Data and ethics

SICSS 2023

Claudia López Moncada

UNIVERSIDAD TECNICA FEDERICO SANTA MARIA
DEPARTAMENTO DE INFORMÁTICA

¿De dónde viene lo que diré?

UNIVERSIDAD TECNICA FEDERICO SANTA MARIA | DEPARTAMENTO DE INFORMÁTICA

UTFSM

Informatics (Eng. and M.Sc).

2010

CSST - Community of sociotechnical researchers

University of Pittsburgh

Ph.D. Information Science and Technology

2015

UTFSM

Informatics

2021

Collaboration with ethicist and computer sciences

CENIA

CENTRO NACIONAL DE INTELIGENCIA ARTIFICIAL

2022

FAIR Futures of Artificial Intelligence Research

Collaboration with sociologists and communication researchers

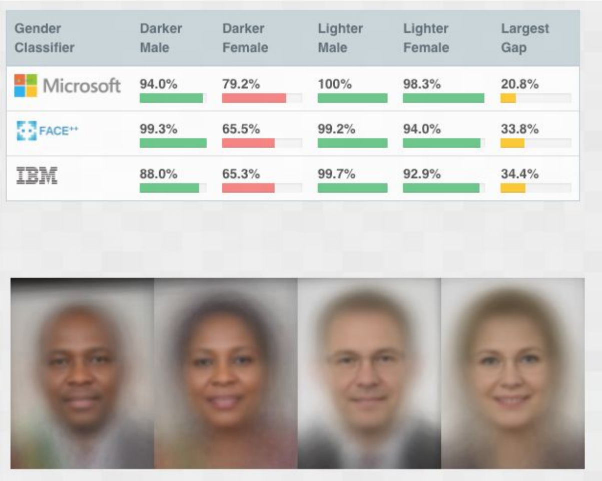
What do I want to communicate today?

- There is evidence that “data-intensive” systems reproduce biases, generating unfair outcomes

- Data codifying the past is part of the problem, but understanding the social context of data is part of the solution

- We need processes to reflect on the characteristics of our data, our goals, and our harm mitigation strategies

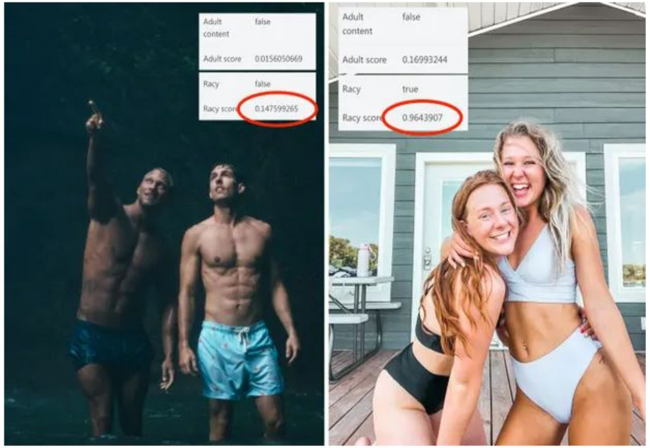
Examples of unfair outcomes of data-intensive



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%

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on FAccT* (pp. 77-91). http://gendershades.org/overview.html

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

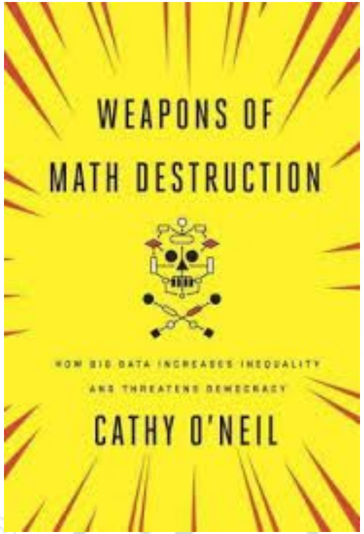


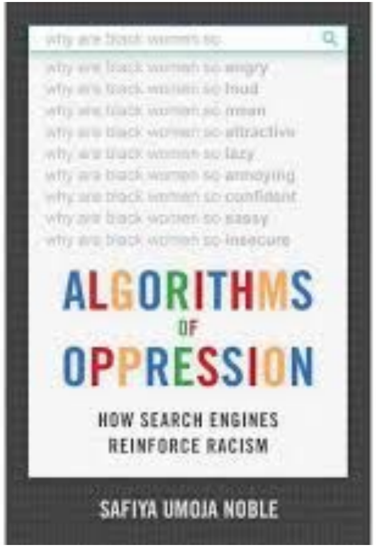
The photo of the women got eight views in one hour, while the picture with the men received 655 views, suggesting the women's photo was either suppressed or shadowbanned. Composite: Gianluca Mauro/The Guardian

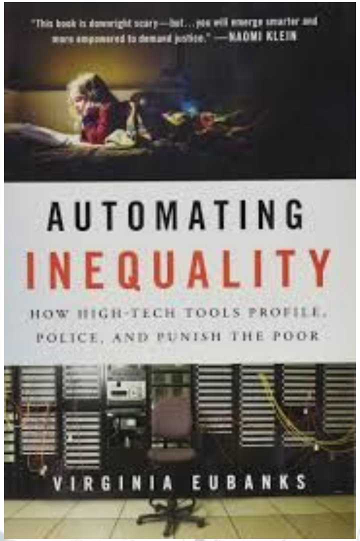
655 views 8 views

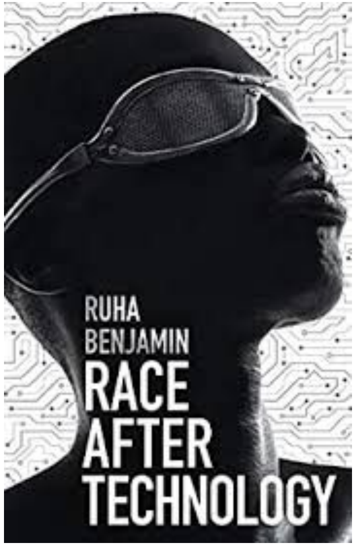
https://www.theguardian.com/technology/2023/feb/08/biased-ai-algorithms-racy-women-bodies

There are more examples in these books



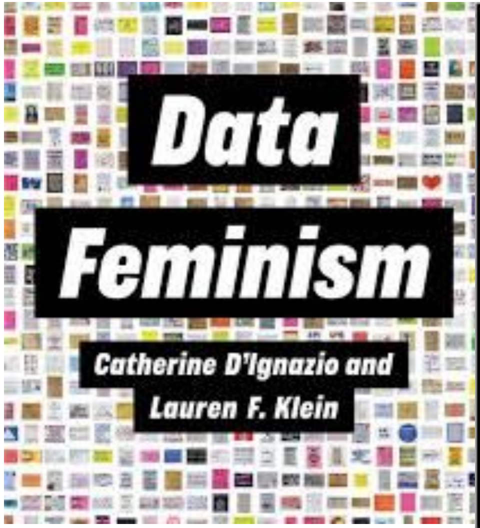


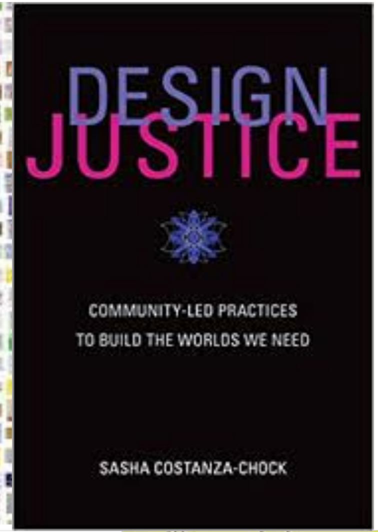


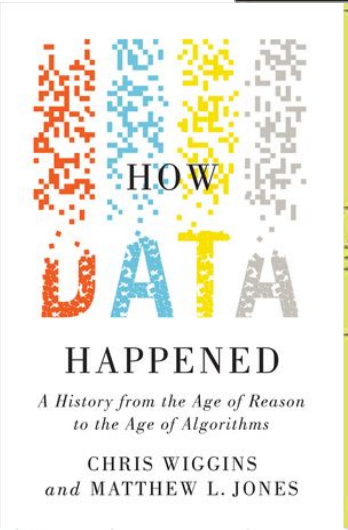


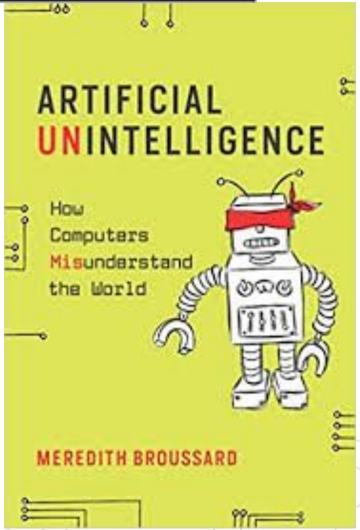
Lots of what I will say is highly influenced by

- [Technology and engineering practice: ethical lenses to look through]
- [Ethics in practice: A toolkit]



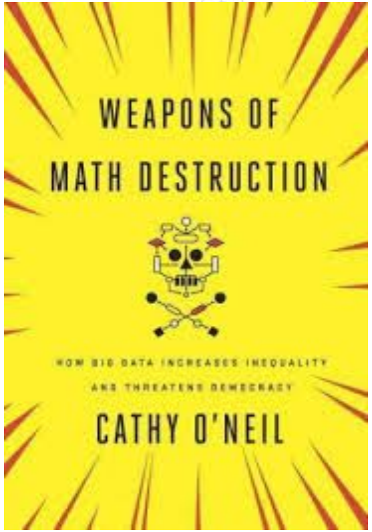






Data codifying the past is part of the problem (not the only one!)

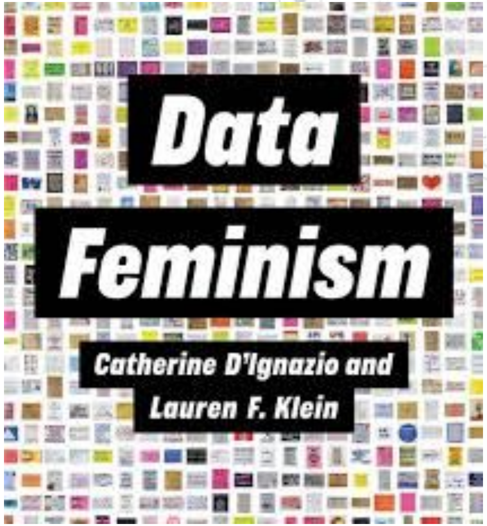
"Big Data processes **codify the past**. They do not invent the future. Doing that requires moral imagination, and that's something only humans can provide. **We have to explicitly embed better values** into our algorithms, creating Big Data models that follow our ethical lead. Sometimes that will mean putting **fairness** ahead of profit."



O'Neil, C. (2017). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

Understanding the social context of data is part of the solution

“The process of **converting life experience into data** always necessarily entails **a reduction of that experience** […]”

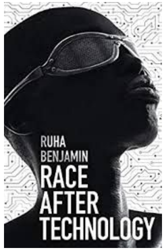


D'Ignazio, Catherine; Klein, Lauren F.. Data Feminism (Strong Ideas) (p. 23). The MIT Press. Kindle Edition

“**before there are data, there are people**—people who offer up their experience to be counted and analyzed, people who perform that counting and analysis, people who visualize the data and promote the findings of any particular project, and people who use the product in the end. There are also, always, people who go uncounted—for better or for worse. And there are problems that cannot be represented—or addressed—by data alone.”

D'Ignazio, Catherine; Klein, Lauren F.. Data Feminism (Strong Ideas) (p. 23). The MIT Press. Kindle Edition

“The **datasets** and models used in these systems **are not objective representations of reality**. They are the culmination of particular tools, people, and power structures that foreground one way of seeing or judging over another. **Without comprehensively accounting for the strengths and weaknesses of technical practices, the work of ethics**—which includes weighing the risks and benefits and potential consequences of an AI system—**will be incomplete.**” (Elish & boyd, 2018)



Don’t Believe Every AI You See. C.N. Elish & danah boyd, 2018. https://ai.shorensteincenter.org/ideas/2018/11/12/dont-believe-every-ai-you-see-1

So, at some extent, it is up to us (those who work with data)

Let me pause to analyze an example of how data issues arise

A study in healthcare in the USA

- Risk prediction to identify patients with complex care needs

  - Percentile 97 and above => assign to care program
  - Percentile 55 and above => physician should evaluate and decide

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

A third-party audit

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Training/evaluation data

- Features do not include race
- They include demographics (e.g., age, sex), insurance type, diagnosis and procedure codes, medications, and detailed costs.

Audit data

- Health: diagnoses, laboratory studies and vital signs capturing the severity of chronic illnesses.
- Cost: insurance claims data on utilization, including outpatient and emergency visits, hospitalizations, and health care costs

At the same predicted risk score, black patients were sicker than white ones

At the 97th percentile, Blacks had 26.3% more chronic illnesses than Whites (4.8 vs. 3.8; $p < 0.001$).

- There is also a gap in hypertension, diabetes, renal failure, and anemia, and higher cholesterol

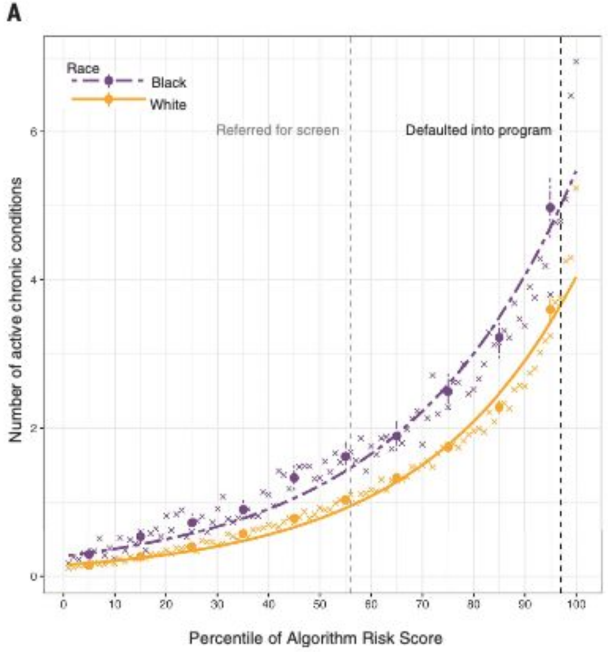


Fig. 1. Number of chronic illnesses versus algorithm-predicted risk, by race. (**A**) Mean number of chronic conditions by race, plotted against algorithm risk score. (**B**) Fraction of Black patients at or above a given risk score for the original algorithm ("original") and for a simulated scenario that removes algorithmic bias ("simulated": at each threshold of risk, defined at a given percentile on the *x* axis, healthier Whites above the threshold are

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

What was the algorithm predicting?

If not illnesses, what do you think the algorithm was predicting?

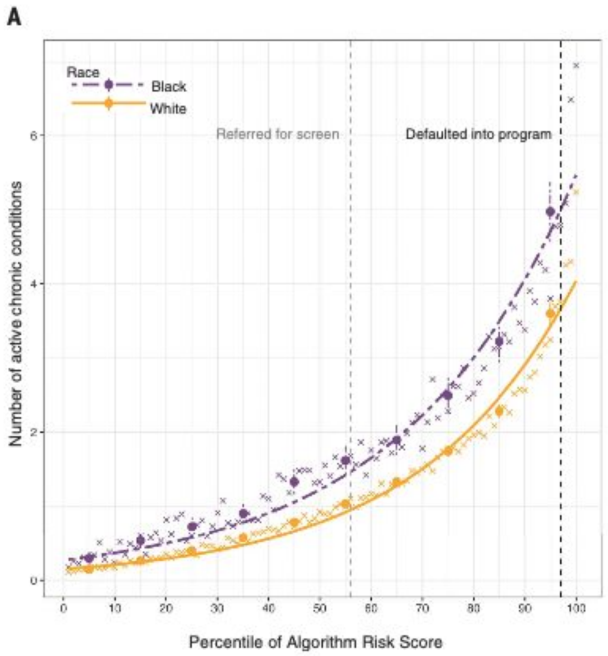
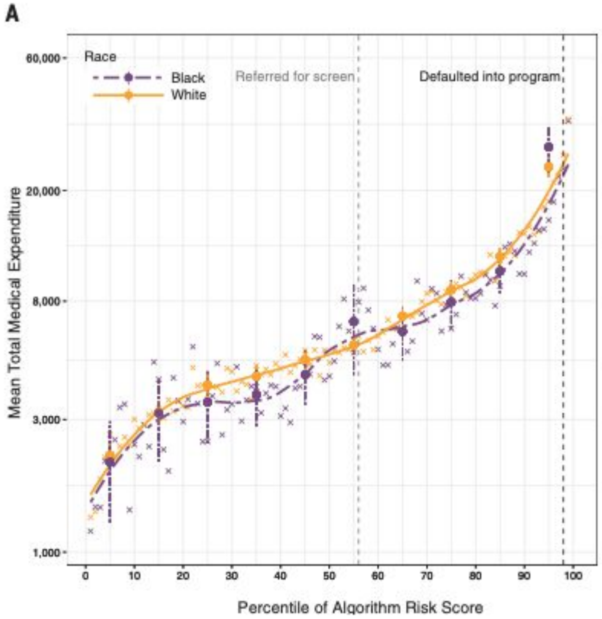


Fig. 1. Number of chronic illnesses versus algorithm-predicted risk, by race. (**A**) Mean number of chronic conditions by race, plotted against algorithm risk score. (**B**) Fraction of Black patients at or above a given risk score for the original algorithm ("original") and for a simulated scenario that removes algorithmic bias ("simulated": at each threshold of risk, defined at a given percentile on the *x* axis, healthier Whites above the threshold are

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

It was predicting total medical expenditure!



Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

It was predicting total medical expenditure!

However, in the USA, Blacks spend less than Whites in healthcare

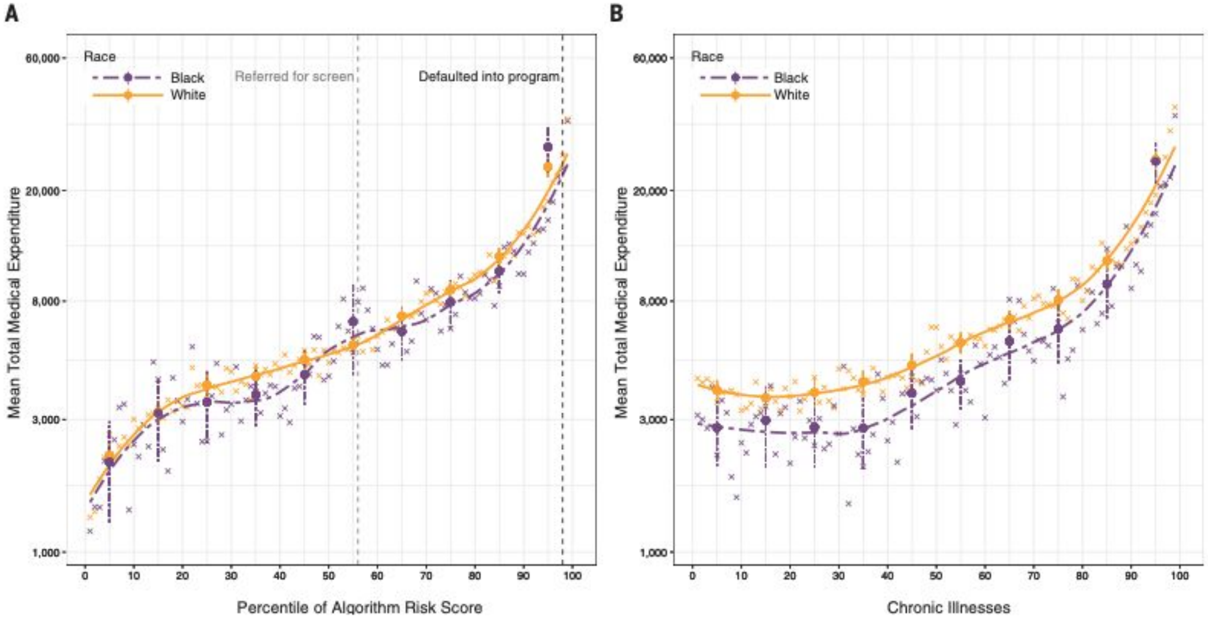


Fig. 3. Costs versus algorithm-predicted risk, and costs versus health, by race. (**A**) Total medical expenditures by race, conditional on algorithm risk score. The dashed vertical lines show the auto-identification threshold (black line: 97th percentile) and the screening threshold (gray line: 55th percentile). (**B**) Total medical expenditures by race, conditional on number of chronic conditions. The × symbols show risk percentiles; circles show risk deciles with 95% confidence intervals clustered by patient. The *y* axis uses a log scale.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

The bias was historical, but also technical

- The source of bias (and harm!) was the selection of target variable
  - It was not a completely illogical decision, but dismissed a well-known social phenomenon in that context

- Different target variables would have increased the number of black patients assigned to the care program

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

There is a gap between learning data science and deploy it in a social context

- "Translating complex objectives into a data mining problem is not self-evident" [Barocas & Selbst, 2016]
  - We learn by using well defined problems and robust datasets (e.g., spam detection) but data science is applied to more complex contexts (e.g., patient risk, creditworthiness)

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. Calif. L. Rev., 104, 671.

My takeaway: we need processes to make us reflect about biases and potential harms

So, what are we doing from here?

[with Gabriela Arriagada, Alexandra Davidoff, et al.]

From Chile, and the Global South

Envisioning and testing processes to include ethical reflection in data science / AI

- Ethics diagnostic @CENIA and map of AI principles (beyond fairness)
- A networked view of biases and mitigations
- An socio-ethical review of AI projects

Uncovering invisible AI in Chile and Latin America

- Observing the investment in AI through open data
- Case studies to identify how people (designers, users and affected people) conceptualize AI and how that affects their interaction with AI

Ethics diagnostic at CENIA

- CENIA: 150+ researchers, engineers, students, staff - interdisciplinary (CS, neuroscientists, math, physics,...)
- Qualitative diagnostic: 24 people (interviews + focus groups
- Conducted by a sociologist + diverse team (ethicist, math, CS, physics) + 2 lawyers

Ethics diagnostic at CENIA

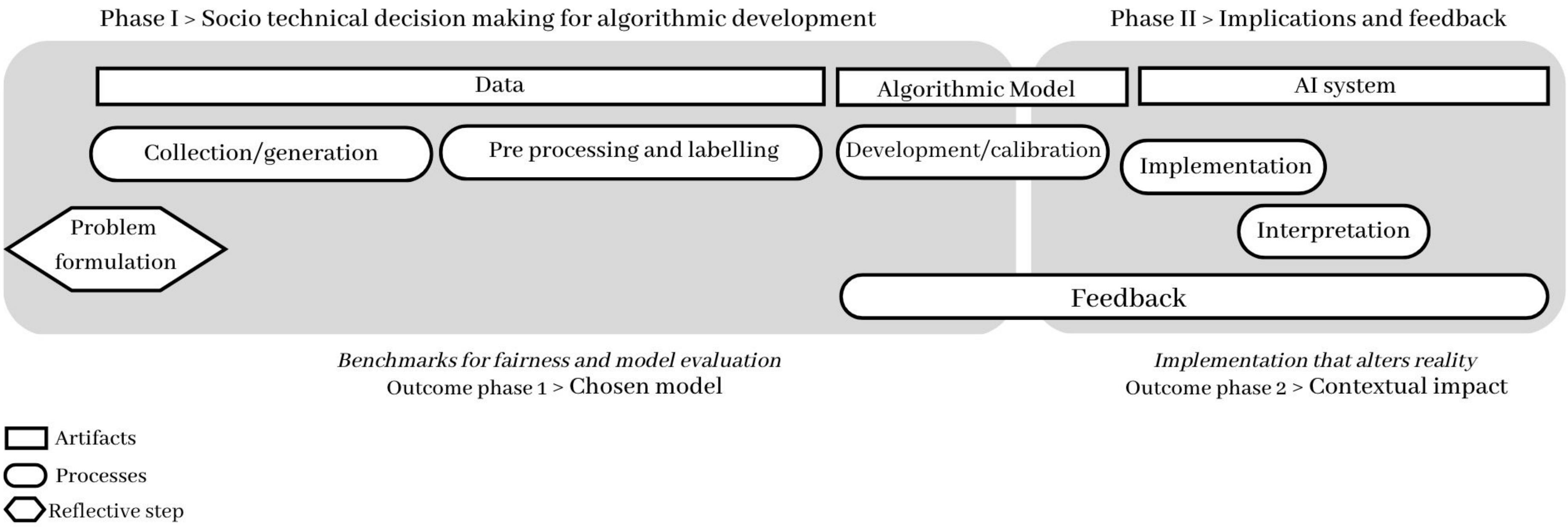
- Scarce formal training in ethics (in AI).
- Heterogeneous informal training

- Main concerns: biases, privacy, discrimination, democratize AI (to users & researchers)
- Key principles: fairness and transparency

Ethics diagnostic at CENIA

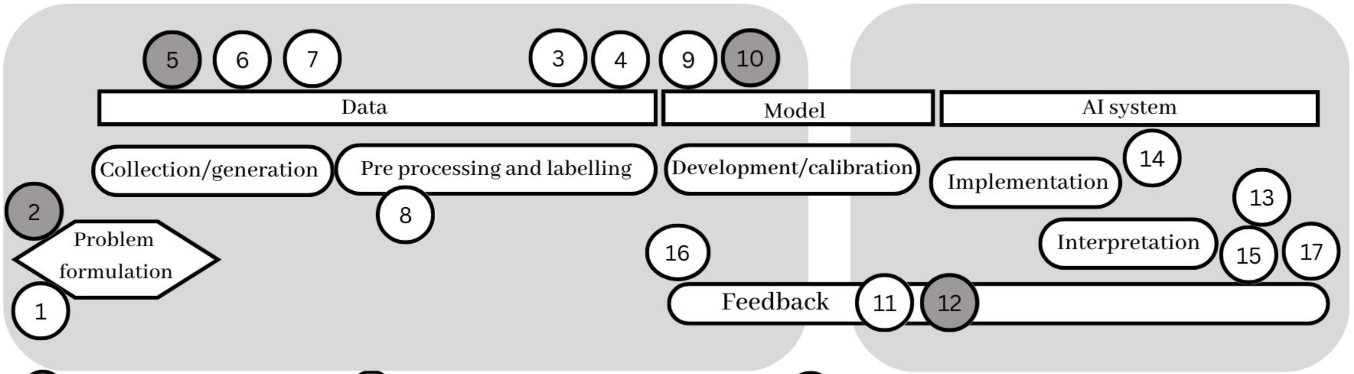
- Transversal relevance of ethics and ethics in AI, but **lack of guidelines/mechanisms**
- **Ethics is seen from a negative point of view (restrictions)**
- Difference in perception of relevance in own work (theoretical work, level of agency)

Pipeline: artifacts and people’s processes



APEC Digital Economic Steering Group (DESG),. Comparative Study on Best Practices to Detect and Avoid Harmful Biases in Artificial Intelligence Systems.; Asia-Pacific Economic Cooperation (APEC)., Forthcoming.

At each step, decisions and biases influence the outcomes



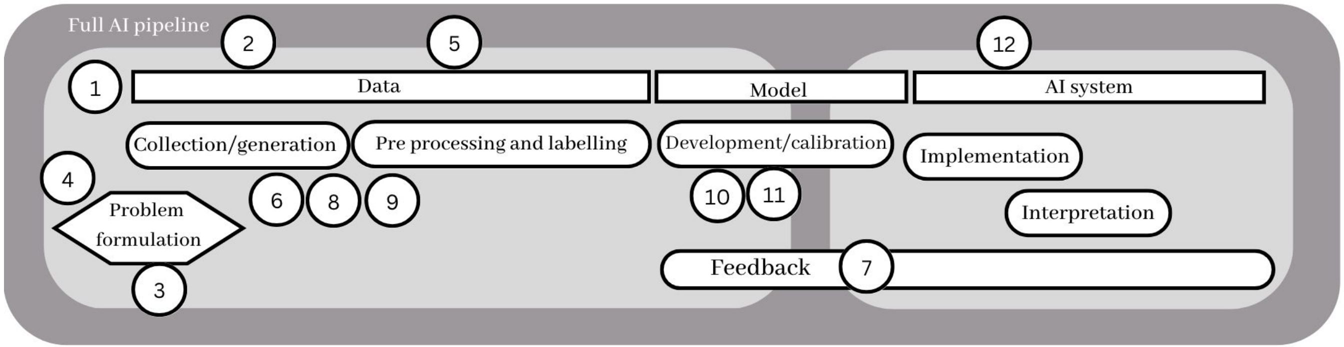
1 Historical bias
2 Lack of clear goals
3 Aggregation bias
4 Representation bias
5 Data quality issues
6 Sampling and self selection biases
7 Measurement bias
8 Labelling bias
9 Algorithmic design bias
10 Inadequate features and correlations
11 Confirmation and recall biases
12 Causation assumption
13 User knowledge bias
14 Inappropriate context reuse
15 Gamification bias
16 Cognitive bias
17 Emerging bias

Artifacts
Procedural steps
Reflective step

“**before there are data, there are people**—people who offer up their experience to be counted and analyzed, **people who perform that counting and analysis**, people who visualize the data and promote the findings of any particular project, and people who use the product in the end. There are also, always, people who go uncounted—for better or for worse. And there are problems that cannot be represented—or addressed—by data alone.”

D'Ignazio, Catherine; Klein, Lauren F.. Data Feminism (Strong Ideas) (p. 23). The MIT Press. Kindle Edition

Mitigation strategies have also been proposed



Mitigation strategies located in the AI pipeline

1 Being aware of lack of representation.

2 Check for the introduction of cognitive biases.

3 Revising the alignment of project values and decisions.

4 Justifications to pursue the project.

5 Check for proxy variables.

6 Implementing retroactive data collection.

7 Involving impacted communities for feedback.

8 Scrutinising existing/available datasets.

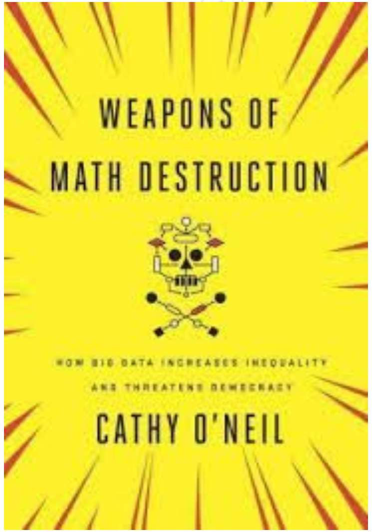
9 Diversification of data sources/type.

10 Dynamic testing, e.g., cryptographic commitment, fair random choices, zero-knowledge proofs .

11 Adversarial debiasing and disparate impact remover.

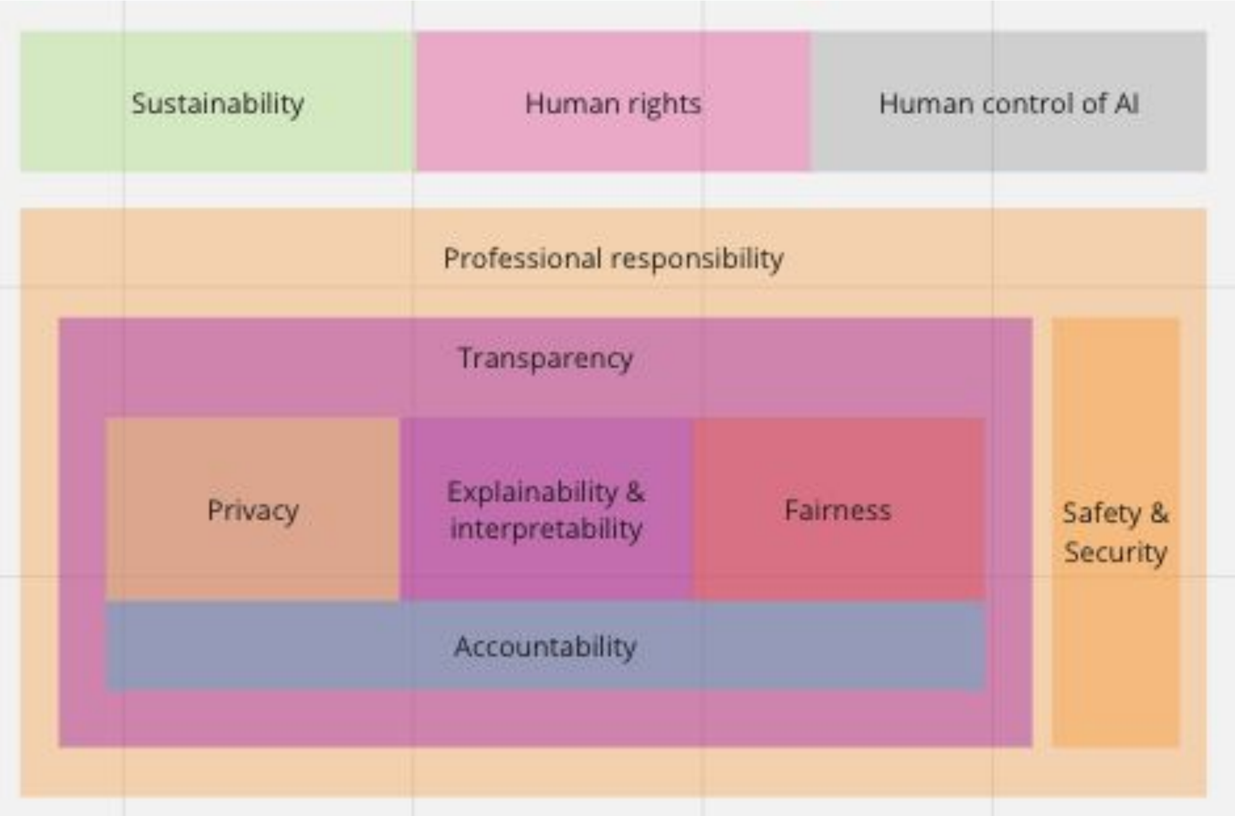
12 Adjusting the model for contextual implementation.

"Big Data processes **codify the past**. They do not invent the future. Doing that requires moral imagination, and that's something only humans can provide. **We have to explicitly embed better values** into our algorithms, creating Big Data models that follow our ethical lead. Sometimes that will mean putting fairness ahead of profit."



O'Neil, C. (2017). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

What are the principles we could prioritize?



With all of that in mind, we are testing two approaches

Biased network approach (BNA)

- As an alternative to isolated mitigation strategies
- Group interview to identify networks of biases at play
- Use it to articulate “threads” of interconnected issues to address

Checklist for socio-ethical reflection

- Problem and goals
- Principles and methods
- Roles and their interest
- Positive and negative impacts
- Data (why this data?)
- Risks and mitigation strategies

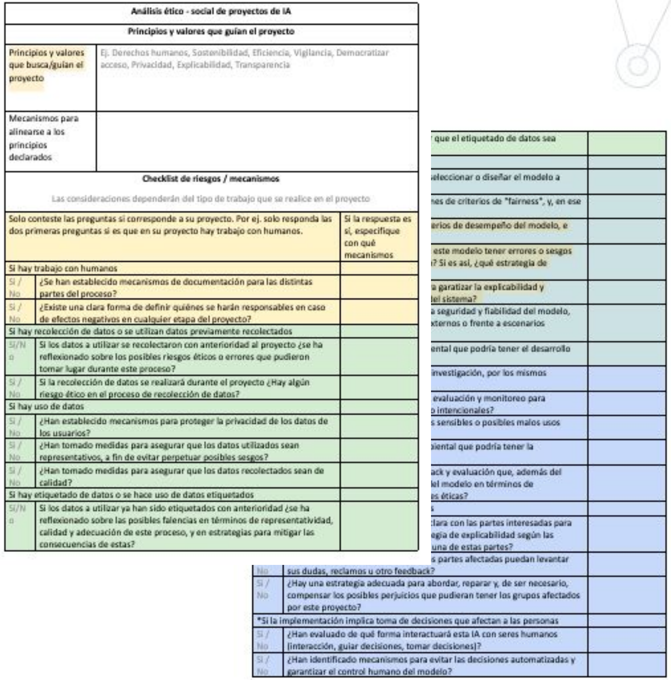
Building up from principles to strategies

Checklist for socio-ethical reflection

Goals: how AI solution helps to achieve the project goal ?

Principles: what are the principles that guide your project?

Methods: how your methods are aligned to the declared principles?



Análisis ético - social de proyectos de IA		
Principios y valores que guían el proyecto		
Principios y valores que busca/guían el proyecto	Ej. Derechos humanos, Sostenibilidad, Eficiencia, Vigilancia, Democratizar acceso, Privacidad, Explicabilidad, Transparencia	
Mecanismos para alinearse a los principios declarados		
Checklist de riesgos / mecanismos Las consideraciones dependerán del tipo de trabajo que se realice en el proyecto		
Solo conteste las preguntas si corresponde a su proyecto. Por ej. solo responda las dos primeras preguntas si es que en su proyecto hay trabajo con humanos.		Si la respuesta es sí, especifique con qué mecanismos
Si hay trabajo con humanos		
Sí / No	¿Se han establecido mecanismos de documentación para las distintas partes del proceso?	
Sí / No	¿Existe una clara forma de definir quiénes se harán responsables en caso de efectos negativos en cualquier etapa del proyecto?	
Si hay recolección de datos o se utilizan datos previamente recolectados		
Sí/No	Si los datos a utilizar se recolectaron con anterioridad al proyecto ¿se ha reflexionado sobre los posibles riesgos éticos o errores que pudieron tomar lugar durante este proceso?	
Sí / No	Si la recolección de datos se realizará durante el proyecto ¿Hay algún riesgo ético en el proceso de recolección de datos?	
Si hay uso de datos		
Sí / No	¿Han establecido mecanismos para proteger la privacidad de los datos de los usuarios?	
Sí / No	¿Han tomado medidas para asegurar que los datos utilizados sean representativos, a fin de evitar perpetuar posibles sesgos?	
Sí / No	¿Han tomado medidas para asegurar que los datos recolectados sean de calidad?	
Si hay etiquetado de datos o se hace uso de datos etiquetados		
Sí/No	Si los datos a utilizar ya han sido etiquetados con anterioridad ¿se ha reflexionado sobre las posibles falencias en términos de representatividad, calidad y adecuación de este proceso, y en estrategias para mitigar las consecuencias de estas?	

No	sus dudas, reclamos u otro feedback?	
Sí / No	¿Hay una estrategia adecuada para abordar, reparar y, de ser necesario, compensar los posibles perjuicios que pudieran tener los grupos afectados por este proyecto?	
*Si la implementación implica toma de decisiones que afectan a las personas		
Sí / No	¿Han evaluado de qué forma interactuará esta IA con seres humanos (interacción, guiar decisiones, tomar decisiones)?	
Sí / No	¿Han identificado mecanismos para evitar las decisiones automatizadas y garantizar el control humano del modelo?	

Checklist for socio-ethical reflection

Roles: who are the active/passive actors who are relevant to the project? What are their interests? What are the positive and negative possible impacts for them? Can they be against the principles?

Data: Why did you choose this dataset instead of the alternatives you had? Who created this dataset? For what purpose? Was there any pre-processing? Who will get benefits from the data analysis?

Additional specific questions

Are you **documenting the decisions** in each step of your project?

Other questions are necessary if there is

- Human subjects
- Data collection or reuse
- Personal data use
- Labelled data

- Implementation beyond research
- Model selection
- Decisions that affect people
- Implementations that affect people

Preliminary results

- It enabled us to identity gaps in documentation and data protection

- It opens new questions about the project’s goals and alignment with initial goals

- But it generates a sense of being overwhelmed by multiple pending issues

Takeaways

Questions I think we should make ourselves

Is our data good enough? Are we doing enough to avoid harms?

There are methods (datasheets for datasets, model cards, fairness metrics, IBM Fairness 360)

What is the purpose of our data analysis? Are people affected by it involved in the process? Who gets the benefits?

There are several approaches, including participatory design.

Thank you!

You can email me at

claudia@inf.utfsm.cl

CENTRO NACIONAL DE INTELIGENCIA ARTIFICIAL

Futures of Artificial Intelligence Research

UNIVERSIDAD TECNICA FEDERICO SANTA MARIA

DEPARTAMENTO DE INFORMÁTICA