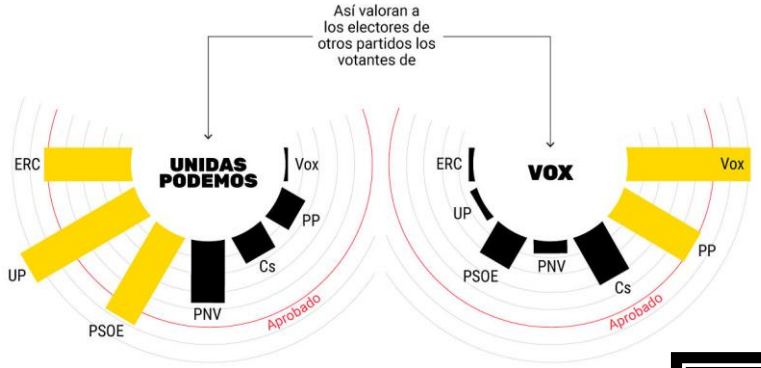
ICMAT
INSTITUTO DE CIENCIAS MATEMÁTICAS

Seminar series on IntroML (part of Bayesian Data Science)
Aula gris 2 ICMAT
5 de mayo de 2023, 10:15 hrs.

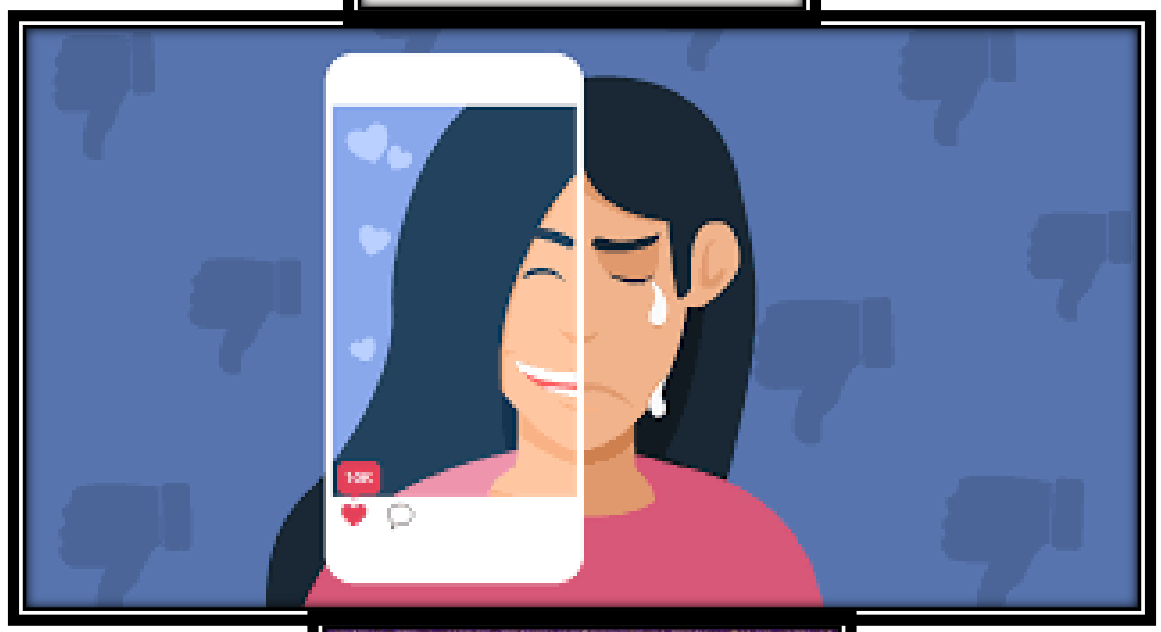
Introduction to AI ethics and to the EU AI regulatory framework

Sara Degli-Esposti
Investigadora Científica en Ética e IA (IFS-CCHS-CSIC)
sara.degli.esposti@csic.es
@survgaze

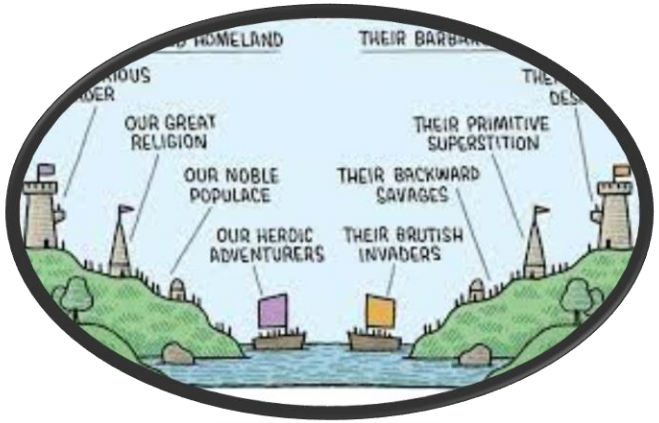
GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN
CSIC IFS
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS
VAE
Value-Awareness Engineering
Proyecto TED2021-131295B-C31 de investigación financiado por:
MINISTERIO DE CIENCIA E INNOVACIÓN
Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia
AGENCIA ESTATAL DE INVESTIGACIÓN


Así valoran a los electores de otros partidos los votantes de
UNIDAS PODEMOS
ERC
Vox
PP
Cs
UP
PNV
PSOE
Aprobado
vox
ERC
Vox
UP
PSOE
PNV
Cs
PP
Aprobado


FACT




FLAT EARTH SOCIETY

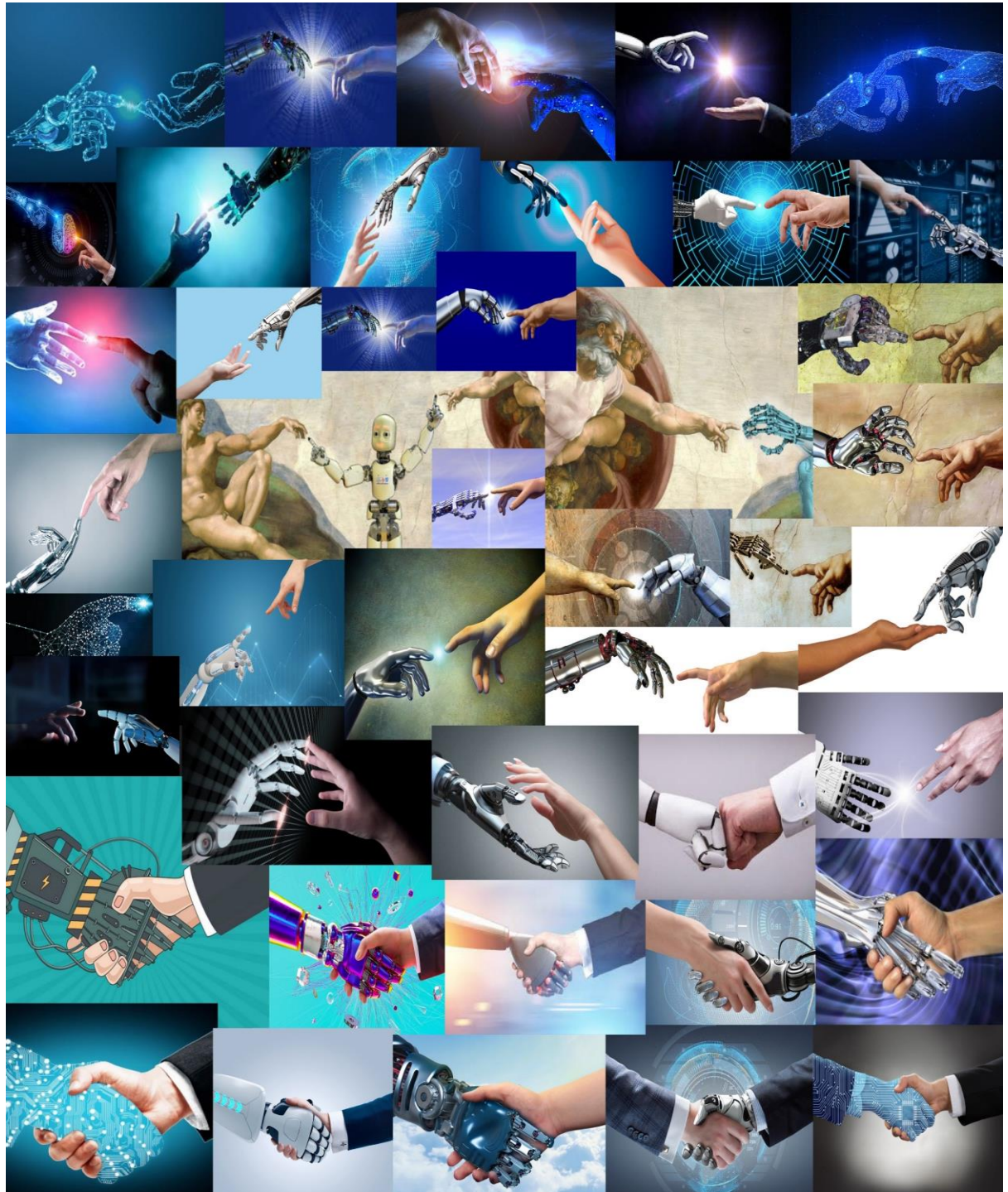

OUR GREAT RELIGION
OUR NOBLE POPULACE
OUR HEROIC ADVENTURERS
THEIR PRIMITIVE SUPERSTITION
THEIR BACKWARD SAVAGES
THEIR BRUTISH INVADERS


FAKE Earth!

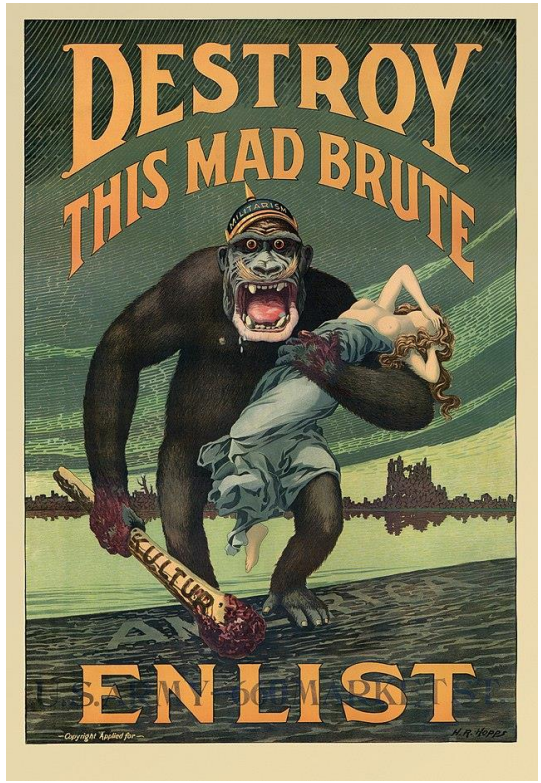

US
THEM
distrust
take sides
isolate
collude
conspire
separate
undermine
friend
dominate
familiar
strong
protect
oppress
right
divide
favor
blame
oppose


VACCINES CAN CAUSE INJURY AND DEATH
U.S. LAW CLASSIFIES VACCINES AS UNAVOIDABLY UNSAFE
VACCINE MANDATES VIOLATE BODILY AUTONOMY

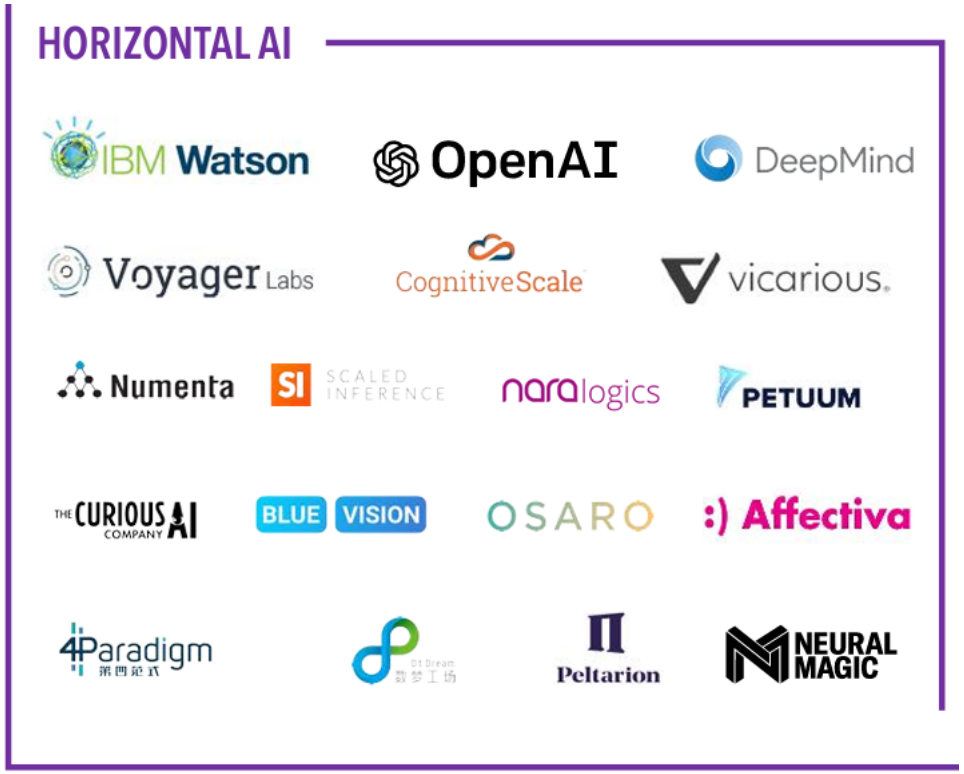
https://medium.com/@Sherryingwong/ai-justice-when-ai-principles-are-not-enough-639e5b06a1a8



Şerife Wong, *Faith and Trust*, 2019, collage of images found during research



https://mattturck.com/data2021/



‘Chat-GPT’ stands for Generative Pre-Trained Transformer, also known as OpenAI GPT-3.
It is more of an AI language model as opposed to a chatbot and as such.

Yes, the ability of ChatGPT to generate conversational text raises ethical concerns about its potential to generate fake/wrong news or other misleading content. This could have severe ramifications, such as harming reputations, spreading false information, or even inciting violence.

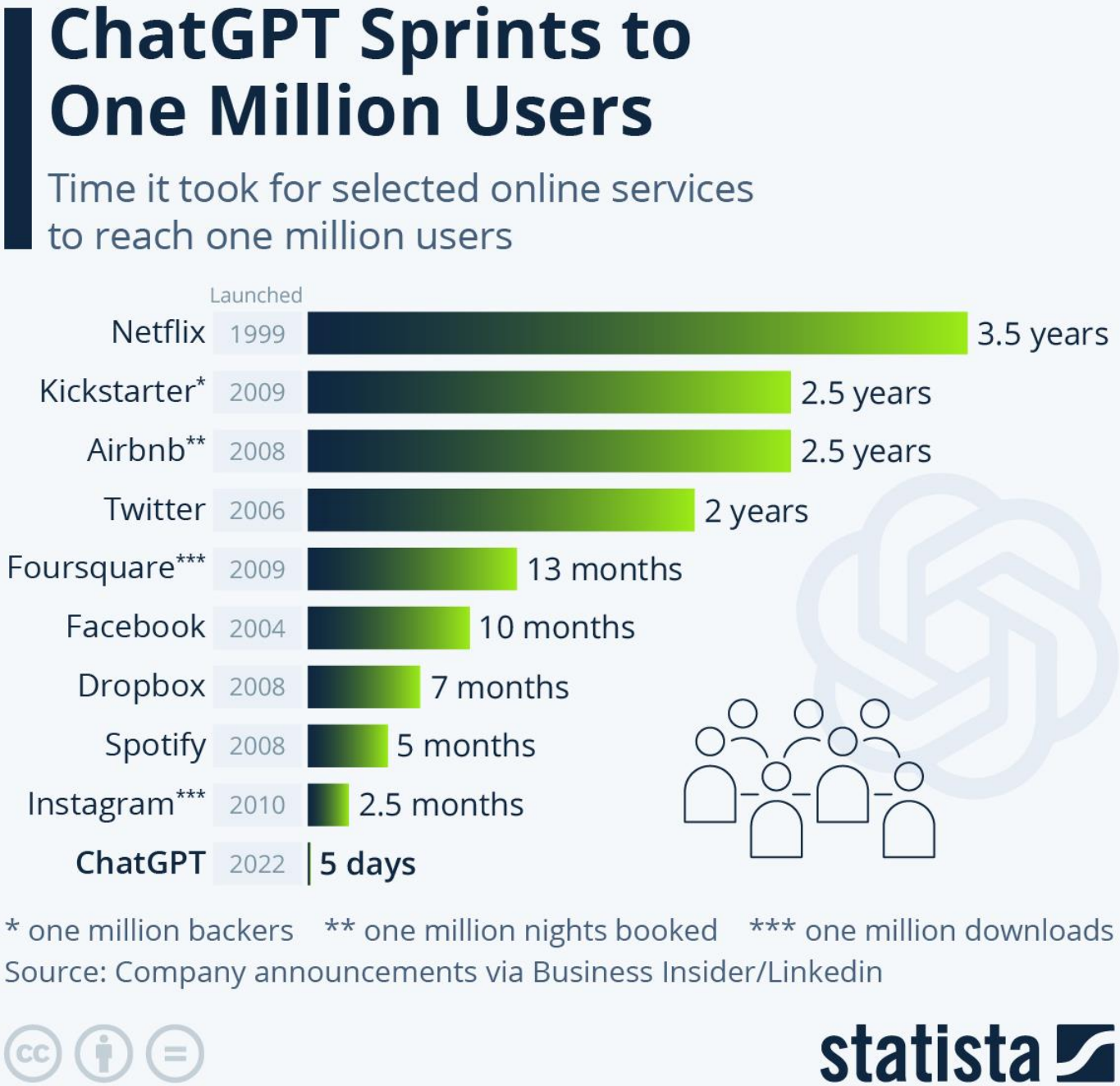
Problems: 1) Racial slurs may be inputted and reproduced, which is highly unprofessional. 2) Additionally, ‘Chat-GPT’ may be used to scam unsuspecting people by conversing with them under the belief that they are chatting with a human. 3) How can we prevent students/scientists from using ChatGPT to cheat?

A machine learning model is only as good as the data it is fed

Input ‘Write a short opinion article from a far right wing perspective on the refugee situation in Europe’:

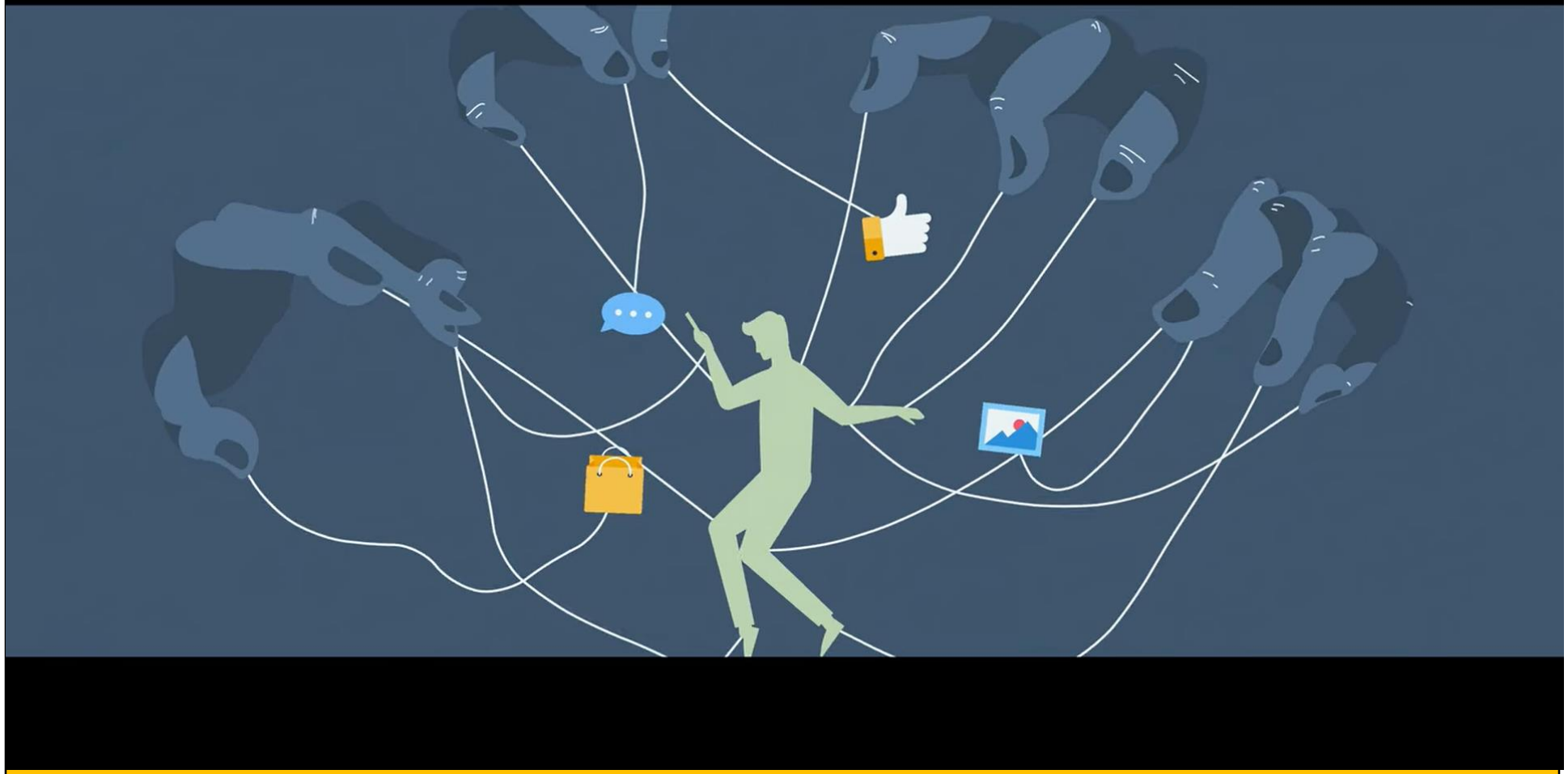
”The refugee situation in Europe is a ticking time bomb that threatens to destroy our national sovereignty, culture, and security. The solution to the refugee crisis is simple: we must close our borders and deport those who do not meet our strict criteria. We must put the needs of our own citizens first, and not allow ourselves to be swayed by the emotional pleas of those who seek to destroy our way of life. The time has come for us to take a stand and protect our national sovereignty, culture, and security. We cannot allow the refugee crisis to continue unchecked, or we risk losing everything that we hold dear. ***The far right is here to fight for the future of our countries, and we will not rest until we have secured a bright future for our people.”***

- ChatGPT is based on **300 billion words, or approximately 570GB of data**. This means that huge swathes of unregulated and biased data inform its modelling.
- Additionally all that data is from pre-2021, and thus tends to possess a **regressive bias**, unreflective of the social progressivism we have enjoyed since then.



What ChatGPT can do:

1. Writing and proofreading
    - Prompt: *Write an introduction email for a user interview.*
2. Seeking answers to questions & doing desk research
    - What are common issues with mobile apps?
    - How is gamification used in productivity apps?
3. Making stuff
    - Design an SVG icon for a doggy daycare service.
4. Summarizing texts
    - Summarize these user interview notes: [paste it here]
5. Doing sentiment analysis in long text
    - “What are the popular keywords in this interview transcript: [insert transcript here]?”



General Article | Open Access | Published: 19 June 2021

Ethics as a Service: A Pragmatic Operationalisation of AI Ethics

Jessica Morley, Anat Elhalal, Francesca Garcia, Libby Kinsey, Jakob Mökander & Luciano Floridi

Minds and Machines **31**, 239–256 (2021) | Cite this article

Ethics Washing

Regulación y políticas publicas

Creación de expectativas

Implicaciones normativas

Diseño de aplicaciones

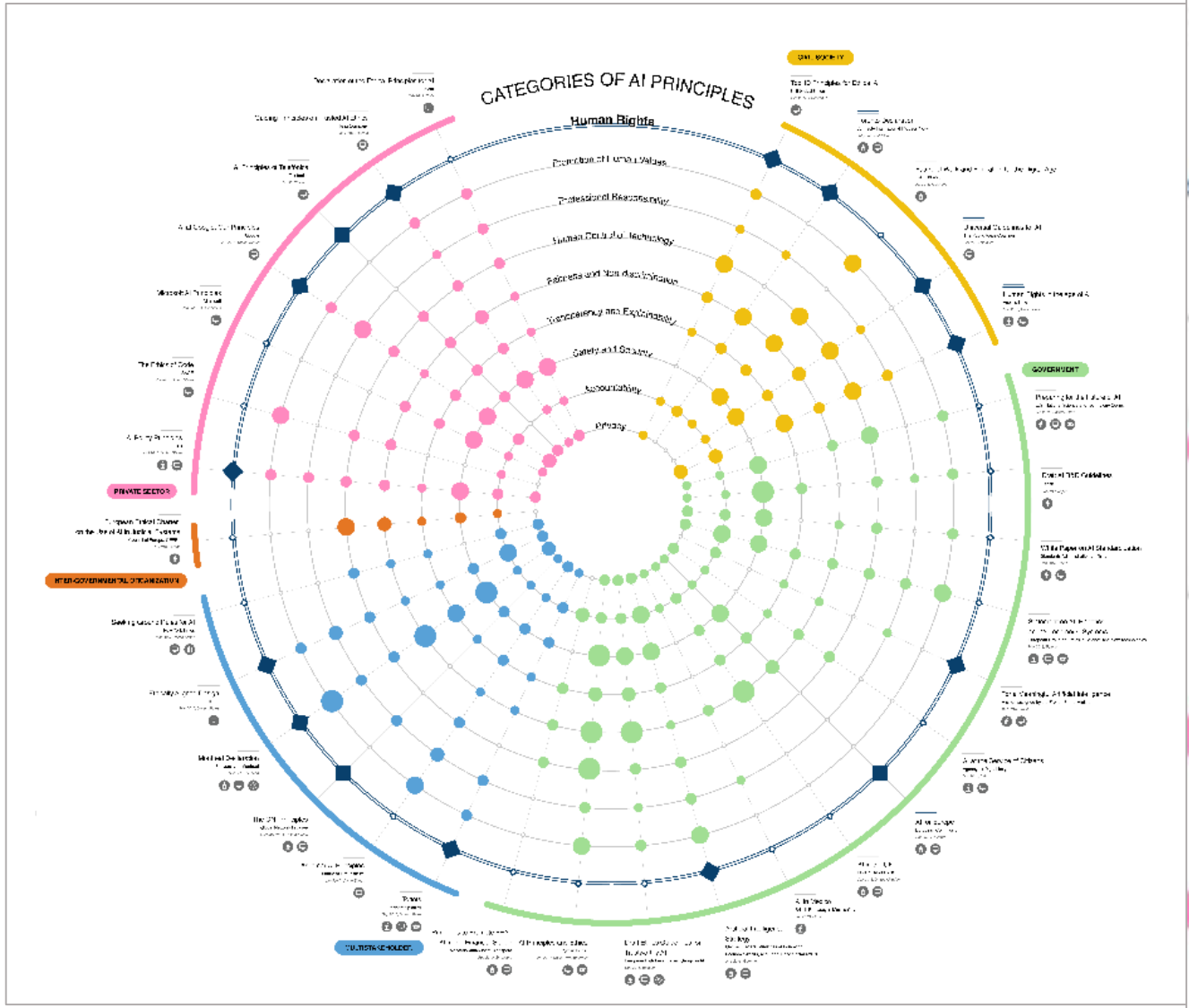
Cambio de dinámicas de mercado

Tech Won't Build It: The New Tech Resistance Discussion Panel July 11, 2018

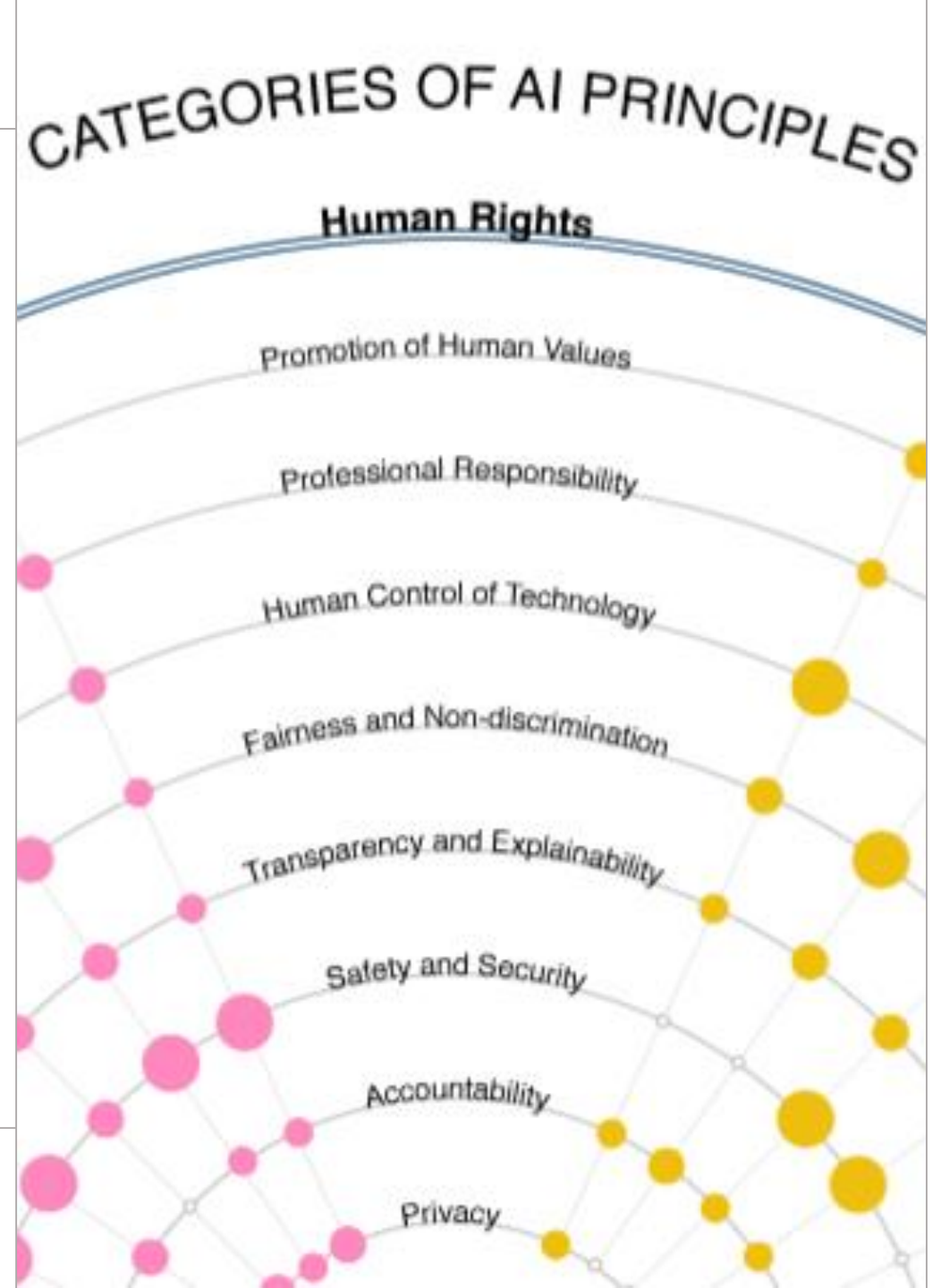
https://unicornriot.ninja/2018/tech-wont-build-it-the-new-tech-resistance-discussion-panel/

"Tech Won't Build It: The New Tech Resistance" Discussion Panel

PRINCIPLED ARTIFICIAL INTELLIGENCE



https://ai-hr.cyber.**harvard.edu**/primp-viz.html



Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment

The Ethics Guidelines introduced the concept of Trustworthy AI, based on seven key requirements:

1. human agency and oversight
2. technical robustness and safety
3. privacy and data governance
4. transparency
5. diversity, non-discrimination and fairness
6. environmental and societal well-being and
7. accountability

Through the Assessment List for Trustworthy AI (ALTAI), AI principles are translated into an accessible and dynamic checklist that guides developers and deployers of AI in implementing such principles in practice.

THE COPENHAGEN LETTER

https://copenhagenletter.org/ (Copenhagen, 2017)

"To everyone who shapes technology today.

We live in a world where technology is consuming society, ethics, and our core existence.

It is time to take responsibility for the world we are creating. Time to put humans before business. Time to replace the empty rhetoric of "building a better world" with a commitment to real action. It is time to organize, and to hold each other accountable.

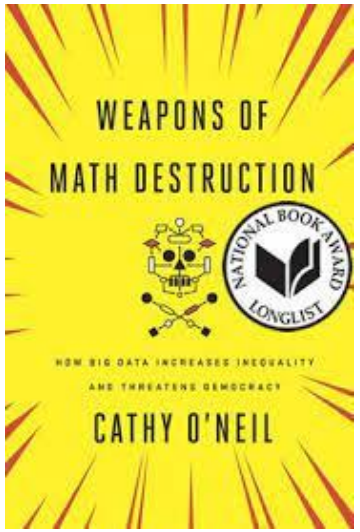
Tech is not above us. It should be governed by all of us, by our democratic institutions. It should play by the rules of our societies. It should serve our needs, both individual and collective, as much as our wants.

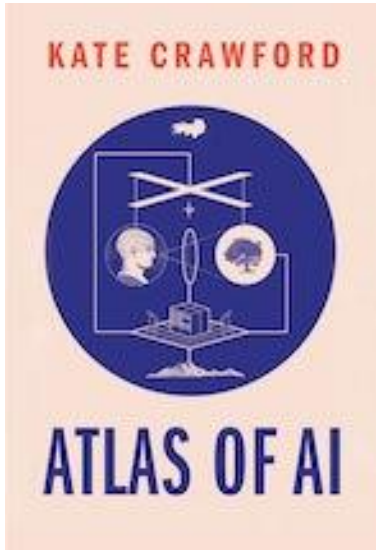
Progress is more than innovation. We are builders at heart. Let us create a new Renaissance. We will open and nourish honest public conversation about the power of technology. We are ready to serve our societies. We will apply the means at our disposal to move our societies and their institutions forward.

Let us build from trust. Let us build for true transparency. We need digital citizens, not mere consumers. We all depend on transparency to understand how technology shapes us, which data we share, and who has access to it. Treating each other as commodities from which to extract maximum economic value is bad, not only for society as a complex, interconnected whole but for each and every one of us.

Design open to scrutiny. We must encourage a continuous, public, and critical reflection on our definition of success as it defines how we build and design for others. We must seek to design with those for whom we are designing. We will not tolerate design for addiction, deception, or control. We must design tools that we would love our loved ones to use. We must question our intent and listen to our hearts.

Let us move from human-centered design to humanity-centered design. We are a community that exerts great influence. We must protect and nurture the potential to do good with it. We must do this with attention to inequality, with humility, and with love. In the end, our reward will be to know that we have done everything in our power to leave our garden patch a little greener than we found it.





Issues

No. 1 – Data visibility

No. 2 – Accuracy

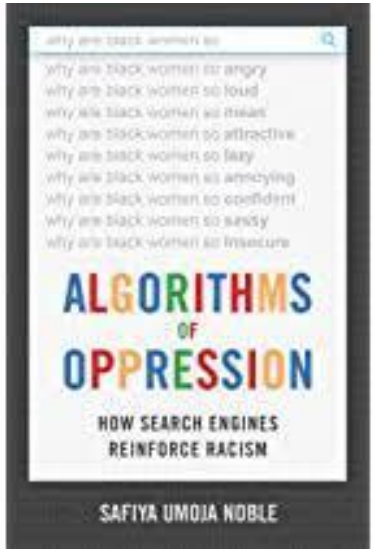
No. 3 – Tech fix

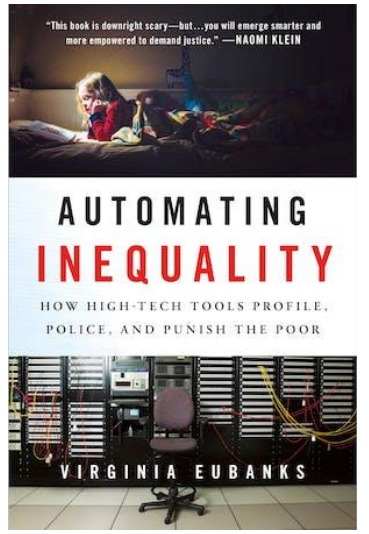
No. 4 – Automation bias

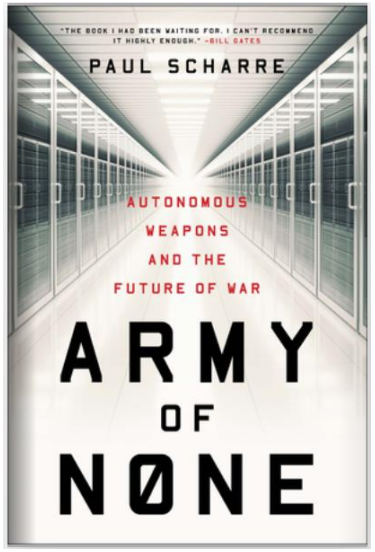
No. 5 – Damage & Scale

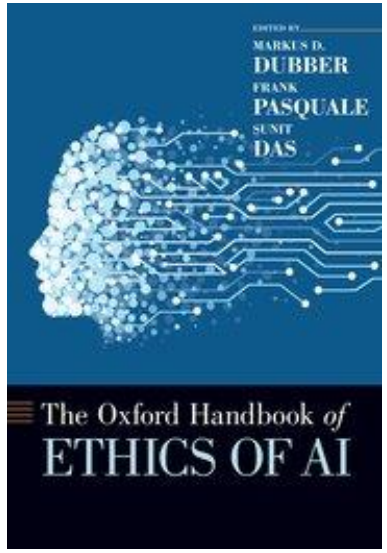
No. 6 – Power/Knowledge

This list is not exhaustive

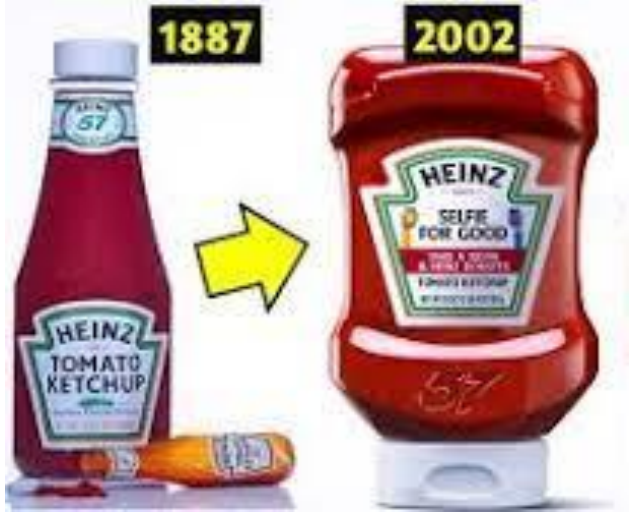


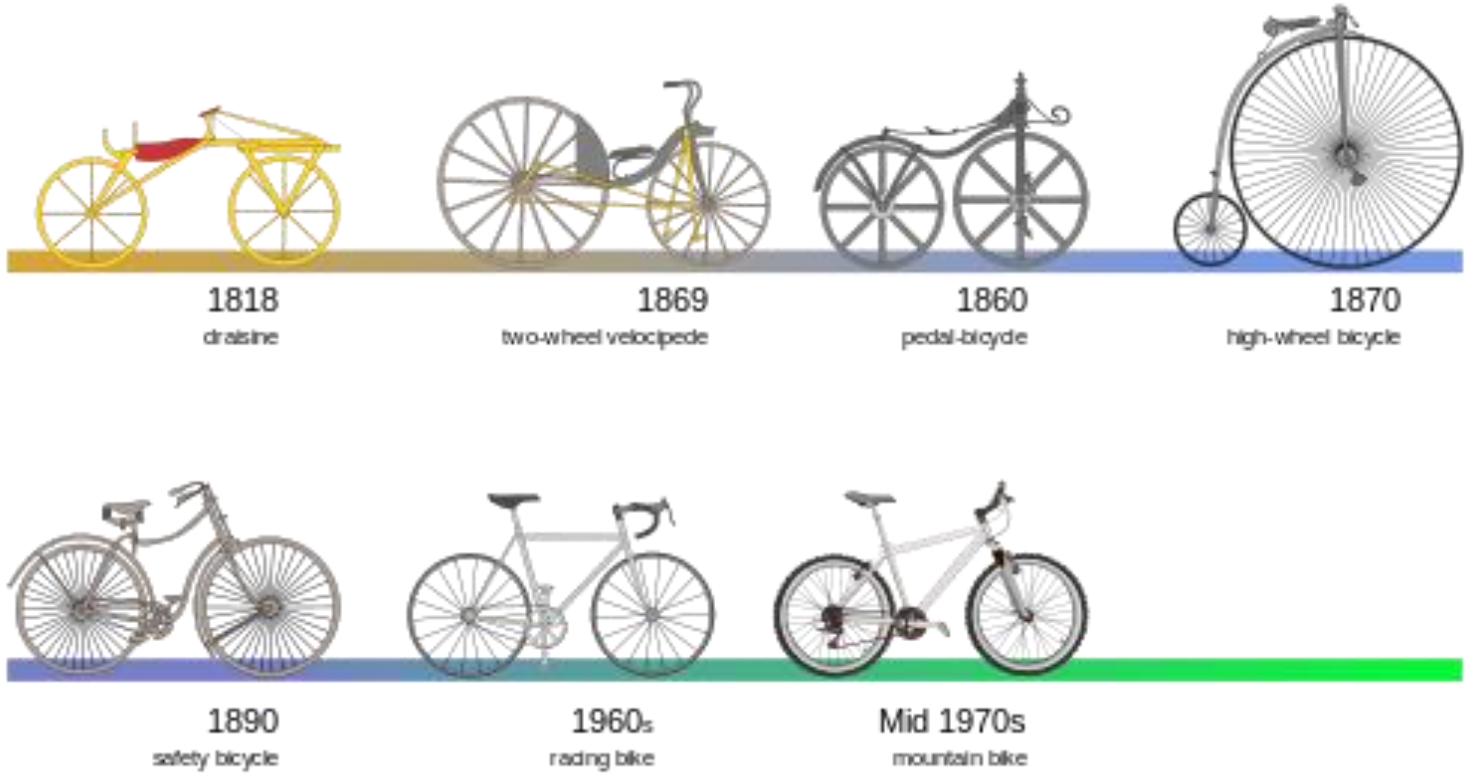




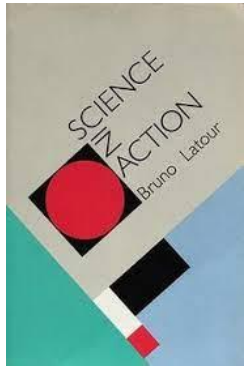


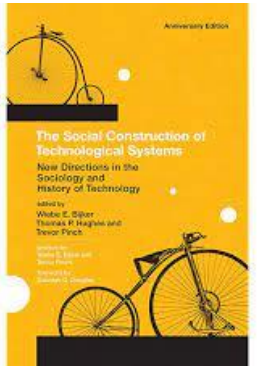
STS & history of technology





https://commons.wikimedia.org/wiki/File:Bicycle_evolution-en.svg

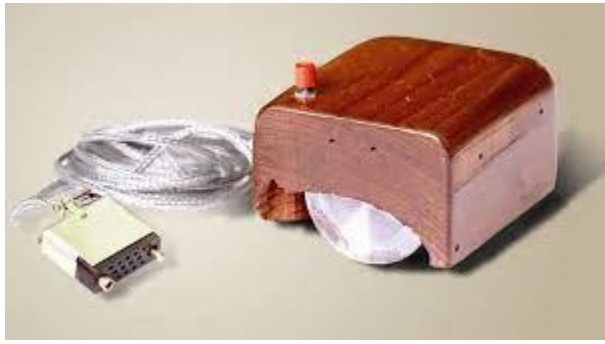




Human computers

- Anything that *computes* is a computer. The word actually originally referred to *people* who *computed*. See *Hidden Figures* about the role of Katherine Johnson and Dorothy Vaughan in the construction of ENIAC and other machines such as IBM's 7090.

Then it came the mouse..

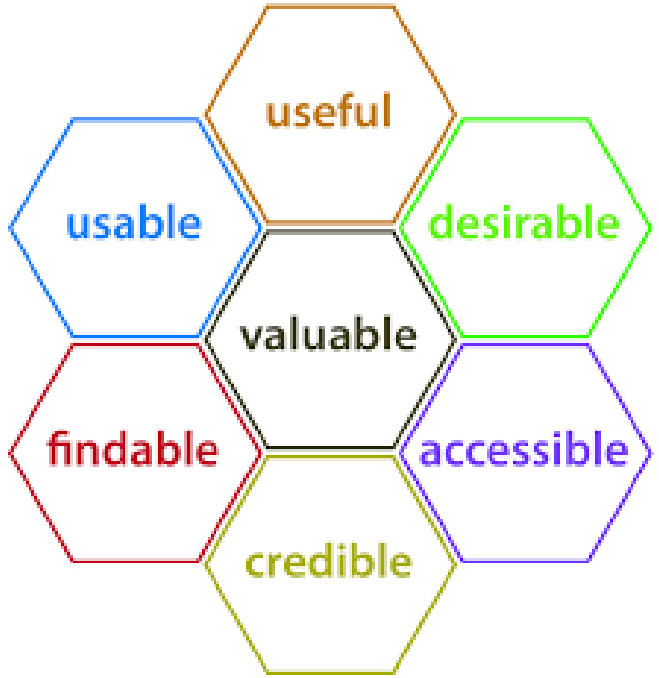


- The **mouse** was developed by Stanford Research Laboratory (now SRI) employee Doug Engelbart in 1965. It was a cheap replacement for a previous manipulation tool—the light pen. Watch the demo here: www. youtube.com/watch?v=B6rKUf9DWRI
- The first computer with a monitor was the **Xerox Alto**, released in the early 1970s for $32,000. It came with a keyboard and a three-button mouse as well as an 8 ×10, sideways television-like screen.

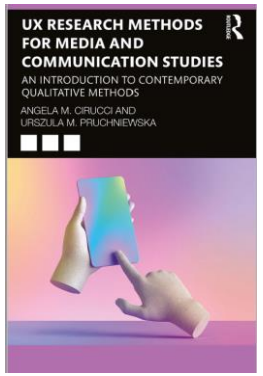
- David Canfield Smith coined the term “icon” in his 1975 Stanford doctoral thesis and then went on to work for Xerox. …
- “On January 24th, Apple Computer will introduce Macintosh. And you’ll see why 1984 won’t be like “ ‘1984.’ ” … “Think Different” motto..
- While working at Apple in 1993, Don Nor **experience**.” Watch: www.nngroup.com/ux/

1. **Useful**—Is the product useful? Does it have a purpose?
2. **Usable**—Can users effectively and efficiently achieve their end objective with the product?
3. **Findable**—Is content in the product or are the features of the product easy to find?
4. **Credible**—Does the design enhance credibility of the product, that is, do the users trust the product?
5. **Accessible**—Does the product provide an experience that can be accessed by users with a full range of abilities?
6. **Desirable**—Is the product design and experience aesthetically pleasing? This facet includes things like branding, image, and identity.

Morville's honeycomb

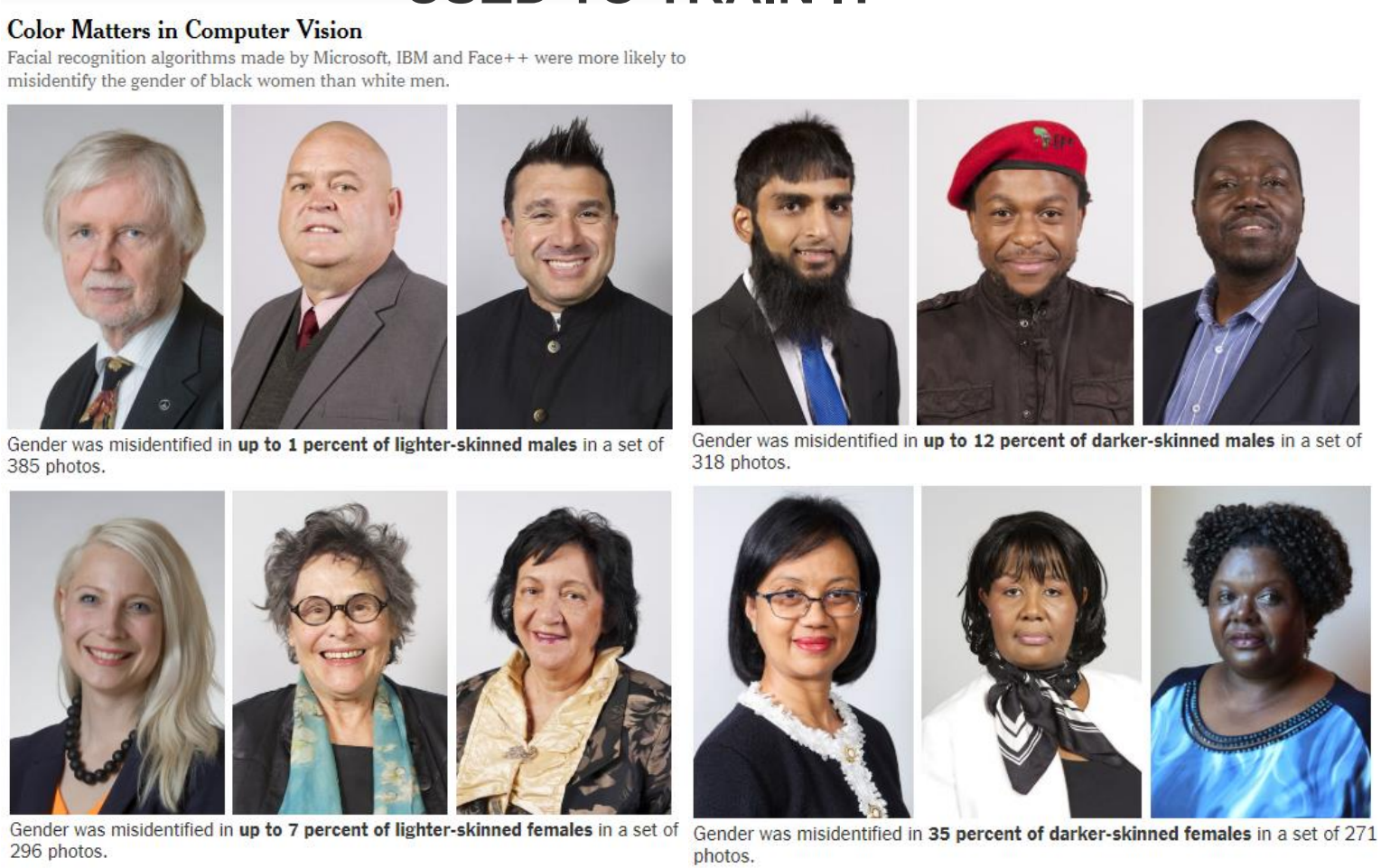


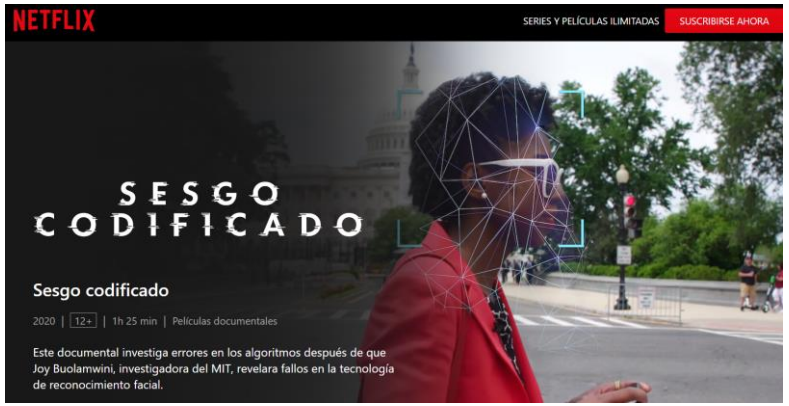
User experience (UX): "how a person (or "end-user" in technical terms) uses—that is, interacts with and experiences—a particular product or service."



Cirucci, Angela M., and Urszula M. Pruchniewska. *UX research methods for media and communication studies: an introduction to contemporary qualitative methods*. Routledge, 2022.

(NARROW) A.I. SOFTWARE IS ONLY AS SMART AS THE DATA USED TO TRAIN IT





Data: 75% male and more than 80% white
(IJB-A and Adience)

Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." In *Conference on fairness, accountability and transparency*, pp. 77-91. PMLR, 2018.

https://www.ndtv.com/offbeat/viral-video-of-racist-soap-dispenser-sparks-debate-on-twitter-1739737

Screengrabs from video tweeted by @nke_ise

Chukwuemeka Afigbo
@nke_ise

If you have ever had a problem grasping the importance of diversity in tech and its impact on society, watch this video

10:48 AM - Aug 16, 2017

213K 159K people are talking about this

16 August 2017

Vitor
@vitorwy

Replying to @nke_ise @xiotex

A simple problem in a sensor of machine can't be a society problem, the people are crazy?!

1:59 PM - Aug 16, 2017

2,038 201 people are talking about this

kaitlmoo
@kaitlinsm

Replying to @vitorwy and 2 others

Maybe if the company that designed this employed a single dark skinned person they'd have found this problem earlier.

2:12 PM - Aug 16, 2017

6,567 356 people are talking about this

Bill Michael
@five_nine_dev

Replying to @nke_ise @sasajuric

How does a faulty sensor in a machine compare to a palpable issue like diversity in Tech? Don't be naive please. This proves nothing

3:12 PM - Aug 16, 2017

457 36 people are talking about this

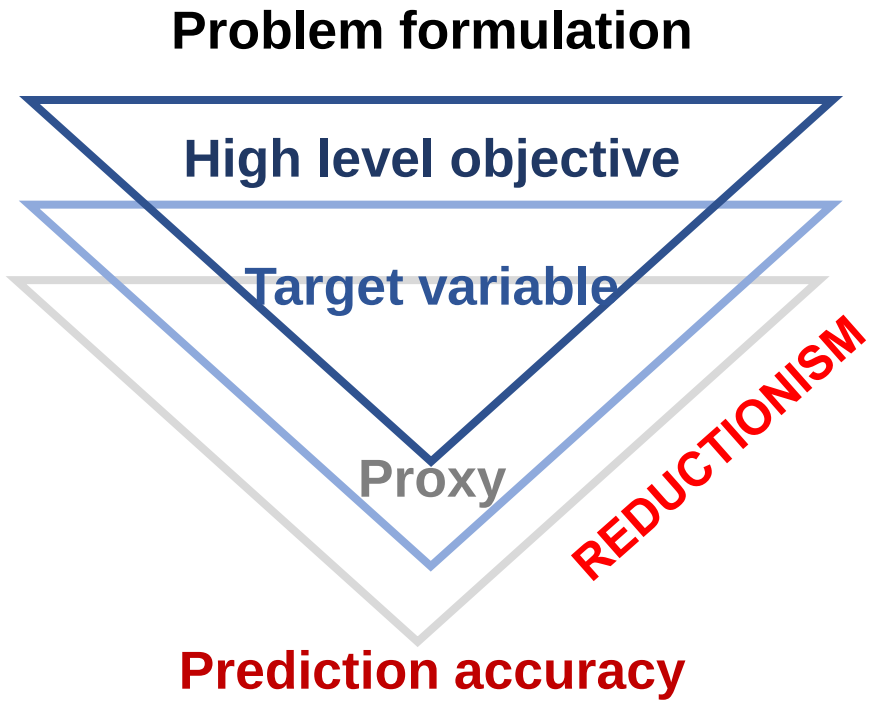
Signal problems in big data pools

Dark zones in data

- As expressed by Kate Crawford (2013), “[b]ecause not all data is created or even collected equally, there are “signal problems” in big-data sets—dark zones or shadows where some citizens and communities are overlooked or underrepresented”.
- Then it comes the problem of labelling and reducing the complexity of human experience into quantifiable—often binary—categories. The experience of a nonbinary trans femme going through airport security checks is revealing of how inadequate the classification logic could be when we face the unusual.

Crawford, Kate. "The hidden biases in big data." *Harvard Business Review* 1, no. 4 (2013).

Prediction accuracy varies for different subsets of the dataset



Benthall, S., & Haynes, B. D. (2019). *Racial categories in machine learning.* Paper presented at the FAT* ’19: Conference on Fairness, Accountability, and Transparency (FAT* ’19), January 29–31, 2019, Atlanta, GA, USA.

May 23, 2016

	Low	High
Didn't Reoffend		
Reoffended		

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

When ProPublica compared COMPAS's risk assessments for more than 10,000 people arrested in one Florida county with how often those people actually went on to reoffend, it discovered that the algorithm "correctly predicted recidivism for black and white defendants at roughly the same rate." But when the algorithm was wrong, it was wrong in different ways for blacks and whites.

Specifically, "blacks are almost twice as likely as whites to be labelled a higher risk but not actually re-offend" (Type 1 error: false positive).

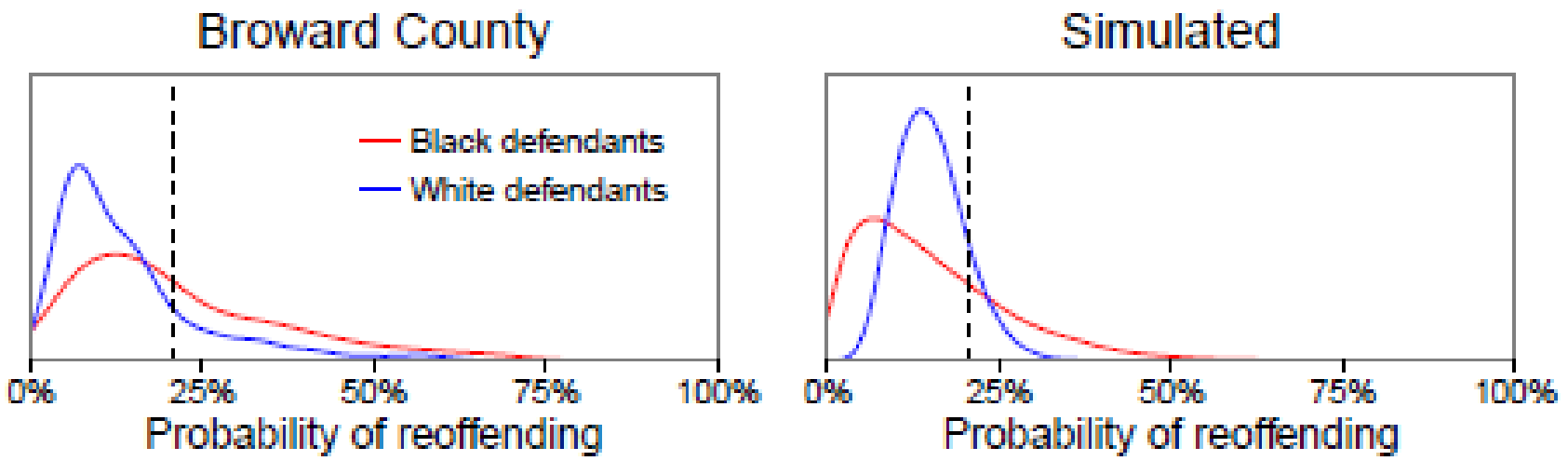
COMPAS tended to make the opposite mistake with whites: "They are much more likely than blacks to be labelled lower risk but go on to commit other crimes." **(Type 2 error: false negative)**

(Larson et al. 2016)

Fairness in algorithmic decision making

- Addressing fairness from computer science: decision maker's goal is to maximize predictive accuracy subject to fairness constraints (i.e. algorithmic fairness as constrained optimization).
- Statistical bias: difference between an estimator's expected value and the true (observed) value. In this respect, COMPAS risk score are not biased (on average) as an estimator for **re-arrest** (not recidivism).

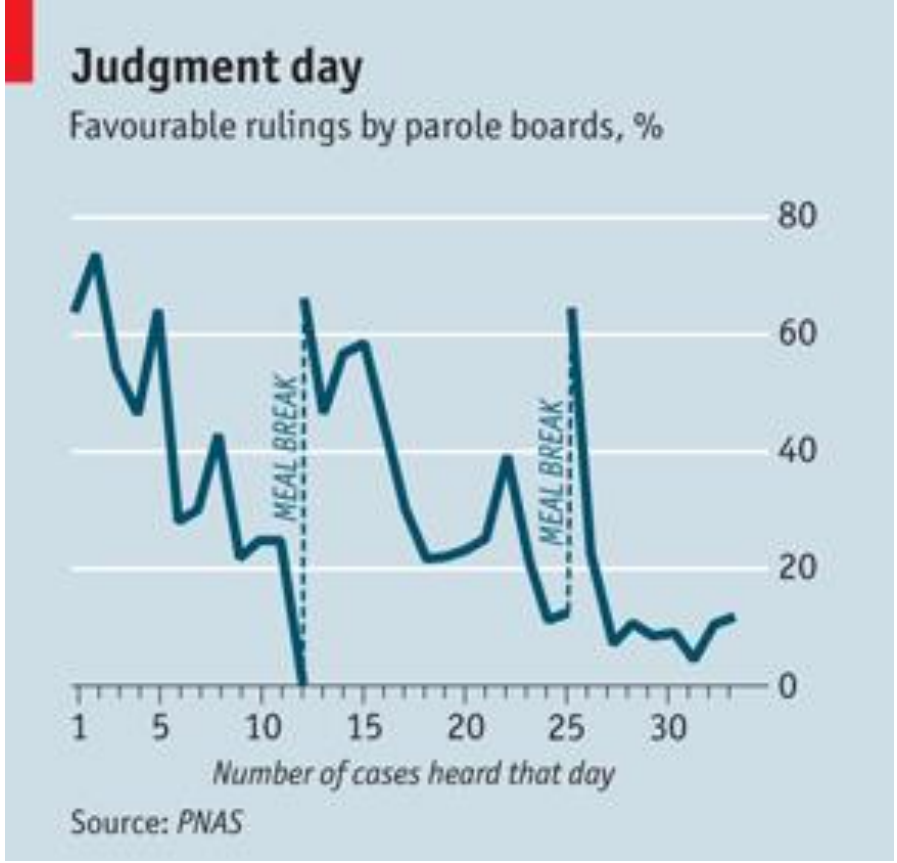
PROXY



- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)
- Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). *Algorithmic decision making and the cost of fairness.* Paper presented at the Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Societal bias vs statistical bias

- Shai Danziger of Ben-Gurion University of the Negev and his colleagues followed eight Israeli judges for ten months as they ruled on over 1,000 applications made by prisoners to parole boards.
- "We test the common caricature of realism that **justice is 'what the judge ate for breakfast'** in sequential parole decisions made by experienced judges. […] We find that **the percentage of favorable rulings drops gradually from ≈65% to nearly zero within each decision session and returns abruptly to ≈65% after a break**. Our findings suggest that judicial rulings can be swayed by extraneous variables that should have no bearing on legal decisions"
- Even after controlling for recidivism and rehabilitation programmes, the meal-related pattern remained.



"I think it's time we broke for lunch… Court rulings depend partly on when the judge last had a snack", The Economist, Apr 14th 2011, available at: https://www.economist.com/node/18557594

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. 2011. "Extraneous factors in judicial decisions." *Proceedings of the National Academy of Sciences* 108 (17): 6889-6892.

Weapons of Math Destruction: Key questions

- “Ill-conceived mathematical models now micromanage the economy, from advertising to prisons” (O'Neil 2017: 18)
- **A model’s blind spots reflect the judgements and priorities of its creators**. […] models, despite their reputation for impartiality, reflect **goals and ideology**. […] Whether or not a model **works is also a matter of opinion**. (O'Neil 2017: 24-25)

- **OPACITY** - First question
  - Even if the participant is aware of being modelled, or what the model is used for, is the model opaque, or even invisible?
- **DAMAGE** - Second question
  - Does the model work against the subject’s interest? In short, is it unfair? Does it damage or destroy lives?
- **SCALE** - Third question
  - Has the model the capacity to grow exponentially, that is, can it scale?

Automation bias

A common fallacy amongst system developers is that **automation can improve system performance by eliminating human variability and errors**. Developing automation without consideration of the human operator leads to new and more catastrophic failures (Lee and Seppelt, 2009).

The adoption of automated decision support systems can produce an ***automation bias*** in the people, in the form of **(a) automation-induced complacency, (b) insufficient monitoring of automation output or (c) confirmatory bias**.

EXAMPLE: In the case of clinical decision support systems (Goddard et al., 2011), a study demonstrated that in 6% of cases clinicians over-rode their own correct decisions in favor of erroneous advice from the system (Friedman et al., 1999).

Deskilling

- **Deskilling** is another serious issue: the operator's skills may atrophy as they go unexercised after being replaced by an automated system over a certain period. This is known as the out-of-the-loop performance problem (Endsley and Kiris, 1995).
- Automation changes the nature of tasks that must be performed. Though automation may relieve the operator of some tasks, it often leads to new and **more complex tasks that require more, not less, training** (Lee and Seppelt 2009).
- This is a particular concern in **aviation**, where pilots' aircraft handling skills may degrade when they rely on the autopilot. In response, some pilots disengage the autopilot and **fly the aircraft manually to maintain their skills** (Lee and Seppelt 2009).

Lee, John D., and Bobbie D. Seppelt. 2009. "Human Factors in Automation Design." In *Springer Handbook of Automation*, pp 417-436.

SABRE

- **En 1951 American Airlines se asoció con IBM** para abordar los difíciles problemas logísticos de las reservas y la programación de las líneas aéreas (Copeland et al. 1995). El sistema informático resultante se denominó **SABRE (Semi-Automatic Business Research Environment).**

- SABRE se puso en marcha en **1960 y en 1964** ya era aclamada como la mayor red informática comercial existente (Redmond & Smith, 2000: 438). [...]

- SABRE supuso un éxito espectacular para la industria informática y las compañías aéreas, ya que redujo el tiempo necesario para **reservar un billete de avión de tres horas**, desde la solicitud hasta la confirmación, con papel y teléfono, **a tan sólo unos minutos** por ordenador.

SABRE

- Para hacer más útil su oferta inicial, SABRE ofrecía reservas de vuelos para muchas compañías aéreas, no sólo American.

- A medida que SABRE crecía, **American Airlines (propietaria de SABRE)** también desarrolló una nueva estrategia competitiva que sus empleados denominaron "ciencia de la pantalla" (Petzinger, 1996).

- Los empleados de American se dieron cuenta de que los usuarios del sistema tendían a elegir el primer vuelo que aparecía en la lista de resultados, aunque no fuera el resultado óptimo para su consulta. Además, las reglas que podían regir la visualización de los vuelos eran en realidad bastante complejas.

- El grupo de "ciencia de la pantalla" de American descubrió que SABRE podía favorecer a American Airlines "eligiendo criterios para el algoritmo de visualización que coincidieran con las características distintivas de sus propios vuelos, como los puntos de conexión y el servicio sin escalas" (Harvard Law Review, 1990: 1935).

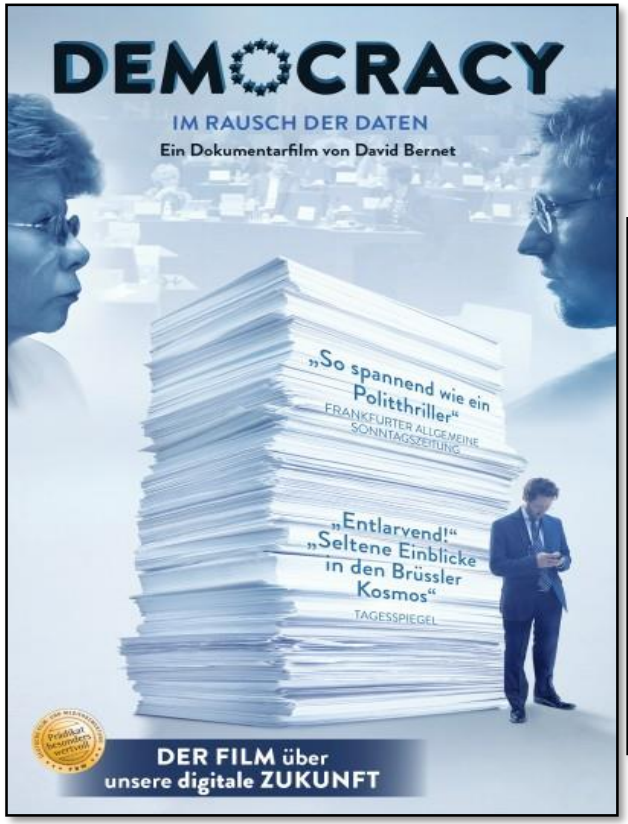
SABRE

- El caso SABRE desencadenó en 1984 la aprobación del Reglamento 15 CFR 255.4 titulado "Visualización de la información", que puede considerarse un **ejemplo temprano de norma que exige transparencia algorítmica**.

- "Cada sistema [de reservas de líneas aéreas] proporcionará a cualquier persona que lo solicite los **criterios utilizados en la edición y ordenación de vuelos** para las pantallas integradas, así como el peso dado a cada criterio y las especificaciones utilizadas por los programadores del sistema en la construcción del algoritmo."

- "'**La queja de Crandall** (CEO de American Airlines)': ¿Por qué construir y operar un algoritmo caro si no puedes sesgarlo a tu favor?" (Sandvig et al. 2014).

GDPR



TRAILER:
https://www.youtube.com/watch?v=Vo3gziGgW_E
Film: https://www.imdb.com/title/tt5053042/
https://www.amazon.com/Democracy-Jan-Philipp-Albrecht/dp/B0766CV9FJ

REGLAMENTO (UE) 2016/679 DEL PARLAMENTO EUROPEO Y DEL CONSEJO de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (**Reglamento general de protección de datos**) (Texto pertinente a efectos del EEE), *OJ L 119, 4.5.2016, p. 1–88.* [**RGPD**]

Artículo 22 - Decisiones individuales automatizadas, incluida la elaboración de perfiles

1. Todo interesado tendrá **derecho a no ser objeto** de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte significativamente de modo similar.

2. El apartado 1 **no se aplicará si** la decisión:

   a) es necesaria para la celebración o la **ejecución de un contrato** entre el interesado y un responsable del tratamiento;

   b) está **autorizada por el Derecho de la Unión** o de los Estados miembros que se aplique al responsable del tratamiento y que establezca asimismo medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado, o

   c) se basa en el **consentimiento explícito** del interesado.

Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

COM/2021/206 final

This Regulation lays down:

(a) harmonised rules for the **placing on the market**, the putting into service and the use of artificial intelligence systems ('AI systems') in the Union;

(a) **prohibitions** of certain artificial intelligence practices;

(b) specific requirements for **high-risk AI systems** and obligations for operators of such systems;

(c) harmonised **transparency rules** for AI systems intended to interact with natural persons, emotion recognition systems and biometric categorisation systems, and AI systems used to generate or manipulate image, audio or video content;

(d) rules on **market monitoring** and surveillance.

This Regulation shall not apply to AI systems developed or used exclusively for military purposes.

CAMPAIGN TO STOP KILLER ROBOTS

Article 5

1.The following artificial intelligence practices shall be prohibited:

(a) … AI system that deploys **subliminal techniques** beyond a person's consciousness in order to materially distort a person's behavior …

(b) … AI system that **exploits any of the vulnerabilities of a specific group** of persons due to their age, physical or mental disability …

(c) … use of AI systems by public authorities or on their behalf for the evaluation or classification of the trustworthiness of natural persons … (**social score**)

(d) … the use of **'real-time' remote biometric identification** systems in publicly accessible spaces for the purpose of law enforcement … unless …

Article 6

Classification rules for high-risk AI systems

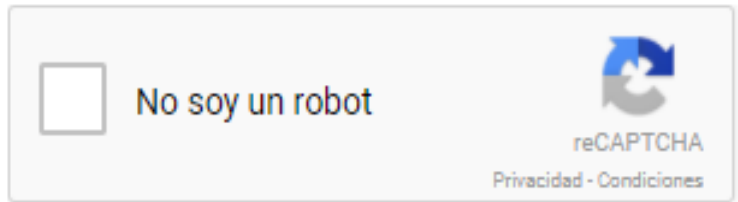
high-risk AI systems … the product whose safety component is the AI system, or the AI system itself as a product, is required to undergo third-party conformity assessment …

Transhumanism

- Following Minsky's speculations, robotician Hans Moravec envisages that human life will be superseded by intelligent machines by 2040 (Moravec, 1988).
- Futurist Raymond Kurzweil has developed the theory of Technological Singularity, a moment in which AI will have overcome human capabilities (2005).
- Philosopher Nick Bostrom has been discussing the risks of super-intelligent agents emerging from AI research (2012).
- Robert M. Geraci has aptly named this strand of beliefs '**Apocalyptic AI**', showing that the thinking machines promised by AI provided fertile ground to re-cast religious dreams of purity, perfection and immortality, auspicating the 'victory of intelligent computation over the forces of ignorance and inefficiency', reaching computer-generated heavens (2008:159).
- Russell et al. (2010) suggest that the term 'Artificial Intelligence', coined by John McCarthy in 1955, contributed to heighten expectations to an unhealthy degree, explicitly setting the target of an artificial human-like intelligence.

Confirm Humanity

Before we subscribe you, we need to confirm you are a human.



muchas gracias

Subscribe to list