INTRODUCTION TO AI ETHICAL PRINCIPLES

Cleber Mikio Ikeda

1. INTRODUCTION

Do you use a banking app to transfer money or pay bills? Do you watch movies and your favorite series on Netflix? Have you shared selfies or liked friends' posts on Facebook? Are you looking for a job? If you answered yes to at least one of these questions, then you have been exposed to **artificial intelligence (AI)**. There are many definitions of AI, so let's make it simple: AI is a computational tool that uses and interacts with data to solve problems and achieve goals. Your bank uses your data to predict your purchasing decisions and then offer you special credit lines; Netflix uses recommendation algorithms to analyze your past preferences and continue delivering the best entertainment possible; Facebook's algorithms analyze your interests to show you ads of products and services you might want to buy; and before your CV reaches recruiters' hands, it might have to be triaged by a Natural Language Processing (NLP) tool.

There are intrinsically ethical, moral aspects of AI. Think about your personal data that has been uploaded into social media platforms over years long. How do companies as Facebook handle and use your data? (You might be thinking: "that is all explained in the terms and conditions I accepted… without reading it!"). Now, think about face recognition being deployed in police surveillance: what if you can be wrongly arrested because state-of-art face recognition technology does not work accurately with darker-skin people? There are also the pressing questions we all make ourselves regarding self-driven cars: whose responsibility is in the event the car hurts or kills somebody? Is the company that owns the car? Is the AI developers' fault? I would not be surprised if someone even blamed the victim. There are countless applications of AI in which ethical dilemmas are present and this is because AI has been so much an integral part of our lives. Inappropriate use of AI can even have drastic consequences to democracy, human rights and, therefore, to the world we want to leave to the next generations.

In this article, I will explore the following **AI ethical principles** and how they can be put into action:

- Fairness
- Accountability
- Human agency
- Transparency
- Privacy
- Respect to human rights

Some disclaimers:
1) I will present each principle individually for pedagogical purposes, but it is important to note there is a lot of interconnectivity amongst the AI principles. Commonly, we find two or more principles operating in real-life AI ethical dilemmas.
2) Many institutions and publications present different lists of AI ethical principles. I made my own list with the principles I deemed more important in terms of applicability to our daily lives.

3) I am an early student and enthusiast of AI Ethics. This article has a single objective: collaborate with the community of professionals who want to spread the knowledge on AI and address its ethical implications to our society. Having that in mind, I will also share with you some good references for further study (my first recommendations are right at the next box), so you can get inspired by great people who have been doing so much in the field of AI Ethics. I hope you enjoy the discussion.

Recommendations
To watch: A quick, though rich conversation with Timnit Gebru on AI Ethics. https://www.youtube.com/watch?v=60nLTtjdiFc.
For a quick reading: He was wrongfully accused by an algorithm. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html.
Looking for a deeper dive? A compilation of thirty-six AI principles documents. https://dash.harvard.edu/handle/1/42160420.
Must-have in your toolbox: A one-stop-shop for references on AI Ethics, by @Merve Hickok. https://www.aiethicist.org/.

2. FAIRNESS

AI has brought many benefits to society. As examples, diseases can be detected quicker and more accurately thanks to the application of computer-aided diagnosis (CAD); algorithms have been applied to public policies aiming to increase public services efficiency and welfare; and judicial systems have been using AI to help judges make fairer and more objective decisions.

What if CAD cannot be as accurate on females as it is in males? What if educational system planners cannot be as successful as possible because their algorithms do not consider race as an attribute? What if judges, under influence of AI, have been taking unfair decisions towards black people? Researchers, journalists and members of civil society have been raising stunning flaws in AI technologies that have a direct impact on people's lives. One of those flaws is known as **AI bias** and it is directly related to the idea of **fairness**. I invite you to analyze such a challenging problem by trying to answer three questions.

Does a more balanced, representative data set improve AI efficiency?

A group of researchers analyzed the performance of CAD systems in detecting several thoracic diseases based on medical imaging datasets.[1] Their main objective was to determine if CAD trained with gender-imbalanced dataset could present poorer performance than CAD trained with gender-balanced dataset.

This is what they found out: **more imbalanced datasets led to statistical implications that worsened pathology classification performance for minority groups** (being that male or female): *"We found that, even with a 25%/75% imbalance ratio, the average performance across all diseases in the minority class is significantly lower than a model trained with a perfectly balanced dataset."* Their results indicated that *"diversity provides additional information and increases the generalization capability of AI systems."*

[1] Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante. 2020. *Gender imbalance in medical imaging dataset produces biased classifiers for computer aided diagnosis*. https://www.pnas.org/content/117/23/12592.

Should attributes as race and gender be present in AI models?

Facts: women make less money than men for the same job position; black people are concentrated in less developed neighborhoods and are poorer than white people; transgender have fewer job opportunities and are stigmatized; immigrants suffer discrimination. People unfairly suffer because of what they are and because of their uniqueness. Despite the increase of initiatives to promote and celebrate diversity worldwide, it continues to be misunderstood and under attack.

Discrimination and unfairness have led many thinkers to propose statistical manipulation or even exclusion of attributes as gender and race from AI algorithms. In the paper *Algorithm Fairness*, researchers tested a predictive model for college success, whose predictions could help a social planner to make admissions decisions.[2] The test consisted of applying three different data sets to the model to determine how the attribute race would impact the maximum anticipated performance of the admitted students: the first data set was blinded for race; the second one was orthogonalized for race; and the third one was completely race-aware. For the sake of simplification, the researchers just focused on two groups: non-Hispanic white students and black students.

Results showed that the race-aware model is always preferable to the orthogonalized and race-blinded models. No matter if you are an efficient planner (preference to maximization of college success regardless of race) or an equitable planner (preference to diversity and college success maximization), the race-aware model presents better anticipated grades.

The researchers conclude that *"Absent legal constraints, one should include variables such as gender and race for fairness reasons"* and *"the inclusion of such variables can increase both equity and efficiency."* By "legal constraints", the researchers are mainly referring to the tradeoff between privacy and fairness.

Is it possible to quantify fairness and achieve AI bias-free outputs?

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is an algorithm-based tool used in the US judicial system that predicts the likelihood of recidivism. The predictions are supposed to support judges' decisions on keeping defendants in jail or releasing them during the pre-trial period. Advocates of COMPAS defend the system is more objective than a judge because it does not "see" attributes like race and religion.

Though COMPAS does not use race as an attribute, ProPublica, an independent organization focused on investigative journalism, argued the tool discriminates against black people: *"Black defendants were (...) 77 percent more likely to be pegged as at higher risk of committing a future violent crime and 45 percent more likely to be predicted to commit a future crime of any kind."* [3]

---

[2] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. *Algorithmic Fairness*. AEA Papers and Proceedings, 108: 22-27.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. *Machine Bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Would it be possible to fix such an unfair algorithm? An article published in the MIT Technology Review demonstrates how hard it is to convey the concept of fairness in AI.[4] Let's say one defines fairness as *"error rates should be the same for both black and white defendants"*. "Error" means either needlessly sending people to jail or erroneously releasing the ones who would be re-arrested. Thus, under that definition, COMPAS should incur in both errors at the same rates for black and white defendants. First problem: to statistically achieve such rate equality, COMPAS would have to have different high-risk score thresholds, around 6-7 for white defendants and 8-9 for black defendants (in a 1-10 scale). That doesn't seem to be a fair algorithm, right?

The authors then analyzed the COMPAS model considering fairness as *"the same risk score threshold applies to all defendants"*. Under this alternative definition, it is statistically impossible to achieve the same error rates for black and white defendants. No matter what the high-risk threshold is, the error rate will always be higher for black defendants. Not fair at all, as already pointed by ProPublica.

Behind the numbers, there are so many stories data has to tell. **Because black and white people have been historically arrested at different rates, predictions on recidivism will invariably reflect the same pattern: black people will receive higher scores than white people, independently of algorithm accuracy.** The reason for black people having been arrested at higher rates is a different, though not a less important issue. The history of US police targeting minorities is only one of the key variables of this intricate equation.

Conclusion

Bias does not have to do only with algorithms; it has mostly to do with structural social bias and prejudice we have inherently carried with us for centuries – many times unconsciously. That's why achieving fairness in AI systems is so difficult: data is inevitably biased, so are the people who work with it.

Bias will always be there and fairness is a concept that varies according to context, culture, judicial systems. **The AI bias problem is not simply a mathematical, statistical problem; it is fundamentally a moral, ethical and political decision the algorithm designer should take while approaching a maximization problem**.

Having that in mind, it is urgent to promote affirmative actions against discrimination and inequalities within the AI field, as they have the power to positively impact algorithms in terms of accuracy and fairness. Also, careful analysis of datasets used to train AI systems plays a big role not only in the efficiency and accuracy of AI systems, but also prevents dissemination and amplification of gender and racial bias.

Recommendations
To watch: What Robin Hauser, a documentary filmmaker (and not a scientist), learned about bias in AI. https://youtu.be/eV_tx4ngVT0.
For a quick reading: More AI bias in predictive policing in the Netherlands. https://thenextweb.com/neural/2020/09/30/dutch-predictive-policing-tool-designed-to-ethnically-profile-study-finds/.
Looking for a deeper dive? Ruha Benjamin. 2019. *Race After Technology*. John Wiley & Sons.

[4] Karen Hao, and Jonathan Strayarchive. 2019. *Can you make AI fairer than a judge? Play our courtroom algorithm game*. MIT Technology Review. https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/.

Must-have in your toolbox: AI Ethics Weekly Newsletter to get updates on the most pressing AI ethics issues. https://lighthouse3.com/newsletter/.

3. ACCOUNTABILITY

6 seconds. That was the time the self-driven car took between recognizing the other "vehicle" and the impact. According to the National Transportation Safety Board (NTSB) report, the Uber autonomous car recognized a "vehicle" 5.6 seconds before the impact; within the next 4.5 seconds it changed to "other", then "vehicle", then "other", then "bicycle", then "other", then "bicycle" once again. Then it held off breaking for one second, even having anticipated the imminent collision. There was not enough time left for breaking and the car hit Elaine Herzberg, a 49-year old woman who was crossing the road on foot while pushing her bike. Elaine died of her injuries and the tragic event became the first known death of a pedestrian involving a self-driven car.[5]

When something goes wrong as a result of an AI system's decision – or lack of it – who is held accountable? The question is more complicated than it seems. Most of the AI systems that operate in circumstances involving risk of life or that potentially pose a significant negative impact on people's lives have a human in the loop. That was the case of the accident that took Elaine Herzberg's life: there was a human, a backup driver who was allegedly watching her cellphone when the car struck the victim. That driver was recently indicted for criminal negligence; Uber, the company that designed the AI system operating the self-driven car, will not face criminal charges.

In the last section on the fairness principle, we saw how gender imbalanced data sets may have a great impact on the accuracy of computer-aided diagnosis of diseases through medical imaging analyses. I'll present another practical example of deployment of imbalanced data in AI systems, but with a focus on how **auditing and impact assessments can increase accountability** of algorithm owners and designers and, as consequence, contribute to addressing the negative impacts of AI systems on society.

Not a coincidence, the European Commission recognizes the AI accountability principle as closely related to the fairness principle. Its High-Level Expert Group on Artificial Intelligence recommends: *"When unjust adverse impact occurs, accessible mechanisms should be foreseen that ensure adequate redress. (...) Particular attention should be paid to vulnerable persons or groups."* It also addresses the importance of impact assessments to mitigate adverse outputs: *"Identifying, assessing, documenting and minimizing the potential negative impacts of AI systems are especially crucial for those (in)directly affected. (...) The use of impact assessments (e.g. red teaming or forms of Algorithmic Impact Assessment) both prior to and during the development, deployment and use of AI systems can be helpful to minimize negative impact."* [6]

[5] Aarian Marshall. *Why Wasn't Uber Charged in a Fatal Self-Driving Car Crash*? Sep/2020. Wired. https://www.wired.com/story/why-not-uber-charged-fatal-self-driving-car-crash/.

[6] European Commission, High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI. 2019*. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top.

Gender, skin and algorithms

In their 2019 paper, Raji and Buolamwini presented the commercial impact of Gender Shades, the first algorithmic auditing on the performance of facial analysis models. Their main achievement was to demonstrate how the auditing results pushed target-companies such as IBM and Microsoft to improve the accuracy of their facial recognition algorithms and reduce bias towards female and darker-skin faces. The study was also able to show that the same advance was not observed for non-target companies such as Amazon.[7]

Gender Shades proposed the utilization of a "user-representative" test, i.e. the test data set is not representative of the demographic distribution of distinct groups, but **incorporates diversity by equal representation of such groups**. Based on this equitable test data set, the auditing determined the error rates for the subgroups males, females, darker skin, lighter skin, and the intersectional subgroups lighter skin males, lighter skin and females, darker skin males and darker skin females. The error rate was the proportion of faces that got wrongly recognized or was not recognized by the facial analysis models. Finally, the researchers run a follow-up auditing a few months later, after the target-companies had updated their algorithms. These were the results of the auditing conducted in August 2018 on the target-companies' facial recognition models:

- The overall error rate ranged from 0.48% to 4.41% across the target-companies.
- The company with the poorest overall performance also presented an error rate for females as high as 9.36% vis-a-vis 0.43% for males; the error rate for darker skin was 8.16%, compared to an error of 1.17% for lighter skin.
- In the intersectional analysis, the algorithmic bias became more obvious: the highest error rate detected for darker skin females was 16.97%, compared to 0.26% for lighter skin males.

In summary, the target-companies' facial analysis models failed to recognize darker skin, female and darker skin female faces with the same accuracy obtained for lighter skin, male and lighter skin male faces.

Gender Shades applied the Coordinated Vulnerability Disclosures (CVD), a strict procedure promoted by the National Computer Emergency Readiness Team (CERT) to inform the target-companies of the auditing results, which had an undeniable impact on them. Seven months later, after the target-companies have examined and updated their facial recognition Application Program Interfaces (APIs), the researchers run a follow-up analysis and the results were rewarding: though the largest error rates persisted for darker skin females, there was a greater reduction in errors for female faces and darker faces, being the darker females subgroup the one with the greatest accuracy improvement.

Those results demonstrated that the **deployment of a user-representative, equitable test data set does not jeopardize the overall performance of facial recognition models**. This methodology, which addresses the bias problem since the early stages of algorithm design, is a promising alternative to the traditional test data subset extracted from training data sets.

[7] Inioluwa Deborah Raji & Joy Buolamwini. *Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products.* 2019. Association for the Advancement of Artificial Intelligence (www.aaai.org).

Another group of researchers stressed the importance of conducting audits before the deployment of AI systems. According to them, *"pre-deployment audit applied throughout the development process enables proactive ethical intervention methods, rather than simply informing reactive measures only implementable after deployment"*.[8]

Inspired by auditing frameworks from other areas such as financial, aerospace and medicine, the researchers propose an original framework for AI systems, which comprises the following 5 phases:

i. Scoping: definition of audit objectives; understanding of AI system's objective and impact; confirmation of principles that will guide the system development; ethical review and social impact assessment.
ii. Mapping: map and interview stakeholders and key audit collaborators.
iii. Artifact Collection: identify, collect and analyze relevant documentation, from design documents to systems architecture diagrams.
iv. Testing: auditors test the AI system based on audit objectives, ethical standards and principles.
v. Reflection: auditors analyze results; confirm if they meet ethical expectations; formalize risk analysis; and recommend design decisions to the AI system owners.

Conclusion

From autonomous cars to face recognition technologies, accountability has the potential to reduce bias and pave the way for more effective redress when something goes wrong. Whenever accountability is clearly established, AI systems' owners successfully build trust with users and promote a more seamless acceptance and adoption of such technologies. Independent auditing and risk assessments represent powerful tools to reach desirable levels of accountability and can serve as indispensable procedures to anticipate bias, gaps and mitigate unintended consequences of autonomous systems' decisions.

As many of you might have already realized, accountability is inherently connected to transparency and human agency, which will be the next AI principles I will explore in this article.

Recommendations

To watch: Get to know more about Joy Buolamwini and her inspiring, impactful work on AI fairness and accountability: https://www.youtube.com/watch?v=UG_X_7g63rY.

For a quick reading: Whom to blame when AI fails? https://www.technologyreview.com/2019/05/28/65748/ai-algorithms-liability-human-blame/.

Looking for a deeper dive? Selbst, Andrew D., *Negligence and AI's Human Users* (2019). 100 Boston University Law Review 1315 (2020), UCLA School of Law, Public Law Research Paper No. 20-01. https://ssrn.com/abstract=3350508.

Must-have in your toolbox: Khari Johnson is Senior AI Staff Writer at VentureBeat. His sharp, well-written articles on how AI impacts society always touch upon the ethical implications of AI. He is also a voice for those who are mostly negatively impacted by AI technologies. You can follow him in the VentureBeat newsletter: https://venturebeat.com/newsletters/.

[8] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*. In Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3351095.3372873.

4. HUMAN AGENCY

Human as agent

Studying the human agency principle, I noticed there is unclarity – and sometimes confusion – on two concepts that are usually present in articles and papers: autonomy and agency. Commonly, both concepts are referred to the same ideas of freedom, independence and ability to perform tasks or make decisions free of external control or influence. That meaning is undoubtedly important to the discussion on human agency when AI systems make decisions that affect people's lives, e.g. going or not going to jail, approving or not bank loans, going to university A or B. As stated by the European Commission, *"Key to this is the right not to be subject to a decision based solely on automated processing when this produces legal effects on users or similarly significantly affects them."* [9] In that sense, the human being is at the center of the problem as a key agent that will ultimately take and/or challenge big decisions. Therefore, **the human agency problem is a moral, ethical problem** in itself: we want people to be free of unfairness and manipulation from AI systems and a human in the loop should play a pivotal role there.

AI as agent

Now, let's place AI at the center of the problem for a moment. From a sole technical perspective, humans have created AI systems to perform tasks quicker, smarter, more accurately and independently, so we could spend our time doing more important things. In that sense, autonomy is a regardable feature, as we want to enable AI to solve many of our modern life problems: from unlocking smartphones to weather forecasts to recommending systems to self-driving cars. Thus, AI as an agent is a technical problem in essence: we want machines and systems to achieve predefined goals without human oversight, review and approval; sometimes, those goals include taking decisions on human's behalf. However, this process of "delegating" tasks and responsibilities to automated systems has a critical consequence to the agency problem: **assigning autonomy to AI systems invariably means giving up and transferring agency to those same systems**. The agency and autonomy concepts when applied to AI agents are not interchangeable ideas. Autonomy has to do with the technical aspects of the AI systems, while agency does embed an ethical aspect. The technical problem of making autonomous systems turns into an ethical problem only because we not only want AI to work autonomously, but also according to the algorithms designed by us. Most importantly, we don't want AI to harm people while operating independently.

The idea of AI as an agent leads us to the principal-agent model so well known in Economics. The principal-agent model refers to the economic relationship between agents that work on behalf of or impacting principals, e.g. business administrators and shareholders, attorneys and clients, brokers and sellers/buyers. The relationship between agent and principal has a particularity: both have self-interests. Considering that agents act on behalf of principals, the latter might be negatively affected by actions and decisions taken by the former if there is no alignment between their interests. That's why the model principal-agent is also known as the agency dilemma.

[9] European Commission, High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. 2019. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top.

Though AI systems don't have self-interests, the agency dilemma is a practical framework for our discussion. In the context of our present discussion, AI systems are the agents and algorithm designers/owners are the principals. In a perfect world, AI would strictly follow directions from the principals without incurring unintended consequences. Unfortunately, that is not what we have been observing and that's one of the main reasons Ethics has become so important for AI. How do we make sure AI works as we intend them to or, in other words, how can AI systems be good agents and behave according to principals' best interests?

The agency dilemma: humans and AI

One of the answers to that question is: **AI systems as agents are not as good as we wish because context matters**; and we know AI is not very good in perceiving and understanding context as we humans do.

Alkhatib & Bernstein explored the idea of street-level bureaucrats and street-level algorithms.[10] Street-level bureaucrats are the persons who bridge the gaps between law, regulations (policy) and decisions, e.g. police officers, judges and teachers. Analogously, street-level algorithms are *"the algorithms that bridge the gaps between policy and decisions about people in a socio-technical system"*. According to the authors, street-level bureaucrats enforce policies through a reflexive decision-making process, particularly when they face a novel situation. An example presented by the authors: a judge assessing a cybercrime fraud perpetrated in a way never seen before. The judge reflects and takes a decision based on his/her professional acumen, analysis of the context and the implications of the decision to the offender and to the law itself. On the other hand, a street-level algorithm operates based on past training data that does not contain the new event or situation, which might lead to erroneous, inappropriate decisions. It is always possible to refine the algorithm, but only after a decision has been made. *"For a bureaucrat, but not an algorithm, the execution of policy is itself reflexive. For an algorithm, but not for a bureaucrat, reflexivity can only occur after the system receives feedback or additional training data"*.

Reflexivity by street-level bureaucrats implies judgment, which implies subjectivity. The level of subjectivity will fundamentally depend on the policy framework that supports the bureaucrat's decision – the less strict, objective and enforceable the framework is, the more subjective decisions might be. Subjectivity will always play a role when humans make decisions, especially on other humans. Objectivity then has been presented as a quintessential feature of AI systems, being free of prejudice and bias so intrinsic to human behavior.

Nyrup, Whittlestone & Cave recognize the important contribution AI may have in reducing human bias in public services due to its objectivity.[11] However, the authors also stress value judgments are essential and integral part of public administration. Decisions on means to achieve certain ends and how to manage competing moral considerations can be observed in healthcare provision as an example: *"(...) one valuable end is to achieve the best overall health-outcomes from the available resources. However, it would not be acceptable to achieve this by (...) means of removing all care from patients less likely to recover, even if this could be shown to maximize overall*

[10] Alkhatib & Bernstein. *Street–Level Algorithms: A Theory at the Gaps Between Policy and Decisions.* In CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3290605.3300760.

[11] Nyrup, R., Whittlestone, J., & Cave, S. (2019). *Why Value Judgements Should Not Be Automated*. https://doi.org/10.17863/CAM.41552.

health outcomes. On the other hand, some prioritization of resources is necessary. Exactly how to (...) balance efficiency and equity of healthcare provision in concrete cases is an example of a value judgment."

Deployment of AI systems in public services as healthcare may thus have a dubious impact on some public life principles such as Openness, Accountability and Leadership. When AI systems take over decisions on citizens' lives, value judgments exerted by humans lack. Even if designers try to capture value judgments in the algorithms, AI-based decisions can be similarly faulty – after all, designers also carry bias in their human nature. The authors conclude: *"All value judgments relevant to the decisions of AI systems should be made explicit and, as far as possible, be validated by relevant human decision-makers."*

AI manipulation and human autonomy

In his book *Thinking, Fast and Slow*, Daniel Kahneman presents the two modes human thought operates: system 1 is intuitive, fast, always willing to save energy; system 2 is logical, slow and deliberative. Think about yourself after a busy, stressful day of work. Your system 1 will tell you: "you deserve some rest, take that ice cream in the refrigerator, lay down on the sofa and let's watch a good movie". Your system 2 will try to convince you of a different program: "you need to lose weight, put your training shoes on and let's go for a run". Your system 1 will not give up that easily: "hey, you can go for a run tomorrow… it is cold outside".
Big tech and social media will always want to explore customers and users' system 1. Advertising and marketing companies have been doing that for decades with less or more use of technology. Today, with the pervasiveness of AI, there have been important discussions on how technology has affected human autonomy. Obvious examples of that are our dependence on smartphones and addiction to social media. More subtle instances are our reliance on search engines (e.g. Google) and other recommendation machines (e.g. Netflix).

Big tech companies use our data to anticipate our future choices and give us more of what we want. You see that logic operating when you caught yourself jumping from video to video on Facebook for an hour or so, without noticing; or when ads pop up on your screen offering exactly the product you are willing to buy – and you click on it. On the other hand, this logic narrows down our options and might limit our ability to perceive a broader and more diverse world. A trade-off is in place: efficiency and convenience by autonomy and diversity.

How strong is your system 1? To what extent are you in control when interacting with AI systems? According to Abeba Birhane of University College Dublin, that is determined by cultural, historical and socio-economical constraints each one of us faces throughout our lives: *"the more privileged we are, the more we are afforded the capacity to overrule algorithmic identification and personalization (or not be subjected to them at all), maximizing our degrees of agency"*.[12]

I am a big enthusiast of AI and optimistic about how autonomous systems can be designed and deployed for the good. I do think we can shape the future of AI in a way that promotes humanitarian values, transforms unfair social structures and empowers human autonomy and agency. As beautifully stated by Mike Zajko, *"we can*

[12] Abeba Birhane. *Human agency in the age of AI*. Essay published as a result of the AI & Agency workshop organized by the Summer Institute on AI and Society in 2019. https://aipulse.org/ai-agency/.

engineer technologies and social systems to enhance human agency, to provide capabilities for transformation of individual or collective conditions; or we can design to preserve and reinforce existing power structures." [13]

Recommendations
To watch: An eye-opening perspective on how we use social media – and how it uses us. https://www.thesocialdilemma.com/.
For a quick reading: The physician who advocates the return of agency to patients through AI. https://www.forbes.com/sites/ashoka/2019/08/20/ai-health-and-the-future-of-human-agency/?sh=35cfba2f25e1.
Looking for a deeper dive? Experts' concerns about the future of AI and its impacts on humankind. https://www.pewresearch.org/internet/2018/12/10/concerns-about-human-agency-evolution-and-survival/.
Must-have in your toolbox: The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems is a thorough, high-quality work on addressing ethics considerations in the design and development of autonomous systems. https://standards.ieee.org/industry-connections/ec/autonomous-systems.html#:~:text=The%20IEEE%20Global%20Initiative's%20mission,for%20the%20benefit%20of%20humanity.%E2%80%9D.

5. TRANSPARENCY

Each AI ethical principle does not stand alone as an isolated concept. In the real world, AI presents challenges that relate to multiple ethical principles at the same time. **Transparency** is one of the principles that have more interconnections with other principles: transparency is essential to determine accountability, understand if there is bias in algorithms or comply with privacy regulations. Transparency is also required to establish user's trust in AI systems. In already commonly used applications of AI as chatbots and recommendation systems, users should be clearly advised they are interacting with AI.

Human-computer interaction gets more complex when AI takes important decisions about human beings. In that scenario, humans should be able to understand the reasoning behind such a decision. If that decision impacts people's lives in a significant way, then people should be able to challenge AI decisions and look for accountability when things go wrong. Sometimes, people must be provided an option to interact directly with a human who will ultimately take the final decision.

According to the European Commission's Ethics Guidelines for Trustworthy AI, explainability (or explicability) is one of the key aspects related to the transparency principle: *"Explicability is crucial for building and maintaining users' trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected. Without such information, a decision cannot be duly contested"*. Moreover, *"Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings."* [14]

Is it even possible to achieve such an explainability? If so, how?

[13] Mike Zajko. *Agency to Change the World*. Essay published as a result of the AI & Agency workshop organized by the Summer Institute on AI and Society in 2019. https://aipulse.org/ai-agency/.
[14] European Commission, High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. 2019. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top.

Hey, AI, you own me an explanation

In an article produced by the Berkman Klein Center Working Group on AI Interpretability, researchers raised the following key question: *"Is it possible to create AI systems that provide the same kinds of explanation that are currently expected of humans, in the contexts that are currently expected of humans, under the law?* [15] Though the question poses several technical challenges, their answer was pretty straightforward: yes, it is possible and we can apply almost the same legal framework to AI as we do to humans.

According to the researchers, *"legally-operative explanations"* from AI systems are feasible and can be obtained through:

- Local explanations: focus on specific outputs and explain them based on the inputs that played a key role for that result instead of explaining the whole model. This method implies acceptance of the fact different factors have different impacts for different instances. The researchers provide us with a simple example: *"(...) for one person, payment history may be the reason behind their loan denial, for another, insufficient income. This notion of locality directly maps to the notion of explaining a specific decision, which is the most common case of when explanation is required under the law."*
- Counterfactual faithfulness: another way to explain an AI model is to verify if decisions change when key inputs change. *"For example, if a person was told that their income was the determining factor for their loan denial, and then their income increases, they might reasonably expect that the system would now deem them worthy of getting the loan."*

The researchers also noted humans and AI systems have **different explanatory capabilities** and this intrinsic factor should be taken into consideration when deciding if we should expect from AI the same level of explainability we look to humans. For example, explainable AI should be designed to provide explanations before any intended or unintended consequences, while humans usually prepare to explain themselves only after their wrong decisions are being contested, e.g. a doctor explaining a wrong diagnosis before the court. Another aspect to consider is the amount of information storage required to enable explainable AI systems. While humans may be legally required to retain and preserve information that supports some of their decisions, under certain circumstances it might not be economically plausible to expect the same type of data retention from AI systems due to the economic burden that would represent. Not to mention some information may not be available due to privacy regulations.

Explaining models and building trust

Researchers of the University of Washington proposed Local Interpretable Model-agnostic Explanations (LIME) as an algorithm that achieves faithful explainability through local explanations.[16] They demonstrated LIME explanations are faithful to the model and good indicators of generalization, and its predictions are trustworthy.

---

[15] Doshi-Velez, Finale, and Mason Kortz. 2017. *Accountability of AI Under the Law: The Role of Explanation*. Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper. https://dash.harvard.edu/bitstream/handle/1/34372584/2017-11_aiexplainability-1.pdf.

[16] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 2016. Pages 1135–1144. https://doi.org/10.1145/2939672.2939778.

Another interesting contribution from those researchers was demonstrating how explainability can impact trust in AI models. They showed subjects of experiment results of an AI model that predicted if a picture represented a wolf or a husky. The model was intentionally trained by the researchers to classify as a wolf if the pictures contained snow in the background and as husky if otherwise. The subjects were presented a data set of 10 pictures, 8 of them correctly classified, 1 incorrectly classified as a wolf (a husky with snow in the background) and 1 incorrectly classified as husky (a wolf without snow in the background). The subjects were not told about the model algorithm nor the incorrectly classified pictures, and were asked if they trusted the model. 37% of them responded positively. Then, the researchers presented to the subjects how the model worked, by showing them a picture in which the snow background was highlighted, and then asked them again if they trusted the model. Trust dropped to only 11%.

The researchers not only demonstrated that **explainability impacts trust, but also that trust impacts user decision on which model to use**. When the subjects of the experiments were provided the models' explanations, they were able to identify the best model "A" even when model "B" presented greater accuracy; and maybe more importantly, the subjects were able to improve the model by observing explanations, which means a great step towards human-AI collaboration.

Human-AI collaboration

Healthcare is one of the areas with the greatest potential for AI assistance. When it comes to severe diseases, AI can augment physicians' capability of doing more accurate diagnosis and consequently enabling more effective treatment. However, there might be resistance from adopters of such technology if information about the system's capability, objectives, design and limitations are not transparent and readily available.

A group of Google researchers demonstrated how important transparency is for collaborative work between AI systems and pathologists in clinical decision support systems (CDSS).[17] The researchers wanted to identify what type of information the pathologists needed when presenting a deep neural network (DNN) predictions for prostate cancer diagnosis. They found out ***"a need for a holistic, global view of the AI Assistant and its capabilities, limitations, and biases, preferably presented in terms relatable to day-to-day practices"***. More specifically, the survey participants desired to know:

- Capabilities and limitations: How does the AI system performance compare to human performance? What is the system performance in edge cases? Is there diversity in the training data?
- Functionality: What are the inputs? How is context taken into consideration by the algorithm? How comparable are the AI and the human decision-making schemas?
- Medical point-of-view: Is subjectivity in clinical practice embedded in the model? How comparable is the AI prediction to the real-life practice of looking for a second opinion?

[17] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. *"Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human–AI Collaborative Decision-Making*. Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 104 (November 2019), 24 pages. https://doi.org/10.1145/3359206.

- Design objective: What is the system's objectives and how those compare to current practices? Is the system set to work more independently or collaboratively with a human?
- Other considerations before adoption, e.g. regulatory approval, legal liability, costs.

According to the researchers, access to that type of information could augment human-AI collaboration in many ways and set user's expectations on the collaboration mode. For examples, by knowing the AI system's capabilities and limitations users could pay more attention to areas of AI weakness; in the case of conflicting opinions, users could look for a better understanding of the inputs for a specific prediction, enabling them to determine a final diagnosis; and users could also take into consideration AI biases to upgrade or downgrade disease severity.

Transparency has not only to do with knowing that we are interacting with AI. Factors such as the level of AI systems' autonomy, users' needs and potential harm to humans may determine the extent explainability, traceability and interpretability are required.

Recommendations
To watch: More on how transparency builds trust on AI. https://www.dotmagazine.online/issues/ai-innovation/the-benefits-of-ai/ai-need-for-transparency
For a quick reading: A great step taken by Amsterdam and Helsinki towards algorithm transparency. https://venturebeat.com/2020/09/28/amsterdam-and-helsinki-launch-algorithm-registries-to-bring-transparency-to-public-deployments-of-ai/.
Looking for a deeper dive? A book on how to make machine learning models interpretable. https://christophm.github.io/interpretable-ml-book/intro.html.
Must-have in your toolbox: Though Big Tech companies have had the greatest challenges to address ethical principles in their AI-based services and products, there are some efforts worth the mention:
- Microsoft FATE (https://www.microsoft.com/en-us/research/theme/fate/)
- IBM Explainable AI (https://www.ibm.com/watson/explainable-ai)
- Google Jigsaw (https://jigsaw.google.com/)

6. PRIVACY

Most of us care about **privacy**. If you use the internet regularly, you know you are being constantly tracked, observed, analyzed: once you are online, companies have the chance to know what you search and how you search; what you read, watch and listen, and how much time you spend doing it; what you like and dislike, and your comments on what you like and dislike; the places you go, the food you eat; how fast your heart beats during your regular runs; who you are connected with, and who you are not. Whatever you do while online generates data that tells who you are, what you do and, most importantly to companies, what you might want to buy and consume. Consciously or unconsciously, we give up data about ourselves in exchange for a more convenient life.

AI systems use data to achieve goals. In general, the more personalized the data, the more accurate the model. Given the increasing concern on the pervasive use of personal data, governments around the world have been promulgating privacy laws that restrict the access and the use of personal data by companies, and give the

consumer or user extended rights to their data. The European General Data Protection (GDPR) and the American California Consumer Privacy Act (CCPA) are examples of how regulations have been reshaping the way companies interact with consumers from a data privacy perspective.

Complying with privacy regulation

GDPR has been the main reference for privacy regulation worldwide.[18] It aims to provide data subjects more protection and ownership of their data, and data controllers more accountability on others' data. It is founded on the following people's privacy rights: to be informed where personal data is being collected; to have access to own data; to rectification and erasure of personal data; to restrict or even object to the processing of personal data; data portability; and rights to automated decision making and profiling.

All those rights are relevant to the development and deployment of AI systems. However, the latter is the most important for our present discussion. According to GDPR's article 22, ***"The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."*** That shall not apply under a few exceptions, including subject's explicit consent. The interconnection of AI principles gets evident once again. GDPR not only touches privacy per se, but also addresses human agency.

Meeting privacy rights is a formidable challenge these days. On one hand, data itself drives business models, impacts profitability and is one of the main fuels of innovation. On the other hand, privacy regulations have been getting stricter, turning profiling and data personalization more difficult. In this context, entire data analytics departments of tech ad companies try to find new ways to extract meaningful data from us and, surprisingly, that does not necessarily mean getting our personal data. With the right data science tools, the screen resolution of your notebook and the model of your smartphone are device-related types of attributes that already allow personalization. This technique is called fingerprinting and in case you don't know (I didn´t), it's been there for seven years or so.[19]

My feelings: your data

Instead of mining on hardware data, other companies rather prefer a more straightforward approach to get a clue of our inner desires and potential of consumption: they look at our faces, infer our emotions and return unique ads in real time. What would be the type of ad that could potentialize your consumption when your face indicates anger or fear? Would this advertising practice be legal?

[18] GDPR.EU. *What is GDPR, the EU's new data protection law?* https://gdpr.eu/what-is-gdpr/#:~:text=The%20General%20Data%20Protection%20Regulation,to%20people%20in%20the%20EU.
[19] Brian X. Cheng. *'Fingerprinting' to Track Us Online Is on the Rise. Here's What to Do.* The New Your Times. July 3, 2019. https://www.nytimes.com/2019/07/03/technology/personaltech/fingerprinting-track-devices-what-to-do.html.

In his 2016 article, Andrew McStay assessed facial coding for emotion detection in public spaces and its application to digital advertising.[20] The facial coding system he analyzed relied on hidden cameras that identified happiness, sadness or neutrality on people's facial expressions, which then triggered changes in the ad being shown. While that digital ad system did not retain or use any identifiable feature as input, McStay questioned how UK people would feel knowing their emotions were being tracked for commercial purposes. His survey found out 50% of the people were not ok with any type of emotion detection technology using facial coding, and 33% were ok with emotion detection since the information was anonymized and without the possibility of personalization.

The current GDPR scope does not apply to situations in which a person is not identifiable from data, a code is not attributed to a person or there is no way to single a person out. That means intimate data (as emotions) that do not allow personalization would not be covered by the data protection framework in Europe. Based on this gap, McStay suggests privacy regulations should also consider intimacy on top identity: *"While emotion detection might not make use of information that is personally identifiable and legally private, it certainly makes use of information that is intimate. Information about emotions feels personal because emotional life is core to personhood and while data may not be identifiable, it certainly connects with a fundamental dimension of human experience."*

How privacy impacts AI

There are many techniques and best practices to protect privacy. In a paper on anonymization for privacy preservation[21], researchers of the Madurai Kamaraj University outline two types of privacy preservation solutions:

- Data modification methods basically identify and modify sensitive data. The most direct method is anonymization through transformation or removal of personally identifiable data. Another example is the condensation technique, which reduces data to groups maintaining some statistical information the same.
- Secure Multiparty Computation (SMC): this type of method is usually very expensive and difficult to put into practice. It relates to techniques like probabilistic distortion, cryptography, clustering and hiding associations.

According to the researchers, one should carefully choose which privacy preservation method to apply as each method might impact data and algorithms in significant ways, from model performance to data loss, or even the possibility that sensitive information can be reconstructed.

Right to privacy

The United Nations (UN) recognizes privacy as *"a gateway to the enjoyment of freedom of opinion and expression"*. Its International Covenant on Civil and Political Rights protects individuals against *"arbitrary or*

[20] Andrew McStay. *Empathic media and advertising: Industry, policy, legal and citizen perspectives (the case for intimacy)*. Big Data & Society, July–December 2016: 1–11.
[21] G. Arumugan, V. Jane Varamani Sulekha. *IMR based Anonymization for Privacy Preservation in Data Mining*. Proceedings of The 11th International Knowledge Management in Organizations Conference on The changing face of Knowledge Management Impacting Society.

unlawful interference with his or her privacy, family, home or correspondence".[22] In a context in which anonymization is still poorly deployed and people have little visibility on how companies acquire and use their data, the UN acknowledges the right to privacy is under significant threat.

AI can do amazing things for humankind. AI can help us to improve our health systems, have more efficient businesses, reduce inequality and protect our planet from deadly exploitation. However, in the wrong hands and without proper guidance, AI can deprive people of basic elements that constitute our human condition – like human rights, which will be the subject of the next section.

Recommendations
To watch: Facebook and labeling of sensitive data. https://cacm.acm.org/magazines/2021/1/249457-does-facebook-use-sensitive-data-for-advertising-purposes/fulltext
For a quick reading: A great example of data encryption on an individual level applied to medical studies. https://venturebeat.com/2020/05/14/researchers-apply-privacy-preserving-ai-to-large-scale-genomic-studies/
Looking for a deeper dive? The California Consumer Privacy Act of 2018 (CCPA) has also become another world-class reference on privacy regulation. https://oag.ca.gov/privacy/ccpa
Must-have in your toolbox: arxiv.org/ is a huge archive of scholarly articles in which you'll find many interesting papers on AI and AI Ethics.

7. RESPECT TO HUMAN RIGHTS

I left what I consider the most important aspect of AI Ethics to the end. Respecting human rights while developing and deploying AI systems is not only appropriate and expected, but also an invitation to get back to the basics. It is discouraging to think that our civilization failed so miserably in addressing foundational, intrinsic concepts of the human condition as dignity, freedom, equality, solidarity, citizens' rights and justice while we have achieved great advances in science and technology – including AI. If we continue failing to observe such basic features when developing and deploying AI systems, the benefits generated by this technology might become soundly questionable.

European Commission's Ethics Guidelines for Trustworthy AI ("Guidelines") is once again a key reference for us because it does affirm fundamental rights as an essential aspect for the deployment of AI systems and highlights the importance of bridging the gap between those rights and ethical principles.[23] Dignity is the first right presented by the Guidelines: *"(...) respect for human dignity entails that all people are treated with respect due to them as moral subjects, rather than merely as objects to be sifted, sorted, scored, herded, conditioned or manipulated"*. Automated algorithms have the potential to threaten people's dignity. One example is algorithms that entrap people in poverty by perpetuating low credit scores and consequently limiting those people access to jobs and housing. When applied to the public sector, AI systems can affect decisions on who receives health care,

[22] United Nations. *Promotion and protection of the right to freedom of opinion and expression.* Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. 2018.
[23] European Commission, High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. 2019. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top.

unemployment aid and child support services.[24] It is a paradox: those more vulnerable and in need are being denied the conditions to escape from harsh situations due to algorithms boosted by vast and easily available internet-based data sources.

Another family of rights mentioned by the Guidelines is **equality, non-discrimination and solidarity**: *"In an AI context, equality entails that the system's operations cannot generate unfairly biased outputs"* and must comprehend *"adequate respect for potentially vulnerable persons and groups, such as workers, women, persons with disabilities, ethnic minorities, children, consumers or others at risk of exclusion."* In the previous section on the fairness principle I gave some examples of how bias may play a big role in healthcare, education services and judicial systems. From a human rights perspective, we have seen alarming examples of how AI technologies have the potential of harming people just because of their identity, culture, condition or ethnicity, i.e. their uniqueness. That becomes more concerning when we find Big Tech companies explicitly offering facial recognition technologies that allow users identification and persecution of minorities.[25]

The Guidelines also touches upon other fundamental rights as freedom of the individual, and respect for democracy, justice and the rule of law. I would like to stress another one that has been subject to scrutiny from society: **citizens' rights**, particularly people's right to vote and how social media impacts it. Social media does have a great power to influence elections. In his article in *The Atlantic*, Alexis C. Madrigal highlights several studies on the impact Facebook had in previous US elections. Incredibly, the social media might have influenced the 2012 US elections through bland design tweaks on functionalities as its "I voted" button, which would have increased youth voter participation in the elections and contributed to the Democrats' victory. However, amplification by social media of the already-overwhelming spreading power of fake news has been one of the biggest concerns to democratic elections in the last decade. Studies have shown top-performing fake election news have better engagement than top-performing election news published by major news companies. And that does not have anything to do with which side of the political spectrum we are: *"[But] the point isn't that a Republican beat a Democrat. The point is that the very roots of the electoral system – the news people see, the events they think happened, the information they digest - had been destabilized."* [26]

Human rights-based approach to AI

There is a thorough legal framework for AI described in the *United Nations' Promotion and protection of the right to freedom of opinion and expression*.[27] In addition to references to other AI principles we already covered (e.g. privacy, obligation of non-discrimination and effective remedy), the document outlines other two fundamental rights and how current AI systems can represent a threat to them:

---

[24] Karen Hao. *The coming war on the hidden algorithms that trap people in poverty*. MIT Technology Review. Dec 4, 2020. https://www-technologyreview-com.cdn.ampproject.org/c/s/www.technologyreview.com/2020/12/04/1013068/algorithms-create-a-poverty-trap-lawyers-fight-back/amp/.

[25] Raymond Zhong. *As China Tracked Muslims, Alibaba Showed Customers How They Could, Too*. The New York Times. Dec 16, 2020. https://www.nytimes.com/2020/12/16/technology/alibaba-china-facial-recognition-uighurs.html?smid=em-share.

[26] Alexis C. Madrigal. *What Facebook Did to American Democracy (And why it was so hard to see it coming)*. The Atlantic. October 12, 2017. https://www.theatlantic.com/technology/archive/2017/10/what-facebook-did/542502/.

[27] United Nations. *Promotion and protection of the right to freedom of opinion and expression.* Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. 2018.

- Freedom of opinion: everyone has the right to hold opinions without any restrictions, including the law. Any type of violence to silence or compel people to change their opinion violates UN human rights. What it comes to the advertising industry and its specific application to elections, AI has worked as a new tool for the old problem of content curation. As stated in the aforementioned UN document, *"The dominance of particular modes of AI-assisted curation raises concern about its impact on the capacity of the individual to form and develop opinions"* – a clear reference to the harmful impact of fake news and hate speech on the internet as weapons of mind manipulation, oppression of opinion and persecution of minorities and its particular political interests.
- Freedom of expression: everyone has the right to express opinions, though this right is not as absolute as the right to hold opinions. According to the UN document, restrictions of legality, legitimacy, necessity and proportionality rule the freedom of expression. Therefore, as an example, advocacy of racial or religious hostility on social media is an evident violation of human rights. The UN also raises another key concern: market concentration and excessive economic power on hands of few social media companies may be pernicious to people's access to diverse sources of information and point-of-views: *"In an AI-governed system, the dissemination of information and ideas is governed by opaque forces with priorities that may be at odds with an enabling environment for media diversity and independent voices."*

Based on this legal framework, the UN suggests a human rights-based approach to AI through substantive standards and processes. That approach can be summarized by this overarching statement: *"Companies should consider how to elaborate professional standards for AI engineers, translating human rights responsibilities into guidance for technical design and operation choices."* The UN relates these suggested standards to specific processes which, to some extent, we touched upon over this article: impact assessments, audits, ensuring of user autonomy, notice and consent, and remedy. Also, when putting the AI engineers at the center of the problem, the UN sheds light on the accountability principle and stresses the importance of addressing ethical aspects right at the AI development phase.

Recommendations
To watch: An enlightening discussion on social media, disinformation and democracy.
https://www.thesocialdilemma.com/event/social-media-manipulation-election-integrity/.
For a quick reading: More on big techs dropping the ball with AI technologies and causing potential harm to human rights.
https://www.washingtonpost.com/technology/2020/12/08/huawei-tested-ai-software-that-could-recognize-uighur-minorities-alert-police-report-says/.
Looking for a deeper dive? Check out how lawyers are defending low-income people to fight unfair automated decisions.
https://datasociety.net/wp-content/uploads/2020/09/Poverty-Lawgorithms-20200915.pdf.
Must-have in your toolbox: the Center for the Governance of AI is a great reference for publications and technical reports on AI Ethics and AI public policy. https://www.fhi.ox.ac.uk/govai/#home.

8. CONCLUSION

AI has changed the way we live, perceive and interact with the world. AI has also altered our perspective of what can be the future of society and economy. AI has shown the world its potential to take our civilization to higher levels through more effective education, healthcare and more efficient businesses. Along with the freshness of a

vibrating, tech-driven world, AI has also exposed some flaws and gaps that harm people's lives. Addressing those issues is an emergent call for action; otherwise, more people will suffer from unethically intended or negligent design of AI. Building a better future for humanity depends on building trust between humans and AI systems.

In this article, I covered some key AI ethical principles, explaining their definition, presenting ways they can be put into practice and how their application into real-live situations can improve people's lives or, at least, avoid harmful outputs to society. We saw how bias is encrusted in our social structure, which makes achieving **fairness** in AI quite a challenging problem; we learned how audits and impact assessments can increase **accountability**, and that balanced data sets do not necessarily jeopardize accuracy and performance of AI models; we saw examples of the importance of **human agency**, i.e. having a human in the loop when AI systems make big decisions, and how critical is that AI systems be not only autonomous, but also effective, good agents; that we can expect **transparency** in AI systems to the same extent we expect transparency from human beings from practical and legal perspectives; we acknowledged that though **privacy** is quite regulated, there are still gaps on law interpretation and enforcement; and finally, we reviewed some concepts and ideas around **human rights** and how negligence on AI development and deployment has the potential of threatening dignity, citizens' rights, freedom of opinion and freedom of expression.

There are other principles cited by the extensive literature available that I didn't cover in my article, which doesn't mean they are less important. Some of them, like **respect to rule of law** and **professional responsibility**, seemed too obvious to me, though I do recognize they might not be that obvious for bad actors and negligent regulatory bodies. We also find studies on **safety**, **security** and **technical robustness**, which I felt I didn't have the proper competence to address them, so I defer to more technical, prepared experts working on those areas.

One of the hottest topics in the current discussion on AI Ethics is **how to make AI ethical principles actionable in corporations and public services**. That has been called the "second wave" and it means a genuine and legitimate call to move AI Ethics from theory to practice. However, I do believe the first wave is still there, is big and keeps its significance these days. The field of AI Ethics is quite recent and there are a lot of people who are still taking the first steps towards a more robust awareness of the ethical sides of AI. And I'd say more: many people are still "blind" about the unethical, immoral and harmful ways AI can affect them.

I have been working with Ethics, Compliance and Corporate Investigations for more than a decade now. About ten years ago, I witnessed the gradual and slow adoption of codes of conduct within corporations, including big, multinational companies. It was never a simple discussion with executives how to implement corporate ethics programs, incorporate them into business strategy, make fraud risks assessments and, especially, manage internal misconduct in a responsible way at the various levels of the corporate hierarchy. I see AI Ethics at the same type of stage today: the discussions that increase awareness on AI ethical principles are still very important and will have an educational role for a few years ahead. The first and the second waves of AI Ethics advance together, reaching people, companies and governments wherever AI changes and impacts our lives in a significant way. The first wave always comes first – it brings the foundational knowledge and the necessary awareness to prepare people, corporations and governments for the second wave. For some of us who think AI ethical principles are something that lives in the past, it is important to remember most of the people who are really harmed by AI do not know much about it yet.

While I wrote this article, it was challenging to keep up with news on unethical, irresponsible use of AI. Though frustrating, that confirmed to me there are many opportunities out there waiting to be addressed. People continue being harmed by intentional and unintentional consequences of AI, and I feel they need people with genuine interest in building a better AI for the current and future generations. There are many people and institutions doing great things and I already mentioned some of them in this work. Aware of the risk of missing many of them (I apologize in advance), I want to take this opportunity to recognize other people who also inspired me in a special way. I'm sure they will inspire you as well: Olivia Gambelin (Founder of Ethical Intelligence), Sofia Trejo (Educational Projects Lead at Alianza en Inteligencia Artificial), Adrian Munguia (Founder & Director at AI Mexico), Murat Durmus (CEO & Founder at AISOMA AG), Sandra Wachter (Associate Professor and Senior Research Fellow bei Oxford Internet Institute), Virginia Dignum (Professor of Ethical and Social Artificial Intelligence), Abhishek Gupta (Founder and Principal Researcher, Montreal AI Ethics Institute), Maria Luciana Axente (Responsible AI & AI for Good Lead at PwC UK), Valerie Morignat (Professor of AI Strategy & Design at aivancity), Shea Brown (Founder & CEO at BABL AI Inc.), Ansgar Koene (Global AI Ethics and Regulatory Leader at EY) and John C. Havens (Executive Director at IEEE Global Initiative on AI Ethics).