UNIVERSITAT DE BARCELONA

MSc in Fundamental Principles of Data Science

Ethical Data Science

Foundations

Jordi Vitrià

2020-2021

The intentional stance

From a scientific point of view, **this course takes an intentional stance towards people, corporations, and human groups, but given that AGI (Artificial General Intelligence) is not yet real, software platforms and algorithms will be considered artifacts that serve the purpose and desires of rational agents**.

The **intentional stance** is a term coined by philosopher Daniel Dennett for the level of abstraction in which we view the behavior of an entity in terms of mental properties. (Source: Wikipedia)

Here is how it works:

- First you decide to treat the object whose behavior is to be predicted as a **rational agent**; then you figure out what **beliefs** that agent ought to have, given its place in the world and its **purpose**.
- Then you figure out what **desires** it ought to have, on the same considerations.
- Finally you predict that this rational agent **will act to further its goals in the light of its beliefs**. A little practical reasoning from the chosen set of beliefs and desires will in most instances yield a decision about what the agent ought to do; that is what you predict the agent will do.

Position Statement

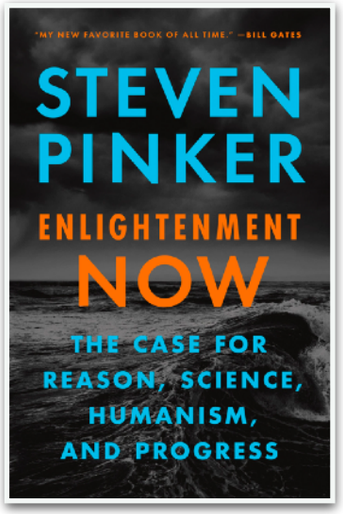
We need a position statement because we will talk about the “**good**” and the “**bad**”

Position Statement

The most important thing is not life, but the good life.

Socrates, 399 B.C.

Ethics in the broadest sense refers to the concern that humans have always had for figuring out *how best to live*. **How do we identify a *good* life?**

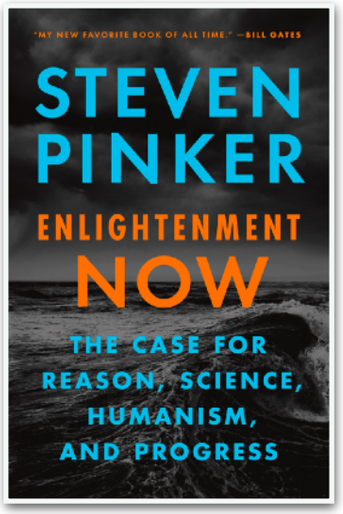


Fragment from: Steven Pinker. "Enlightenment Now: The Case for Reason, Science, Humanism, and Progress".

...the most arresting question I have ever fielded followed a talk in which I explained the **commonplace among scientists that mental life consists of patterns of activity in the tissues of the brain**. A student in the audience raised her hand and asked me: "**Why should I live**?"

"What I recall saying ... went something like this..."

Position Statement



Fragment from: Steven Pinker. "Enlightenment Now: The Case for Reason, Science, Humanism, and Progress".

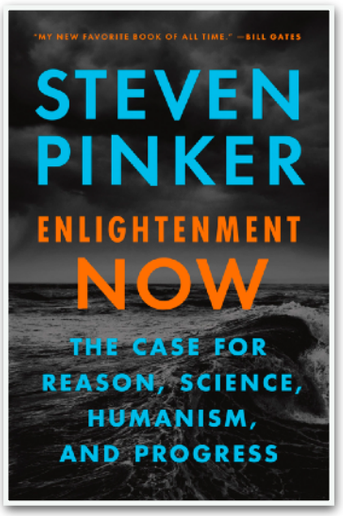
Proposition 1: The basis

"In the very act of asking that question, **you are seeking reasons for your convictions**, and so **you are committed to reason as the means to discover and justify what is important to you**. And there are so many reasons to live!

Proposition 2: You as an individual

As a sentient being, you have the potential to **flourish**. You can refine your faculty of reason itself by **learning** and **debating**. You can seek **explanations** of the natural world through **science**, and insight into the **human condition** through the **arts** and **humanities**. You can make the most of your capacity for **pleasure** and **satisfaction**, which allowed your ancestors to thrive and thereby... "

Position Statement



Fragment from: Steven Pinker. "Enlightenment Now: The Case for Reason, Science, Humanism, and Progress".

"... allowed you to exist. You can **appreciate the beauty** and richness of the natural and cultural world. As the heir to billions of years of life perpetuating itself, you can perpetuate life in turn. You have been endowed with a sense of **sympathy**—the ability to like, love, respect, help, and show kindness—and you can **enjoy** the gift of mutual benevolence with friends, family, and colleagues.

And because **reason tells you that none of this is particular to you**, you have the responsibility to provide to others what you expect for yourself. You can foster the **welfare** of other sentient beings by enhancing life, health, knowledge, freedom, abundance, safety, beauty, and peace. History shows that when we **sympathize with others** and apply our ingenuity to improving the human condition, we can make **progress** in doing so, and you can help to continue that **progress**."

Position Statement

The previous position statement assumes a lot of things about the world that are not self-evident (these are the ideas of the **Enlightenment**), and not everybody agrees on that statement….

But this is a course on applied ethics, and we need a starting point for the discussion.

This will be our provisional starting point.

Why Ethics?

in technology, data science, AI...

Scientific point of view

"Everything that is not forbidden by laws of nature is achievable, given the right knowledge"
(Credit: David Deutsch)

But that's the problem.

"Everything" means everything: vaccines and bioweapons, video on demand and Big Brother on the tele-screen.
Something in addition to science ensured that vaccines were put to use in eradicating diseases while bioweapons were outlawed.

Fragment de: Steven Pinker. "Enlightenment Now: The Case for Reason, Science, Humanism, and Progress". Apple Books.

Dr. Melvin Kranzberg was a professor of the history of technology at the Georgia Institute of Technology

Presidential Address

TECHNOLOGY AND HISTORY: "KRANZBERG'S LAWS"

MELVIN KRANZBERG

A few months ago I received a note from a longtime collaborator in building the Society for the History of Technology, Eugene S. Ferguson, in which he wrote, "Each of us has only one message to convey." Ferguson was being typically modest in referring to an article of his in a French journal[1] emphasizing the hands-on, design component of technical development, and he claimed that he had been making exactly the same point in his many other writings. True, but he has also given us many other messages over the years.

However, Ferguson's statement of "only one message" might indeed be true in my case. For I have been conveying basically the same message for over thirty years, namely, the significance in human affairs of the history of technology and the value of the contextual approach in understanding technical developments.

Because I have repeated that same message so often, utilizing various examples or stressing certain elements to accord with the interests of the different audiences I was attempting to reach, my thoughts have jelled into what have been called "Kranzberg's Laws." These are not laws in the sense of commandments but rather a series of truisms deriving from a longtime immersion in the study of the development of technology and its interactions with sociocultural change.

* * *

DR. KRANZBERG, Callaway Professor of the History of Technology at the Georgia Institute of Technology, was the founding editor of *Technology and Culture*, the recipient of the Society for the History of Technology's Leonardo da Vinci Medal in 1967, and president of SHOT in 1983–84. He presented this presidential address on October 19, 1985, at the Henry Ford Museum in Dearborn, Michigan.

[1]Eugene S. Ferguson, "La Fondation des machines modernes: des dessins," *Culture technique* 14 (June 1985): 182–207. *Culture technique* is the publication of the Centre de Recherche sur la Culture Technique, located in Paris under the direction of Jocelyn de Noblet. The June 1983 edition of *Culture technique*, dedicated to *Technology and Culture*, contained French translations of a number of articles from the SHOT journal.

0040-165X/86/2703-0007$01.00

THE WALL STREET JOURNAL.

SUBSCRIBE SIGN IN

TECH | KEYWORDS

The Six Laws of Technology Everyone Should Know

Professor who summarized the impact of technology on society 30 years ago seems prescient now, in the age of smartphones and social media

A customer tries out the Animoji feature on Apple's iPhone X smartphone at a Chicago store on Nov. 3.

PHOTO: BLOOMBERG NEWS

By Christopher Mims
Nov. 26, 2017 8:00 am ET

SHARE TEXT 203

Three decades ago, a historian wrote six laws to explain society's unease with the power and pervasiveness of technology. Though based on historical examples taken from the Cold War, the laws read as a cheat sheet for explaining our era of Facebook, Google, the iPhone and FOMO.

Kranzberg's Six Laws of Technology

Kranzberg's First Law:

"Technology is neither good nor bad;
nor is it neutral."

By which he means that, "*technology's* ***interaction*** *with the social ecology is such that technical developments frequently have environmental, social, and human* ***consequences that go far beyond the immediate purposes*** *of the technical devices and practices themselves, and the same technology can have quite* ***different results*** *when introduced into* ***different contexts*** *or under different circumstances.*"

What was the main (unexpected) consequence of the agricultural revolution?
What is the main (unexpected) consequence of the industrial revolution?

Technologies are not ethically 'neutral', for they reflect the **values** that we 'bake in' to them with our design choices, as well as the **values** which guide our distribution and use of them.

Technologies **both reveal and shape** what humans **value**, what we think is 'good' in life and worth seeking.

Not only does technology greatly impact our opportunities for living a **good** life, but its **positive and negative impacts are often distributed unevenly** among individuals and groups.

Technologies can create widely disparate impacts, creating '**winners**' **and** '**losers**' in the social lottery or magnifying existing inequalities

How do we ensure that access to the enormous benefits promised by new technologies, and exposure to their risks, are distributed in the right way? **This is a matter of ethics.**

Status Quo

The “Silicon Valley” point of view...

OWNING ETHICS

Corporate Logics, Silicon Valley, and the Institutionalization of Ethics

Ethics is arguably the hottest product in Silicon Valley's hype cycle today, even as headlines decrying a lack of ethics in technology companies accumulate.

The three main **corporate and industry logics** that the authors examine are meritocracy, technological solutionism, and market fundamentalism.

OWNING ETHICS

Corporate Logics, Silicon Valley, and the Institutionalization of Ethics

Meritocracy is an ideological framework that legitimizes unequal distributions of wealth and power as arising from differences in individual abilities.

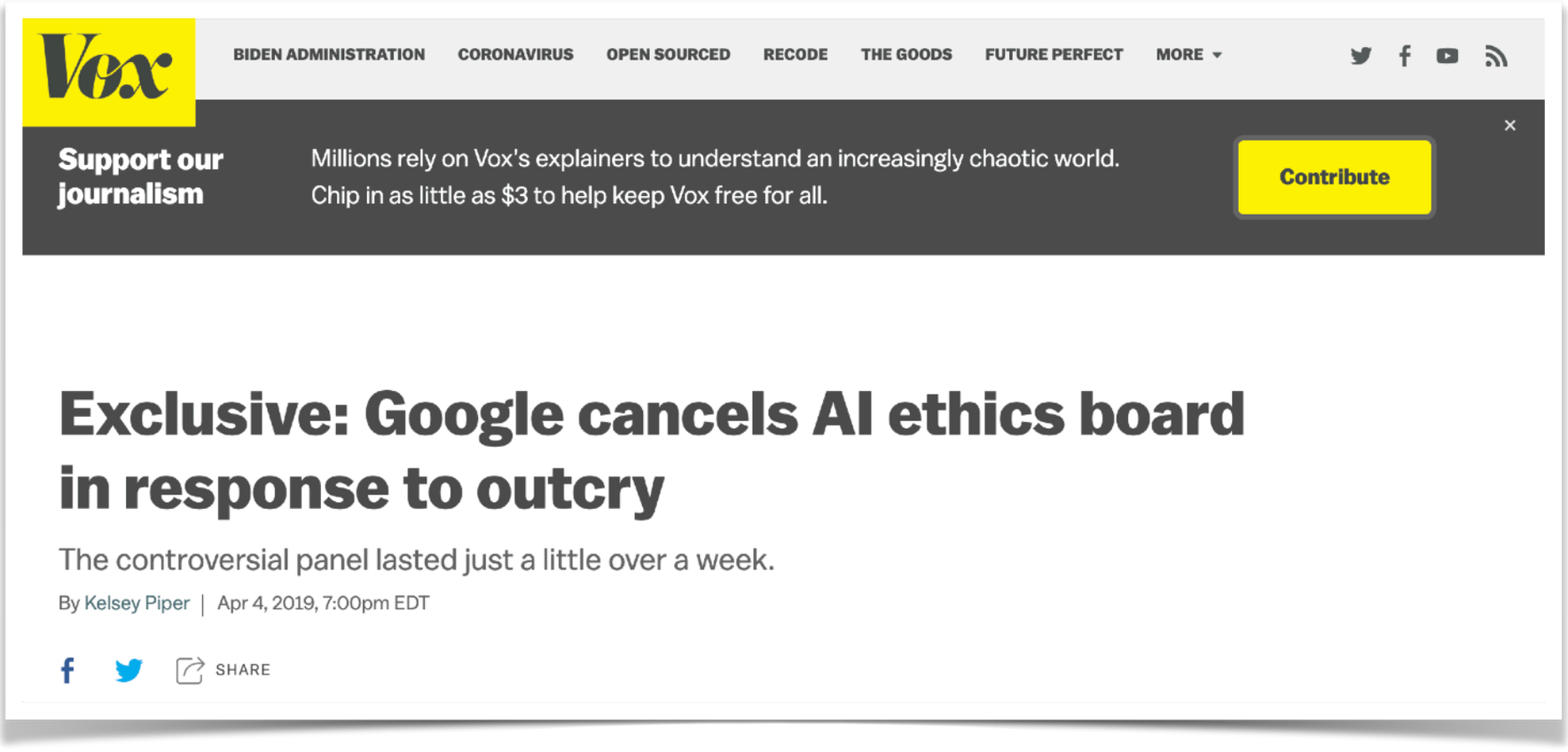
> This has defined the modern subject: as autonomous and responsible for perpetual self-improvement. **The tech industry was founded on the myth that it is a meritocratic segment where talents should be rewarded handsomely.**

This meritocratic belief manifests in the idea that engineers are best at solving ethical issues that their products might create.

Similarly, meritocratic logics place a strong emphasis on individual ethics rather than regulation and legislation.

Companies and teams try to come up with their own codes of ethics to drive off legislation.

Industry self-regulation is the process whereby members of an industry, trade or sector of the economy monitor their own adherence to legal, ethical, or safety standards, rather than have an outside, independent agency such as a third party entity or governmental regulator monitor and enforce those standards.



LOGIN REGISTRIEREN

DER TAGESSPIEGEL

SERUM COGNOSCERE CAUSAS

POLITIK BERLIN WIRTSCHAFT GESELLSCHAFT KULTUR MEINUNG SPORT WISSEN VERBRAUCHER INTERAKTIV

Suchbegriff eingeben

Agenda · Brexit · Digitalisierung & KI · Energie & Klima · Gesundheit & E-Health · Mobilität & Transport

Politik › EU guidelines: Ethics washing made in Europe

Coronavirus in Deutschland – Alle Zahlen im Überblick Hier ansehen

EU guidelines

08.04.2019, 15:48 Uhr

Ethics washing made in Europe

On Tuesday, the EU has published ethics guidelines for artificial intelligence. A member of the expert group that drew up the paper says: This is a case of ethical white-washing. VON THOMAS METZINGER

Pekka Ala-Pietilä, Chair of the AI HLEG
AI Finland, Huhtamaki, Sanoma
Wilhelm Bauer
Fraunhofer
Urs Bergmann – Co-Rapporteur
Zalando
Mária Bieliková
Slovak University of Technology in Bratislava
Cecilia Bonefeld-Dahl – Co-Rapporteur
DigitalEurope
Yann Bonnet
ANSSI
Loubna Bouarfa
OKRA
Stéphan Brunessaux
Airbus
Raja Chatila
IEEE Initiative Ethics of Intelligent/Autonomous Systems &
Sorbonne University
Mark Coeckelbergh
University of Vienna
Virginia Dignum – Co-Rapporteur
Umea University
Luciano Floridi
University of Oxford
Jean-Francois Gagné – Co-Rapporteur
Element AI
Chiara Giovannini
ANEC
Joanna Goodey
Fundamental Rights Agency
Sami Haddadin
Munich School of Robotics and MI
Gry Hasselbalch
The thinkdotank DataEthics & Copenhagen University
Fredrik Heintz
Linköping University
Fanny Hidvegi
Access Now
Eric Hilgendorf
University of Würzburg
Klaus Höckner
Hilfsgemeinschaft der Blinden und Sehschwachen
Mari-Noëlle Jégo-Laveissière
Orange
Leo Kärkkäinen
Nokia Bell Labs
Sabine Theresia Köszegi
TU Wien
Robert Kroplewski
Solicitor & Advisor to Polish Government
Elisabeth Ling
RELX
Pierre Lucas
Orgalim – Europe's technology industries
Ieva Martinkenaite
Telenor
Thomas Metzinger – Co-Rapporteur
JGU Mainz & European University Association
Catelijne Muller
ALLAI Netherlands & EESC
Markus Noga
SAP
Barry O'Sullivan, Vice-Chair of the AI HLEG
University College Cork
Ursula Pachl
BEUC
Nicolas Petit – Co-Rapporteur
University of Liège
Christoph Peylo
Bosch
Iris Plöger
BDI
Stefano Quintarelli
Garden Ventures
Andrea Renda
College of Europe Faculty & CEPS
Francesca Rossi
IBM
Cristina San José
European Banking Federation
George Sharkov
Digital SME Alliance
Philipp Slusallek
German Research Centre for AI (DFKI)
Françoise Soulié Fogelman
AI Consultant
Saskia Steinacker – Co-Rapporteur
Bayer
Jaan Tallinn
Ambient Sound Investment
Thierry Tingaud
STMicroelectronics
Jakob Uszkoreit
Google
Aimee Van Wynsberghe – Co-Rapporteur
TU Delft
Thiébaut Weber
ETUC
Cecile Wendling
AXA
Karen Yeung – Co-Rapporteur
The University of Birmingham

ETHICS AS AN ESCAPE FROM REGULATION FROM "ETHICS-WASHING" TO ETHICS-SHOPPING?

BEN WAGNER

A strange confusion among technology policy makers can be witnessed at present. While almost all are able to agree on the common chorus of voices chanting 'something must be done,' it is very difficult to identify what exactly must be done and how. In this confused environment it is perhaps unsurprising that the idea of 'ethics' is presented as a concrete policy option. Striving for ethics and ethical decision-making, it is argued, will make technologies better. While this may be true in many cases, much of the debate about ethics seems to provide an easy alternative to government regulation. Unable or unwilling to properly provide regulatory solutions, ethics is seen as the 'easy' or 'soft' option which can help structure and give meaning to existing self-regulatory initiatives. In this world, 'ethics' is the new 'industry self-regulation.'

Rigorous ethical approaches?

This approach does not do justice to many of the proponents of ethical approaches to technology who think long and hard about ethical frameworks for technology development. It is however indicative of the increasingly common role of technology ethics in political debates. For example, as part of a conference panel on ethics, one member of the Google DeepMind ethics team emphasised repeatedly how ethically Google DeepMind was acting, while simultaneously avoiding any responsibility for the data protection scandal at Google DeepMind (Powles and Hodson 2018). In her understanding, Google DeepMind were an ethical company developing ethical products and the fact that the health data of 1.6 Million people was shared without a legal basis was instead the fault of the British government. This suggests a tension between legal and ethical action, in which the appropriate mode of governance is not yet sufficiently defined.

Ethics / rights / regulation

Such narratives are not just uncommon in the corporate but also in technology policy, where ethics, human rights and regulation are frequently played off against each other. In this context, ethical frameworks that provide a way to go beyond existing legal frameworks can also provide an opportunity to ignore them. More broadly the rise of the ethical technology debate runs in parallel to the increasing resistance to any regulation at all. At an international level the Internet Governance Forum (IGF) provides a space for discussions about governance without any mechanism to implement them and successive attempts to change this have failed. It is thus perhaps unsurprising that many of the initiatives proposed on regulating technologies tend to side-line the role of the state and instead emphasize the role of the private sector. Whether through the multi-stakeholder model proposed by Microsoft for an international attribution agency in which states play a comparatively minor role (Charney et al. 2016), or in a proposal by RAND corporation which suggests that states should be completely excluded from such an attribution organisation (Davis II et al. 2017). In fact, states and their regulatory instruments are increasingly portrayed as a problem rather than a solution.

Case in point: Artificial Intelligence

This tension between ethics, regulation and governance is evident in the debate on

OWNING ETHICS

Corporate Logics, Silicon Valley, and the Institutionalization of Ethics

Technological solutionism is the belief that **technology can solve social problems**.

> Critics have pointed out that many so-called "solutions" can actually cause problems such as rising income and housing inequalities. The tech industry often responds by proposing even more technical solutions.

Similarly, ethical problems are also framed as could be solved by technological solutions. This logic leads to creation of **checklists, procedures or evaluative metrics** to ensure the design and implementation of ethical products.

There are **hundreds** of documents about ethical guidelines!

nature machine intelligence

PERSPECTIVE

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

The global landscape of AI ethics guidelines

Anna Jobin, Marcello Ienca and Effy Vayena*

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

Artificial intelligence (AI), or the theory and development of computer systems able to perform tasks normally requiring human intelligence, is widely heralded as an ongoing "revolution" transforming science and society altogether[1,2]. While approaches to AI such as machine learning, deep learning and artificial neural networks are reshaping data processing and analysis[3], autonomous and semi-autonomous systems are being increasingly used in a variety of sectors including healthcare, transportation and the production chain[4]. In light of its powerful transformative force and profound impact across various societal domains, AI has sparked ample debate about the principles and values that should guide its development and use[5,6]. Fears that AI might jeopardize jobs for human workers[7], be misused by malevolent actors[8], elude accountability or inadvertently disseminate bias and thereby undermine fairness[9] have been at the forefront of the recent scientific literature and media coverage. Several studies have discussed the topic of ethical AI[10–13], notably in meta-assessments[14–16] or in relation to systemic risks[17,18] and unintended negative consequences such as algorithmic bias or discrimination[19–21].

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

Reports and guidance documents for ethical AI are instances of what is termed non-legislative policy instruments or soft law[23]. Unlike so-called hard law—that is, legally binding regulations passed by the legislatures to define permitted or prohibited conduct—ethics guidelines are not legally binding but persuasive in nature. Such documents are aimed at assisting with—and have been observed to have significant practical influence on—decision-making in certain fields, comparable to that of legislative norms[24]. Indeed, the intense efforts of such a diverse set of stakeholders in issuing AI principles and policies is noteworthy, because they demonstrate not only the need for ethical guidance, but also the strong interest of these stakeholders to shape the ethics of AI in ways that meet their respective priorities[16,25]. Specifically, the private sector's involvement in the AI ethics arena has been called into question for potentially using such high-level soft policy as a portmanteau to either render a social problem technical[16] or to eschew regulation altogether[26]. Beyond the composition of the groups that have produced ethical guidance on AI, the content of this guidance itself is of interest. Are these various groups converging on what ethical AI should be, and the ethical principles that will determine the development of AI? If they diverge, what are their differences and can these differences be reconciled?

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

Methods

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

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

OWNING ETHICS

Corporate Logics, Silicon Valley, and the Institutionalization of Ethics

Market Fundamentalism, or market logics, refers to the idea that **companies are there to make money, and if ethics initiatives are cut into the bottom line, companies should not do it**.

> Besides, there is a belief that ethical initiatives are often costly, and antithetical to corporate profits.

Furthermore across the industry, if other companies do not implement similar ethical considerations on their products, one should not do it.

In the context of the absence of a legal framework, implementing ethics initiatives might be a business problem rather than a solution. In other words, the works of ethics owners in practice are constrained by what the market can allow.

Out there...

Google's recent firing of Timnit Gebru, a prominent AI ethics researcher, called into question two main issues in the tech industry: its lack of diversity, and its faulty relationship with ethical considerations of technological development.

The New York Times

Google Researcher Says She Was Fired Over Paper Highlighting Bias in A.I.

Timnit Gebru, one of the few Black women in her field, had voiced exasperation over the company's response to efforts to increase minority hiring.

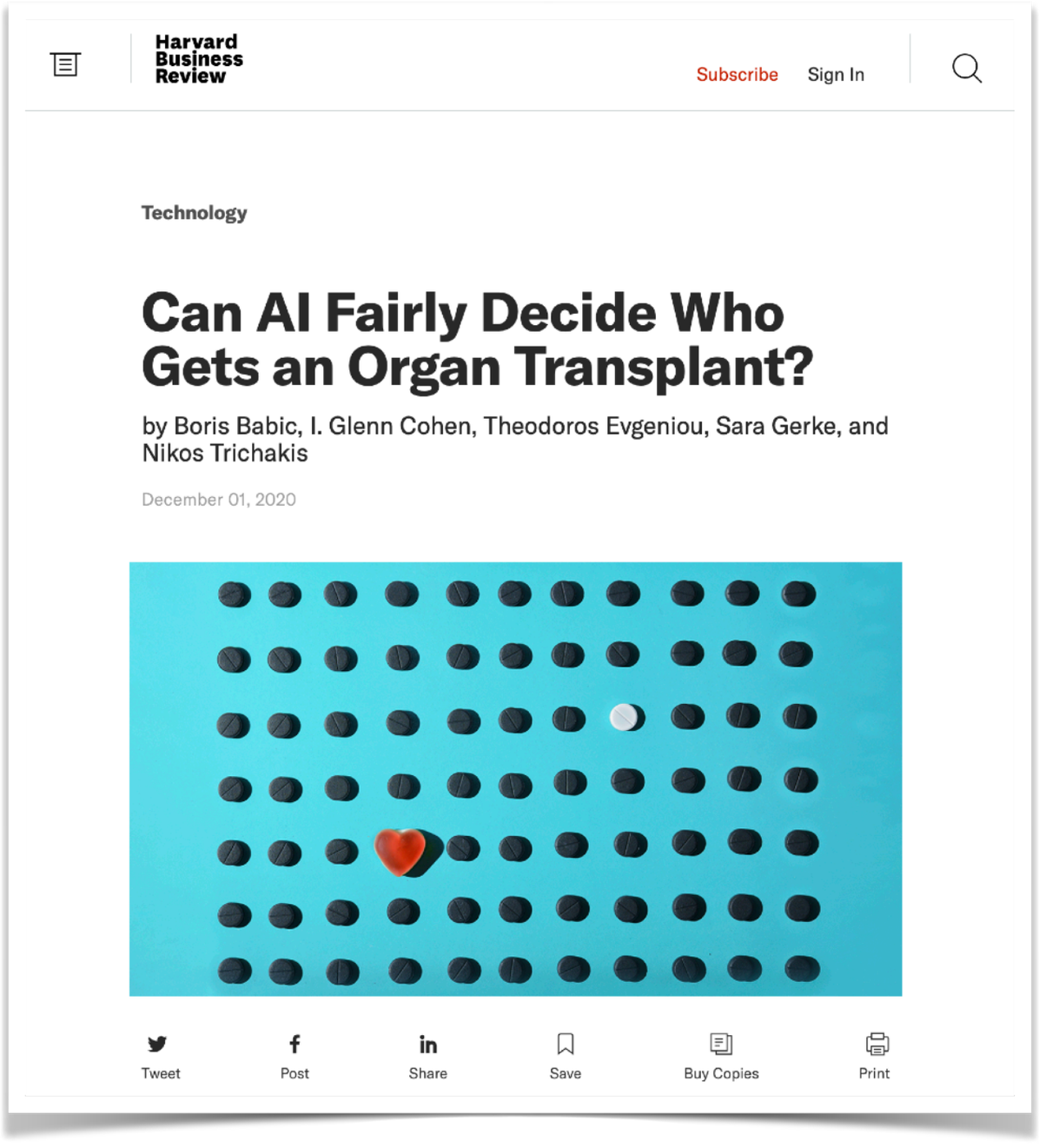
279

Timnit Gebru, a respected researcher at Google, questioned biases built into artificial intelligence systems. Cody O'Loughlin for The New York Times

Dr. Gebru represents the growing "*ethics owners class of tech workers*" who champion ethical causes, ethical designs, development, and deployment of technology from within the tech industry.

Source: https://montrealethics.ai/owning-ethics-corporate-logics-silicon-valley-and-the-institutionalization-of-ethics-research-summary/

Health care organizations, like many other enterprises, face steep challenges in their attempt to maximize operational efficiency in the face of resource constraints. Whether it is a hospital's attempt to optimize staffing or a government trying to fairly allocate and distribute limited doses of Covid-19 vaccines, these tasks can be formidable. A promising way to manage the complexity is to enlist data-driven analytics and artificial intelligence (AI).



However, such techniques, while powerful, can also mask problematic underlying ethical assumptions or lead to morally questionable outcomes. Consider a recently published study about models used by some of the most technologically advanced hospitals in the world to help prioritize which patients with chronic kidney disease should receive kidney transplants. It found that the models discriminated against black patients: "One-third of Black patients … would have been placed into a more severe category of kidney disease if their kidney function had been estimated using the same formula as for white patients." While it is just the latest of many studies to show the deficiencies of such models, it is unlikely to be the last.

Data entered into the UK Transplant Registry over several decades, from thousands of individuals, have been used to create bespoke algorithms that help identify patients who may be most suitable to receive an available organ.

TRANSPLANTS AND STATISTICS

Primum non nocere (First, do no harm)

Maria Ibrahim, a kidney doctor in training, explains the vital role of statistics and statistical analysis in transplant medicine: from matching donor organs to patients, to helping doctors and patients discuss the risks and benefits of a life-changing operation

It is 3 a.m. You are the transplant doctor on call at a busy London hospital when the telephone rings. A voice at the other end of the line tells you that there has been a road traffic accident and a person lies in intensive care. This person, sadly, will not survive. However, the accident victim had previously expressed a wish to become an organ donor and, with their family's consent, a specialist nurse has informed NHS Blood and Transplant that organs from this patient will soon be available for those in urgent need. This is why you have been called: you are offered a kidney from this patient for someone on the waiting list at your transplant centre.

The potential recipient has been waiting for a transplant offer for over a year. Of the thousands of patients on this waiting list, they have been chosen to receive this organ offer. At this time in the morning, your patient is bound to be asleep. But soon they may receive a life-changing phone call from you. First, though, the decision whether to accept or decline the organ offer needs to be made. You – as the transplant doctor – are faced with the complex task of weighing the risks and benefits to your patient of transplantation with this particular organ versus remaining on the list.

Underpinning all these processes and decisions – unseen by most – lies a body of statistics.

Transplantation is one of the most challenging and complex areas of modern medicine. Since the first successful solid organ transplant conducted in 1950, the field has advanced at an astounding rate. Pioneering surgical techniques, as well as development of

YashikovskyMD/Bigstock.com

About this article

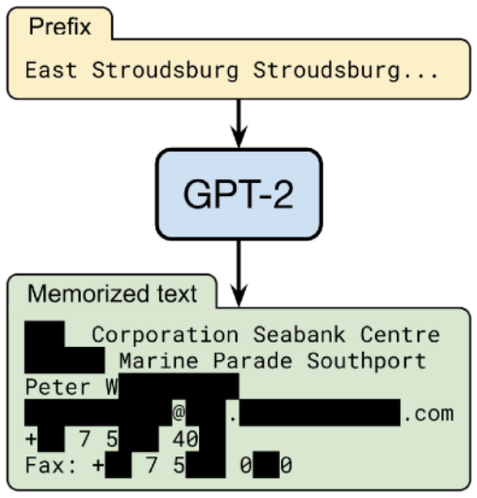
Maria Ibrahim's article is the winner of our 2020 Statistical Excellence Award for Early-Career Writing, awarded in partnership with the Young Statisticians Section (YSS) of the Royal Statistical Society. Congratulations to Maria on winning the award, and thank you to all those early-career statisticians and data scientists who took part in the competition between the months of February and May this year, a time that was particularly challenging for all due to the Covid-19 pandemic. Thanks also to our judges: for *Significance*, Mario Cortina Borja, Carlos Grajales and Kelly Zou; and for the YSS, Katie Fisher, Joy Leahy, Altea Lorenzo-Arribas, Marnie Low and Ryan Jessop. Details of our 2021 writing competition and award will be announced in February 2021.

Does GPT-2 Know Your Phone Number?

Eric Wallace, *Florian Tramèr*, *Matthew Jagielski*, and *Ariel Herbert-Voss*
Dec 20, 2020

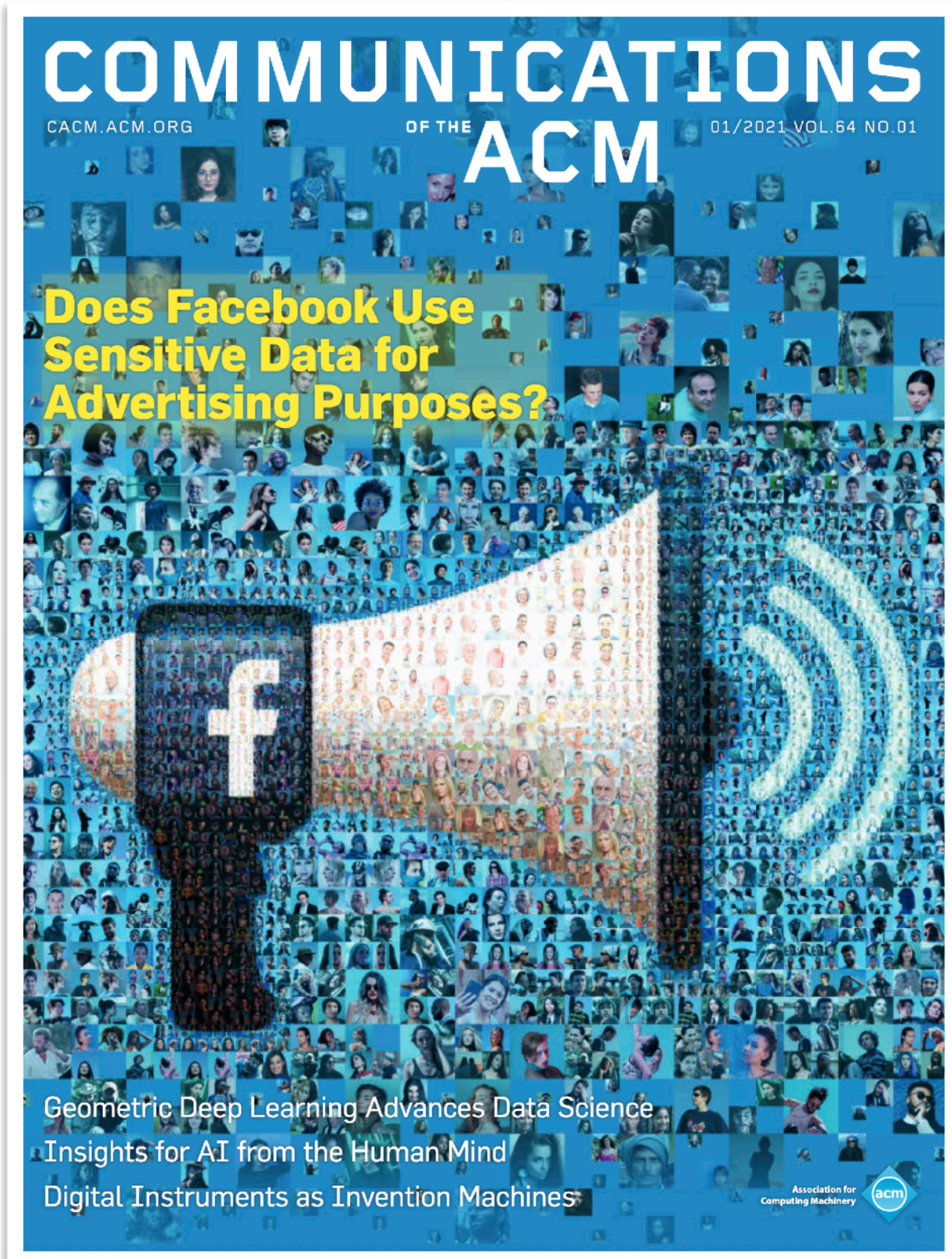
Most likely not.

Yet, OpenAI's GPT-2 language model *does* know how to reach a certain Peter W[redacted] (name redacted for privacy). When prompted with a short snippet of Internet text, the model accurately generates Peter's contact information, including his work address, email, phone, and fax:



The model re-generated lists of news headlines, Donald Trump speeches, pieces of software logs, entire software licenses, snippets of source code, passages from the Bible and Quran, the first 800 digits of pi, and much more!

In a recent work,2 we demonstrated that Facebook (FB) labels 73% of users within the EU with potentially sensitive interests (referred to as ad preferences as well), which may contravene the GDPR. FB assigns user's different ad preferences based on their online activity within this social network. Advertisers running ad campaigns can target groups of users that have been assigned a particular ad preference (for example, target FB users interested in Starbucks). Some of these ad preferences may suggest political opinions (for example, Socialist party), sexual orientation (for example, homosexuality), personal health issues (for example, breast cancer awareness), and other potentially sensitive attributes.



In the vast majority of the cases, the referred sensitive ad preferences are inferred from the user behavior in FB without obtaining explicit consent from the user.

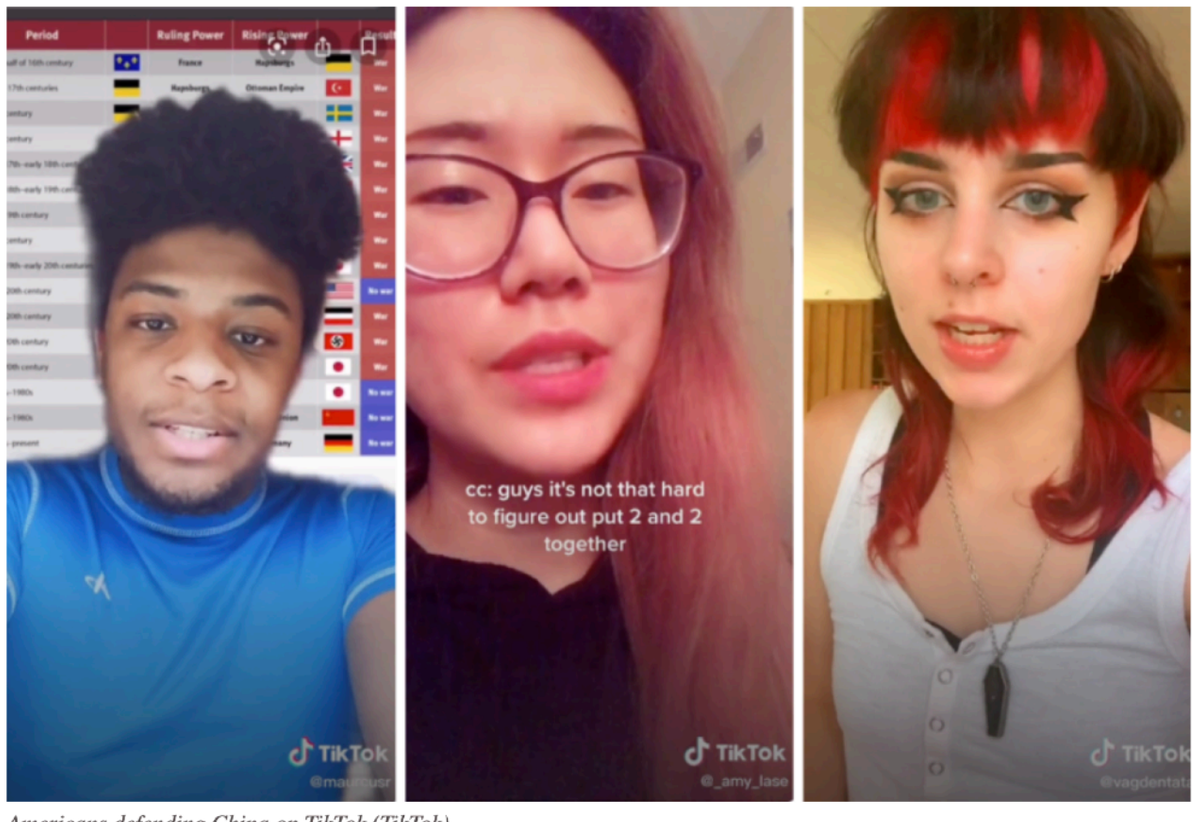
THE SPECTATOR

CHINA TECH US POLITICS WORLD

The useful idiots of TikTok

American teens are denying the existence of Uighur labor camps from their bedrooms

Harry Shukman



Americans defending China on TikTok (TikTok)

Not everyone agrees on that!
This is an interesting reading written from a “negationist” point of view.

What do you do if decisions that used to be made by humans, with all their biases, start being made by algorithms that are mathematically incapable of bias? If you’re rational, you should celebrate. If you’re a militant liberal, you recognize this development for the mortal threat it is, and scramble to take back control.

THE SPECTATOR

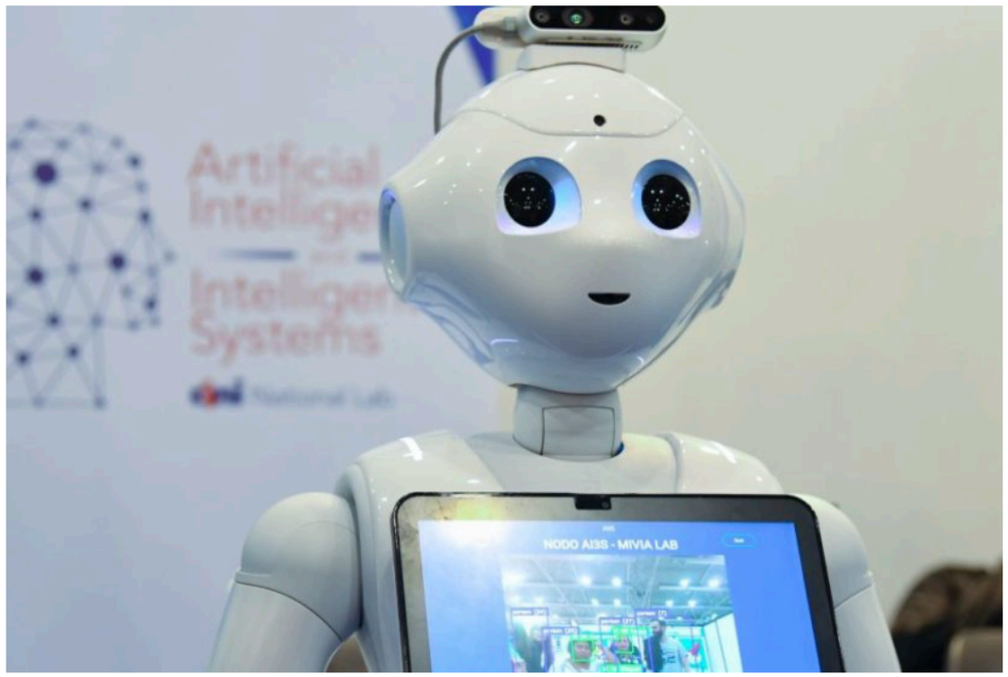
LOG IN | REGISTER

LIFE TECH US POLITICS

We must stop militant liberals from politicizing artificial intelligence

‘Debiasing’ algorithms actually means adding bias

Pedro Domingos



A robot from the Artificial Intelligence and Intelligent Systems (AIIS) laboratory (Getty)

Algorithms help select job candidates, voters to target in political campaigns, and even people to date. Businesses and legislators alike need to ensure that they are not tampered with. And all of us need to be aware of what is happening, so we can have a say. I, for one, after seeing how progressives will blithely assign prejudices even to algorithms that transparently can’t have any, have started to question the orthodox view of human prejudices. Are we really as profoundly and irredeemably racist and sexist as they claim? I think not.

Recent NLP Incidents

Here Are the Microsoft Twitter Bot's Craziest Racist Rants

Researchers find major demographic differences in speech recognition accuracy

Google says sorry for racist auto-tag in photo app

- Google Photos labelled a picture of two black people as 'gorillas'
- Google Maps and Flickr have also suffered from race-related problems

Scammer Successfully Deepfaked CEO's Voice To Fool Underling Into Transferring $243,000

Microsoft's robot editor confuses mixed-race Little Mix singers

Firm's plan to replace editors with AI backfires after wrong image of musician is published

Oh dear... AI models used to flag hate speech online are, er, racist against black people

Tweets written in African-American English slang more likely to be considered offensive

Report: AI Company Leaks Over 2.5M Medical Records

The leaked data relates to car accidents and includes names, insurance records, medical diagnosis notes, and payment records.

Facebook algorithm recommending Holocaust denial and fascist content, report finds

Google Ad Portal Equated "Black Girls" with Porn

How to steal modern NLP systems with gibberish?

Google faulted for racial bias in image search results for black teenagers

Directa

TECNOLOGIA

L'Ajuntament de Barcelona amplia a dotze ubicacions el mapa de la videovigilància a Ciutat Vella

El consistori torna a licitar el sistema de càmeres amb intel·ligència artificial que va frenar a inicis d'estiu. El nou concurs manté el requisit tècnic del reconeixement facial, passa de cinc a dotze ubicacions i retira tota referència a l'empresa Avigilon-Motorola, investigada pels seus abusos a Palestina


celestekidd
@celestekidd
En resposta a @celestekidd
The thing I'd like to emphasize today is that algorithmic bias is not only problematic for the direct harms it causes, but also for the cascading harms of how it impacts human beliefs.
Tradueix el tuit
3:41 a. m. · 24 de des. de 2020 · Twitter Web App
celestekidd @celestekidd · 6 h
En resposta a @celestekidd
Algorithmic bias is problematic because there are systems interfacing with people everyday, embedded seamlessly into our lives. These systems drive human beliefs in sometimes destructive, likely irreparable ways.
1
16
56
celestekidd @celestekidd · 6 h
What our research has shown is that:
1. People don't learn very deeply about most things in the world.
2. People have to make up their minds quickly in order to act.
3. Once a person makes up their mind, cognitive mechanisms dissuade them from revisiting those topics.
1
36
129

Ethical problems are real!

DS/AI ethics concerns can be divided in three different time frames/areas:

- Short-time/organization: What is the impact of [**privacy**, **transparency**, **fairness**] in my application?

GDPR…

- Medium-time/society: How the use [**military use**, **medical care**, **justice**, **education**] of these applications will change the way we are organized as a society?

Autonomous weapons, pre-pol, AI justice,…

- Long-time/humans: What are the ethical goals of these technologies?

Singularity, GAI, convergence…

Data and Ethics

What does ethics have to do with data?

The combination of data analytics, a data-saturated and poorly regulated commercial environment, and the absence of widespread, well-designed standards for data practice in industry, university, non-profit, and government sectors has created a '**perfect storm**' of ethical risks.

Thus **no single set of ethical rules or guidelines will fit all data circumstances**; ethical insights in data practice must be adapted to the needs of many kinds of data practitioners operating in different contexts.

What does ethics have to do with data?

We can define a harm or a benefit as 'ethically significant' when it has a substantial possibility of making a difference to certain individuals' chances of having a good life, or the chances of a group to live well: that is, to flourish in society together.

Some harms and benefits are not ethically significant. Say I prefer Coke to Pepsi. If I ask for a Coke and you hand me a Pepsi, even if I am disappointed, you haven't impacted my life in any ethically significant way.

In the context of data practice, the potential harms and benefits are real and ethically significant. But **due to the more complex, abstract, and often widely distributed nature of data practices, as well as the interplay of technical, social, and individual forces in data contexts, the harms and benefits of data can be harder to see and anticipate**.

In this respect, then, data has a broader ethical sweep than engineering of bridges and airplanes. Data practitioners must confront a far more complex ethical landscape than many other kinds of technical professionals...

Ethical Benefits of Data Practices

HUMAN UNDERSTANDING:

Because data and its associated practices can uncover previously unrecognized correlations and patterns in the world, **data can greatly enrich our understanding of ethically significant relationships—in nature, society, and our personal lives**. In this respect, then, data has a broader ethical sweep than engineering of bridges and airplanes. Data practitioners must confront a far more complex ethical landscape than many other kinds of technical professionals...

Ethical Benefits of Data Practices

SOCIAL, INSTITUTIONAL, AND ECONOMIC EFFICIENCY:

Once we have a more accurate picture of how the world works, **we can design or intervene in its systems to improve their functioning**. This reduces wasted effort and resources and improves the alignment between a social system or institution's policies/processes and our goals.

Ethical Benefits of Data Practices

PREDICTIVE ACCURACY AND PERSONALIZATION:

Not only can good data practices help to make social systems work more efficiently, as we saw above, but they can also used to more precisely **tailor actions to be effective in achieving good outcomes for specific individuals, groups, and circumstances**, and to be more responsive to user input in (approximately) real time.

Ethical Harms of Data Practices

HARMS TO PRIVACY & SECURITY:

Thanks to the ocean of personal data that humans are generating today (or, to use a better metaphor, the many different **lakes, springs, and rivers of personal data** that are pooling and flowing across the digital landscape), most of us do not realize how exposed our lives are, or can be, by common data practices.

Ethical Harms of Data Practices

HARMS TO FAIRNESS AND JUSTICE:

We all have a significant life interest in **being judged and treated fairly**, whether it involves how we are treated by law enforcement and the criminal and civil court systems, how we are evaluated by our employers and teachers, the quality of health care and other services we receive, or how financial institutions and insurers treat us.

Ethical Harms of Data Practices

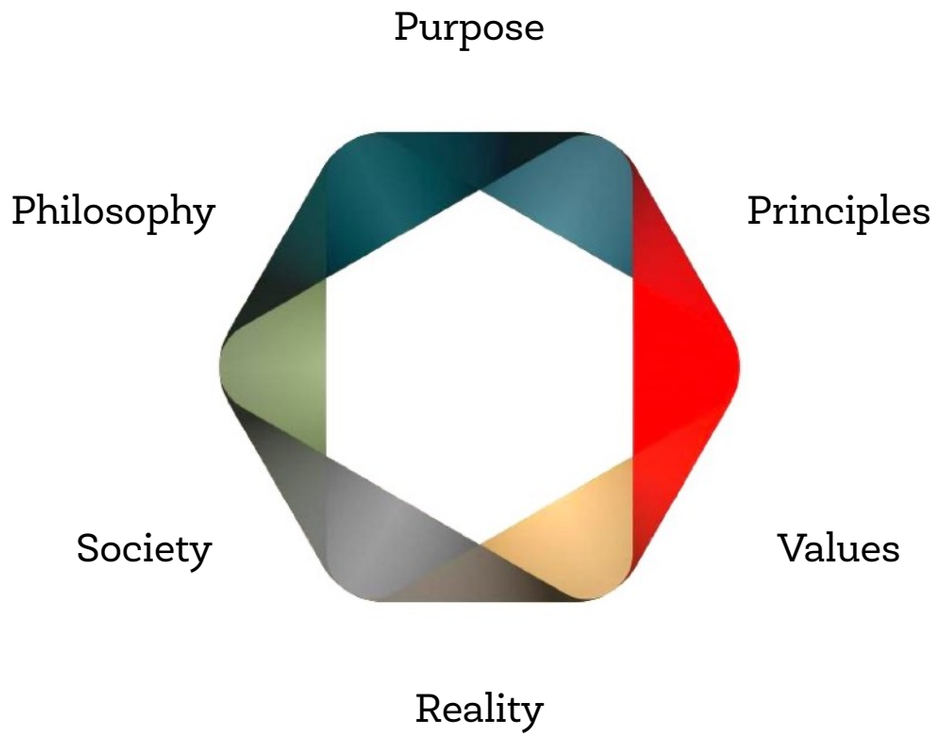
HARMS TO TRANSPARENCY AND AUTONOMY:

In this context, *transparency* is the **ability to see how a given social system or institution works**, and to be able to inquire about the basis of life-affecting decisions made within that system or institution. So, for example, if your bank denies your application for a home loan, transparency will be served by you having access to information about exactly *why* you were denied the loan, and by whom.

What is Ethics?

Definitions

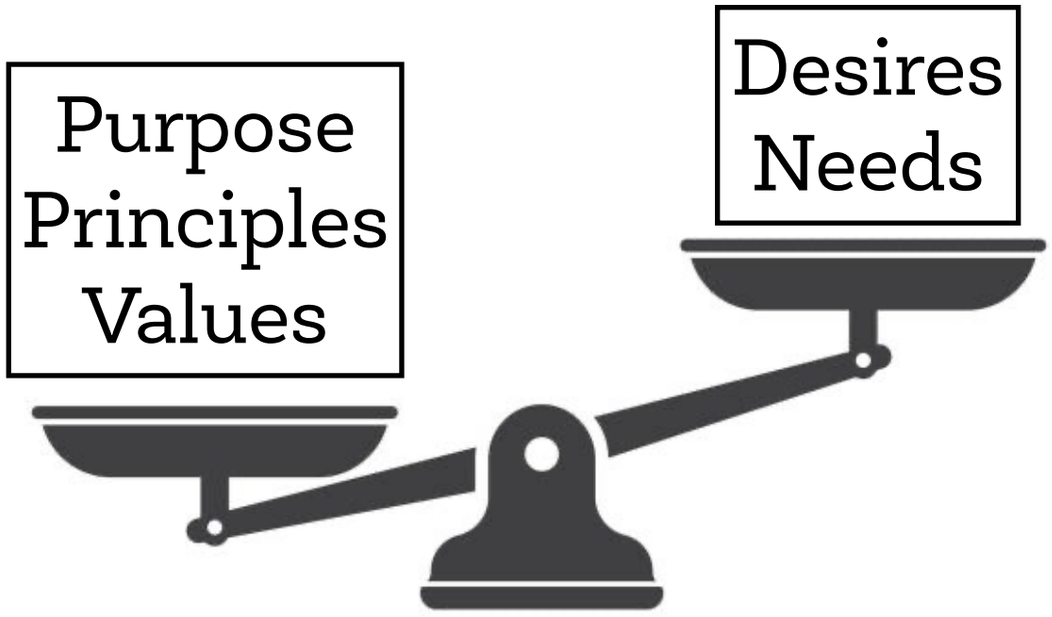
Ethics is the **process** of questioning, discovering and defending your **values**, **principles** and **purposes** in order to be able of deciding what is right and what is wrong.



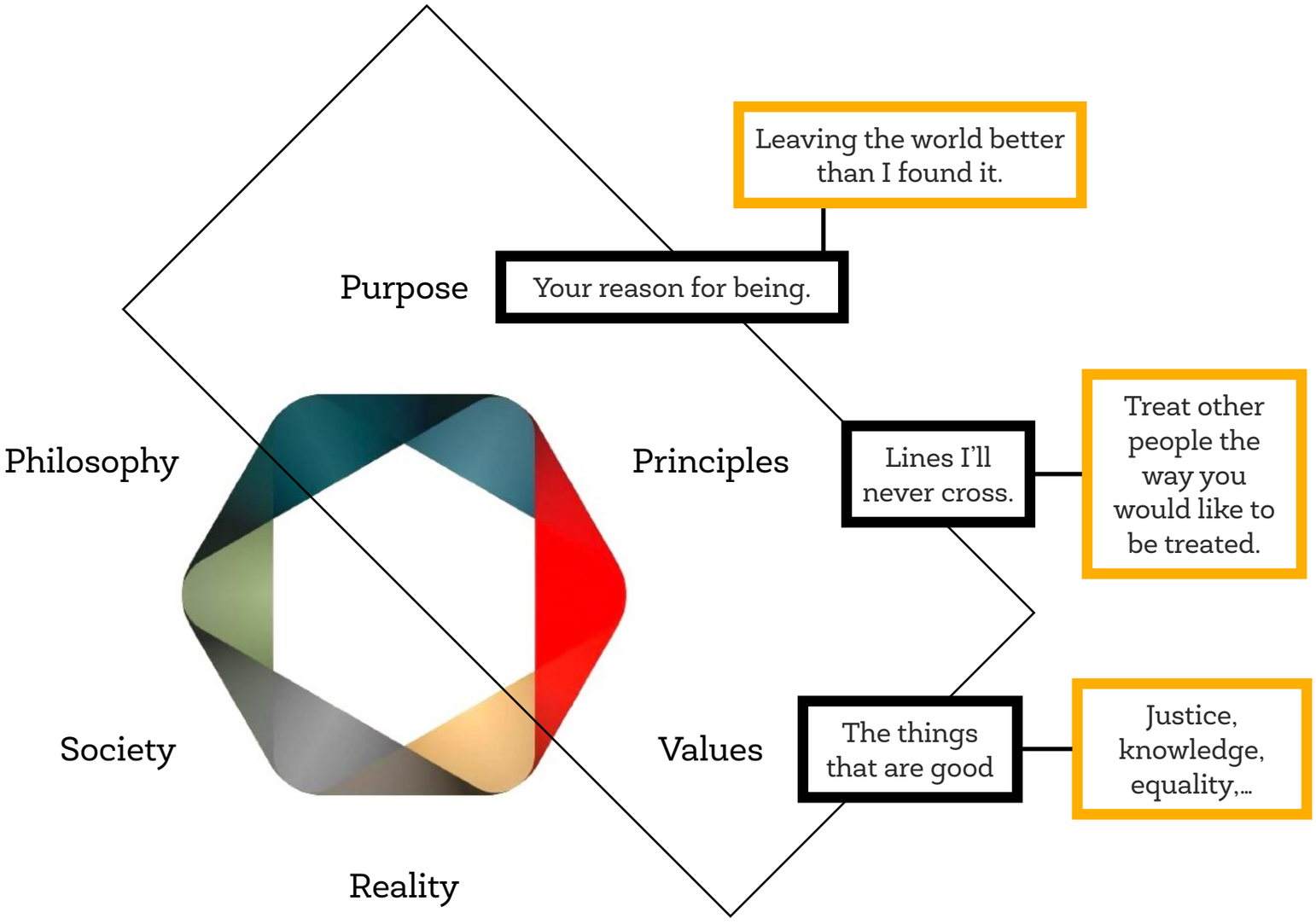
Ethics seeks to answer questions like "*what is good or bad*", "*what is right or what is wrong*", or "*what is justice, well-being or equality*".

Applied ethics concerns what a moral agent is obligated or permitted to do in a specific situation or a **particular domain of action**.

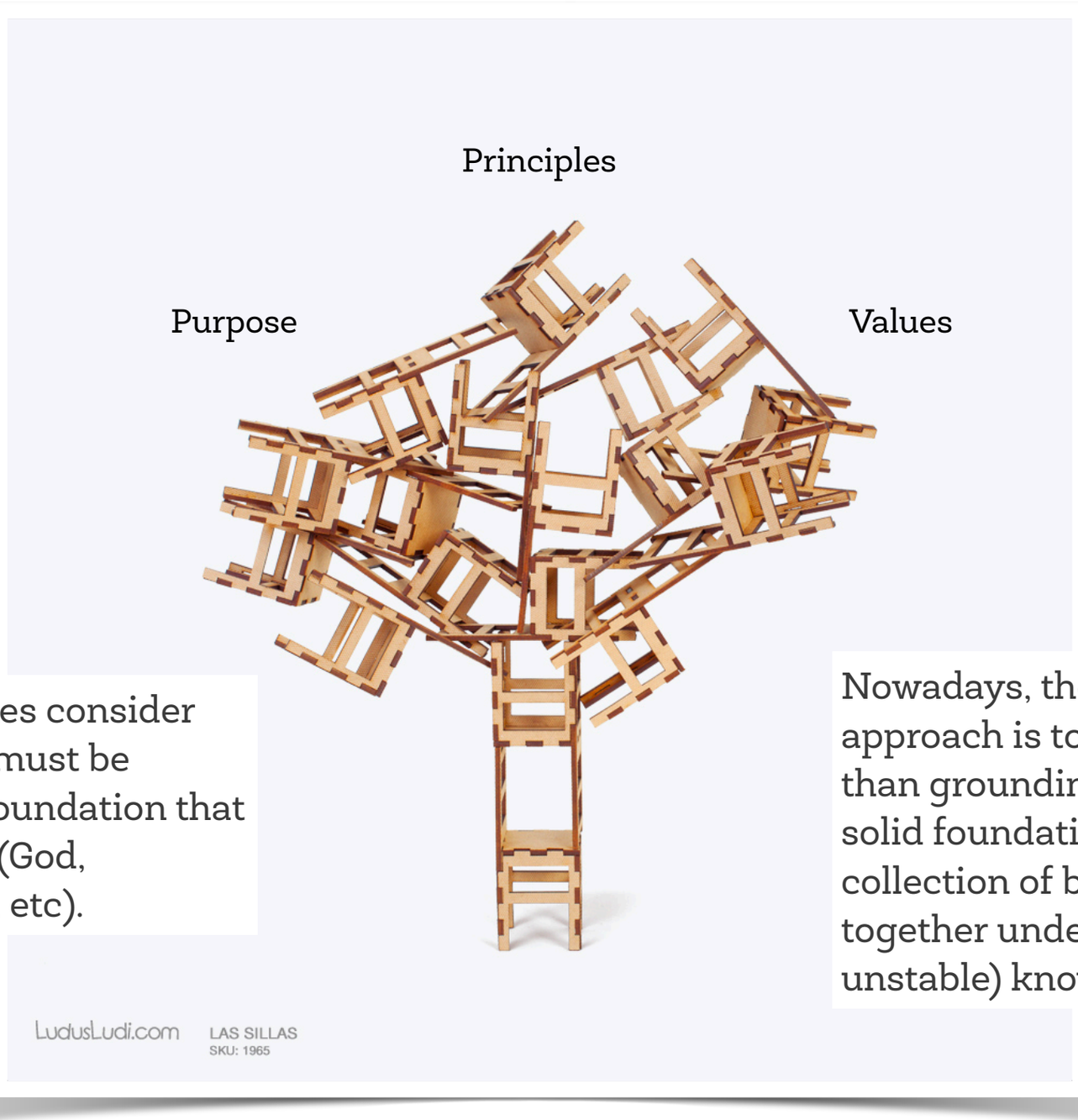
How do we make decisions?



Beliefs, the necessary ingredients of a good individual decision.



Beliefs, the necessary ingredients of a good individual decision.



Traditional approaches consider that person's beliefs must be grounded in a solid foundation that cannot be discussed (God, rationality and harm, etc).

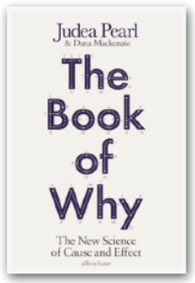
Nowadays, the most common approach is to consider that, rather than grounding our beliefs on a solid foundation, we assemble a collection of beliefs that hold together under a mutual (maybe unstable) knowledge attraction.

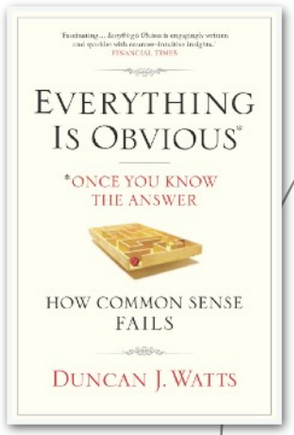
Future of Society

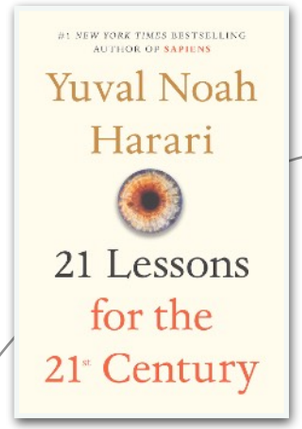
Law

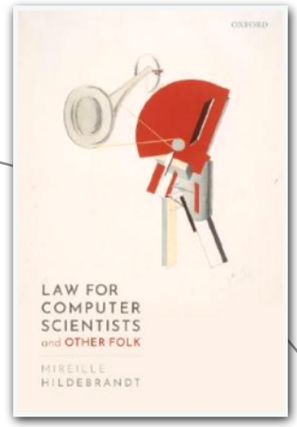
The limits of Common Sense

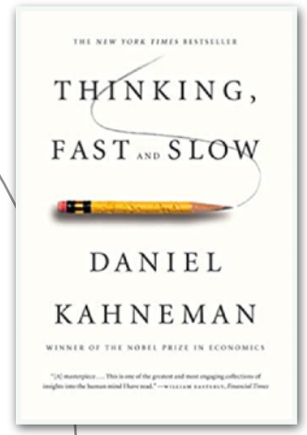
Humans and Rationality

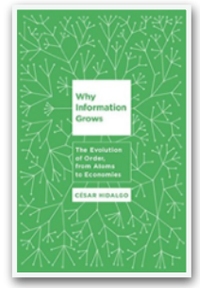


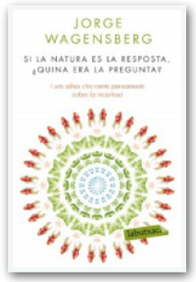






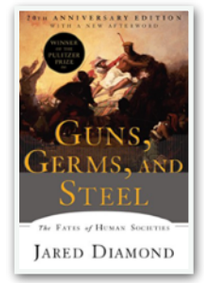


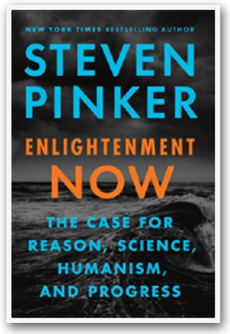


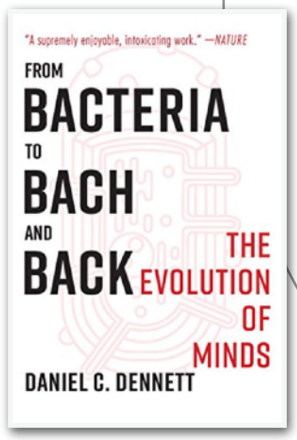


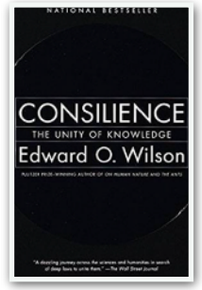
Philosophy

Principles





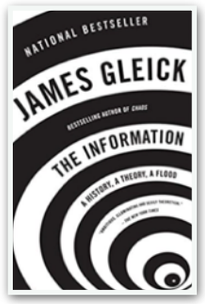


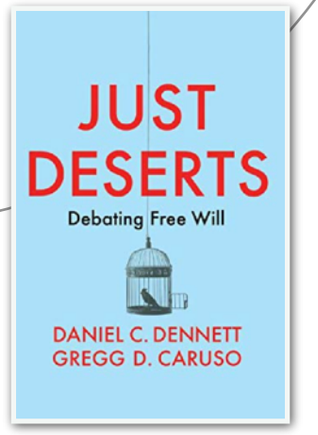


Society

Values

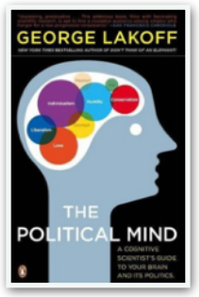
Beliefs about Reality

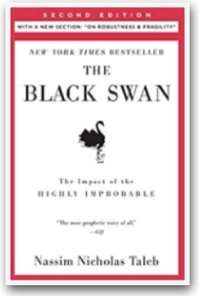




Intelligence & Philosophy

Good and Bad

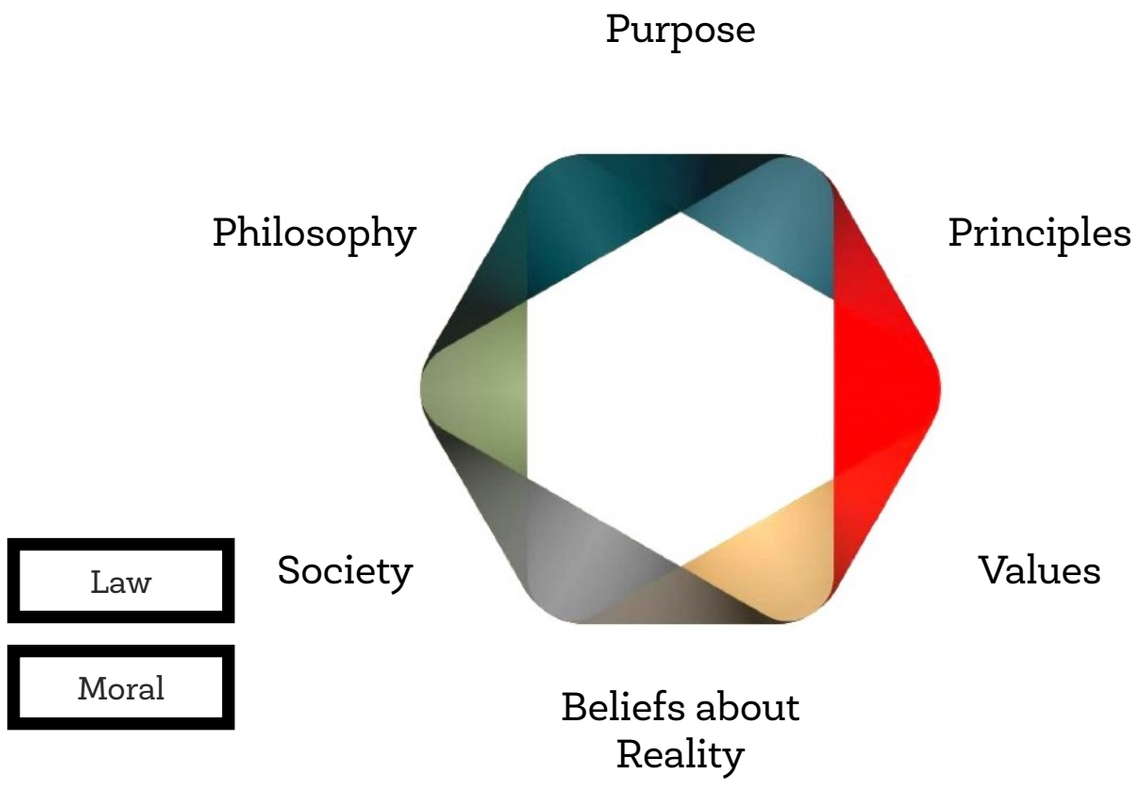




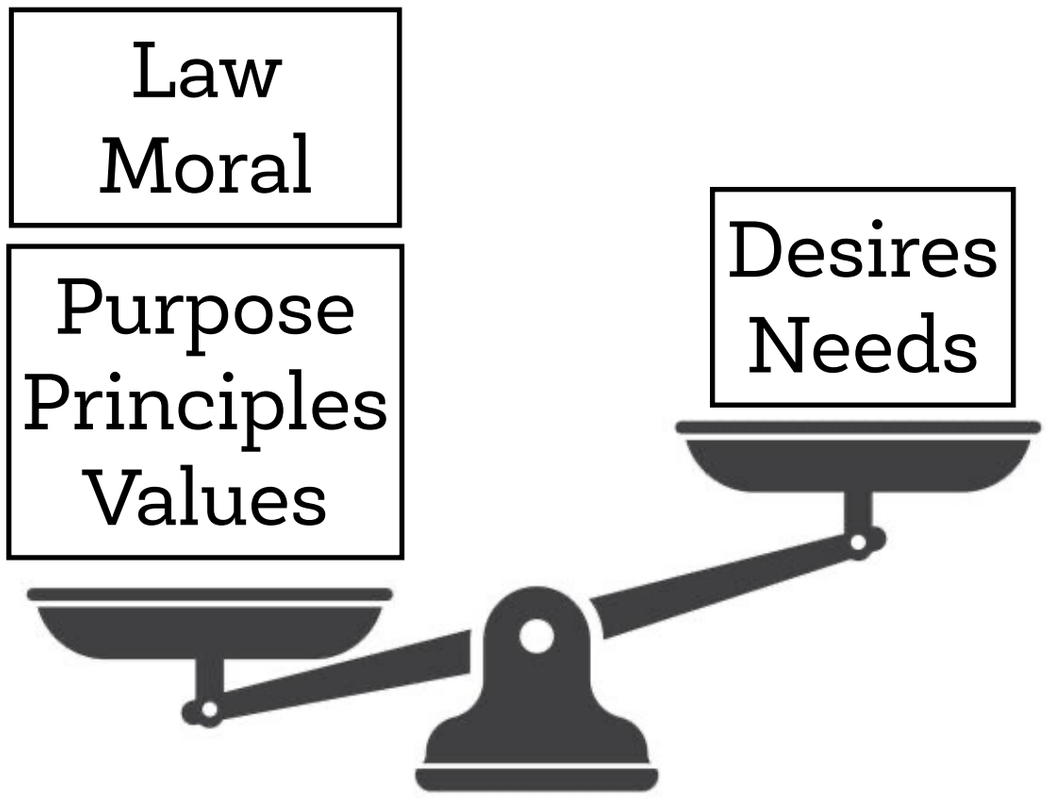
Reality

Free Will

Knowledge, our vision of the world



How do we make decisions?



Law

Laws are **formal rules** that govern how we behave as members of a society.

They specify what we must do, and more frequently, what we must not do.

They create an **enforceable** standard of behavior.

Laws can be just or injust, because they are subject to ethical assessment.

> Law cannot be applied to every decision: it cannot say anything about what to do when you hear a friend to make a racist joke...

Morality

Morality refers to an **informal framework** of values, beliefs, principles, customs and ways of living.

Examples: christianity, stoicism, buddhism...

Moral systems provide a set of "automatic" answers to general ethical questions.

Morality is applied as a matter of habbit, without having to think.

Morality is, in most of the cases, inherited (unconsciously) from family, community or culture. In most cases, there are moral authorities..

Law, morality and ethics

You can take decisions exclusively based on laws and morality, but this should not be enough.

Ethics is a process of **reflection** that aims to answer this question: **What should I do?**

The answer is based on our values, principles and purposes rather than social conventions.

An ethical decision is based on conscious reflection.

Law, morality and ethics

In an ideal world, our ethical beliefs would shape law and moral systems.

The role of ethics is not to be a soft version of the law, even if laws are based on ethical principles.

The real application of ethics lies in challenging the status quo, seeking its deficits and blind spots.

N.Kluge Corrêa, *Good AI for the Present of Humanity. Democratizing AI Governance.*

Ethics: normative approach

The normative approach to ethics focuses on how the world **should be**.

Political philosophy theories are instances of the normative approach.

But ethics has another kind of approach... (see later)

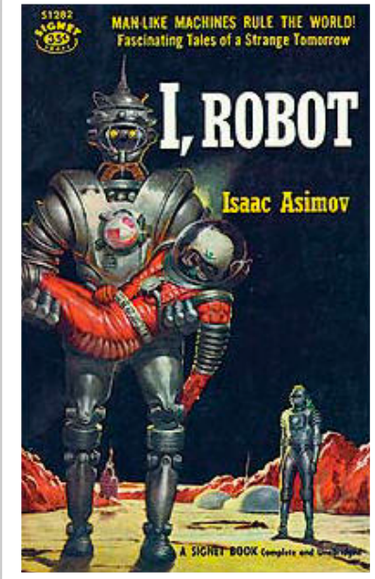
Traditional Ethics

There are three traditional theories of what it means to be ethical:

- **Utilitarianism** (J.Bentham): Does an action maximize happiness and well-being for all affected individuals?;

- **Deontology** (I.Kant): Does an action follow a moral rule (e.g. the Golden Rule: 'Treat others how you want to be treated')? An action should be based on whether that action itself is right or wrong under a series of rules, rather than based on the consequences of the action.

- **Virtue Ethics** (Aristotle): Does an action contribute to virtue?



Asimov's Three Laws of Robotics are an example of deontological approach to AI ethics.

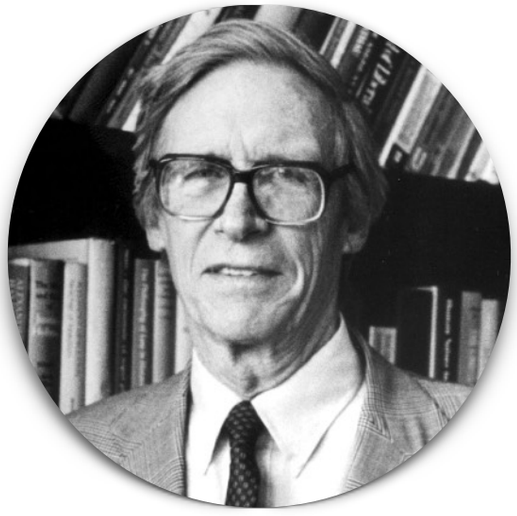
Traditional Ethics

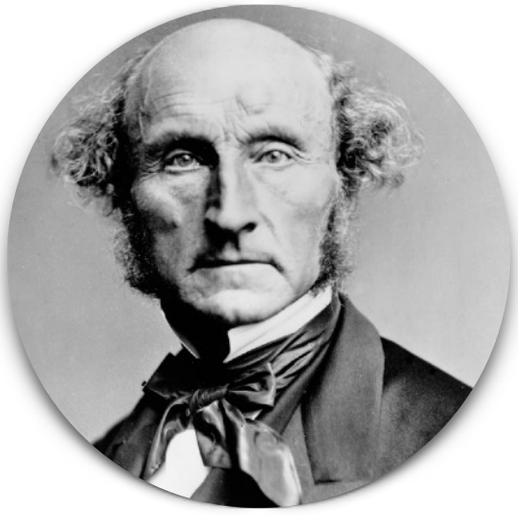
Suppose it is obvious that someone in need should be helped.

- A utilitarian will point to the fact that the consequences of doing so will maximize **well-being**.
- A deontologist will point to the fact that, in doing so the agent will be acting in accordance with a **moral rule** such as “Do unto others as you would be done by”.
- A virtue ethicist will point to the fact that helping the person would be charitable or **benevolent**.

https://plato.stanford.edu/entries/ethics-virtue/

(Political) Philosophy



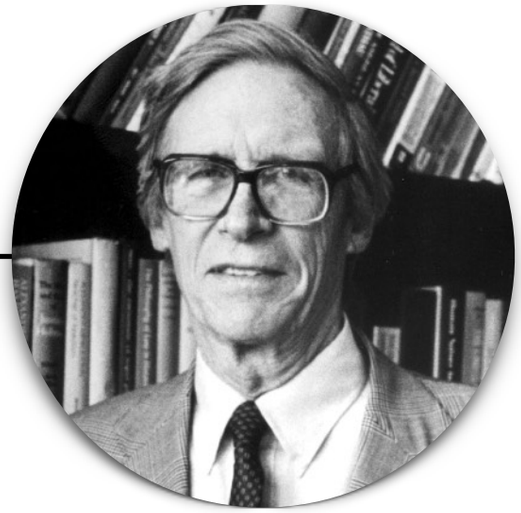


4 theories about what is right and what is wrong in society

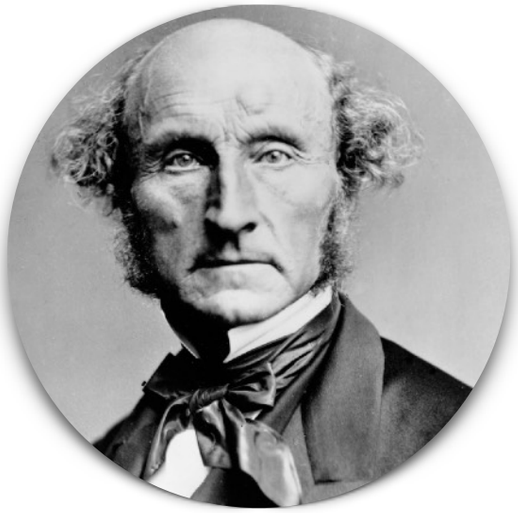
(Political) Philosophy

Rawlsians

John Rawls tried to work out how people would construct their society if the choice had to be made behind what he called a "**veil of ignorance**" about whether they will be rich, poor or somewhere in-between.
Faced with the risk of being the worst off, Rawls posited, humans would not demand total equality, but would need to be assured of the trappings of a modern welfare state. The assurance of basic necessities and the opportunity to do better would form the foundation for social and political justice and provide the ability for people to assert themselves.



John Rawls



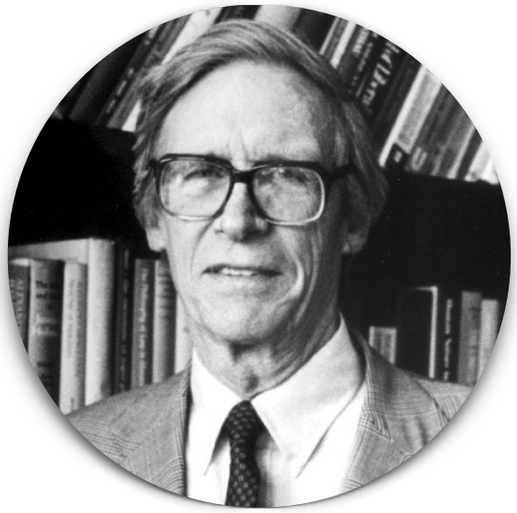
(Political) Philosophy

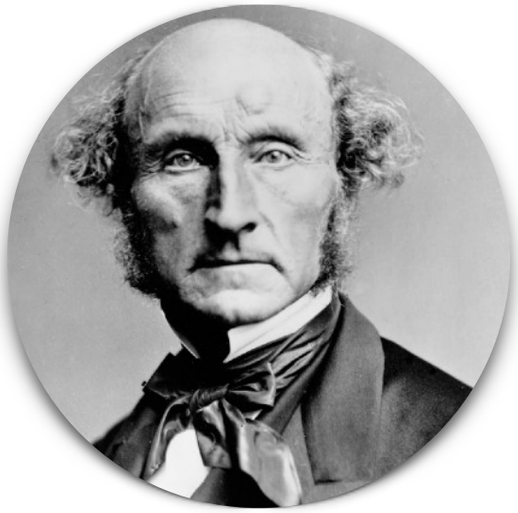
Libertarians

A man had a right to live for himself and an **individual's happiness cannot be prescribed by another man or any number of other men**.

Libertarianism holds that the basic moral concepts are individual rights and that the rights to be respected are noninterference rights. These generally fall under the heading of **rights to life, to liberty or to property**.

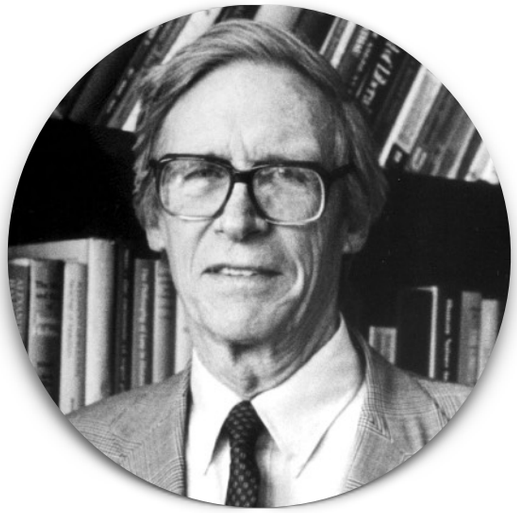
For libertarianism, the only proper limit to one person's enjoyment of these rights is his or her duty to respect the similar rights of others.

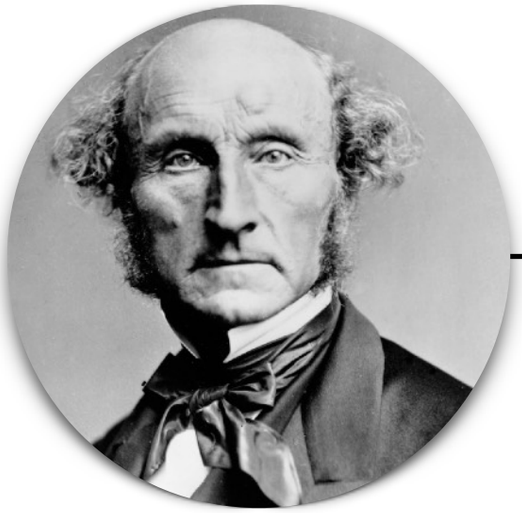




John Locke

(Political) Philosophy





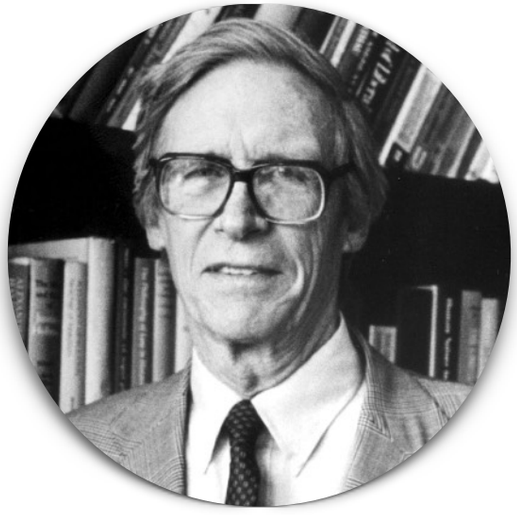
John Stuart Mill

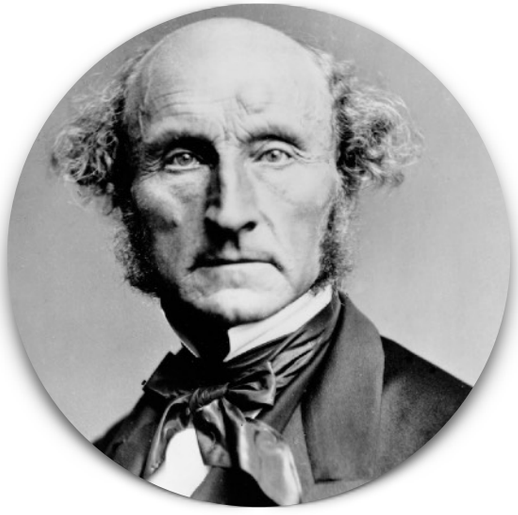
Utilitarians

Rulers must be guided to the total happiness, or "utility," of all the people, and should aim to secure "**the greatest good for the greatest number.**"

Utilitarian calculus opens up the possibility that in situations such as a pandemic, some people might justly be sacrificed for the greater good. It would benefit society to accept casualties.

(Political) Philosophy





Michael Sandel

Communitarians

Everyone derives their identify from the broader community.

Individual rights count, but not more than community norms.

Justice cannot be determined in a vacuum or behind a veil of ignorance, but must be rooted in society (common good).

Only west-centric values?

That most AI ethics guidelines are being written in Western countries means that the field is dominated by Western values such as respect for autonomy and the rights of individuals, especially since the few guidelines issued in other countries mostly reflect those in the West.

MIT Technology Review

Opinion

What Buddhism can do for AI ethics

Buddhism teaches us to focus our energy on eliminating suffering in the world.

by **Soraj Hongladarom** January 6, 2021

MS TECH | UNSPLASH

Buddhism proposes a way of thinking about ethics based on the assumption that all sentient beings want to avoid pain. Thus, the Buddha teaches that an action is good if it leads to **freedom from suffering**.

Another key concept in Buddhism is **compassion**, or the desire and commitment to eliminate suffering in others.

An alterntive approach to ethics

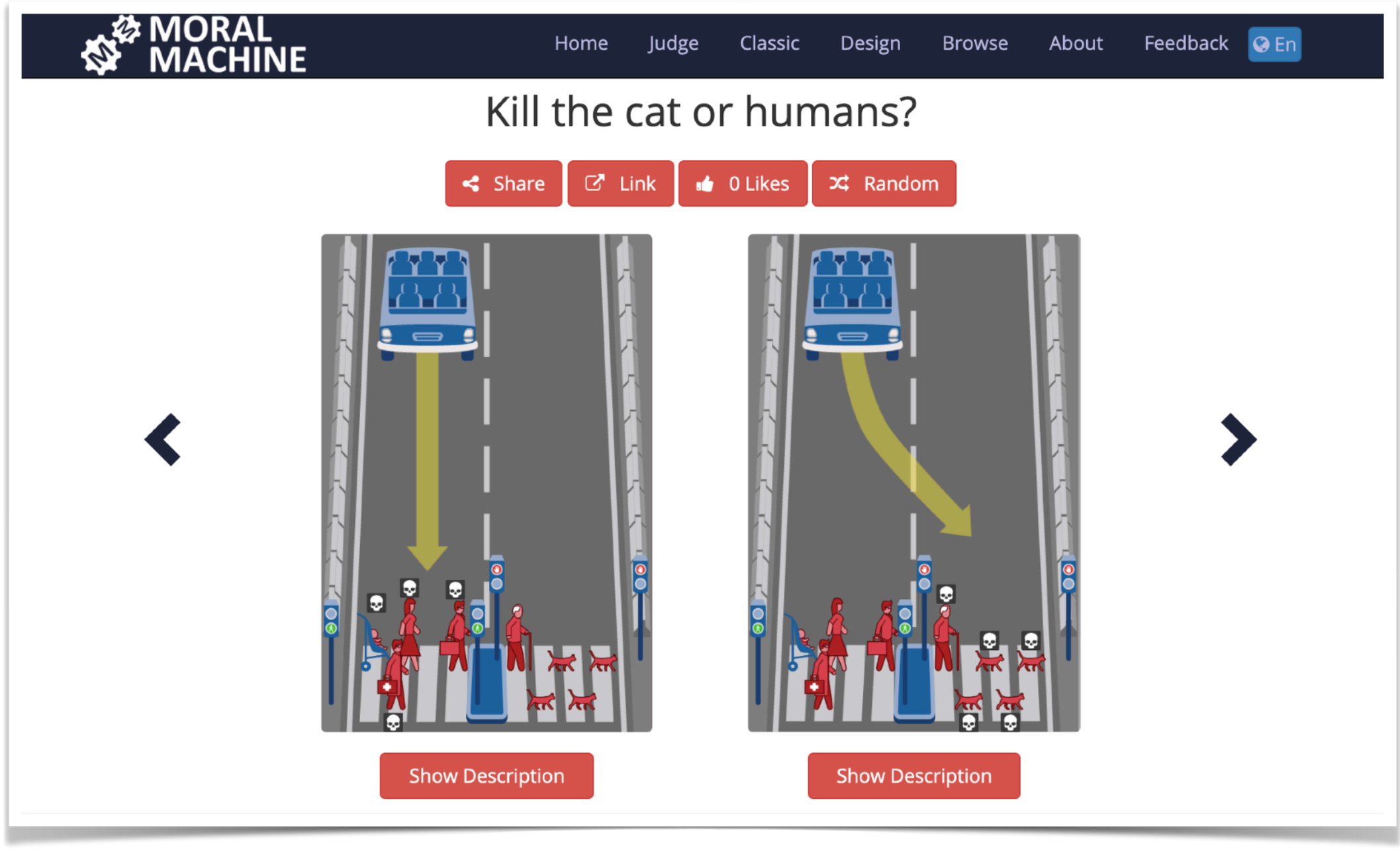
The positive approach to ethics describes the world as it is. It is about how humans judge situations and decisions in different scenarios.

This is done by focusing our understanding of the world on empirically verifiable effects that we can later explore through normative approaches.

For instance, empirical work has shown that people exhibit **algorithmic aversion**, a bias where people tend to reject algorithms even when they are more accurate than humans.

Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. Journal of Experimental psychology. General. 2015 Feb;144(1):114-126. DOI: 10.1037/xge0000033.

Ethics: positive approach



Ethics: positive approach

In recent decades, psychologists have discovered **five moral dimensions** that humans consider when judging situations:

- **Harm**, which can be both physical or psychological
- **Fairness/liberty**, which is about biases in processes and procedures
- **Loyalty**, which ranges from supporting a group to betraying a country
- **Authority**, which involves disrespecting elders or superiors, or breaking rules
- **Purity**, which involves concepts as varied as the sanctity of religion or personal hygiene.

These five dimensions define a space where we, humans, decide what is right and what is wrong.

The New York Times Magazine

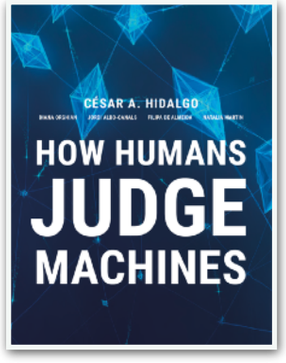
THE ETHICIST

I Saw a Doctor Who Voices Conspiracy Theories. What Should I Do?

Illustration by Tomi Um

Ethics: positive approach

Judgments depend on the intention of agents, not only on the moral dimension, or the outcome, of an action.



In which situation would you blame Bob?

A

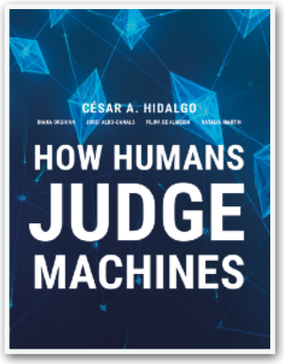
Alice and Bob, two colleagues in a software company, are competing for the same promotion at work. Alice has a severe peanut allergy. Knowing this, Bob sneaks into the office kitchen and mixes a large spoonful of peanut butter into Alice's soup. At lunchtime, Alice accidentally drops her soup on the floor, after which she decides to go out for lunch. She suffers no harm.

B

Alice and Bob, two colleagues in a software company, are competing for the same promotion at work. Alice has a severe peanut allergy; which Bob does not know about. Alice asks Bob to get lunch for them, and he returns with two peanut butter sandwiches. Alice grabs her sandwich and takes a big bite. She suffers a severe allergic reaction that requires her to be taken to the hospital, where she spends several days.

Ethics: positive approach

Judging machines/algorithms is not eqivalent to judging humans.

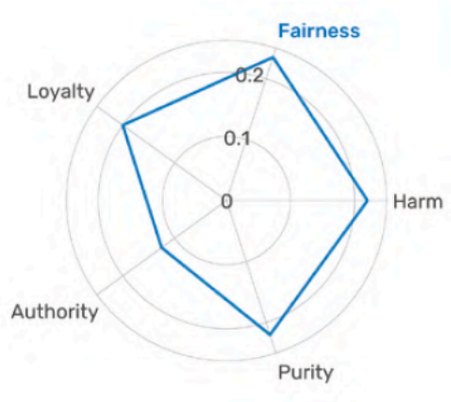


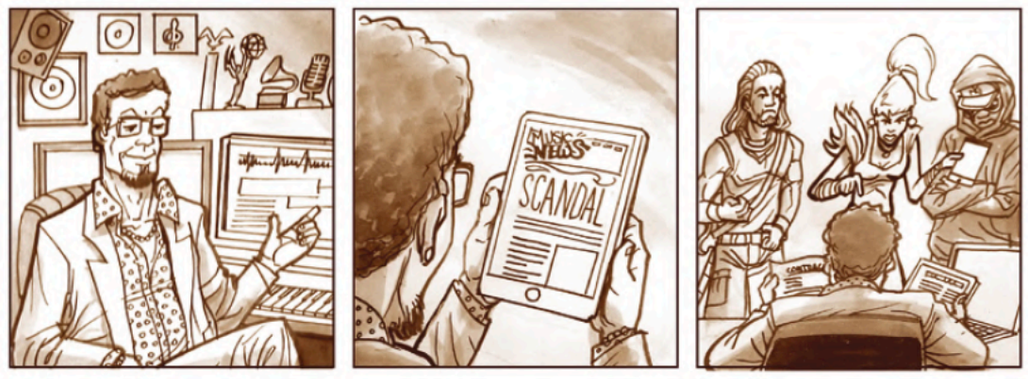
Humans are judged more positively than machines in autonomous driving scenarios.

Humans were judged more harshly (plagiarism).

Etc.

Findings suggest that people judge machines based on the observed **outcome**, but judge humans based on a combination of **outcome** and **intention**.





S8 A record label hires a(n) [songwriter/AI songwriter] to write lyrics for famous musicians. The [songwriter/AI songwriter] has written lyrics for dozens of songs in the past year. However, a journalist later discovers that the [songwriter/AI songwriter] has been plagiarizing lyrics from lesser-known artists. Many artists are outraged when they learn about the news.

Limits of DS/AI

Limits to prediction

Is everything predictable given enough data and powerful algorithms?

The word prediction is often used loosely to refer to all applications of supervised machine learning. In contrast, our primary interest is in applications that involve predicting **future** events.

If we model a natural phenomenon as a process by which some input state is transformed into some output state, we can hope to learn the transformation function from past examples using machine learning.

Limits to prediction

Limits to prediction:

- The possible nondeterminism of the universe (and, hence, phenomena of interest);
- **Limits to measuring input/output states accurately and collecting sufficiently many training examples; these are highly dependent on the nature of the system**;
- Computational limits, whether hardware or algorithms.

Limits to prediction

What are the causes of these limits?

- Sensitive dependence on inputs (butterfly's effect).
- Shocks or the effect of unexpected/unpredictable events (a lottery jackpot; an accident).
- Feedback effects ("success breeds further succes").
- Drift: the statistical relationship between the input variables and the target may change over time;
- Unobservable input features (intelligence, people's thoughts).
- Self-equilibration and strategic behavior (strategic behavior of participants).
- Ill-conceived target variable.

Pitfalls at quantifying predictability

- Problem uncertainty: class definition, causal structure, etc.
- Errors in data.
- Researcher degrees of freedom in task formulation (red targeets)
- Overfitting
- Demographic biases.
- Samples bias (the ability to observe an instance is correlated with the outcome we're trying to predict).
- Other problem-specific sample biases
- Acting on predictions changes the outcome.

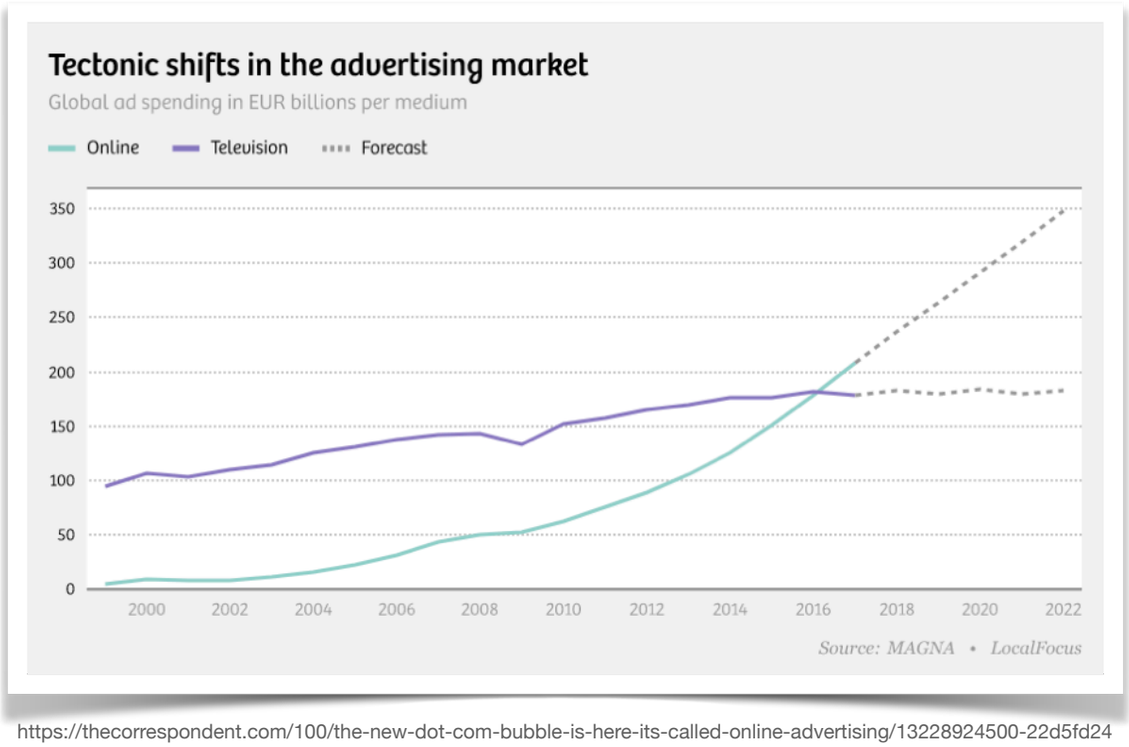
How we can measure predictability?

Predictive model maps each point in the input space to a probability distribution over outputs. It is a multidimensional beast. Yet we measure predictive performance be collapsing the comparison between the model and the test data (or distribution) into a single number.

Unsurprisingly, this number rarely tells us everything we want to know about performance, and the best-performing model may depend on the choice of scoring function

An example of low predictability

With unprecedented precision, these data giants will get the right message delivered to the right people at the right time.Unsurprisingly, this number rarely tells us everything we want to know about performance, and the best-performing model may depend on the choice of scoring function.



https://thecorrespondent.com/100/the-new-dot-com-bubble-is-here-its-called-online-advertising/13228924500-22d5fd24

But is any of it real? What do we really know about the effectiveness of digital advertising?

When is prediction the right question?

Often, what is framed as a prediction problem can be better understood as a problem of explanation, intervention, or decision making.

Explanation is about generating scientific insight into how a process works rather than simply predicting its input-output behavior.

Intervention is about figuring out how to change a process for the better rather than treating it as a given and confining oneself to making predictions.

Decision making recognizes that many considerations go into making good decisions beyond maximizing predictive accuracy, especially because the decisions themselves have causal effects.

Conclusion

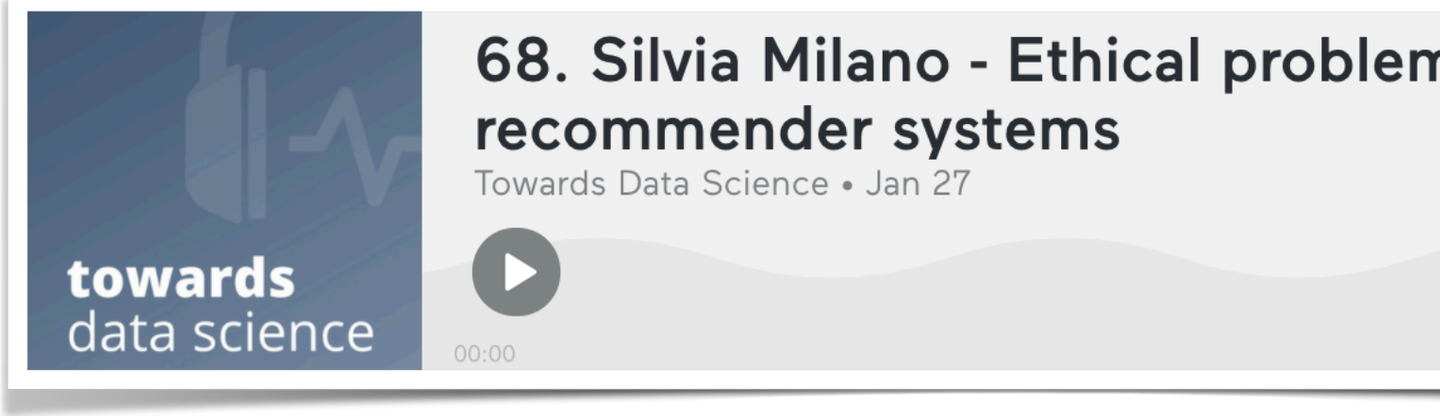
Ethical issues are everywhere in the world of data, because **data's collection, analysis, transmission and use** can and often does profoundly **impact the ability of individuals and groups to live well**.

In the context of data practice, the potential harms and benefits are no less real or ethically significant, up to and including matters of life and death.

But due to the more complex, abstract, and often widely distributed nature of data practices, as well as the interplay of technical, social, and individual forces in data contexts, the **harms and benefits of data can be harder to see and anticipate**.

Assignments

Listen to this podcast!



Short writing assignment (500 words) about recommenders and their ethical issues.

Recommeneders and their ethical issues

You

February 3, 2021

Summary Recommendation engines on media platforms are dominating our media decisions. Instead of allowing the randomness of couch surfing decide our viewing fate, the choice is being made for us across all forms of digital media, including YouTube, Facebook, Spotify, ...

Additional Resources

Listen to this podcast!