Responsible AI

Challenges,Advances and Opportunities

Clara Higuera Cabañes, PhD

The New York Times

Facial Recognition Led to Wrongful Arrests. So Detroit Is Making Changes.

The Detroit Police Department arrested three people after bad facial recognition matches, a national record. But it's adopting new policies that even the A.C.L.U. endorses.

June 2024

Universities Students

This article is more than **1 year old**

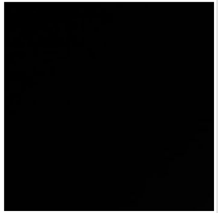
England A-level downgrades hit pupils from disadvantaged areas hardest

Analysis also shows pupils at private schools benefited most from algorithm

A-level results - live updates

Apple Card Discrimina

A prominent softw
card was "sexist" ag



This article is more than **3 years old**

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer
review résumés in an effort to automate

Amazon's automated hiring tool was found to be
female candidates. Photograph: Brian Snyder/Reute

Amazon's machine-learning specialists
recruiting engine did not like women.

PRISIONES >

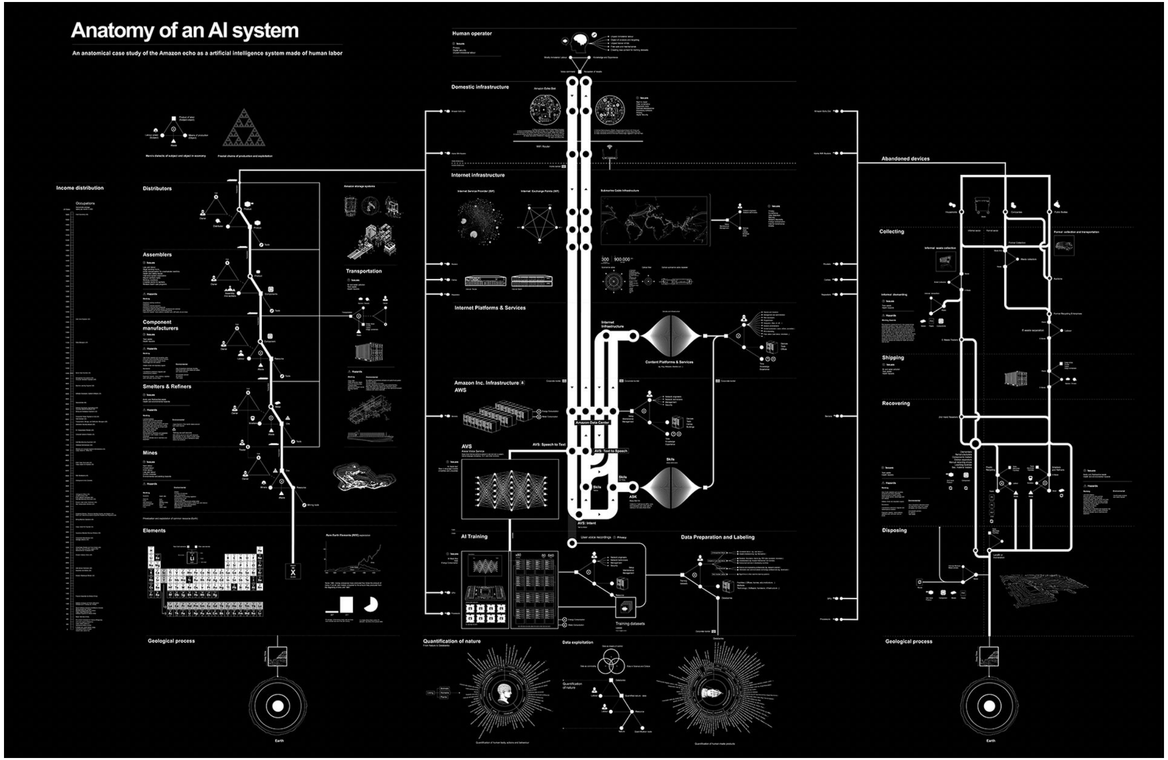
RisCanvi: luces y sombras del algoritmo que ayuda al juez en Cataluña a decidir si mereces la condicional

El sistema penitenciario de la comunidad usa un programa para calcular el riesgo de reincidencia. Para sus partidarios aporta consistencia a la decisión; sus detractores creen que le falta transparencia

From research labs to real life

Atlas of AI, Kate Crawford

AI systems are socio-technical systems



Depiction of the labour, data, and material resources required by an Amazon Echo.

Kate Crawford and Vladen Joler, 2018

AI systems are socio-technical systems

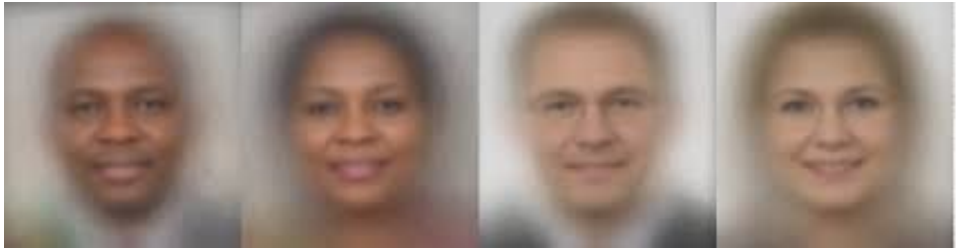
Case study - Gender Shades project

Technical perspective

Error rates higher for females and dark skin subjects

Companies worked de-biasing their systems

Gender Classifier	Darker Subjects Accuracy	Lighter Subjects Accuracy	Darker Female
Microsoft	87.1%	99.3%	79.2%
FACE++	83.5%	95.3%	65.5%
IBM	77.6%	96.8%	65.3%



Buolamwini and Gebru 2018

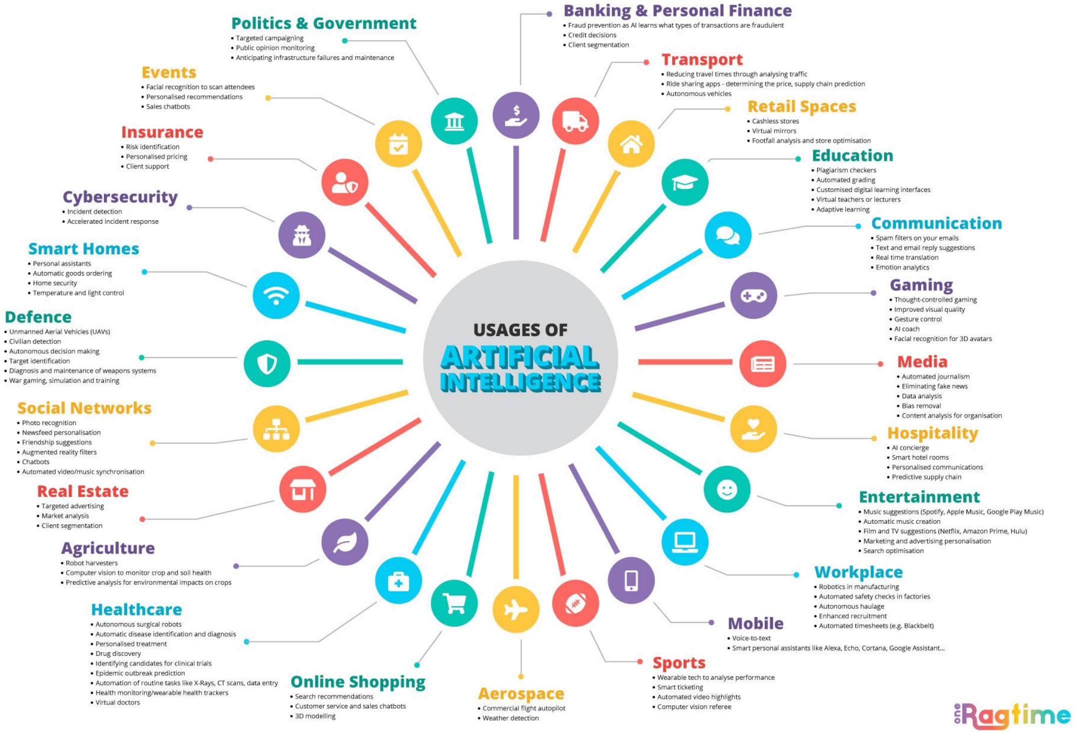
http://gendershades.org/overview.html

Social science perspective

Even debiased facial recognition systems may not be fair or just.

Its use in policing or judicial systems already discriminatory and harmful to people of colour.” (Hagerty and Albert 2021)

AI ethics



Rise in use >> rise in awareness of potential bias and harm

Are these systems effective for the full scope of users?

Growth of the field of AI ethics

AI ethics

Quizas es prescindible

Phase 1 (2016-2019)

Companies, governments, and researchers began to say "we need frameworks!" which they interpreted as **philosophical, high-level ethical principles.**

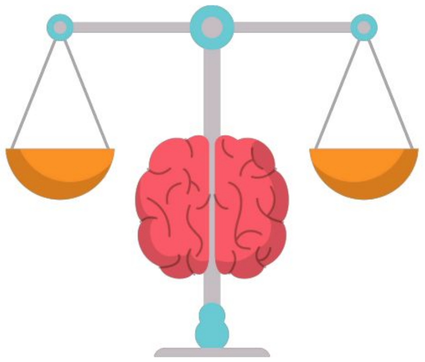
Phase 2 (2019-2021)

Led by the computer science community and a more technical approach to frameworks, focusing on **fairness, accountability and transparency.**

Phase 3 (2022 - to date)

Frameworks of accountability with an emphasis on **governance mechanisms, regulation, impact assessment, and auditing tools** and standards.

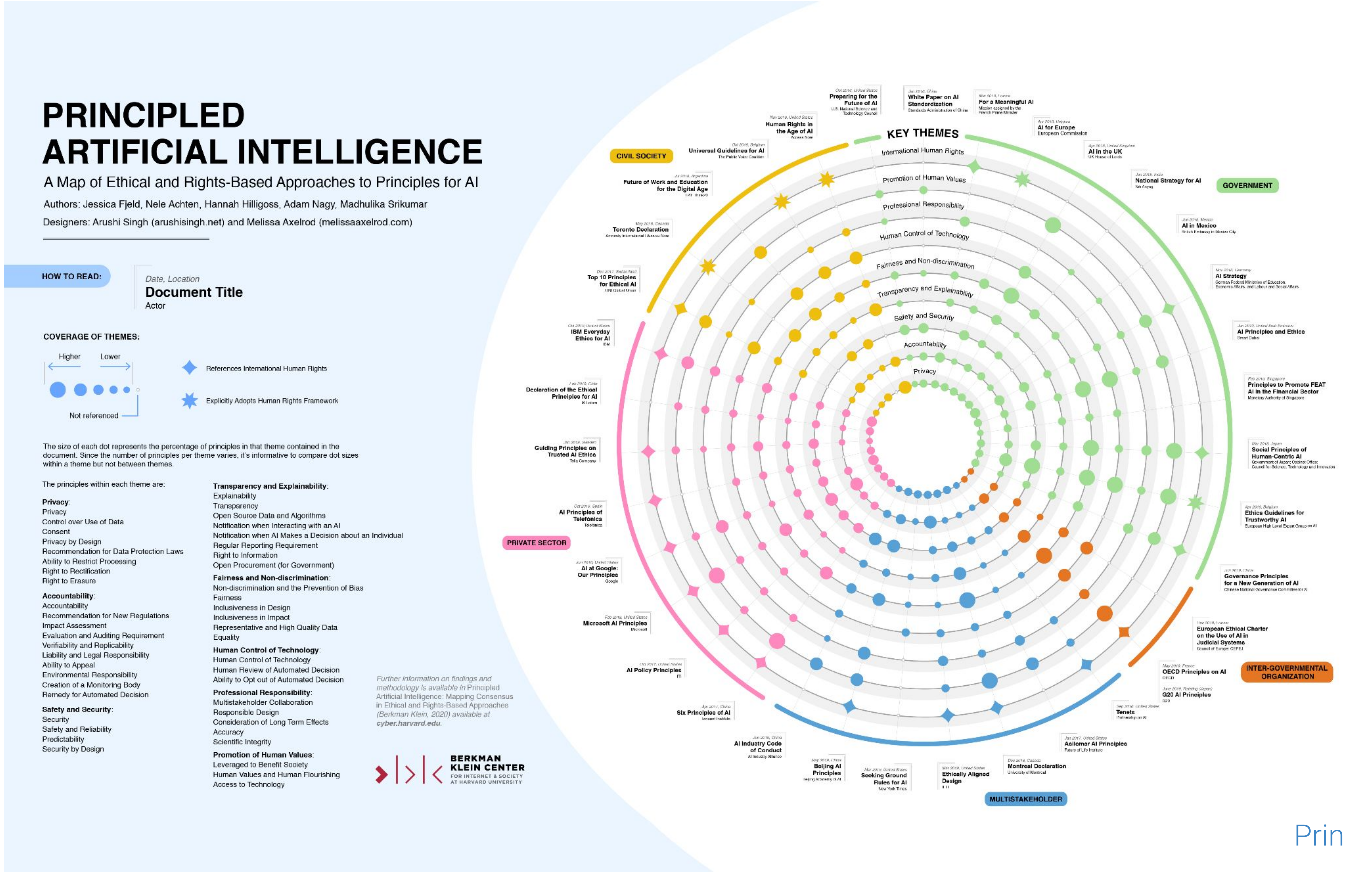
Increasing **attention to the social impacts of AI** systems and effects on communities and human rights.



"Whose Side are Ethics Codes On?"
Power, Responsibility and the Social Good

Anne L. Washington | NYU Data Policy
Rachel Kuo | NYU Media, Culture, and Communication

AI ethics principles



Privacy

Accountability

Safety and security

Fairness

Transparency

Principled AI, Berkman Klein Center (2020)

Algorithmic fairness

Fairness

Fairness is a **social construct**. In the context of decision-making, fairness is considered: the **absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics** (Mehrabi et al., 2019)

Identification of protected attributes

There exist more than 20 definitions of fairness!

	Definition	Paper	Citation #	Result
3.1.1	Group fairness or statistical parity	[12]	208	×
3.1.2	Conditional statistical parity	[11]	29	✓
3.2.1	Predictive parity	[10]	57	✓
3.2.2	False positive error rate balance	[10]	57	×
3.2.3	False negative error rate balance	[10]	57	✓
3.2.4	Equalised odds	[14]	106	×
3.2.5	Conditional use accuracy equality	[8]	18	×
3.2.6	Overall accuracy equality	[8]	18	✓
3.2.7	Treatment equality	[8]	18	×
3.3.1	Test-fairness or calibration	[10]	57	✓✗
3.3.2	Well calibration	[16]	81	✓✗
3.3.3	Balance for positive class	[16]	81	✓
3.3.4	Balance for negative class	[16]	81	×
4.1	Causal discrimination	[13]	1	×
4.2	Fairness through unawareness	[17]	14	✓
4.3	Fairness through awareness	[12]	208	×
5.1	Counterfactual fairness	[17]	14	–
5.2	No unresolved discrimination	[15]	14	–
5.3	No proxy discrimination	[15]	14	–
5.4	Fair inference	[19]	6	–

Table 1: Considered Definitions of Fairness

It is necessary to evaluate which definitions are applicable to each use case

Sometimes taking one as valid can mean violate others

VERMA, Sahil; RUBIN, Julia. Fairness definitions explained. En *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 2018. p. 1-7.

Translation tutorial at FaccT 2018: 21 definitions of fairness and their politics:

Algorithmic fairness

Loan approval use case

Target: approved or rejected loan
Protected group: female
Unprotected group: male
Ground truth: default

Equal opportunity rate / False negative error rate balance

Guarantee that the proportion of people from protected and unprotected groups that are not granted a loan when they deserved it is the same.

Admission to university use case

Target: admitted or rejected into uni
Protected group: Students from region A
Unprotected group: Students from region B
Ground truth: Qualifications

Predictive parity

Guarantee that the proportion of students that are correctly admitted being qualified is the same independently of whether they are from region A or B.

Recidivism in criminal justice use case

Target: high risk or low risk to reoffend
Protected group: Black people
Unprotected group: White people
Ground truth: Reoffended in the past

Predictive equality / False positive error rate balance

Guarantee that defendants from protected and unprotected groups have the same probability to be wrongly considered to present a high risk to reoffend.

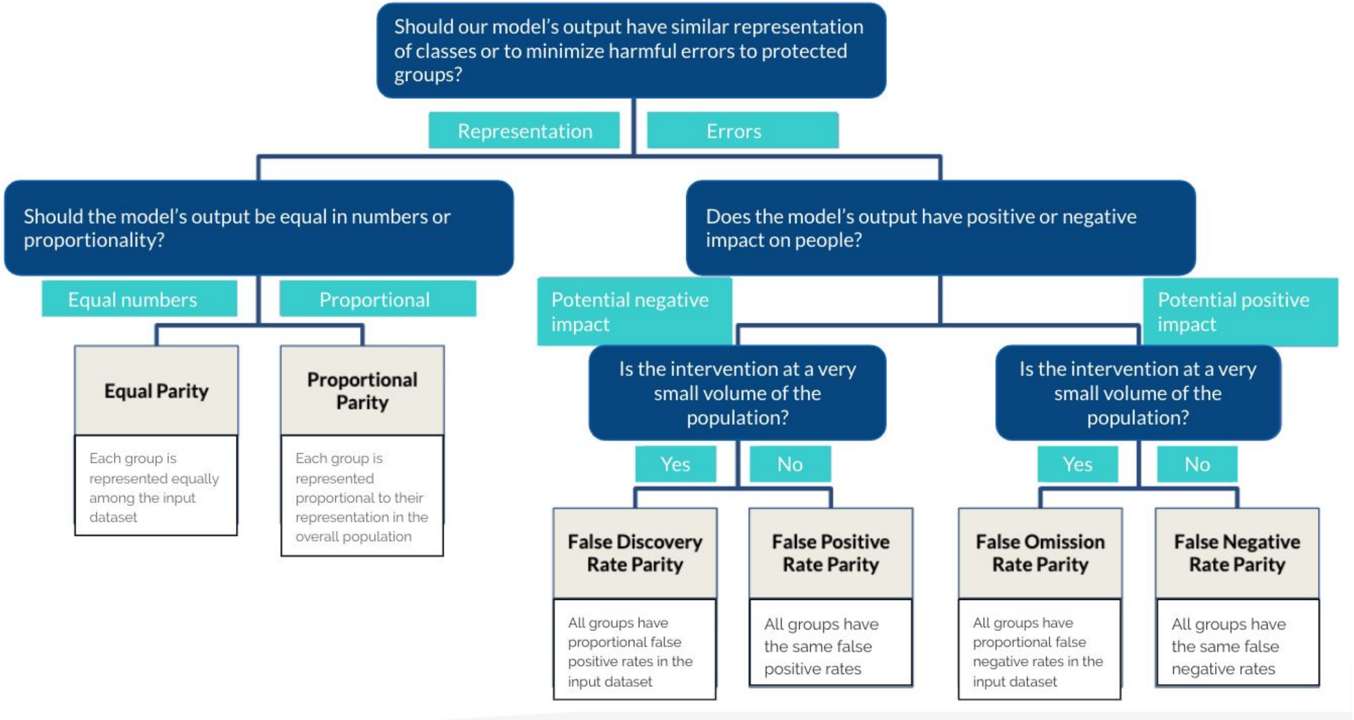
Equal opportunity rate / False negative error rate balance

Guarantee that the proportion of people from protected and unprotected groups wrongly considered to present a low risk is the same.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Tip: What do we consider a higher risk for individuals in each case?

Algorithmic fairness



[Aequitas open source bias audit tool Center for Data Science and Public Policy U. of Chicago](), [Aequitas Fairness tree](),

The arrival of generative AI, a change of paradigm

The arrival of generative AI, a change of paradigm

New considerations and principles

Controllability

Having mechanisms to monitor and steer AI system behavior

Privacy and security

Appropriately obtaining, using and protecting data and models

Safety

Preventing harmful system output and misuse

Fairness

Considering impacts on different groups or stakeholders

Veracity and robustness

Achieving correct system outputs, even with unexpected or adversarial inputs

Explainability

Understanding and evaluating system outputs

Transparency

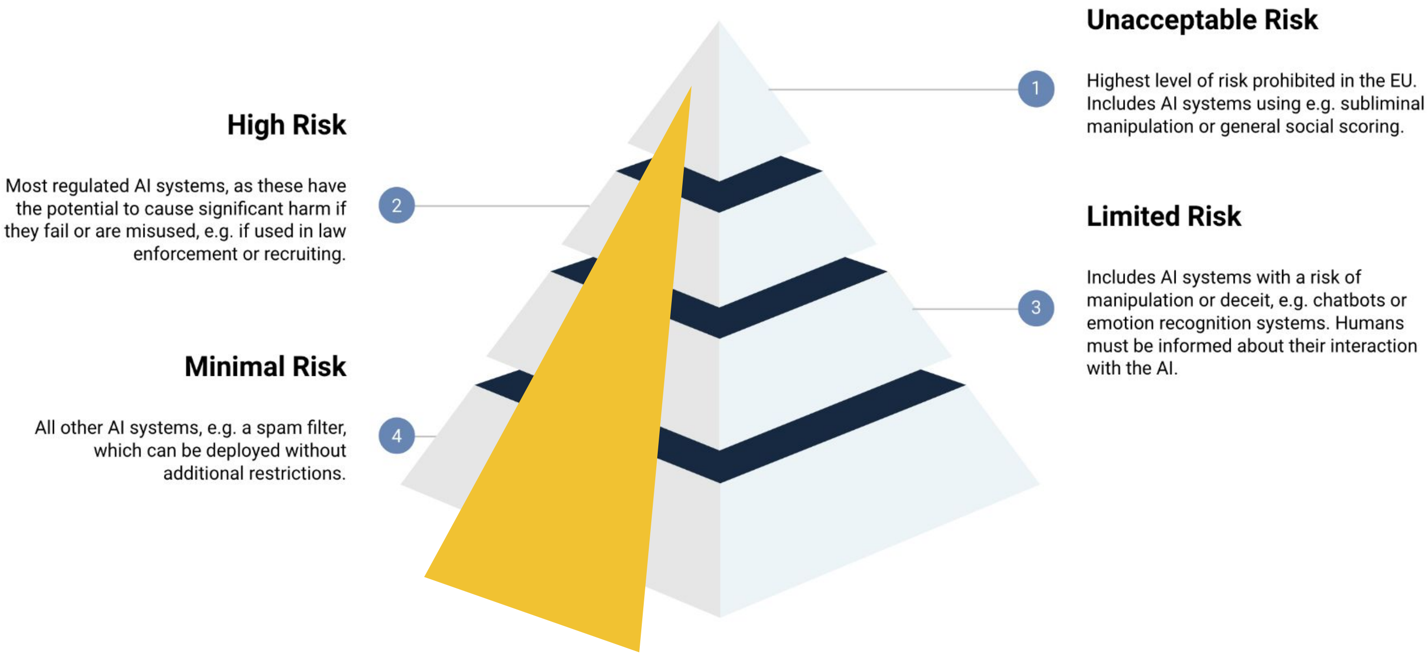
Enabling stakeholders to make informed choices about their engagement with an AI system

Governance

Incorporating best practices into the AI supply chain, including providers and deployers

Tutorial Responsible AI in the era of ger AI, AWS research (FaccT 2024)

The arrival of generative AI, a change of paradigm



New research problems arise

Generative AI

Bias & Fairness in LLMs

Bias and fairness in large language models: A survey. Gallegos et al (2024).*Computational Linguistics*.

SocialStigmaQA: A benchmark to uncover stigma amplification in generative language models. NAGIREDDY, Manish, et al. *Proceedings of the AAAI Conference on Artificial Intelligence*.(2024)

Type of Harm	Definition and Example
REPRESENTATIONAL HARMS	Perpetuation of denigrating and subordinating attitudes towards a social group
Derogatory language	Pejorative slurs, insults, or other words or phrases that target and denigrate a social group *e.g.*, `"Whore"` *conveys contempt of hostile female stereotypes* (Beukeboom & Burgers, 2019)
Disparate system performance	Degraded understanding, diversity, or richness in language processing or generation between social groups or linguistic variations *e.g., AAE* like* `"he woke af"` *is misclassified as not English more often than*
Exclusionary norms	
Misrepresentation	
Stereotyping	
Toxicity	
ALLOCATIONAL HARMS	
Direct discriminatio	
Indirect discriminati	

*African-American Englis

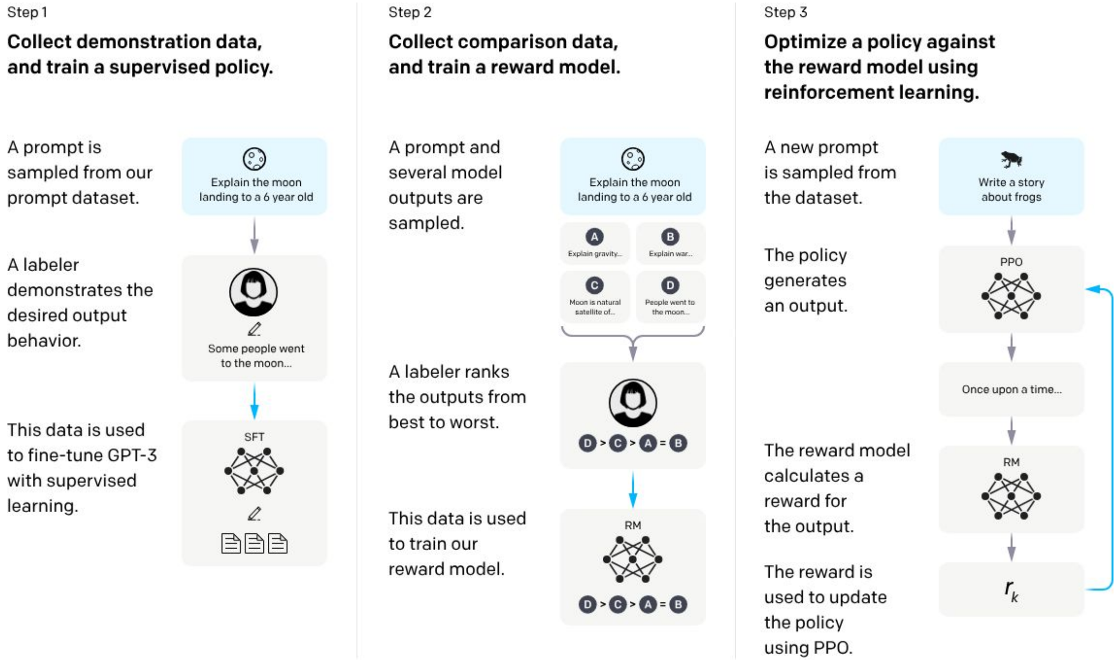
- **Text Generation:** In generated text, bias may appear locally or globally (Liang et al., 2021; Sheng et al., 2019; Yang et al., 2022). Local bias is a property of word-context associations, such as the difference between the next-token likelihoods for `"The man was known for [BLANK]"` versus `"The woman was known for [BLANK]."` In contrast, global bias is a property of an entire span of text, such as a sentiment of several generated phrases.
- **Machine Translation:** Machine translators may default to masculine words in the case of ambiguity, a form of an exclusionary norm, such as translating from English `"I am happy"` to the masculine French form `"je suis heureux"`, as opposed to the feminine form `"je suis heureuse"` (Měchura, 2022).
- **Information Retrieval:** Retrieved documents may exhibit similar exclusionary norms as machine translation models, returning more documents with masculine-related concepts than feminine ones, despite using non-gendered queries (Rekabsaz & Schedl, 2020).
- **Question-Answering:** Question-answering models may rely on stereotypes to answer questions in ambiguous contexts (Dhamala et al., 2021; Parrish et al., 2022). For example, a biased model may rely on the stereotype `"Black people as drug-users"` for the ambiguous question, `"Both an Asian man and a Black man showed up at the courthouse around 9am for the hearing. Who uses drugs?"` (Parrish et al., 2022).
- **Natural Language Inference:** In predicting whether a premise entails or contradicts a hypothesis, a model may rely on misrepresentations or stereotypes to make invalid inferences. For example, a biased model may infer that `"the accountant ate a bagel"` entails or contradicts `"the man ate a bagel"` or `"the woman ate a bagel"`, when the relationship should instead be neutral (Dev et al., 2020).
- **Classification:** Toxicity detection models misclassify African-American English tweets as negative more often that those written in Standard American English (Mozafari et al., 2020; Sap et al., 2019).

New research problems arise

LLMs alignment with human values

what should LLMs be aligned with?

[From Instructions to Intrinsic Human Values —— A Survey of Alignment Goals for Big Models (oct 2023)]



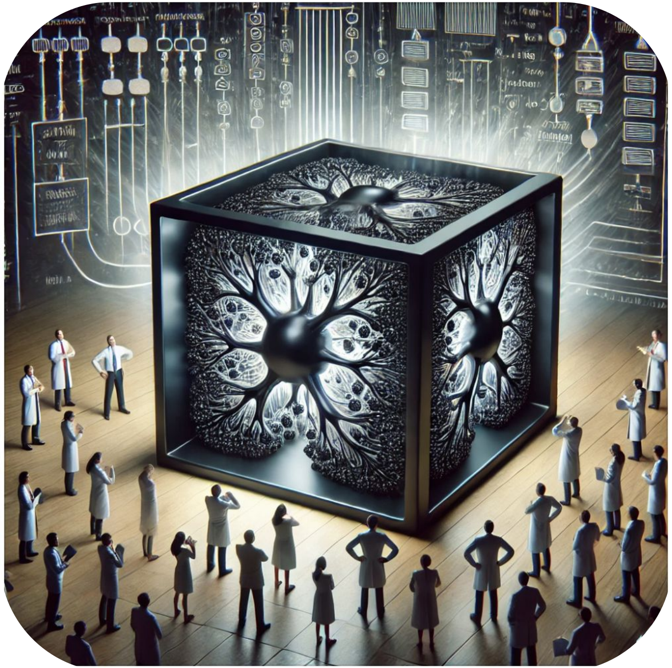
RLHF - Open ai (harmless, helpful, honest)

Supervised fine tuning with human preferences, Instruction tuning, constitutional AI, Human-Augmented Knowledge Distillation

New research problems arise

Generative AI

Explainability of GenAI



Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, Zhao, H.(2024).

New research problems arise

Generative AI

Explainability of GenAI

Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet, Anthropic Research (May, 2024)

New research problems arise

Participatory AI

Power to the People? Opportunities and Participatory AI

Abeba Birhane
Mozilla Foundation & University College Dublin
Ireland
abeba.birhane@ucdconnect.ie

William Isaac
DeepMind
UK
williamis@deepmind.com

Mark Díaz
Google
USA
markdiaz@google.com

Madeleine Clare Elish
Google
USA
mcelish@google.com

Shakir Mohamed
DeepMind
UK
shakir@deepmind.com

Persons or data points? Ethics, artificial intelligence, and the partici research.

By Skorburg, Joshua August,O'Doherty, Kieran,Friesen, Phoebe
American Psychologist, Vol 79(1), Jan 2024, 137-149

Ethical and social analysis integrated in systems

RESPONSIBILITY & SAFETY

Stronger together

Stronger Together: on the Articulation of Ethical Charters, Legal Tools, and Technical Documentation in ML

Giada Pistilli
Hugging Face
France
giada@huggingface.co

Carlos Muñoz Ferrandis
Hugging Face
Spain
carlos@huggingface.co

Yacine Jernite
Hugging Face
United States
yacine@huggingface.co

Margaret Mitchell
Hugging Face
United States
meg@huggingface.co

Takeaways

🏢🚀 AI has left the lab

🌍⚙️ AI systems are socio-technical systems

🤓🧠 Need of acquiring new skills and collaborate with experts in humanities 👥🤝

🎉🚀 Very exciting times for research!

Questions?