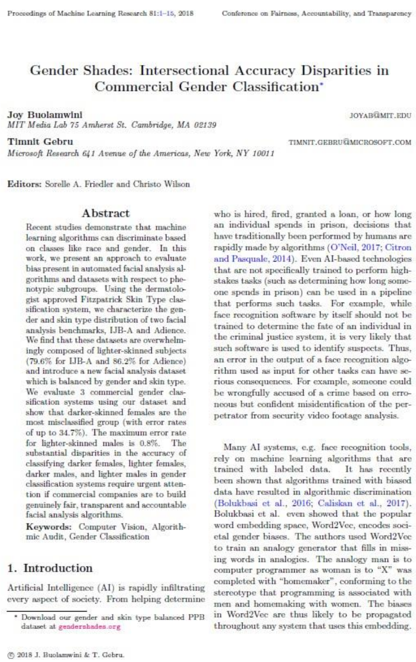
Model debiasing

- Corrective actions on the model, e.g. on word embeddings used in NLP



THE PROJECT *GENDER SHADES*

- Contrast to intersectional bias
- Delivery of a dataset representative of the wide variety of human faces



Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139

Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

Recent studies demonstrate that machine learning algorithms can discriminate based on classes like race and gender. In this work, we present an approach to evaluate bias present in automated facial analysis algorithms and datasets with respect to phenotypic subgroups. Using the dermatologist approved Fitzpatrick Skin Type classification system, we characterize the gender and skin type distribution of two facial analysis benchmarks, IJB-A and Adience. We find that these datasets are overwhelmingly composed of lighter-skinned subjects (79.6% for IJB-A and 86.2% for Adience) and introduce a new facial analysis dataset which is balanced by gender and skin type. We evaluate 3 commercial gender classification systems using our dataset and show that darker-skinned females are the most misclassified group (with error rates of up to 34.7%). The maximum error rate for lighter-skinned males is 0.8%. The substantial disparities in the accuracy of classifying darker females, lighter females, darker males, and lighter males in gender classification systems require urgent attention if commercial companies are to build genuinely fair, transparent and accountable facial analysis algorithms.

Keywords: Computer Vision, Algorithmic Audit, Gender Classification

1. Introduction

Artificial Intelligence (AI) is rapidly infiltrating every aspect of society. From helping determine who is hired, fired, granted a loan, or how long an individual spends in prison, decisions that have traditionally been performed by humans are rapidly made by algorithms (O'Neil, 2017; Citron and Pasquale, 2014). Even AI-based technologies that are not specifically trained to perform high-stakes tasks (such as determining how long someone spends in prison) can be used in a pipeline that performs such tasks. For example, while face recognition software by itself should not be trained to determine the fate of an individual in the criminal justice system, it is very likely that such software is used to identify suspects. Thus, an error in the output of a face recognition algorithm used as input for other tasks can have serious consequences. For example, someone could be wrongfully accused of a crime based on erroneous but confident misidentification of the perpetrator from security video footage analysis.

Many AI systems, e.g. face recognition tools, rely on machine learning algorithms that are trained with labeled data. It has recently been shown that algorithms trained with biased data have resulted in algorithmic discrimination (Bolukbasi et al., 2016; Caliskan et al., 2017). Bolukbasi et al. even showed that the popular word embedding space, Word2Vec, encodes societal gender biases. The authors used Word2Vec to train an analogy generator that fills in missing words in analogies. The analogy man is to computer programmer as woman is to "X" was completed with "homemaker", conforming to the stereotype that programming is associated with men and homemaking with women. The biases in Word2Vec are thus likely to be propagated throughout any system that uses this embedding.

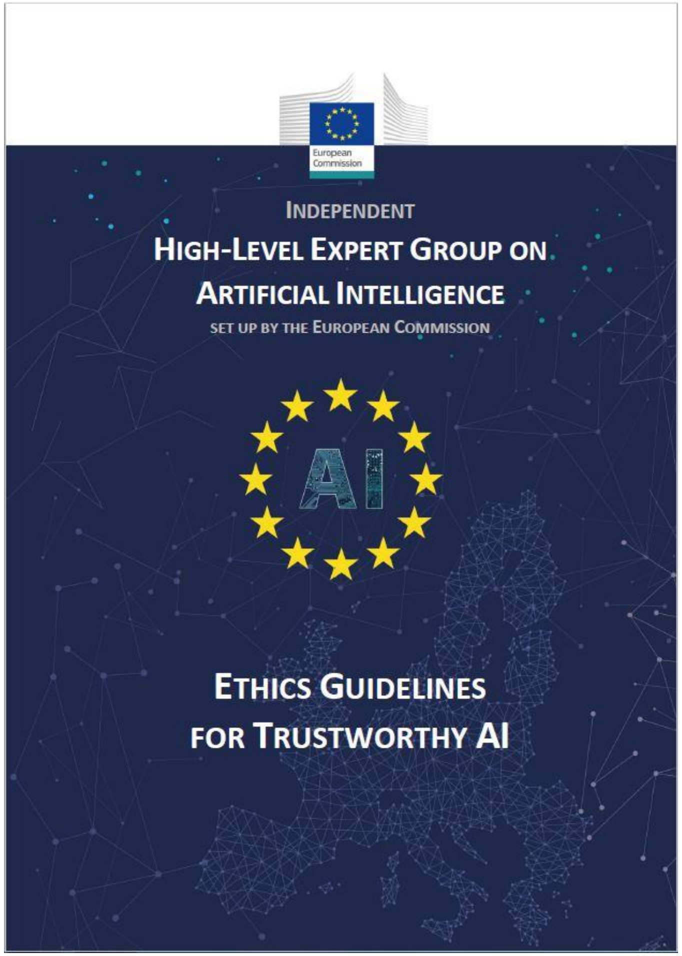
* Download our gender and skin type balanced PPB dataset at gendershades.org

TRUSTWORTHY AI

Ethics and Law for trustable AI applications

TRUSTWORTHY AI

1) lawful, complying with all applicable laws and regulations
2) ethical, ensuring adherence to ethical principles and values
3) robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm.



TRUSTWORTHY AI: REQUIREMENTS

- human agency and oversight
- technical robustness and safety
- privacy and data governance
- transparency
- diversity, non-discrimination and fairness
- environmental and societal well-being
- accountability

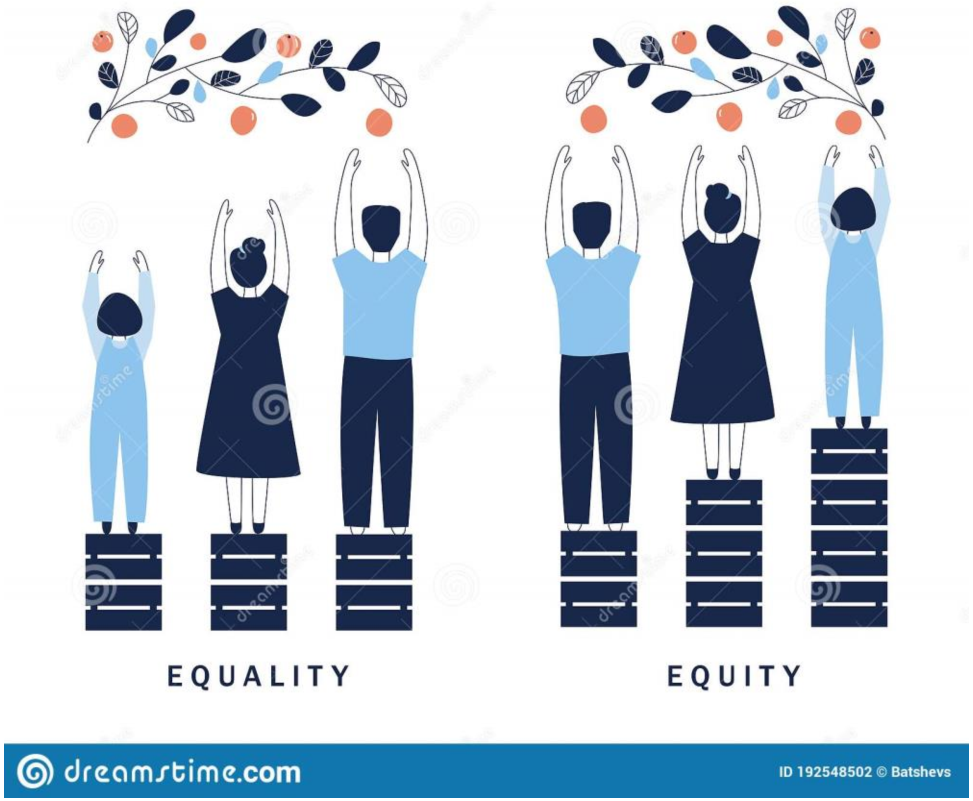
THE MANY FACETS OF DIVERSITY

- race
- ethnicity
- gender
- age
- religion
- disability
- sexual orientation
- socioeconomic status
- cultural background

FAIRNESS AND ITS PARADOXES

Algorithms should incorporate principles such as equality and equity

Note that equality =/= equity



TRUSTWORTHY AI: REQUIREMENTS

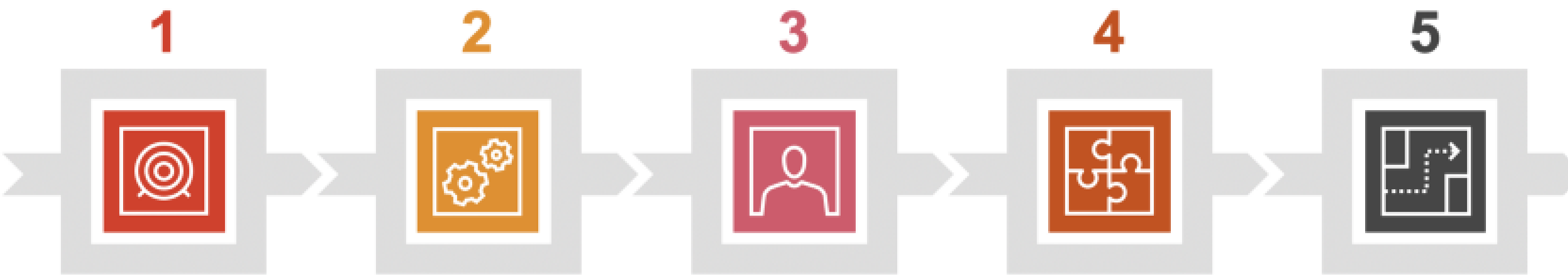
- human agency and oversight
- technical robustness and safety
- privacy and data governance
- transparency
- diversity, non-discrimination and fairness
- environmental and societal well-being
- accountability

TRANSPARENCY

- Traceability mechanisms can help achieving this for both the data, the system and the AI business models.
- AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned.
- Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations.

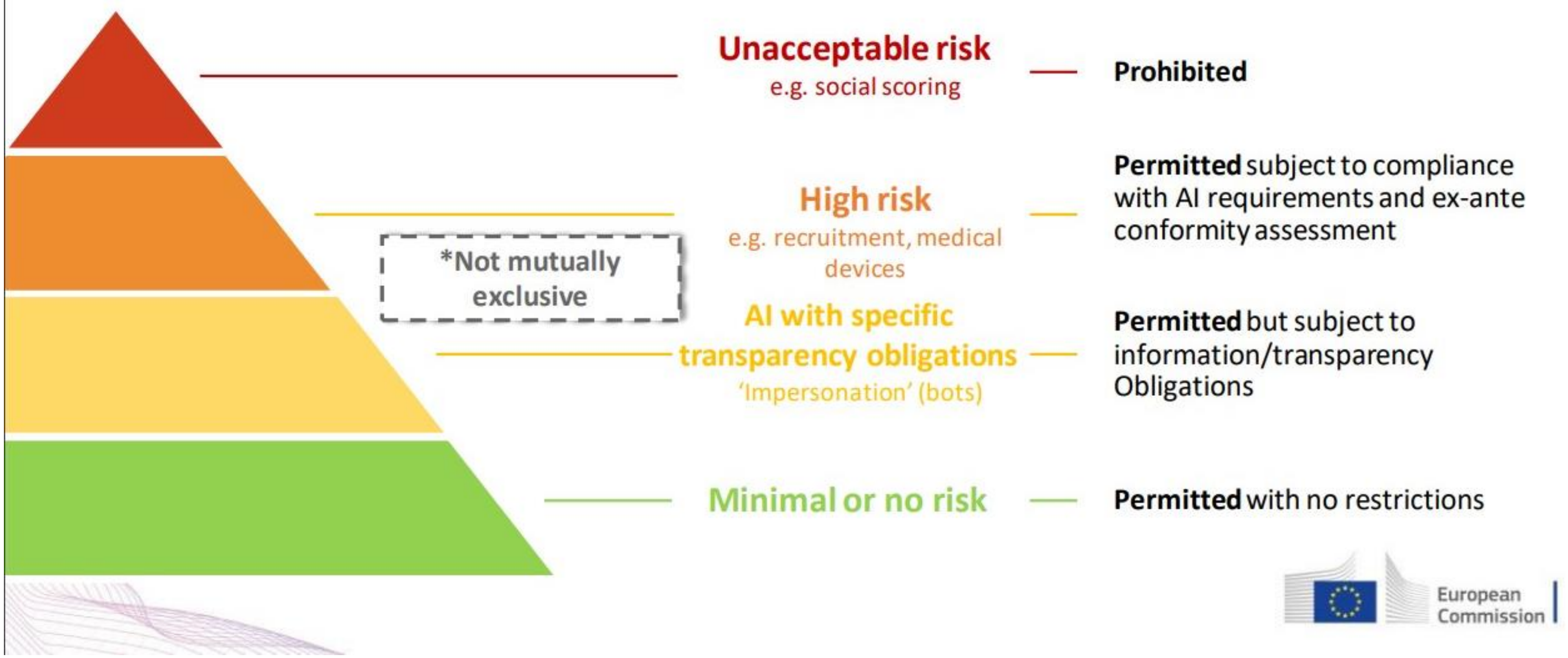
AI & THE RIGHT TO AN EXPLANATION

Five key questions to answer when building Explainable AI



1. **Explain to whom?** Understand the different **stakeholders**
2. **Why explain?** List the objectives and **reasons** for the explanation
3. **How to explain?** Explore the different **methods** of explanation
4. **When to explain?** Understand the need for explanations **before, during, after** the model is built
5. **What are the explanation techniques?** **Taxonomy** of different techniques of explanation

Author and Copyright: Anand Rao, Global Artificial Intelligence Lead, PWC, via towardsdatascience.com



A RISK-BASED APPROACH TO AI

Rules aimed at the promotion of human-centric and trustworthy AI, and the protection of the health, safety, fundamental rights and democracy from harmful effects of AI.

AI ACT

- Once approved, it will be the first law on AI by a major regulator anywhere!
  - On Wednesday June 14th, the European Parliament adopted its negotiating position on the AI Act with 499 votes in favour, 28 against and 93 abstentions ahead of talks with EU member states on the final shape of the law.
- Bans on real-time biometric surveillance, emotion recognition, predictive policing AI systems
- Tailor-made regimes for general-purpose AI and foundation models like GPT (Generative Pre-trained Transformer)
- The right to make complaints about AI systems

RISK OF MASS SURVEILLANCE

The new rules would ban AI systems for social scoring, biometric categorisation and emotion recognition

GENERAL-PURPOSE AI

- Providers must guarantee robust protection of fundamental rights, health and safety and the environment, democracy and rule of law.
- Compliance with additional transparency requirements, like:
  - disclosing that the content was generated by AI,
  - designing the model to prevent it from generating illegal content and
  - publishing summaries of copyrighted data used for training.

FUTURE AI RESEARCH

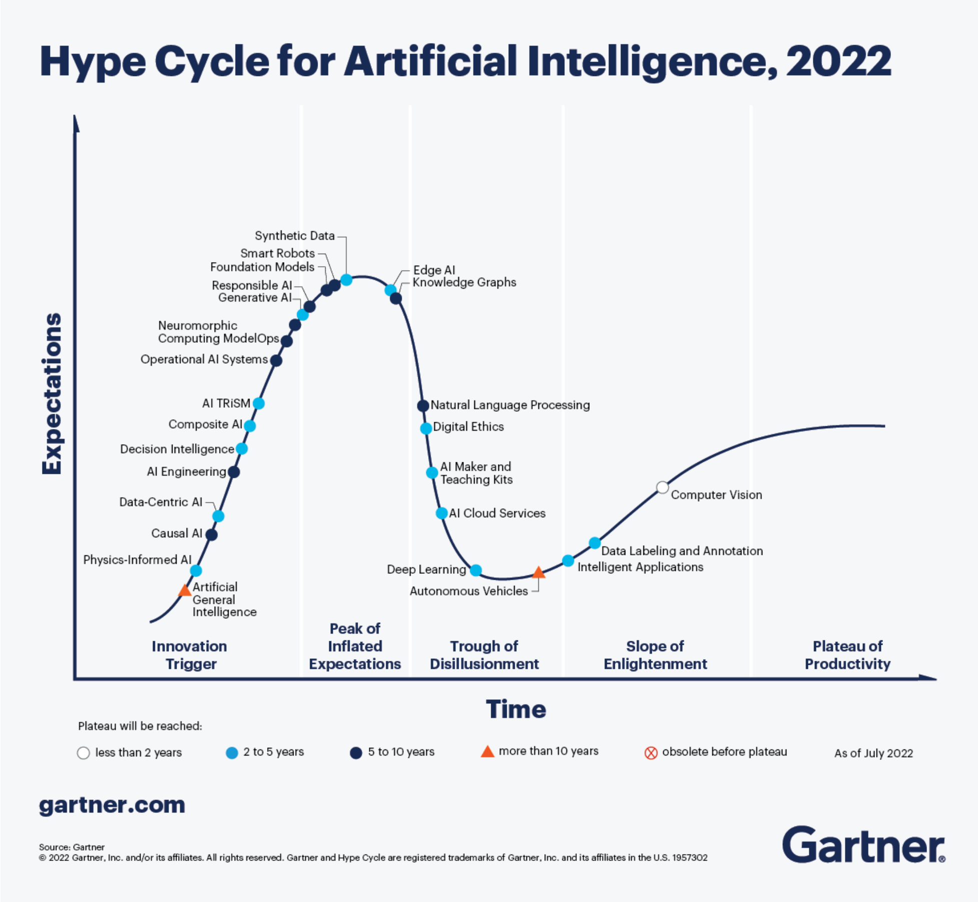
What are the challenges?

In which domains?

HYPE CYCLE FOR AI

The AI innovations on the Hype Cycle reflect complementary and sometimes conflicting priorities across four main categories:

- data-centric AI;
- model-centric AI;
- applications-centric AI;
- human-centric AI.



DIGITAL FORENSICS

- Focus on the phase of Evidence Analysis:
  - Examination and aggregation of evidence, collected from various electronic devices, about crimes and criminals in order to reconstruct events, event sequences and scenarios related to a crime.
  - Results are then made available to law enforcement, investigators, intelligence agencies, public prosecutors, lawyers and judges

EVIDENCE ANALYSIS: CHALLENGES FOR AI

FRAGMENTED KNOWLEDGE

COMPLEX SCENARIOS (SPACE, TIME, CAUSALITY, UNCERTAINTY, ETC.)

BIG DATA

TRANSPARENCY AND EXPLAINABILITY

DIGFORASP

COST Action CA17124: “Digital Forensics: Evidence Analysis via Intelligent Systems and Practices”

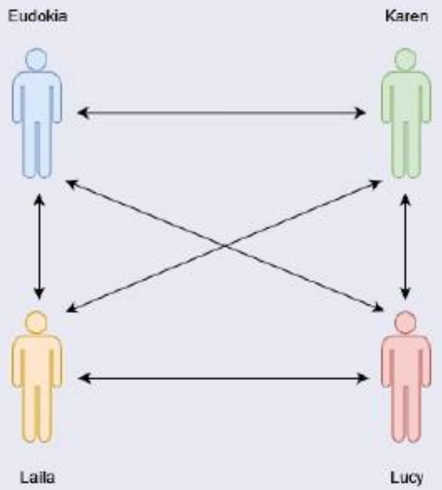
- Formal and verifiable AI methods and techniques for Evidence Analysis [Costantini et al., 2019b]
- Preference for logic-based AI methods for explainability reasons, e.g., nonmonotonic reasoning with Answer Set Programming (ASP) [Costantini et al., 2019a]
- Several interesting problems, e.g., phone call analysis

Incoming SMS	Andrea Levy	Eudokia Makrembolitissa	Alder Road		07:58:33.000	00:00:00.000	2040-12-29
Incoming SMS	Andrea Levy	Eudokia Makrembolitissa	Alexander Muir Road		20:00:51.000	00:00:00.000	2041-01-01
Incoming SMS	Andrea Levy	Eudokia Makrembolitissa	Alhart Drive		22:04:29.000	00:00:00.000	2041-01-02
Incoming SMS	Andrea Levy	Eudokia Makrembolitissa	Assiniboine Road		19:11:43.000	00:00:00.000	2041-01-05
Incoming SMS	Andrea Levy	Eudokia Makrembolitissa	Assiniboine Road		12:52:13.000	00:00:00.000	2041-01-06
Incoming SMS	Andrea Levy	Eudokia Makrembolitissa	Assiniboine Road		13:02:11.000	00:00:00.000	2041-01-06
Incoming SMS (/	Andrea Levy	Eudokia Makrembolitissa	Athletic Avenue		12:57:29.000	00:00:00.000	2041-01-06
Incoming call	Andrea Levy	Eudokia Makrembolitissa	3420 St Clair Avenue East	3420 St Clair Aven	14:35:05.000	00:00:25.000	2040-12-28
Incoming call	Andrea Levy	Eudokia Makrembolitissa	Amarillo Drive	Amarillo Drive	12:12:34.000	00:00:20.000	2041-01-01
Incoming call	Angela Rawlings	Eudokia Makrembolitissa	Alder Road	Alder Road	23:01:30.000	00:00:02.000	2041-01-11
Incoming SMS	Angela Topping	Eudokia Makrembolitissa	Abilene Drive		22:13:08.000	00:00:00.000	2041-01-18
Incoming call	Anita Brookner	Eudokia Makrembolitissa	21st Street	31st Street	20:49:30.000	00:00:10.000	2040-12-18
Incoming call	Anita Brookner	Eudokia Makrembolitissa	Abbottswood Road	Abbotsfield Gate L	19:34:40.000	00:00:48.000	2040-12-18
Incoming SMS	Ann Kiessling	Eudokia Makrembolitissa	Alexander Muir Road		10:20:44.000	00:00:00.000	2040-12-07
Incoming call	Ann Taylor	Eudokia Makrembolitissa	Alanbury Crescent	Alanbury Crescent	14:30:26.000	00:00:13.000	2041-02-11
Incoming SMS	Anna Bijns	Eudokia Makrembolitissa	Alcorn Avenue		12:57:32.000	00:00:00.000	2041-02-05
Incoming SMS	Anna Eliza Bray	Eudokia Makrembolitissa	Alder Road		19:48:47.000	00:00:00.000	2040-10-21
Incoming SMS	Anna Eliza Bray	Eudokia Makrembolitissa	Aldergrove Avenue		15:46:00.000	00:00:00.000	2040-10-20
Incoming call	Anna Eliza Bray	Eudokia Makrembolitissa	Alexander Muir Road	Alder Road	15:12:01.000	00:00:19.000	2040-10-20
Incoming SMS	Anna Zahorska	Eudokia Makrembolitissa	Addison Crescent		23:46:45.000	00:00:00.000	2040-11-01
Incoming SMS	Anna Zahorska	Eudokia Makrembolitissa	Adelaide Street East		23:37:16.000	00:00:00.000	2040-11-01
Incoming SMS	Anna Zahorska	Eudokia Makrembolitissa	Advance Road		23:42:33.000	00:00:00.000	2040-11-01
Incoming SMS	Anna Zahorska	Eudokia Makrembolitissa	Alder Road		11:31:11.000	00:00:00.000	2040-11-04
Incoming SMS	Anne Askew	Eudokia Makrembolitissa	Alder Road		15:11:59.000	00:00:00.000	2040-11-02
Incoming SMS	Anne Askew	Eudokia Makrembolitissa	Alder Road		22:56:36.000	00:00:00.000	2040-11-18
Incoming SMS	Anne Bishop	Eudokia Makrembolitissa	Alder Road		21:52:27.000	00:00:00.000	2040-09-05
Incoming SMS (/	Anne Bradstreet	Eudokia Makrembolitissa	Assiniboine Road		16:59:47.000	00:00:00.000	2041-01-06
Incoming SMS	Anne de Marquets	Eudokia Makrembolitissa	Alder Road		10:01:47.000	00:00:00.000	2040-09-21
Incoming SMS	Anne Elliot	Eudokia Makrembolitissa	Adler Street		14:49:41.000	00:00:00.000	2040-10-19
Incoming SMS	Anne Hébert	Eudokia Makrembolitissa	27 S Eglinton E Ramp		23:02:31.000	00:00:00.000	2041-02-01
Incoming SMS	Anne Hébert	Eudokia Makrembolitissa	Alameda Avenue		22:33:46.000	00:00:00.000	2041-01-20

MOBILE PHONE RECORDS

Four excel files with structure (type, caller, callee, street, time, duration, date)

The problem of phone call analysis



1. From the Eudokia Makrembolitissa dataset, would it be possible to find her accomplices Karen Cook McNally or/and Laila Lalami?
2. From the Eudokia Makrembolitissa, Karen Cook McNally and Laila Lalami dataset, would it be possible to find Lucy Delaney?
3. Do same people gather physically often?
4. **When X calls Y, do always Y calls Z shortly afterwards?**
5. At the time of the crime, who was at the same location, or called by Eudokia Makrembolitissa?
6. The day before, who spoke with Eudokia Makrembolitissa? Or any other suspect?

Università degli Studi di Bari Aldo Moro

SEQUENCE MINING IN PHONE CALLS WITH ASP

Pre-processing: From records to communication sequences

ASP encoding of the data and the problem

Declarative pattern mining:

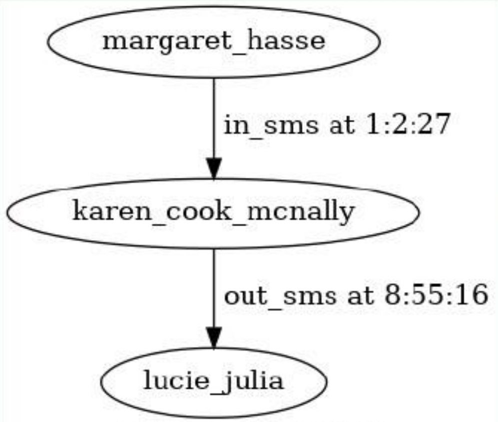
Use of a solver to find frequent patterns (answers are solutions)

An example of communication sequences

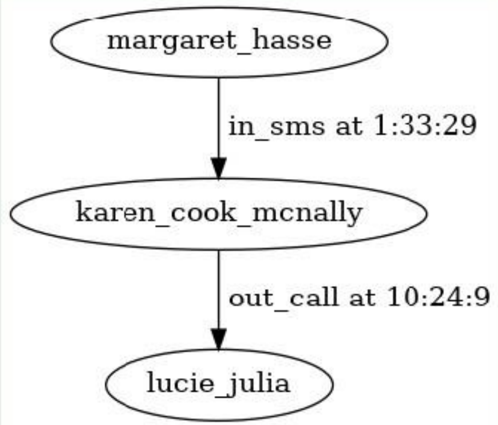
```
avg_len_sequences(53).
number_of_sequences(164).
max_len_sequences((1,2,2041),129).
seq((1,9,2040),1,(eudokia_makrembolitissa,florence_violet_mckenzie)).
seq((1,9,2040),2,(eudokia_makrembolitissa,florence_violet_mckenzie)).
seq((1,9,2040),3,(florence_violet_mckenzie,eudokia_makrembolitissa)).
.
.
seq((2,9,2040),1,(annie_dillard,eudokia_makrembolitissa)).
seq((2,9,2040),2,(eudokia_makrembolitissa,irena_jordanova)).
seq((2,9,2040),3,(eudokia_makrembolitissa,irena_jordanova)).
.
.
```

An example of sequential pattern

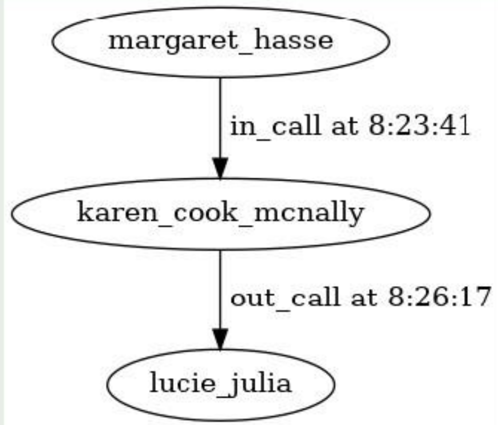
```
Answer: 1
pat(1,(margaret_hasse,karen_cook_mcnally))
pat(2,(karen_cook_mcnally,lucie_julia))
support((8,9,2040)) support((9,9,2040)) support((12,9,2040))
pat_information((8,9,2040),(1,(margaret_hasse,karen_cook_mcnally)),in_sms(simple),(1,0,55))
pat_information((8,9,2040),(1,(margaret_hasse,karen_cook_mcnally)),in_sms(simple),(1,2,27))
pat_information((8,9,2040),(2,(karen_cook_mcnally,lucie_julia)),out_sms(simple),(8,55,9))
pat_information((8,9,2040),(2,(karen_cook_mcnally,lucie_julia)),out_sms(simple),(8,55,16))
pat_information((9,9,2040),(1,(margaret_hasse,karen_cook_mcnally)),in_sms(simple),(1,33,29))
pat_information((9,9,2040),(2,(karen_cook_mcnally,lucie_julia)),out_call(simple),(10,24,9))
pat_information((12,9,2040),(1,(margaret_hasse,karen_cook_mcnally)),in_call(simple),(8,23,41))
pat_information((12,9,2040),(2,(karen_cook_mcnally,lucie_julia)),out_call(simple),(8,26,17))
len_support(3)
```



Pattern of 8-9-2040



Pattern of 9-9-2040



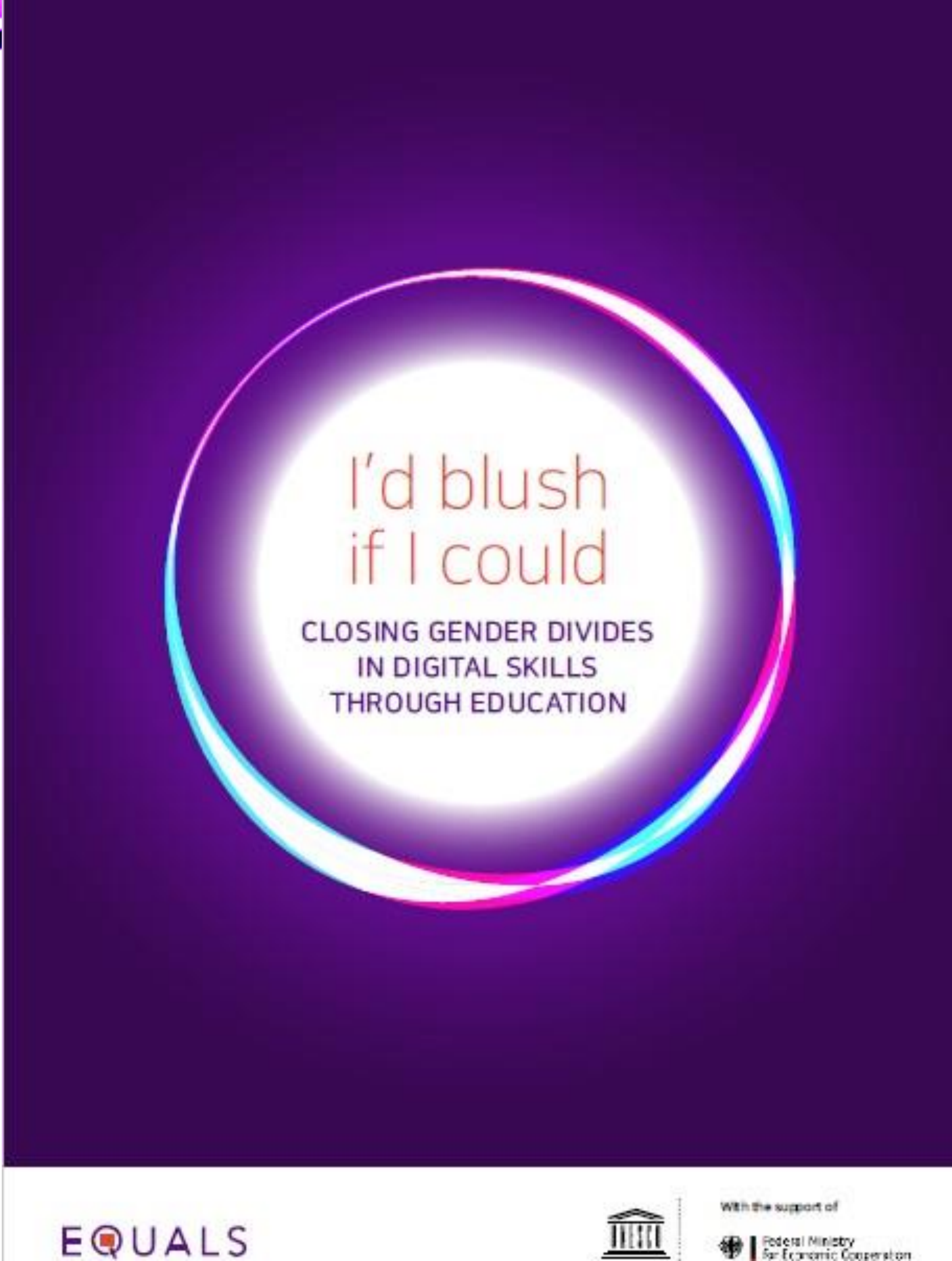
Pattern of 12-9-2040

Spoke 6: “Symbiotic AI” (led by University of Bari) - 2023

- Symbiotic AI (SAI) aims to boost human-machine collaboration by augmenting human cognitive abilities rather than replacing them.
- Main scientific question: to design AI systems according to a human-centered approach, as developed within the Human-Computer Interaction (HCI) community, in order to foster human-AI symbiosis.
- In particular, how to improve the understandability, acceptability and sustainability of SAI systems?

ETHICAL CONCERNS OF CHATBOTS

- Trust and Transparency
- Privacy
- Agent Persona
- Anthropomorphism and
- Sexualization



AN EXAMPLE IN CUSTOMER SERVICE

A CUSTOMER CONTACTS THE CUSTOMER SERVICE OF SOME COMPANY ASKING FOR A PARTICULAR PRODUCT OF THE COMPANY.

THE EMPLOYEE ILLUSTRATES THE PRODUCT CHARACTERISTICS AND TRIES TO CONVINCE THE CUSTOMER TO BUY THE PRODUCT.

(S)HE MENTIONS THAT THE PRODUCT IS ENVIRONMENTALLY FRIENDLY (WHICH IS IRRELEVANT IN THIS CASE), AND THIS IS AN ADVANTAGE OF THEIR PRODUCT OVER ANALOGOUS PRODUCTS FROM OTHER COMPANIES.

IS IT ETHICAL FOR THE EMPLOYEE TO SAY THAT?

ETHICS OF DIALOGUES IN AI-BASED CHATBOTS

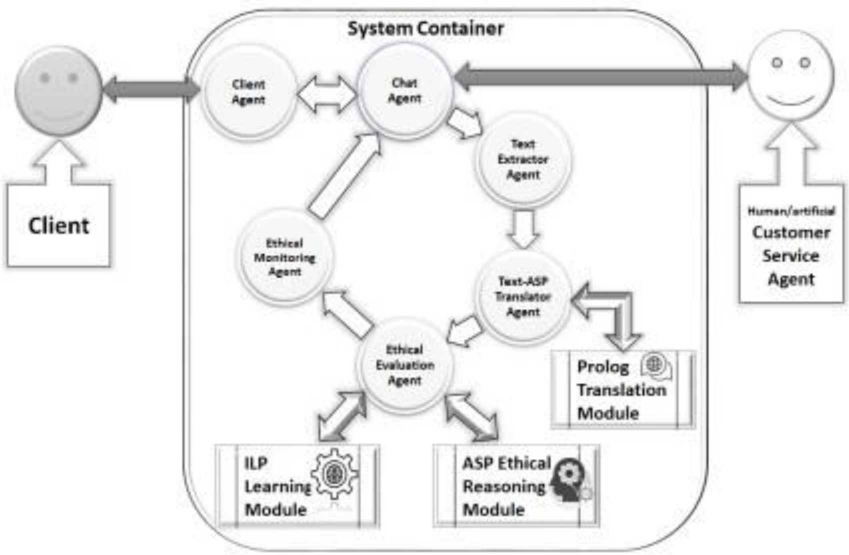


Fig. 1. EthicalEvalMAS Architecture

A Logic-based Multi-agent System for Ethical Monitoring and Evaluation of Dialogues

Abeer Dyoub
DISIM, University of L'Aquila, Italy
abeer.dyoub@univaq.it

Stefania Costantini
DISIM, University of L'Aquila, Italy
stefania.costantini@univaq.it

Ivan Letteri
DISIM, University of L'Aquila, Italy
ivan.letteri@univaq.it

Francesca A. Lisi
DIB & CILA, University of Bari "Aldo Moro", Italy
FrancescaAlessandra.Lisi@uniba.it

Dialogue Systems are tools designed for various practical purposes concerning human-machine interaction. These systems should be built on ethical foundations because their behavior may heavily influence a user (think especially about children). The primary objective of this paper is to present the architecture and prototype implementation of a Multi Agent System (MAS) designed for ethical monitoring and evaluation of a dialogue system. A prototype application, for monitoring and evaluation of chatting agents' (human/artificial) ethical behavior in an online customer service chat point w.r.t their institution/company's codes of ethics and conduct, is developed and presented. Future work and open issues with this research are discussed.

1 Introduction

Machine Ethics is an emerging field concerning itself with the ethical behavior of autonomous intelligent agents. Concerns about the ethical behavior of such machines is growing, especially with the increasing autonomy, and with agents 'invading' our everyday life and starting to perform many tasks on our behalf. Engineering machine ethics, or building practical ethical machines is not just about traditional engineering. With machine ethics, we need to find ways to practically build machines that are ethically restricted, and can also reason about ethics. This involves philosophical aspects, even though the problem has a non-trivial computational nature.

Chatbots are tools aimed at simplifying the interaction between humans and computers, typically used in dialogue systems for various practical purposes including customer service or information acquisition. From a technological point of view, a chatbot represents the natural evolution of question-answering system leveraging Natural Language Processing. Today, most chatbots are either accessed via virtual assistants such as Google Assistant and Amazon Alexa, or via messaging apps such as Facebook Messenger or WeChat, or via individual organizations' apps and websites. Business activities are rapidly moving towards the adoption of chatbots and other self-service technologies. This in order to automate basic communications and customer service, to reduce the call center costs and to provide advanced services to users. However, chatbots raise many ethical concerns. Unethical Artificial Intelligence and bots are a big concern for many consumers. The chatbot should be built on ethical foundations because its behavior influences the company's image, and unethical behavior will lead to mistrust from the client-side.

In previous works [1, 2, 3], a hybrid logic-based approach was proposed for ethical evaluation of chatbots' behavior, concerning online customer service chat points, w.r.t institution/company's codes of

A. Formisano, Y.A Liu, et al. (Eds.): International Conference on Logic Programming (Technical Communications) 2021 (ICLP 2021)
EPTCS 345, 2021, pp. 182–188, doi:10.4204/EPTCS.345.32

LEARNING & REASONING

- Integrating learning and reasoning constitutes one of the key open questions in AI
- It holds the potential of addressing many of the shortcomings of contemporary AI approaches, e.g.
  - the black-box nature and the brittleness of deep learning
  - the difficulty to revise knowledge bases in the light of new data.
- It calls for approaches that combine knowledge representation and automated reasoning techniques with algorithms from the fields of neural, statistical and relational learning.


IJCLR 2023
Home
Call for Papers
Important Dates
Programme
Speakers
Organisation
Registration
Venue & Local information
Attending
LEARNING & REASONING
3rd International Joint Conference on Learning & Reasoning 13-15 November 2023
ILP 2023
AAIP 2023
CogAI 2023

THANKS FOR THE ATTENTION

Prof. Francesca A. Lisi
Dipartimento di Informatica
Università degli Studi di Bari «Aldo Moro»
francesca.lisi@uniba.it