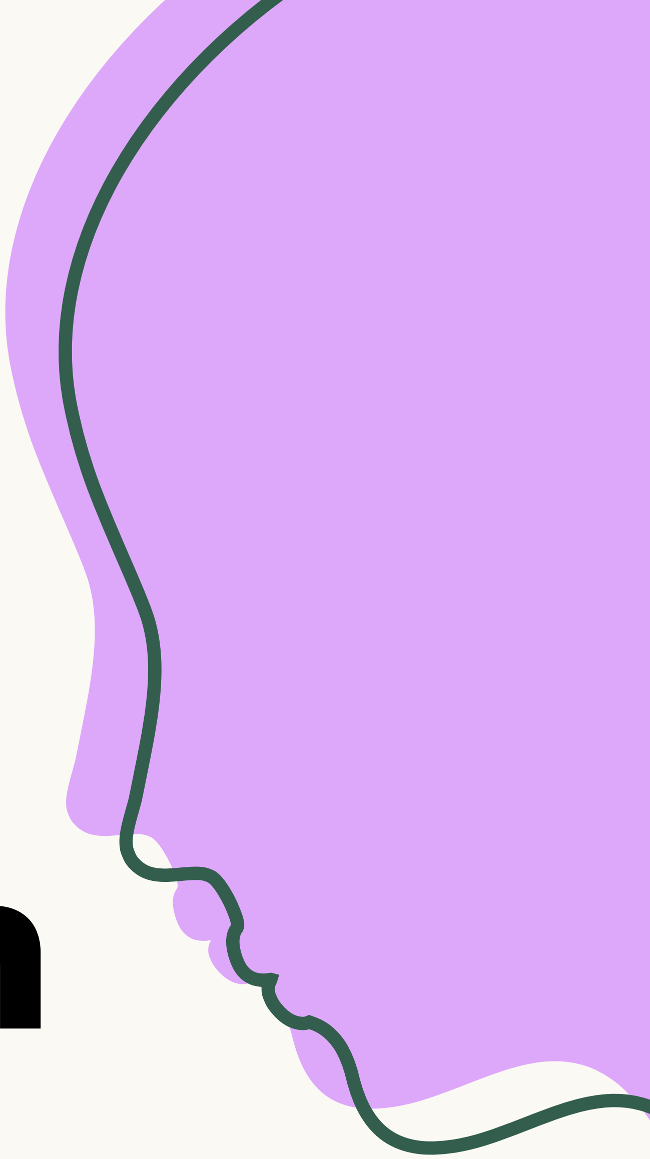
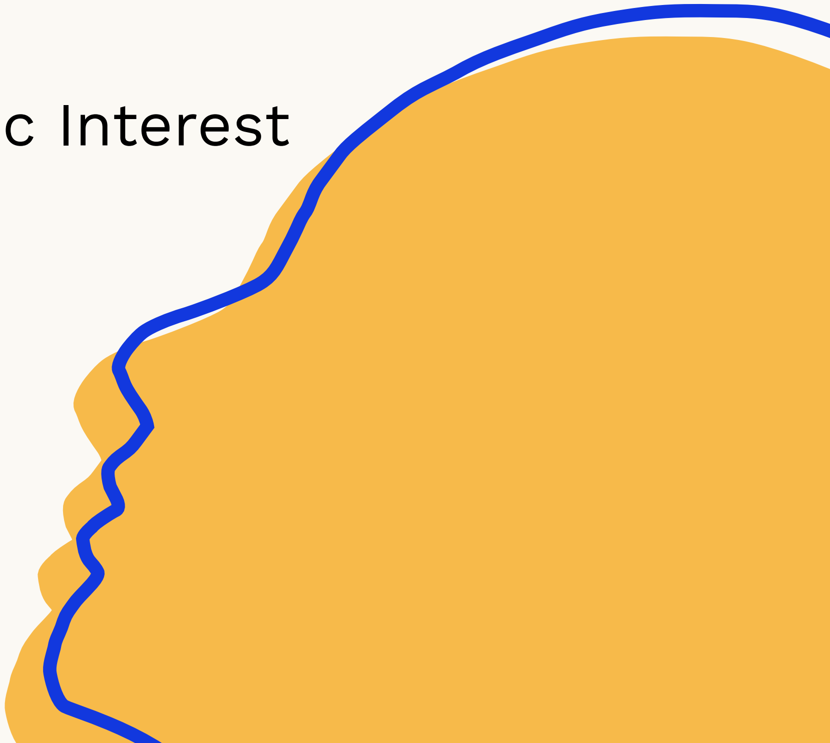


AI and Human Rights

Building a Tech Future
Aligned With the Public Interest

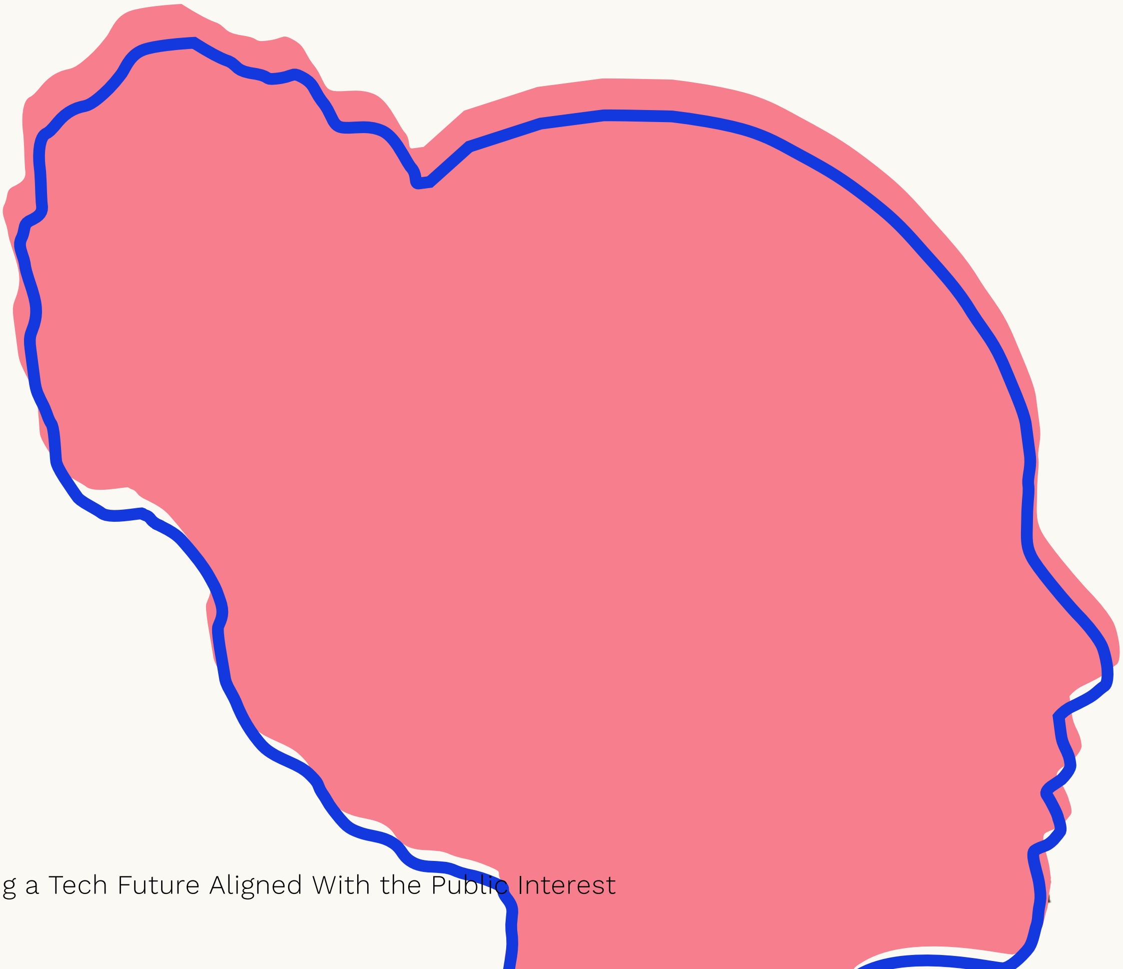


all tech is
human

AI and Human Rights:
Building a Tech Future Aligned With the Public Interest

To read and download the latest version of this report, please visit AIHumanRightsReport.com

27 June 2022
All Tech Is Human
New York, United States

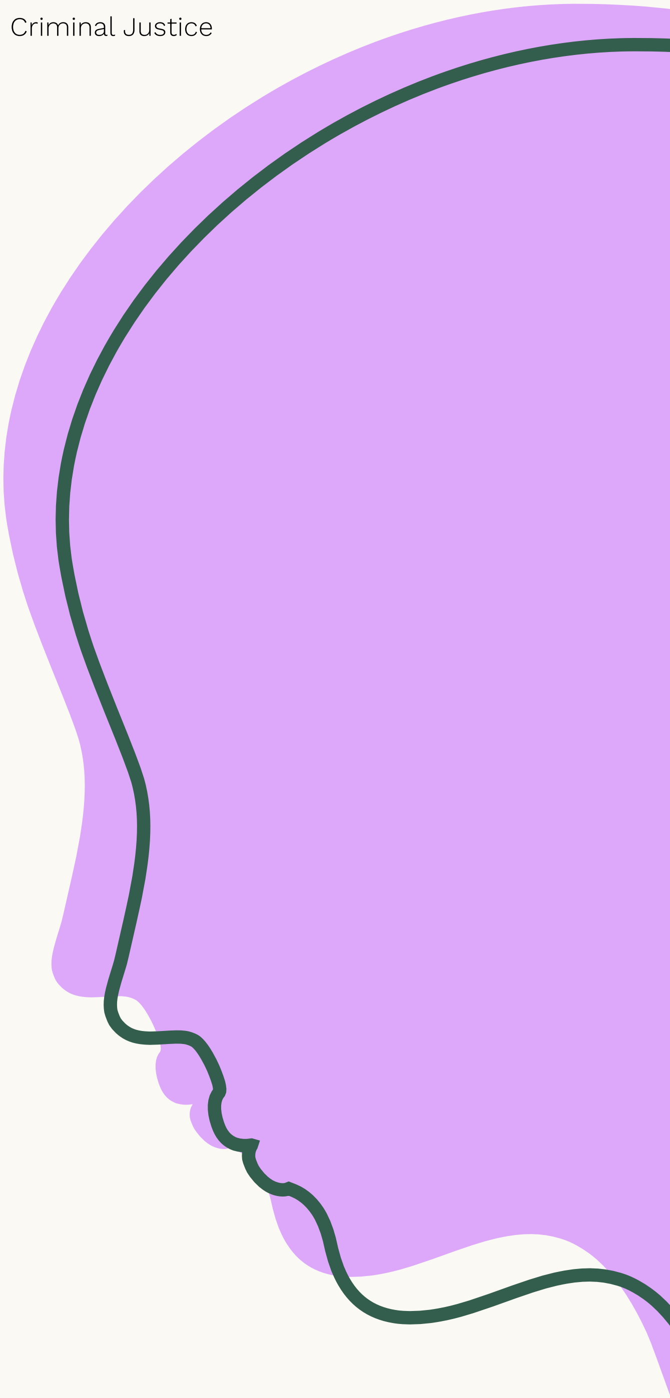


all tech is
human

Table of Contents

7 Key Areas of AI and Human Rights

Profile Interviews



Resources & Organizations

Welcome Letter

By David Ryan Polgar

The future of technology is intertwined with the future of democracy and our human condition. In particular, the rapid advancement of artificial intelligence brings forward thorny questions related to privacy, fairness, and human agency. While it can be difficult to slow down innovation, I deeply believe that there is a considerable opportunity to massively speed up society's ability to grapple with how we consider the social impact of emerging technologies.

If this is your first time engaging with All Tech Is Human, welcome. If you have been with us since our founding in 2018, thank you for the support that has allowed us to grow our activities focused around multidisciplinary education, diversifying the tech pipeline, and uniting a broad range of stakeholders across civil society, government, industry and academia. The mission of our non-profit is to co-create a tech future that is aligned with the public interest. This report aims to help surface the incredible range of people, organizations, and ideas at the intersection of AI and human rights with the goal of promoting knowledge-sharing and collaboration.

We not only welcome, but encourage feedback about this report as our goal is for a highly-participatory design where our work is not only helping inform the Responsible Tech community, but perpetually being informed by the community. After all, you can't align the tech future with the public interest unless you actively understand the needs and values of the public.

By working together we have the power of community, the power of collective intelligence, and the power to change systems.

Let's co-create a better tech future,

David

David Ryan Polgar

Founder & Director of All Tech Is Human

Executive Summary

Artificial Intelligence (AI) and other digital technologies are tools. As such, they can be used to create unprecedented opportunities and advancement, and can also be wielded in ways that oppress and harm. Applied unconsciously, they will continue to recreate existing social norms, biases and power structures. Developed thoughtfully, they can help transform society for the better.

While the breadth of applications for AI are enormous, the ways in which these technologies intersect with human rights are no different from previous technologies. In other words, AI can be leveraged to advance human knowledge and capacity in a range of fields - including, but not limited to, medicine, engineering, energy, finance, law, communication and security. In the process, they can also enable dangerous violations of human rights in areas of digital identity, privacy, healthcare and civil rights, to name a few.

This report outlines key issues and opportunities related to the intersection of AI and Human Rights, across seven key areas - Automated Decision Making Systems and Civil Rights; Data Privacy; Synthetic Media and Information Integrity; Content Moderation; Healthcare; Surveillance Technology, Predictive Technology and Criminal Justice; and Cybersecurity and Autonomous Weapons. These topics are not intended to be comprehensive. Rather, they explore some of the key issues surfaced by our team of 93 co-authors. Each section provides an overview of the topic, and lists key barriers and challenges, next steps and proposed solutions.

Best practices related to AI and Human Rights, such as those developed by the UN, and those that appear as recurring themes in this report, are similar to the best practices we have identified in previous Responsible Technology reports, including our 2021 report on Improving Social Media, and our 2022 report on HX - Human (vs User) Experience. These themes and values include:

- Transparency
- Explainability
- User Notification & Consent
- Oversight & Accountability
- Due Process & Redress
- Privacy by default
- Participant centered
- Conducting impact assessments
- Creating standards, regulation and legislation

The report also includes profile interviews of over 40 community leaders, to provide readers with examples of the people and possibilities in the space. We have also curated a list of over 100 Organizations addressing challenges in Human Rights and AI.

Like previous All Tech is Human collaborative publications, this report is as much about the process as it is about the final outcome. Each section was co-created with the input of close to a hundred participants over the course of three months. The interactions, relationships and collaborations that helped build this report are essential aspects of the growing field of Responsible Technology. We at All Tech is Human are grateful for all of these contributions. We hope that this collection of key issues, key people, and key organizations in AI and Human Rights will help shape the field of Responsible Technology, and advance our cause of aligning tech with the public interest.

Acknowledgments

All Tech Is Human would like to express its sincere gratitude to the 93 collaborators who have dedicated their time, energy and expertise to co-creating this report in a record 3 months. You are the backbone of this publication, and we could not have done it without you! Another special thanks to Mustafa Gökberk Tektek, Sukhi Chuhan, Bintu Alkassoum, and Adamma Izuegbunam Chau for designing the report.

We would also like to thank each of the 43 thought leaders profiled in this report for contributing their time and insights. We are particularly grateful to our Editorial Team for their time and subject matter expertise: - Karina Alexanyan, Renée Cummings, Cate Gwin, Andrew McAdams and Rebecca Wright. And finally, a special thank you to Cate Gwin for producing this report, and shepherding it from inception to completion. We are grateful for, and inspired by, the efforts of people like you in helping us co-create a Responsible Tech Future. A complete list of contributors to this report is found below.

Adam Yasmin
Adamma Izuegbunam Chau
Alexandra Robinson
Alison Safran
Aman Khera,
Andrew McAdams
Angela Garabet
Audrey Gu
Ayca Ariyoruk
Ayesha Abduljali
Ben Grazda
Bintu Alkassoum
Bobby Zipp
Bryanne Gilkinson
Carey Crooks
Cari Miller
Carla Bell
Cate Gwin
Chhavi Chauhan
Chiara Arena
Chidinma Umahi Nwankwo
David Ryan Polgar
Deborah Ehling-Gwin
Duduetsang Mokoele
Elijah Otor
Emily Clay
Eric Wamugu
Farah Bogani
Frank H. Nein, Jr.
Freida Siregar
Gabrielle Lisk
Gökçe Çobansoy Hızel
Gwyn Glasser
Hana Gabrielle Bidon
Helin Yontar
Isabell C. Camillo
Joan Edwards
Joi Podgorny
Karina Alexanyan
Kashyap Rajesh
Katie Mansfield
Keerti Rajagopalan
Kimberly Wright
Lama Mohammed
Lauran Howard
lavina ramkissoon
Lili Siri Spira
Linda Shum
Lisa Hehnke
Lucy McGrath
Lukas Madl
Mallika Dargan
Manojit Nandi
Marilena Velonia-Bellonia
Marisa Dinko
Martina Howard
Melissa Guzman
Mert Sarıkaya
Mustafa Gökberk Tektek
Nayyara Rahman
Neil Davis
Nell Watson
Nipuna Ambanpola
Nyle Hamidi
Ori Freiman
Paul Guinnessy
Phil Kennedy
Pujitha Gullapalli
Rachel Lewis
Rebecca Wright
Rebekah Tweed
Renée Cummings
Rhonda Moore
Roberto Lopez-Davila
Sai Bourothu
Savannah Thais
Scott Trump
Sedona Thomas
Seth Villegas
Shallie The god
Sinan Robillard
Spencer Gurley
Sukhi Chuhan
Thabang Mmese
Tilottama Pal
Tina Lassiter
Tom Thorley
Tomer Poran
Tosin Fatoyinbo
William McSweeney
Yasmin Al-Douri
Zac Lovell
Zoe Viner

About All Tech Is Human

All Tech Is Human is a non-profit committed to co-creating a tech future aligned with the public interest. Based in New York City with a global audience and lens, we have a wide-range of activities focused around three key workstreams: multi-stakeholder convening & community building, multidisciplinary education, and diversifying the traditional tech pipeline with a broad range of backgrounds and lived experiences.This holistic, multi-prong approach allows us to grow and support the overall Responsible Tech ecosystem and movement around community values.

The report you are reading stems from our commitment to multidisciplinary education and community building. Previous working groups and reports released by our organization includes the HX Report: Aligning Our Tech Future With the Human Experience, Improving Social Media: The People, Organizations and Ideas for a Better Tech Future, The Business Case for AI Ethics: Moving From Theory to Action, and our flagship resource, the Responsible Tech Guide, which has a major revision released every September since its inaugural version in 2020. You can see the over 400 individuals we have had as previous interviewees in our reports, speakers at our gatherings, or panelists for our livestreams in this growing list.

If you are interested in getting more involved with the Responsible Tech movement and ecosystem, you have come to the right location. Feel free to join our community Slack group of over 3.1k members from across the globe, get involved in a future working group, check out our Responsible Tech Job Board, participate in our mentorship program or university ambassadors program, subscribe to our newsletter, and attend a future mixer or summit (both online and in person). We also have office hours if you are looking to discover additional ways to get involved in the community, and our organization has volunteer opportunities. You can find all of our projects and links at https://linktr.ee/AllTechIsHuman

Funding for our organization derives from the support of Ford Foundation, the Patrick J. McGovern Foundation, Susan Crown Exchange, Unfinished, and Pivotal Ventures, the investment arm of Melinda French Gates. Our organization also recently started a Responsible Tech Talent Matchmaking Service for companies and organizations looking to find talented individuals related to Responsible AI, Trust and Safety, public interest technology, and more.

Stay in touch at AllTechIsHuman.org

Introduction

Issues related to human rights and Artificial Intelligence (AI) technologies stem from how computational power can impact an individual or group's essential human rights. As the UN Human Rights Office of the High Commissioner notes "digital technologies provide new means to advocate for, defend, and exercise human rights... they are equally used to suppress, limit and violate rights, for instance through surveillance, censorship, online harassment, algorithmic bias and automated decision-making systems. The misuse of digital technologies also disproportionately affects marginalized individuals and groups, leading to inequality and discrimination - both online and offline."

Rapidly developing technologies such as AI pose new challenges to established frameworks like the UN's Universal Declaration of Human Rights (UDHR). There is a need to identify and address gaps between current human rights frameworks and the impacts of AI technologies. For instance, how are human rights operationalized in the development and deployment of AI technologies? How can we mitigate negative human rights impacts, and seek redress for harmful human rights impacts when they do occur? How do we measure and quantify abuses?

AI technologies have an immense impact on people and the environment beyond the end product. From natural resource extraction, to the energy costs of large server farms, to the mental health impacts of content moderation and reliance on cheap labor for data labeling, AI has many hidden impacts on human lives.

Public awareness of the potential harms of AI has grown in recent years. Books like Algorithms of Oppression by Dr. Safiya U. Noble and the documentary Coded Bias featuring Dr. Joy Buolamwini raise awareness of discriminatory algorithms. Civil society organizations have produced documents such as The Toronto Declaration on protecting the right to equality and non-discrimination in machine learning systems. In the United States, the public backlash against the use of facial recognition by the Internal Revenue Service (IRS), resulting in the plan's withdrawal. At the same time, corporations have increased their focus in the development of ethics boards and principles documents, sometimes leading to high-profile internal conflicts. In government legislation, the EU's General Data Protection Regulation (GDPR) has had global effects, serving as a template for data privacy legislation in other nation states. Additionally, the EU is currently drafting the AI Act, a risk-based, regulatory framework that would protect the fundamental rights outlined in the EU Charter of Fundamental Rights.

All Tech is Human created this report on AI and Human Rights to highlight key issues and opportunities at the intersection of these two important topics. Co-authored by close to 100 people, the report covers seven key areas: Automated Decision Making Systems and Civil Rights; Data Privacy; Synthetic Media and Information Integrity; Content Moderation; Healthcare; Surveillance Technology, Predictive Technology and Criminal Justice; and Cybersecurity and Autonomous Weapons. Each section provides an overview, and highlights key barriers and challenges, as well as next steps and proposed solutions.

Systems that leverage AI to help automate decisions are growing in use, many without human oversight. These applications directly affect our lives – in areas from personal finance to home ownership to hiring and job performance, to criminal law, to insurance, and many others. Data Privacy deals with our ability to control the automated collection and use of our personal data. Synthetic Media addresses issues of information integrity and trust in an evolving audio visual ecosystem. Content Moderation touches on the automation of complex decisions regarding freedom of expression, as well as the human rights of the moderators themselves. Healthcare involves how AI can help advance medical decision making - as well as impact data privacy related to the use of digital health apps and tools, algorithmic bias in healthcare, and even the use of AIs to interpret emotional health and wellbeing. Surveillance Technology, Predictive Technology and Criminal Justice describes the myriad of ways in which AI intended to "protect" can impinge on human rights - due to flaws in design and application, and even intentional abuse. Cybersecurity and Autonomous Weapons covers information safety and attacks - including the controversial and chilling developments surrounding the use of AI in automating weapons.

These topics raise many concerns and troubling applications:

- AI used to screen individuals for jobs, houses, college admissions, or loans using biased data that reinforces systemic patterns of discrimination like racism, sexism, ageism, and ableism
- The abuse of image generation and/or language models to promote misinformation for political manipulation or even to harm an individual
- Political authorities and law enforcers using surveillance technology and facial recognition to target specific minority groups, journalists or activists
- Algorithms that foster polarization or radicalization in online platforms, with terrible real-world consequences
- The rise of "Ethics Washing" - companies utilize ethical vocabulary without addressing any of the actual issues present in their organizations
- The increasing AI digital divide, as countries utilizing AI become more advanced while others are left behind
- The lack of diversity among the creators, developers and deployers of AI technologies.

This report also surfaces a series of best practices—recurring themes that touch on ways to ensure that these technologies align with the public interest. Similar to the best practices we have identified in previous Responsible Technology reports, including our 2021 report on Improving Social Media and our 2022 report on HX - Human (vs User) Experience, these overlapping themes and values include:

- Transparency
- Explainability
- User notification & consent
- Oversight & accountability
- Due process & redress
- Privacy by default
- Participant centered
- Conducting impact assessments
- Creating standards, regulation and legislation

The report includes profile interviews of over 40 community leaders to provide readers with examples of the people and possibilities in the space. We have also curated a list of over 100 Organizations addressing challenges in Human Rights and AI. This overview of key issues, key people and key organizations in AI and Human Rights help shape the field of Responsible Technology, and advance our cause of aligning tech with the public interest.

7 Key Areas of AI & Human Rights

#1
Automated Decision-Making Systems and Civil Rights

#2
Data Privacy

#3
Synthetic Media and Information Integrity

#4
Content Moderation

#5
Healthcare

#6
Surveillance Technology, Predictive Technology, and Criminal Justice

#7
Cybersecurity and Autonomous Weapons

#1 Automated Decision-Making Systems and Civil Rights

An Automated decision system (ADS) uses programmatic algorithms to either help a human decision-maker or make decisions autonomously, without human oversight. These algorithms are designed to recognize patterns in historical data and use those patterns to facilitate decisions descriptively and predictively. In an ideal scenario, an ADS facilitates the consistent evaluation of decision criteria, resulting in a more accurate and efficient decision-making process.

Automated decision systems demonstrate varying degrees of risks and rewards. Some of these systems are largely advantageous, like those that monitor our credit card usage to detect fraudulent purchases or our self checkout scanning to detect theft. Others are riskier because they intersect with our fundamental civil rights and freedoms, like those used to determine such high-stakes outcomes as eligibility for a loan, a mortgage, a job, a commuted prison sentence, healthcare, insurance, and welfare services. Both advantageous and risky ADS are increasingly being used by government agencies, corporations, and other organizations.

Unfortunately, careless usage of automated decision systems have been shown to perpetuate or exacerbate existing societal biases, prejudices, and inequalities, delivering outcomes that result in disproportionate, negative impacts on marginalized individuals and groups. For example, predictive policing algorithms and criminal risk assessment scores have reinforced and embedded the hidden biases within historical crime data by turning correlations, like low income and crime, into causal scoring factors. Faulty automated background checks have frozen out renters and prevented them from getting loans. In schools, poorly designed facial recognition software have caused proctoring issues for students of color, and online activity monitoring software may inadvertently out LGBTQ+ students. Targeted advertising has resulted in housing discrimination, as users are not shown certain ads because the algorithm is programmed to maximize clicks and deems certain demographics as less likely to engage. When job searching on Google, male users are shown more high-paying job ads than female users. These are unfortunate, yet real, outcomes of risky ADS.

Biases in society are not new. However, biases and faults in ADS can scale these harms at unprecedented rates. As historian Mar Hicks writes, “technologies are never neutral. This is because the data that powers AI doesn’t exist in a vacuum; it reflects existing social, political, and economic values, including biases.” Organizations such as the National Institute of Science and Technology have identified that biases in ADS largely stem from three sources—data, models (or algorithms), and users. It is not uncommon for data to reflect the inequities and societal biases of the past and present. Studies have also shown that the algorithms or models that developers create can perpetuate unfair results due to either the nature of the model itself, unconscious biases introduced into the models by programmers and developers, or a combination of both. Finally, misuse, disuse, and abuse by end users may also introduce harms for individuals from historically marginalized communities when users lack proper training and governance oversight.

Challenges & Barriers

- **Lack Of Diversity in Tech**
  The tech industry is currently predominantly composed of individuals from similar socio-economic, cultural, geographical, and academic backgrounds, which limits the sharing of diverse perspectives during ADS development cycles to explore potential use cases, flaws, and ramifications.

- **Biased and Unrepresentative Training Data**
  Due to inequalities in access to information and technical services, marginalized populations like unhoused individuals, Native Americans living on reservations, refugees and immigrants, people with disabilities or language barriers, and many others, are less likely to be accurately represented in training datasets.

- **Limited Public Literacy**
  Limited technical literacy in the general public, particularly among historically marginalized individuals and groups, makes it more difficult for impacted individuals and groups to advocate for their own best interests, demand inclusion in datasets, understand when an ADS has erred, and request redress.

- **Limited Governmental Literacy**
  Limited technical literacy among politicians and government officials and employees often results in governmental ADS being deployed without rigorous testing and continuous monitoring for fair and equitable outcomes. When legislators lack technical literacy, it makes it difficult to develop effective regulation for public sector ADS.

- **Exploitative Data Collection and Ads Development**
  The data brokerage market, a lack of oversight and cohesive regulation for data and ADS, and the promise large financial gains have created a culture that often rewards the aggressive collection of personal data and subsequently harmful shortcuts in ADS development and usage.

- **Unexplainable Automated Decision Systems And Outcomes**
  ADS are often referred to as black boxes because it's difficult or impossible to articulate how they arrived at a decision. Although these models may provide improved accuracy according to some metrics, they sacrifice the ability of their users (and even their developers) to understand and explain the model mechanics and the outcomes as they relate to particular individuals and groups. This makes challenging the results nearly impossible for those who have been harmed.

- **Punitive Automated Decision Systems**
  Not all harmful or civil rights violating ADS implementations are due to data biases or other systemic flaws. ADS can just as easily be developed for purposes of targeted control. For example, some ADS technologies target overpoliced low income and minority communities with intentionally punitive predictive policing systems. ADS have made access to services like immigration or asylum seeking more difficult for specific groups, and have been used to surveil and oppress specific minority populations, monitor and suppress protesters, and quash unionization efforts.

Proposed Solutions/Next Steps

- **More Diversity In Tech**
  We must increase the diversity of ADS developers across all disciplines so that stakeholders are accurately and consistently represented in the design and implementation process.

- **Data Accuracy, Reliability, and Robustness**
  Work must be done to create training datasets that are representative of the populations an ADS will be applied to through improved collection or development of synthetic data.

- **Explainability and Redress**
  Monitoring and governance oversight does not end after model development. When an ADS is deployed in sensitive situations, the impacted individuals and groups of these systems are owed an explanation of how the system arrived at its decision. Any ADS used in sensitive or public sector domains should be equipped with post-hoc explainability mechanisms and a mechanism for an individual to request a redress of a negative outcome.

- **ADS Impact Assessments**
  A framework must be adopted for measuring the impact of an ADS across different populations and community types before it is deployed. This will require participation from technical developers, social scientists, and community activists. Impact assessments should focus particularly on historically marginalized populations as they are the most vulnerable to exploitation.

- **Independent Audits**
  Much like publicly traded companies are required to have annual financial audits and pharmaceutical companies are required to submit to trials that are overseen by an independent authority, an ADS should also be subject to specific civil rights laws and regulations. A fully compliant system must be audited for use case ideation, dataset collection, management, data efficacy, model parameters and design choices, security, training procedures, and evaluation metrics. Developers and legislatures can draw inspiration from existing work around transparent dataset and model documentation.

- **Enforceable Local and National Legislation**
  Governments have a duty to ensure all ADS deployments respect civil rights, and should hold developers and operators accountable. Legislators must take a decisively broad and highly participatory approach when developing these regulations by soliciting input from technical researchers and developers, civil rights activists, social scientists, impacted populations and other diverse stakeholders.

- **Evaluation of System Necessity**
  If an ADS cannot be designed such that it meets the appropriate standard for fairness, equality, transparency, and accountability, then there should be a critical conversation across all stakeholders as to whether or not the system should be deployed at all.

#2

Data Privacy

Privacy is essential to human dignity and autonomy. It is subjective, a dynamic concept informed by culture and context. It underpins the rights of humans to be free, to make choices about our bodies, who we are, what we think, and how we live, and how information about ourselves is shared with others. Data privacy, information privacy, and data protection are intertwined. All have to do with an individual's ability to own and control the collection, sharing, and use of one's personal data, and related legal, commercial, social, and political issues.

Unsurprisingly, this amorphous concept can be difficult to code into technology ecosystems, including AI systems. However, even where the tools exist, the practical implementation of privacy by design is far from universal. There are significant financial and power-based incentives for companies and governments to gather, retain, and exploit as much data as possible, despite the significant impact on humans' privacy. The efficacy and enforcement of laws and ethical mandates aren't sufficient to prevent serious harms due to Automated Decision Systems and other AI technologies.

The capacity, complexity, and power of AI systems present novel challenges to data privacy. Constant data collection coupled with AI-based prediction methods enables corporations and governments access to all areas of life.The privacy impact of AI systems is universal across all walks of life, but a greater burden falls on populations with less power and agency—children, women, people of color, people experiencing humanitarian distress, and other historically marginalized and vulnerable populations.

AI systems' negative impacts on data privacy include the following:

- Subjecting people to constant and pervasive surveillance. This can be harmful in itself, and also the data gathered can be used to reveal deeply personal information.
- Lack of informed, meaningful consent from individuals for widespread collection and use of data and metadata — used to train AI systems and to target individuals and groups for algorithmic analysis and manipulation.
- Deliberate lack of transparency, denying individuals clear information and ability to control the usage of AI systems that may cause them harm
- Algorithms using online behavior to tailor content that can impact individuals' rights to make decisions and hold opinions about who they vote for and date. These methods also enable inferences about sensitive individual characteristics such as political orientation and sexual orientation (even if sometimes erroneously).
- Data from online mental health services being repurposed for advertising and profit.
- Corporate and government surveillance of children during online learning.
- Reidentifying people through purportedly de-identified clickstream data.

Determined individuals, academics, regulators, non-profits, standards, groups, and organizations are working hard to create sustainable change and ensure that AI systems reflect the values of a society that protects human privacy. A range of mathematical and computer science techniques have been developed that can reduce the privacy violations of AI systems, including efforts in academia, the public sector, and emerging privacy technology start-ups. Some important techniques are differential privacy, federated learning, and secure multiparty computation. Socio-technical approaches encourage designers and operators to consider human impact, as well as to be more conscious of the people whose data is used to train the systems. NIST's draft AI Risk Management Framework notes privacy as a characteristic that must be considered in deploying AI systems, along with explainability, interpretability, safety, and managing bias.

Many existing laws[1], regulations, and frameworks that protect data privacy are already applicable to most AI systems and their use, and new additional instruments are being developed. Despite variation across jurisdictions, industries, and cultures, there are shared principles that are clear beacons to guide AI systems development. Key shared principles include: using data for fair and lawful reasons, data minimization (collecting only minimum data and deleting data once no longer required), transparency for individuals, accountability, and oversight, limiting use of data to the purposes collected, ensuring security of data, enabling individuals to control their data. However, compliance with existing principles and privacy law remains poor and enforcement has not yet proved a significant barrier to privacy invasion, particularly in "big tech". Some key new and emerging laws address the specific increased risks associated with AI systems, including the EU's Artificial Intelligence Act, as well as the EU's Digital Markets Act (DMA) and Digital Services Act (DSA). Recently, a draft federal US privacy bill with bi-partisan support has been issued with specific requirements to assess privacy impact of AI Systems. Furthermore, enforcement of some existing laws is becoming more common — e.g. the FTC demanding destruction of AI algorithms and the sensitive children's data on which they were trained, as well as images misused for facial recognition. multi-million dollar fines and deletion orders by multiple regulators against Clearview AI (and here and here),and fines against the use of AI systems in the gig economy in Italy and the Netherlands. In the US, the Computer Fraud and Abuse Act

Challenges & Barriers

- Anti-privacy practices are profitable, and often the incentive to overlook privacy is strong and systemic.
- Ineffective, inconsistent laws and regulations and enforcement.
- Lack of fairness, transparency, and explainability of AI systems inputs, model behavior, and outcomes; lack of transparency and agency over data use, "Black box" AI.
- AI training data sets are not representative of populations impacted and embed existing bias.
- Lack of meaningful individual access / control, Power asymmetry between individuals and corporations / entities reduces individual agency and embeds bias.
- Lower digital literacy, high burden & cost of accessing privacy protections which unfairly impacts populations with less power and agency.
- Poor privacy and data governance literacy and education, particularly in computer science curricula.

[1] e.g. US Constitution, Europe's GDPR, California's CCPA, HIPAA, China's PIPL, FTC requirements (which have recently been very powerful in demanding destruction of algorithms and other laws) - computer, wiretap, healthcare and common laws mandate requirements that apply to AI systems.

Proposed Solutions/Next Steps

- Take a measured, considered approach to implementing AI Systems. Be aware that AI systems will not be appropriate in some (or many) cases.
- Support Privacy by Default and Privacy byDesign as the standard. A thorough approach is essential. This requires evaluation, implementation and monitoring throughout the development and operation of AI Systems. Using strong data management practices is key. Four actions to mitigate risk include: delete data more regularly and effectively; minimize use and generation of personal data in training algorithms as much as possible; create clear, consistent documentation about AI Systems and component parts, and address context and user-level issues - e.g. children's design practices from UK/Ireland**,** Australia, and California (upcoming).
- Support the efforts to improve transparency. Continue innovation and research to improve privacy-preserving tools and frameworks, e.g. Plot4AI *Privacy Library of Threats*. Implement common sense and effective Solutions such as those espoused by IEEE or emerging frameworks such as: NIST AI RMF, Transparency - Lingua Franca: Principles, IEEE 7001:2021 Standard for Transparency of Autonomous Systems, Australia's AI Ethics Framework and UK's Standard for Algorithmic Transparency.
- Support development and use common privacy iconography and tools to enable consistency (e.g. Privacy Patterns).
- Clear, consistent and actionable regulation would benefit AI Systems operators as well as humans whose privacy may be impacted.
- Improve data literacy for all, ensure simplicity and consistency of concepts. Integrate data privacy into core school and college curricula, and educate existing engineering and AI development teams on privacy risks and technologies. Provide technology ethics education with cross-cultural focus. Ensure targeted support for groups as required - e.g., Cyber Collective is a primarily POC women-run nonprofit that empowers people to think critically about their relationship with technology and inspire a more socially responsible future concerning data privacy and cybersecurity.

#3 Synthetic Media and Information Integrity

Synthetic media are digital media that are manipulated and modified using digital technologies - often AI - sometimes without disclosure and with the purpose of changing the original meaning and misleading people. “Deepfakes” are a type of synthetic media, where the “deep” refers to “deep learning.” Deepfakes leverage AI to create realistic simulations of someone’s face, voice or actions, so that the person in the video can seems to say or do something they didn’t, or an event can appear to happen that never occurred.

Deepfake technologies can have beneficial applications. They can be used in movies, fashion, entertainment, video games, art, education (imagine having a historical figure as your teacher), and medicine (such as showing the reconstruction of a face after surgery). They have been used to protect a person's privacy, or even preserve historical memory. When their inauthentic nature is not disclosed, however, deepfakes can and have been used to mislead, manipulate public opinion, provoke gendered harassment, and generally undermine confidence in evidence or other verification processes.

As deepfakes are getting easier to make, they raise concerns of the potential for these technologies to negatively impact society, business and politics. Deepfakes can have major impacts on human rights - especially privacy - as well as on information integrity, journalism and trust in institutions. Their convergence with other technologies (such as video conferencing or live-streaming) leads to new possibilities and challenges. As the European Parliament concluded in a 2021 Report on deepfakes: "The increased likelihood of deepfakes forces society to adopt a higher level of distrust towards all audio-graphic information. Individuals and institutions will need to develop new skills and procedures to construct a trustworthy image of reality."

Challenges & Barriers

Some of the dangers and threats associated with deepfakes include::

- Manipulating public opinion and public discourse for political gain.
- Causing electoral disruption or damage to international diplomatic relations.
- Impersonating a public figure causing reputation or brand damages.
- Undermining evidence in the legal system.
- Commiting a range of frauds (e.g. insurance, financial, and electoral frauds), such as passing the 'liveness tests' used by banks for verification.
- Establishing the formation of a new kind of identity theft, creating a totally untraceable identity, falsifying identity or fooling security cameras with facial recognition, recreating the identity of a deceased individual, opening questions about consent.
- Defamation, intimidation, and extortion, including sextortion.
- Deepening gender gaps, as almost all of the sexual deepfakes are of women who did not consent to their images being used.

Regulators and policy makers face complicated decisions on issues such as:

- Which deepfake practices should be illegal?
- Should deepfakes be identified, monitored, moderated and/or labeled - and how?
- What kind of international collaboration is needed?
- How should questions of responsibility and blame be dealt with? Who along the creation pipeline should be liable? How can regulators address questions of diffused responsibility?

Proposed Solutions/Next Steps

- Education for the general public, policy makers and technologists alike will be necessary to help promote best practices in identifying and mitigating the harms of deepfakes.
- As deepfakes proliferate, so do deepfake detection tools. Research at Partnership on AI has concluded that, “while imperfect and susceptible to adversarial dynamics, with responsible deployment and adequate training and support for users, detection tools can contribute to the realization of a healthier online information ecosystem that supports truthful claims and the certification of current events.”
- Partnership on AI’s focus group on AI & Media Integrity has published 12 design principles for labeling manipulated media as well as recommendations for the automated categorization of manipulated media. The group is currently working to develop a Synthetic Media Code of Conduct.
- The DeepTrust Alliance, produced a report outlining the threats, costs, and potential interventions surrounding deepfakes. Their aim is to build technology standards, best practices and collaboration tools to power the fight against disinformation.
- The European Parliament produced a policy report on tackling deepfakes, with in depth analysis of varying factors, and a long list of policy recommendations along five key dimensions — 1. Technology, 2. Creation, 3. Circulation, 4. Target, 5. Audience. They conclude that a combination of measures will likely be necessary to limit the risks of deepfakes, while harnessing their potential.

#4 Content Moderation

Content moderation is understood as “the organized practice of screening user-generated content (UGC) posted to Internet sites, social media and other online outlets, in order to determine the appropriateness of the content for a given site, locality, or jurisdiction.”

As social media companies navigate the challenges of complying with speech-related legal frameworks while also trying to promote safety, positive user experiences, and free expression, they are increasingly automating the process of content moderation–using AI and machine-learning tools to help curate, organize, filter, and classify the information we see online. While AI and ML can enable more rapid decision making, they also lack transparency, with design flaws and limitations that can serve to reproduce existing biases and exacerbate harms.

In the context of human rights, there are three aspects of automated content moderation to consider:

1. the organizations and actors utilizing AI for content moderation
2. the criteria by which content may be judged for automated moderation or removal
3. the impacts of automated content moderation on the human rights of the creators and consumers of social media, as well as on the content moderators themselves

1. Organizations and Actors

The process of content moderation using AI is usually performed by private companies, such as tech companies and content publishers, moderating content on their platforms. Government agencies and regulators have utilized AI to censor the content that can be posted and accessed in their jurisdiction.

2. Criteria for Moderation

Content is generally moderated with concerns for the following qualities:

- **Authenticity.** This includes identifying instances, where an author or creator misrepresents who they are or what they understand to be factual information related to a particular issue. Examples include impersonation, misinformation and disinformation.
- **Safety.** Some content may be deemed to risk harming individuals or groups or to threaten national security or public order. Examples include bullying, doxxing, threats of harm, content that incites violence, hate speech, and harassment.
- **Sensitive content.** This includes content that may be offensive or upsetting to some. It may include vulgar language, violent or graphic content, nudity, and sexually-explicit material; content around suicide, self-harm, and disordered eating; minor content (like sexual exploitation or grooming behavior); and dangerous acts and challenges.
- **Illegal activity and regulated substances.** This covers content that is explicitly illegal or restricted by law, such as drug dealing, fraud, scams, gambling, or piracy. (Note, laws vary in different regions and territories.ome laws in one country may violate what are basic human rights in another country.)
- **General content and format.** Many platforms have additional content moderation policies that serve to control for quality and user interest. For example, Tiktok's For Your Feed algorithm does not promote "unoriginal, low-quality, and QR code content". Many also moderate other **content deemed unwanted** by the user, advertisers, or the platform—like spam, or external links.

3. Impacts of content moderation human rights

Automated content moderation decisions face similar challenges to human ones. These decisions involve not only removing or downgrading the visibility of content, but also leaving content available.. Each of these can violate an individual or group's human rights in complex ways. At the simplest level, content removal can impact the individual's right to freedom of expression, or impact the rights of historically vulnerable groups. Conversely, deciding to leave potentially damaging content up online could incite violence or lead to other harms. At the same time, automated AI content moderation decisions can have the negative consequence of removing valuable evidence of human rights violations. Users have called for platforms to build the procedures that would instead archive and share this evidence.

Challenges & Barriers

- Key issues with automated, AI and ML based moderation tools are a lack of accuracy, reliability, transparency and accountability. They also encode creator and dataset bias.
- Shifting legal regulations around what is moderated and what speech (or other medium) is permitted on platforms, and how that speech may harm communities or create risk at scale also present challenges. Regulations often lack understanding of how speech or platforms operate, as well as how automated decisions are made.
- A shadowy network of third-party contractors does the majority of moderation and there is virtually no oversight over how effectively they do their job and how fairly they treat their employees.
- Content moderators themselves lack training and support. Unattainable quotas force split-second judgment on content, often without proper context. More employees are needed, but those employees must have proper training, fair pay, and psychological support to do their job.
- AI decisions often lack necessary contextual understandings of human speech. This challenge is exacerbated by the fact that languages and cultures are continuously evolving.
- There is a lack of investment in resources for marginalized languages. For example, Facebook's investment in English language content is 87% of its entire safety budget, and there is currently no means for communities in other countries to argue for more investment.

Proposed Solutions/Next Steps

- More fairness, accountability & transparency around the processes for the development and use of AI in content moderation is needed. This includes transparency around content moderation methodologies and their impact. Transparency could be augmented by audits, public reports to regulators, investment in industry watchdogs, etc. Expansion of feedback mechanisms could enable users to address errors and biases, and promote accountability.
- Policy makers, technologists and citizens need better education on the limitations of automated moderation tools.
- More agency and control to Individuals (users) to decide the level of content exposure they would like in digital spaces.
- Automated content moderation tools should supplement and not supplant human decision majors. Human content moderators need improved training and mental health support.
- A catalog of resources related to automated content moderation would be useful. These could include:
  - Reports on limitations and key issues
  - Aggregation of existing guidelines
  - List of groups and other resources
  - Cautionary examples and Case Studies
- Cultural and linguistic representation should be more equitable across a platform's user base. Transparency in how language investment decisions are made in the area of automated content moderation would advance this, as would providing an avenue for advocacy.
- A human rights-based approach implies greater attention should be paid to impacts of automated content moderation decisions on vulnerable or marginalized groups.

#5

Healthcare

In this report, we define healthcare as the physical, emotional and mental well-being provided by trained and licensed professionals. The Universal Declaration of Human Rights (UDHR) claims that healthcare is a fundamental human right, stating “everyone has the right to a standard of living adequate for the health and well-being of himself and of his family, including food, clothing, housing and medical care...” The World Health Organization (WHO) clarifies that, “health is...not merely the absence of disease,” but also a balanced state of mind and body.

AI and related technologies are intertwined in our everyday lives and our healthcare is no exception. AI and ML have great potential for assisting clinical decision-making and revolutionizing the field of healthcare. These technologies help analyze and act on medical data, improve diagnostics and predict outcomes, review and analyze medical literature, interpret medical images, and help provide virtual patient care, including telehealth and mental health apps, among others. Since the COVID-19 pandemic in particular, the use of virtual technology as a tool to deliver care has substantially increased and may be one of the most enduring transformations arising from the pandemic. Expanded remote access to clinical healthcare services has already had a tremendous effect on helping certain populations with overcoming access challenges. Beyond accessibility, key issues at the intersection of AI, healthcare and human rights touch on data privacy, systemic racism, mental health and even emotional computing. Solutions involve fostering transparency and explainability, accountability and oversight, equity, inclusion and user consent and control.

Challenges & Barriers

- **Data Privacy in Digital Healthcare.**
  Healthcare data tends to be sensitive and highly regulated. In the United States, for instance, the Health Insurance Portability and Accountability Act (HIPAA) is a federal law that required the creation of national standards to protect sensitive patient health information from being disclosed without the patient's consent or knowledge.

  However, mobile apps, including mental health apps, face additional issues. Between the user and their app lies a layer of agents - the mobile phone manufacturer (Apple, Google, etc.), the app maker, and a variety of thi rd party service providers. In a 2021 study on mobile health and privacy, for instance, researchers found that as much as 87% of data collection practices were carried out by third party services. There are a number of industry gray areas and large gaps between what HIPAA covers and the types of user data these companies are able to collect and share, and the types of notifications they are required to provide.

  Health-tech sectors affected by digital technology, (i.e. digital care, period tracking apps, Fitbits, etc.) all have differences in their security settings, types of data being collected, and opt-in features. This often makes it difficult to regulate across, and even within industries, especially in a world where digital products are progressing faster than the regulation. In addition, privacy policies across all industries are notorious for being difficult to understand - encouraging users to accept terms rather than read the fine print.

- **Algorithmic Bias in Healthcare.**Technologies are not “race neutral” - rather, they have been shown to reproduce and propagate existing societal biases. Researchers and medical health professionals are raising awareness of the dangerous ways in which health algorithms can encode and reinforce racial health inequities. Bias in big data and AI in healthcare is a result of unconscious and conscious bias, data bias or inaccuracy in the biomedical datasets, and bias in the algorithms themselves. Specifically, if AI training data does not accurately represent a population, the algorithms will be less accurate for misrepresented groups, which can lead to misdiagnosis, lack of generalization, and even fatal outcomes. When such biases are embedded in AI systems they can lead to advanced disparity in underrepresented communities such as Black, Latinx, Women, LGBTQIA+.

- **Algorithms, Healthcare and MIsinformation.** During the Covid-19 pandemic the public increasingly turned to social media and other online information sources for guidance on Covid-19 precautions, treatments and vaccines, as well as for mental health support. Augmented by algorithms that encourage engagement over fact checking, health misinformation has emerged as a leading issue worldwide, with the WHO producing a report on the misinformation “infodemic” surrounding Covid-19 in particular. The proliferation of mental health “experts” on social media, offering questionable guidance, is also a major issue.

- **Emotional AI and Affective Computing in Healthcare.** Another recent trend that bears further scrutiny is the application of AI technologies to learn and recognize human emotions. “Emotional AI,” also known as “affective computing,” leverages computer vision, natural language processing (NLP) and sensors to measure and evaluate mood and emotional states. As research at Partnership on AI notes, this application of AI has the potential to help individuals better understand and control their own emotional and affective states, with enormous potential for good. At the same time, by automating the ability to read or impact others emotions, this technology also has substantial implications for economic and political power, as well as human rights.

  Emotional AI is already being used to enhance common business functions in many industries such as in advertising, marketing, and even human resources. The field is increasingly gaining traction in the healthcare industry and mental health space in particular. The difficulty is that human emotions are notoriously complex and challenging to gauge. Unsurprisingly, studies have shown that AI-based emotion recognition systems are known to be unreliable predictors of human emotion. This is because these systems rely on standard taxonomies to create a dictionary of responses that point to an assumed universality of human feeling – often disregarding cultural differences in how emotions are expressed and de-coded. The risk of deploying emotion-based systems in health care cannot be understated, especially when considering the pervasiveness of a closely related ill: entrenched racism in health and health-related research.

Proposed Solutions & Next Steps

- There must be an expectation and enforcement of **transparency** when AI is used in a healthcare context, and in any context with health implications on users.
- **Participatory Science & Equity Centered Inclusion**. Participant-centered development of AI algorithms should be standard best practice. AI-enabled healthcare Solutions should be imagined, designed, delivered, and continuously improved **in collaboration with those who will be impacted** by the proposed technology. This should include people who may not be expected to be users of the technology, but who may experience its impacts. Strive for representation across all dimensions of human diversity, and adopt an equity-centered approach.
- Similarly, patients, health care practitioners, and others impacted by an AI decision should have easy access to clear **explanations** of how the algorithms were developed, what variables were incorporated and why, what data from the user is collected to feed the algorithm, and how that data is protected (for example: de identification, encryption). Nutritional labels for artificial intelligence systems is one mechanism to improve transparency.
- Ensure meaningfully engaged **interdisciplinary expertise** throughout the design and development of healthcare Solutions, particularly at the early stages of discovery and exploration. Lean on sociologists, psychologists, policy makers, healthcare practitioners, historians and others to think critically about *whether* a particular product should be built, and if so, *how* best to do it to ensure the greatest benefit is captured while preventing unintended consequences.
- Enable **users to control data collection** - so individuals can selectively provide their **informed consent** and **opt into** sharing specific data with a company. People should be empowered to **remove** collected data of their choice, whenever they want.
- Healthcare services that are prescribed and/or delivered via AI require **independent oversight** by a public body to ensure equitable access, quality of care, and health as a human right. This includes automated and/or human augmented diagnostic services.
- Future research needs to focus on developing standards for AI in healthcare that enable transparency and data sharing, while at the same time preserving patients' privacy.
- Education in how AI development and deployment can perpetuate bias, as well as explorations on how to develop AI in healthcare technologies in a conscious manner, mitigating harms to the greatest degree.

#6
Surveillance Technology, Predictive Technology, and Criminal Justice

From social media platforms predicting which posts and advertisements users will engage with to smartphones utilizing facial recognition to allow users to access their devices, surveillant and predictive technologies are embedded within our everyday lives. They include CCTVs and cameras on public buildings and algorithms that monitor our online behavior, as well tools that track our interactions with personal devices, smart speakers, and other Internet of Things (IoT) technologies. On a less mundane level, surveillance technologies are also used for predictive policing and in criminal justice decision making. The stated purpose of such systems is to promote public safety, however many civil society organizations and other groups argue that these systems negatively impact privacy and human rights.

Core issues with surveillance tech involve data privacy, consent, and accountability. Key questions are: Whose data is being collected and what is that person or group's level of awareness, consent, and control? Why is the data being collected and what the data is being used for? What happens when the decision made by an AI system is wrong? Designers and users are encouraged to ask if this AI surveillance tool is actually necessary—with "privacy by default" as the ideal standard.

As Harvard's Carr Center explains, "From a practical perspective, technology can help move the human rights agenda forward. For instance, the use of satellite data can monitor the flow of displaced people; artificial intelligence can assist with image recognition to gather data on rights abuses; and the use of forensic technology can reconstruct crime scenes and hold perpetrators accountable. Yet for the multitude of areas in which emerging technologies advance the human rights agenda, technological developments have equal capacity to undermine efforts. From authoritarian states monitoring political dissidents by way of surveillance technologies, to the phenomenon of "deepfakes" destabilizing the democratic public sphere, ethical and policy-oriented implications must be taken into consideration with the development of technological innovations."

The flaws of applying AI and machine learning for criminal justice decision making, including perpetuating long-standing racial biases in the carceral system, are becoming better understood. More recently, human rights watchers have noted the impacts of increased digital public health surveillance as a response to the Covid-19 pandemic. The intersection of surveillance technology, immigrant policing, and refugee rights is another important area of concern. Some of the most egregious violations of human rights involve China's deployment of facial recognition surveillance software to monitor and control Uyghur Muslims. Those most negatively impacted by surveillance tech are usually those with the least power and autonomy.

Challenges & Barriers

- Algorithms used in surveillance tools are created without transparency and are often built from structural inequality that could continue to perpetuate harm, racism, misogynoir, xenophobia, and more.
- New surveillance technologies—such as those used in education, policing, health care, and the workplace—disproportionately harm disabled people.
- Navigating the complex trade offs involving issues of security, safety, privacy and human rights and creating a consensus on the ethical considerations of using surveillance technology is a key challenge—especially when there is a tendency to over rely on technology to solve social problems.

Proposed Solutions & Next Steps

- Implement privacy by design and privacy by default.
- Support work towards equitable and accountable AI, by organizations such as Algorithmic Justice League, to check companies' use of AI tools.
- Support work that fosters digital inclusion and digital equity, empowering those whose data is being used to have autonomy over that data.
- Co-create systems that center the most vulnerable groups for decision making at each stage of design.
- Promote human rights data literacy for governments, policy makers, and technology companies to ensure better data management and collection practices.

#7

Cybersecurity and Autonomous Weapons

Cybersecurity addresses the protection of online information, as used by individuals, the private sector, civil society, companies, organizations, institutions and governments, with both civilian and military implications. Cybersecurity today is struggling to keep up with cyber attacks that are increasing in sophistication and scale. Cyber threats range from relatively innocuous spam, to more dangerous computer viruses, data breaches, identity theft, cybercrime and denial of service attacks, to major threats to infrastructure and national security. Various countries such as the U.S., Japan, Kenya, Russia and European Union (EU) member states, have declared cyberattacks and cyberwarfare as a national security threat. The use of AI in cyberattacks can make them even more rapid and dangerous. AI-driven malware targeting critical infrastructure, such as power grids in Iran in 2010 and Ukraine in 2016, can cause serious physical damage, put lives at risk, and cut off essential resources.

The term "cybersecurity" can also be code for censorship, prompting growing concern around the potentially negative impact on human rights that overarching and broad cybersecurity laws may have - impinging on freedom of expression, freedom of speech, the right to privacy, freedom of opinion, and freedom of association online. For example, Vietnam's cybersecurity law in 2019 mandated internet companies to remove content that opposed the state and prohibit users from spreading anti-government information online. A report by GP Digital notes that cybersecurity surveillance within the context of international laws and human rights is possible, as long as "necessary and proportionate principles" are applied, including proper public oversight, due process, and a system for user notification.

Autonomous weapons systems (AWS) are another, far more dangerous and controversial, application of AI for security, defense - and attack. The exact definition of what an autonomous weapon is can vary, depending on how key elements of autonomy, human control, capabilities and purpose of use are understood. International Committee of the Red Cross' (ICRC) asserts that an autonomous weapon system must be used in accordance with International Humanitarian Law. rights law, including key rules of distinction (between civilian and military agents), proportionality and precautions in attack. Proponents of autonomous weapons believe that they will make war more accurate, at a reduced cost to human lives. Opponents believe autonomous weapons are profoundly dangerous and incapable of complying with international human rights. They argue that "technologies that apply force without real human control provide a fundamental test for our relationship with AI and new technologies across all areas of society. Rejecting digital dehumanization and ensuring meaningful human control over the use of force are key steps to building a more empowering relationship with technology for all people now and in the future.".

Challenges & Barriers

- Ransomware attacks on critical public infrastructure like hospitals, water treatment facilities, and gas pipelines could lead to dire international and national security, economic, and safety consequences – including death.
- Government hacking of journalists, activists, peaceful protestors, and human rights defenders infringes on freedom of expression and the right to privacy, while also targeting them for surveillance.
- Cybersecurity companies, like the NSO Group who built Pegasus spyware, are non-state actors that can support repressive governments by developing, marketing, and selling products contributing to human rights violations.
- Data breaches and doxxing of personal information, photos, or videos put women at risk of online and offline gender and sexual-based violence, such as revenge porn, identity theft, stalking, harassment, abuse, and rape.
- Autonomous weapons pose a host of issues, including, and not limited to, lack of human judgment and understanding, lack of accountability and lack of explainability. The existence of autonomous weapons lowers the threshold to war, can promote a destabilizing arms race, and exacerbates digital dehumanization - treating human beings as data points and perpetuating algorithmic bias.
- Autonomous weapons exacerbate a litany of social ills, such as gender and sexual-based violence, domination, dehumanization and marginalization and racism.
- Autonomous weapons could be used in other circumstances outside of armed conflict, such as in border control and policing. They could fall into the hands of insurgent groups and terrorists or be be used to suppress protest or to prop up regimes.

- Machines and systems are not immune to hacking. Software, target profiles, sensors, and other related components of autonomous weapons systems could be hacked or attacked by a malicious state or non-state actors; for example, changing a weapon's location on GPS. In addition, autonomous weapons can make errors, and hit the wrong targets.
- The technologies in "high tech weapons" and fully autonomous lethal weapons overlap, creating a slippery slope and challenging regulation.

Proposed Solutions & Next Steps

- Promoting international cooperation on policy, regulation and standards around the development - or restriction- of autonomous weapons systems. Stop Killer Robots, a global coalition of 180+ organizations that work to ensure human control in the use of force, as well as groups such as Human Right Watch and Harvard Law School's International Human Right Clinic recommend that states adopt laws, policies and other internationally legally binding agreements that prohibit the development, production and use of fully autonomous weapons.
- At the very least, the general public should be working towards education and awareness that helps ensure meaningful human control of weapons - ensuring that there is a human in charge of understanding the technologies we use, understanding where we are using them, and being fully engaged with the consequences of our actions.
- Raising awareness among technology companies, tech workers, scientists, academics, and others involved in developing artificial intelligence or robotics on how their work can contribute to the development of fully autonomous weapons. Employees at Google and Microsoft have already protested against any such involvement.
- Council on Foreign Relations argues that understanding how systemic racism influences cybersecurity is integral to protecting the American people, deterring adversaries, and defending American businesses.
- Civic Tech Field Guide has produced a useful list of digital security and privacy resources.
- Building community and facilitating dialogue around alternatives of better, more humane and ethical defense and security measures. This includes engagement with civil society to ensure a wide range of voices and perspectives are heard. An intersectional feminist, anti-racist, decolonization lens would balance out traditional approaches to the development, regulation and use of these technologies.

Profile Interviews

“

Institutions that use AI have an affirmative obligation to make their AI systems as transparent and readily-understood as their analog predecessors. Anything less will structurally disempower the very people that AI vendors claim to help.

”

Albert Fox Cahn

Executive Director

S.T.O.P. - the Surveillance Technology Oversight Project

What are key factors in the way AI is developed/deployed that have the most impact on protecting human rights?

All too often, AI has been developed and wielded by the powerful against those they seek to better control. Employers use AI to reduce employees to crude productivity metrics, Companies use AI to turn consumers into forecasting models, and the government uses AI to more efficiently jail and deport members of our communities. Even as AI is used in the name of protecting and empowering the public, it frequently is doing just the reverse.

How can we strengthen interdisciplinary engagement in technology development, deployment, governance, regulation and oversight?

We have to stop treating AI as just being a privileged domain for the expert class. Members of the public don't need to understand the details of an AI model or the higher-level math that makes it possible to have an indispensable perspective on the impact and unintended consequences of these systems. Institutions that use AI have an affirmative obligation to make their AI systems as transparent and readily-understood as their analog predecessors. Anything less will structurally disempower the very people that AI vendors claim to help.

What are the areas of AI and Machine Learning development whose potential social implications are currently being overlooked in the general discourse? Which areas get comparatively too much attention relative to their potential impact?

We have two tiers of public engagement on government AI policy, depending on which members of the public are impacted. When the IRS sought to deploy facial recognition for a narrow range of identity verification functions, the public rightfully pushed back. But the scale of the outrage was orders of magnitude larger than when the exact same tools are used as public benefits programs. Increasingly, we act as if only the middle class and wealthy deserve to be protected from AI harms, not those with lower socio-economic status. If this trend is unchecked, AI protections will become only the domain of privileged members of the public who are typically benefited by discriminatory AI.

What emerging regulatory frameworks are having the greatest impact on AI development at the present time?

Complete, categorical bans. As AI firms pour millions of lobbying dollars into convincing lawmakers to develop light-touch regulatory schemes that serve the interests of industry, not communities, a growing national movement is pushing back, calling for complete bans on the most risky applications of AI: criminal justice, military, and hiring. All-too-often, audit standards and other transparency tools are just an invitation for self-regulation by profit-maximizing actors who have proven unwilling and unable to deploy technology in the public interest.

How can we equip policymakers to better keep pace with the speed of AI and Machine Learning development?

The United States needs a massive investment in digital civics, including a better understanding of novel deployments of AI. For years, experts have evaluated AI in silos, excluding the public from discourse over the technology that increasingly controls their lives. But the truth is that it's relatively easy to engage in broad-based community education on AI, it just takes a n investment in time and money. Members of the public don't need to know every nuance of an AI model to understand the threats and benefits it poses, the types of errors it can create, and the incentives of the institutional actors who are creating or validating the technology.

What models and frameworks do you use to examine the assets and/or predict the impact of an emerging application of Machine Learning and AI? How well have these models worked for you?

We don't focus on algorithms and models when evaluating the social impact of AI systems. As a human rights organization, we have limited to zero visibility in the precise operation of the systems we evaluate. And what insights we do receive often come from interested parties with an incentive to positively spin the broader impact of the relevant AI system. Instead, we take a broader socio-technical examination of the context in which an AI system is operating, the incentives of the parties involved, and the likely points of failure in deployment. This has allowed us to make prescient predictions about even opaque AI systems based on stakeholder incentives and past AI failures, not training data and source code.

How can individuals be empowered to know and protect their rights in the face of potential threats to privacy and civil liberties linked to use of AI?

The key to effective AI public education is storytelling, relating the goals and potential benefits (and harms) of an AI system to existing analog systems. Through community-centered, not technology-centered AI education, we can easily educate large numbers of residents on the impact AI tools are having on their lives. And we can use this community-centered approach to help the public exercise their rights in the most high-tech settings. Too many technologists and experts focus on the items that are within their expertise and domain when developing AI training courses. But the goal isn't to train new technologists, it is to create new forms of digital civics that can empower everyone, irrespective of technical capacity, to understand the most pressing questions of public life.

How can we ensure that the use of AI for "public safety" does not reinforce existing patterns of discrimination?

AI used in public safety contexts WILL reinforce existing patterns of discrimination. That's why we must ban it. AI is far too error prone to be permitted in a setting where the costs of errors are as high as they are with public safety. As Professor Ruha Benjamin and countless others have written, even a technically non-biased system will perpetuate bias when used in a biased societal context, and there simply is no societal context as biased as policing in America. We must not allow this technology, protected by a fig leaf of nondiscrimination auditing, to become part of how Americans are arrested, tried, and wrongly convicted… well, not become even more of a part. The technology is far too prominent already in the public safety field, and the only answer is a ban.

> “
> **We need to diversify not just the datasets that are used by current AI models, but the teams that develop them. This is a long process that includes reforming the training pipeline in the schools and universities that will bring us the next generation of technologists.**
> ”

Alberto Rodriguez

Senior Program Manager,
Public Interest Technology
New America

What are key factors in the way AI is developed/deployed that have the most impact on protecting human rights?

I would say two factors are the most important toward its development: The data that AI is using to train, and the question, it is optimized to solve. Both of these are human in its core, and end up deciding how an AI is going to perform the tasks it is asked.

On the other hand, the key factor for its deployment is even more human-based. There should be a clear division on what social tasks are given to an AI, and which should be always human-based. Context ALWAYS matters when it comes to ethical considerations, and whenever it is needed (at least in the near future) should remain outside AI implementation.

How can we strengthen interdisciplinary engagement in technology development, deployment, governance, regulation and oversight?

Imbuing DEI [Diversity, Equity, and Inclusion] from design to deployment. Using a clear set of rules for its use & transparency (whenever it is possible) on the algorithms and datasets used. We need to diversify not just the datasets that are used by current AI models, but the teams that develop them. This is a long process that includes reforming the training pipeline in the schools and universities that will bring us the next generation of technologists.

We need to accept the fact that AI development is too cheap, dispersed and nimble to effectively regulate. This leaves us with the need to regulate WHERE AI is used and what kind of decisions it should be used for. Separating it from the civic realms, where cultural and social complexities are paramount, to ensure ethical outcomes.

But above all, we need to educate - particularly decision makers - on what AI (and other emerging technologies) can and can't do, how it works, and how it can be exploited. This is the only way to ensure that leaders in the public and private sector can make the right decisions around AI implementation.

What are the areas of AI and Machine Learning development whose potential social implications are currently being overlooked in the general discourse? Which areas get comparatively too much attention relative to their potential impact?

The area most overlooked is the lack of enforceability of any AI governance model. Governments and international organizations are working to create a set of rules and limits for AI design and deployment, but ethical systems (particularly on a global scale) are by nature complex and sometimes confusing. This might be a problem that we are not equipped to solve, as it relies on solving the perennial societal questions we haven't solved for ethics and morality.

What we need to double down is on the political and government infrastructure that will provide a correct enforcement of the rules we do develop, and the people that are trained to enforce it. Policymakers and public servants today are notoriously not prepared to handle technical issues, and less so the complexities of AI design and implementation. We need to train and educate the public sector in these topics.

How can we equip policymakers to better keep pace with the speed of AI and Machine Learning development?

Educating and mythbusting around the nature, limitations and vulnerabilities of Artificial Intelligence and Machine Learning models for policymakers. This needs to particularly focus on the dangers of algorithmic bias and privacy which are currently the crux of the worst mistakes on public sector AI implementation.

It is also providing good case studies on how AI has been used as an effective tool to help guide policy decisions, instead of supplanting the decision making process.

Lastly, is supporting new policymakers and future leaders that do have the technical training and expertise to join public service. We need technologists at every level of the government.

How can we ensure that the use of AI for "public safety" does not reinforce existing patterns of discrimination?

Constant oversight and iteration, transparency, as well as the inclusion of human centered design techniques for its development and deployment.

We need to normalize design research whenever AI is being implemented, so that we can understand its impact.

We should also commit to transparency and responsible disclosure regarding AI systems, and the datasets it uses for training. This can foster a general understanding of AI systems and to enable those that could possibly be adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information.

How do we ensure that new and emerging technologies and systems do not reproduce and augment existing inequalities?

We need to avoid attributing agency and free will to digital systems such as AI and ML. We need to make sure that human interaction is not stripped away from the decision making process, particularly around the provision of services and access to rights.

The biggest danger of allowing complex black box systems to make societal decisions is not just how it could augment the inequalities that we suffer today, but that it could create new clusters that lead to new inequalities. It is not crazy to think how an identity checking AI can inadvertently strip human beings from their fundamental rights by incorrectly denying access based on it not recognizing you.

esponsible Tech ecosystem?

deral government in Mexico where I had the rs, developers and policymakers who were It was through their understanding that digital access their rights that all clicked.

nd particularly digital technology) as a way to ivate sector. But just as those designers had izens' needs, I've placed particular emphasis n our policy problems.

c Interest Technology (PIT) team ecompasses the PIT field with the public good as the key

“We need to identify the person who will be held responsible if these AI systems ‘mess up’, or otherwise don’t perform as expected in ways that have negative human impact.”

Alexa Koenig

Executive Director
Human Rights Center,
UC Berkeley
School of Law

What are key factors in the way AI is developed/deployed that have the most impact on protecting human rights?

One key factor is the quality of the data used to train the AI system(s). Unfortunately, the biases embedded in existing data can be deeply problematic for the deployment of these systems in ways that advance human rights. Another challenge is the difficulty auditing or assessing the quality of the work produced by AI, given the black box nature of machine learning. One option, of course, is to compare before and after outputs and use those outputs as a proxy for the quality of the work. Another key issue is how humans will be trained to use these AI systems so that they are empowered to know when the systems aren't helpful, or whether humans are trained for other jobs when replaced by AI. Finally, we need to identify the person who will be held responsible if these AI systems "mess up," or otherwise don't perform as expected in ways that have negative human impact.

How can we strengthen interdisciplinary engagement in technology development, deployment, governance, regulation and oversight?

Find more common projects of interest. For example, if we are collectively focused on documenting conflict in Ukraine, embedding computer scientists and engineers on investigation teams can be a way to effectively identify pain points and design to overcome those pain points. In my opinion, this is the only way to force the translation between disciplines in a way that can have a positive human impact. Disciplinary silos are deeply problematic for addressing social needs with tech.

What are the areas of AI and Machine Learning development whose potential social implications are currently being overlooked in the general discourse? Which areas get comparatively too much attention relative to their potential impact?

In my field of practice, there is currently a lot of focus on object detection, less so on natural language processing, especially in diverse languages. I'd love to see more of the latter as that will support greater cross-national or international work. The applications of AI / ML are also skewed, in human rights towards documentation efforts, as opposed to ways to protect the public from harm (for example, psychological harm from being bombarded with graphic material).

How does your team make decisions around integrating AI and Machine Learning into your product? How do you handle data collection, management, and model optimization? Who is at the table in these conversations? What are you optimizing for?

We make decisions based on pain points in our workflow, and based on the passions, interests, and focus of our students. As for data collection and management, we try not to do that portion in house but to rely on existing data sets that have already been vetted by Berkeley, and to follow internal university protocols, which of course align with state and federal law. As for who is at the table, it would be leadership in our organization (including me), the program director who is considering an AI application to resolve a pain point, the students who are designing that intervention, their CS / EE faculty, and of course our human rights partners/clients, who represent the most affected populations. We are optimizing for greater efficiency and ethics in data management, so that we can meet the human rights concerns of our various partners.

How can we ensure that the use of AI for "public safety" does not reinforce existing patterns of discrimination?

This is where independent research is so critical - as well as access to underlying datasets and outputs. While much of this information may be protected under existing corporate frameworks, I do think there's the potential to have trusts and other closed environments in which researchers should be able to get access to even proprietary data (and held to strict ethical and legal standards) in order to best protect the public interest.

What can we do to help technologists and policymakers from multiple intersectional social identities contribute to human-centered AI without feeling like they're the token minority?

By truly understanding and communicating the business case for and social value of that diversity. I find that having an incredibly diverse team, which speaks multiple languages and comes from very disparate social contexts makes our work stronger than that of our competitors, because we are better able to identify potential downstream problems before they happen. We may not always work as fast, but our work is ultimately stronger - it's about playing a long game versus quick ROI. We can also "vet" each new application of technology against various demographic profiles, in order to understand who may be made less secure or otherwise endangered by that application.

What specific structural changes to incentives and business models are needed in order to prioritize user well being and human rights? What kinds of incentive systems (companies, people, regulators, societal) need to be tapped and how?

Have Solutions be driven by problems and not designing Solutions based on assumptions, and then try to think what that solution "solves." We also need to value all people in the creation to application timeline and better credit them for their labor, instead of spotlighting a very narrow band of CS "rockstars."

“

NLP [natural language processing] is not equally accurate across languages or cultural contexts and puts particular groups at disproportionate risks of being both falsely identified as offenders and un-detected/under-served as victims

”

Alexandra Robinson

Head of Data Ethics and Social Impact

threshold.world

How can we ensure that the use of AI for "public safety" does not reinforce existing patterns of discrimination?

The EU Commission's recently proposed child sexual abuse material (CSAM) regulation exemplifies how AI could meaningfully support public safety, while also risking magnifying existing discrimination and civil harms.

Many tech companies already use ML/AI to scan for, remove, and report CSAM images at faster speeds and volumes than ever before. The EU proposal, however, requires tech companies (under detection orders) to scan user text data - including private messages - to detect grooming intent. This implicitly requires NLP [natural language processing] (or some kind of ML/AI) at scale. A basic key word search, however, shows the proposal only explicitly mentions algorithms once, and doesn't outright mention AI, ML, or bias at all.

The draft mentions safeguards to prevent reporting 'obviously false positives', but doesn't discuss real world drivers or consequences of false positives (or negatives). The EU represents incredible linguistic, ethnic, and cultural diversity. Race, ethnicity, immigration status, and language leaves certain populations more vulnerable to being both persecuted and neglected by law enforcement. The lack of access, for example, to translators, often hinders victims of domestic and sexual violence from accessing justice.

NLP [natural language processing] is not equally accurate across languages or cultural contexts and puts particular groups at disproportionate risks of being both falsely identified as offenders and un-detected/under-served as victims.

The EU must go farther to explicitly address these risks. In addition to requiring tech companies to report false positive rates, both tech companies and EU oversight/enforcement bodies should report rates disaggregated by language and race/ethnicity. Embedding translators and cross-cultural experts into EU CSAM oversight mechanisms will be crucial to equitably protect victims and justly identify perpetrators.

What technical safeguards, oversight, and best-practices are needed to ensure safety by design and protection of human rights while mitigating harms?

Implementing Institutional Review Board (IRB)-style AI oversight models comes up a lot in conversations with peers across academia, the public sector, and tech. IRBs provide a helpful conceptual model, but should not be replicated without carefully navigating their strengths and common pitfalls.

Certain features of IRBs may be particularly beneficial to safeguarding in AI investments:

1. IRBs are conducted at the proposal stage, presenting a proactive, rather than reactive approach to risk reduction.
2. IRBs are empowered to reject or request modifications to proposals, presenting an alternative 'business model' where ethics (ostensibly) have the final say in investment decisions.
3. IRBs acknowledge the role of power dynamics in data collection and include special precautions for children, prisoners, and vulnerable groups who cannot give unambiguous consent.

There are other issues and pitfalls, however, that we should challenge:

1. Widespread DEI barriers in the sciences affect IRB diversity, with serious implications for what/whose/how AI research is conducted. Establishing IRBs must be part of larger organizational efforts to prioritize inclusion, representation, and human well-being across decision-making/makers.
2. IRBs are long processes, and may not be suited to evolve in-step with technology, geopolitical, and human rights contexts.
3. The scope of scientific IRBs is too narrow to support human rights safeguarding in AI R&D and may exempt critical tech investments from oversight. Additionally, IRBs assess risk to research subjects, not to broader populations. Developing AI based on the data representing a particular demography/geography - when that tech will be widely used outside of it - requires a significantly broader examination of risks and harms.

What can we do to help technologists and policymakers from multiple intersectional social identities contribute to human-centered AI without feeling like they're the token minority?

I wrestle with this question daily, both as a woman in tech and as a manager and leader. But I want to take that question farther; what we can do to enable policy makers and technologists from intersectional identities to *lead*, not just contribute?

The broader AI/tech sector - across academia, policy, and corporate – desperately needs diverse decision-makers and boardrooms to set AI research agendas, drive inclusive company cultures, and shape business models.

What can “we” do depends on who “we” are. I am a cis-white woman in tech and I lead remote, cross-functional, multi-racial, cross-cultural tech teams. For me personally, preventing tokenization and being a good manager/leader requires:

1. Continuously learning about drivers, conditions, and consequences of inequity – systemically and in tech – and applying vetted practices to promote equitable hiring, advancement, and retention.
2. Advocating for visibility, growth, and managerial opportunities that gear women and staff from traditionally under-represented groups in tech to become decision-makers and executives. I benchmark my own performance (in my OKRs) by the opportunities I facilitate for staff (speaking engagements, training opportunities, etc.).
3. Shutting down harmful biases in real-time: When someone refers to a male engineer as “the” engineer on a project/team, I correct them in real time and follow-up privately one-on-one.
4. Modeling empathy and building relationships so I not only understand how each team member’s social identity affects their experiences, but also what they need to succeed in and outside of work.

What specific structural changes to incentives and business models are needed in order to prioritize user well being and human rights? What kinds of incentive systems (companies, people, regulators, societal) need to be tapped and how?

The lingering optimist in me wants to believe that tech can reform itself from within and that good, ethical leaders, held accountable by diverse, empowered workforces, can achieve change. But my inner nagging tech skeptic begs to differ. News of Elon Musk’s planned purchase and privatization of Twitter is a reminder that a single Big Tech boardroom can unilaterally alter the landscape of user well-being. When primary accountability is to shareholders, good, ethical leaders aren’t enough. The US needs coherent national-level data privacy and AI regulation that protects the rights, privacy, and dignity of individuals and demographically-identifiable groups. Such regulation must be enforced through meaningful penalties that deter abuses. Infringing human rights, privacy, and user-well being must viably impact profit margins in order to incentivize meaningful change.

Tell us about your role:

My role at an early-stage start-up involves wearing many hats, but at the heart of what I do, I work to ensure we responsibly design, use, & build tech that helps non-profits/ CSR teams achieve social impact in ethical, privacy-affirming ways. I work internally at the organizational level, on specific tech products, and with social impact clients. At the org-level, I lead data policy, social impact strategy, breach preparedness and response, and strategic positioning on data ethics & emerging tech. Within the company, I'm deeply invested in building a culture where women in tech can thrive; I co-founded and co-lead our Women's Affinity Group alongside my incredible colleague Maritere Martinez, who was previously one of two female software engineers on a team of 50. I also lead cross-functional teams developing tech for social impact customers, and advise non-profits and CSR teams on social impact strategy development & measurement, responsible technology, and privacy risk management.

How did you carve out your career in the Responsible Tech ecosystem?

As a kid, I dreamed of being a human rights lawyer and prosecuting gender-based violence internationally. I worked in the court and prison systems with both SGBV victims and perpetrators throughout college. After graduation, I learned Nepali and moved to Nepal for a human rights legal internship. I brought a duffel full of LSAT books and planned to start law school that year. When I arrived in 2010, there was anecdotal evidence of labor trafficking/exploitation of Nepali women in the Gulf. These cases didn't neatly fit the dominant narratives of sex-trafficking to India and law enforcement was turning women away.

I ended up leading research to investigate national patterns of women's labor exploitation and law enforcement responses, and then developed digital forensic systems for intelligence-led trafficking investigations. I discovered I loved analytics and could use tech to pursue justice. I stayed in Nepal for two and a half years and abandoned law school all-together.

After two years working directly with victims of trafficking and child sexual exploitation in Nepal, I burned out. I left for grad school in the UK and studied emerging tech policy, analytics, criminology, and human rights. Since then, I've worked at the intersection of technology, social impact, ethics, and human rights in contexts like Ebola-affected West Africa, the Nepal-Gulf labor corridor, and the Thai Seafood industry. I've never second guessed my career path, but my early career irrevocably shaped me into a 'skeptical-tech-optimist'. The most rewarding moments in my career involve helping organizations identify their blindspots, re-examine assumptions, and re-group to build sounder tech.

“

The future of humane AI in social media is achieving a balance between customization and ease of use. Platforms need to think about their community members as whole, unique humans with their own individual needs, not just ‘users’.

”

Alice Hunsberger
VP, Customer Experience
Grindr

How can we ensure that the use of AI for "public safety" does not reinforce existing patterns of discrimination?

There is always a balance between safety and free speech. Some platforms err on the side of caution and moderate heavily so that users are kept as safe as possible. However, that can stifle free expression and can reinforce discrimination when it's too heavy-handed. Having frameworks in place that balance these tradeoffs is important, as is getting input from the communities that are served.

Another key aspect is making sure that there is oversight by a group of humans who are using an ethical framework and have the public's best interests in mind. It's also imperative that the team working on these issues has a seat at the table with key executives and decision makers. Safety-related AI can be hugely beneficial to online communities, but it can sometimes be at odds with metrics like engagement or user growth — although arguably a healthy community is one that is going to lead to long-term growth.

Finally, transparency and communication can help here. Platforms often keep trust and safety issues close to the vest so that bad actors can't take advantage of that information to harm the business or the users. But a lack of solid information can make users jump to conclusions that may not be based in actual fact.

What technical safeguards, oversight, and best-practices are needed to ensure safety by design and protection of human rights while mitigating harms?

It's important to make sure that the most marginalized or vulnerable communities are considered when reviewing the impact of AI or automated systems. The exact Solutions will vary depending on the platform, the communities using it, and the goals of the business, but this is not a "one size fits all" issue.

For example - power dynamics between two people can have a massive difference in how a statement can be interpreted. The same message could be perfectly fine between two "in group" members, or completely unacceptable if it's being said to someone from a group that has been historically discriminated against by someone outside of that group. For this reason, AI must be combined with additional data points, nuanced policy, and human moderation teams.

One key best practice is to have a robust appeals system. Every user should have the right to appeal an automated decision, and to have that looked at by a trained human. AI can be great to get harmful content off a platform quickly, but mistakes can happen, so appeals are key.

Looking at your crystal ball of a positive tech future. Let's say we have human-centered AI. What would humane AI look like? What would its pillars be?

The future of humane AI in social media is achieving a balance between customization and ease of use. Platforms need to think about their community members as whole, unique humans with their own individual needs, not just "users." Platforms are working on ways to allow people to make their own decisions about what they do or don't want to see. The problem here is that most users want an easy experience on social media sites. Most people don't spend a lot of time adjusting their account settings, and giving people choices that are too granular could be counterproductive. It has to be easy and intuitive.

One of the great things about social media is the ability to see content from people all over the world and from all walks of life. Allowing users to set an algorithm that shows them a bubble of hateful content is a real concern, so platforms must be able to set intelligent guardrails and make sure that the customizations that people are making are beneficial and helping to create a positive community.

How did you carve out your career in the Responsible Tech ecosystem?

Almost 20 years ago (pre-social media!), I founded an online community forum, which is where I started to learn about community building and online moderation. I've also co-founded two in-person community spaces – a meditation center and a community gaming space. When I came back to online community building and the tech world, it was as head and founder of the community and customer experience team at OkCupid, a global dating app, where I worked for a decade. Two years ago I took on the role of head of Customer Experience at Grindr, the world's biggest LGBTQ+ social networking site.

In all the work I've done, I've seen how important it is to be intentional with community building – whether in a physical space or online – and how "open and free spaces" where people can do anything they want often devolve into chaos and drama over time. I've learned that it's critical to design spaces from the very beginning with a core set of values, and to not be afraid to enforce those values when necessary. It's also important to think about the community as a whole vs. individual freedoms (what's right for one person may not be right for the group), and to think about your community members as whole, unique humans, not just "users."

“

Companies deploying AI need to assume that their products will produce biased results if those products are left on their own. Companies shouldn’t assume that math solves racism or sexism.

”

Anupam Chander

Scott K. Ginsburg Professor of Law and Technology

Georgetown University

What are the areas of AI and Machine Learning development whose potential social implications are currently being overlooked in the general discourse? Which areas get comparatively too much attention relative to their potential impact?

The black box statistical approach of contemporary AI means that decisions are often defensible only as a matter of statistics, not as a matter of logic. Reasoned decisions might be replaced by statistical analysis. The possibility of errors is significant. Of course, human decision-making is not without its biases and errors as well. But we should pay attention to biases and errors in both human and algorithm domains.

What emerging regulatory frameworks are having the greatest impact on AI development at the present time?

In the U.S., the interpretation of the Computer Fraud and Abuse Act will be particularly important to the data gathering used by AI systems. China has broad new regulations of automated algorithms. And the EU is planning a very broad AI Act, which will be particularly important when it is finalized.

How can we ensure that the use of AI for "public safety" does not reinforce existing patterns of discrimination?

When police rely on AI methods, they need to ensure due process, including transparency and the ability to challenge the method. Trade secrets can be a roadblock to the due process that is needed in such uses of AI. Right now, government agencies are signing up for AI tools with non-disclosure agreements, preventing them from disclosing information about them—which makes it hard for those who feel wronged to challenge the accuracy of those tools or to uncover any biases in those tools.

How do we ensure that new and emerging technologies and systems do not reproduce and augment existing inequalities?

In my paper, "*The Racist Algorithm?*", in the Michigan Law Review, I have argued for "algorithmic affirmative action"—borrowing from our experience with affirmative action to respond to biases that persist in society. Companies deploying AI need to assume that their products will produce biased results if they are left on their own. They shouldn't assume that math solves racism.

“

AI is not just accentuating the negative impacts of globalization but also creating new challenges, for example, as seen in how internet intermediaries are wielding political power. If this power is combined with new data management arrangements, a new world order is emerging in terms of which a few technology corporations wield tremendous powers including making decisions of social, economic, and political consequences.

”

Arthur Gwagwa

Doctoral Researcher

Utrecht University

Ethics Institute

What are the areas of AI and Machine Learning development whose potential social implications are currently being overlooked in the general discourse? Which areas get comparatively too much attention relative to their potential impact?

AI is creating a new world order. Globalization is in a state of atrophy, but in its wake, it is leaving a trail of intractable problems that unfairly distribute benefits and burdens/harms in favour of the Global North, for example, unfair value chains. AI is not just accentuating the negative impacts of globalization but also creating new challenges, for example, as seen in how internet intermediaries are wielding political power. If this power is combined with new data management arrangements, a new world order is emerging in terms of which a few technology corporations wield tremendous powers including making decisions of social, economic, and political consequences.

Currently most attention is being paid to local national challenges - how AI discriminates within national borders, yet AI is a cross border technology and as the Russia-Ukraine war or the pandemic have shown us all, a problem in one part of the world is a problem for us all, particularly when it comes to AI.

What emerging regulatory frameworks are having the greatest impact on AI development at the present time?

The UK Centre for Data Ethics and Innovation's (CDEI) AI assurance ecosystem roadmap recently published by the UK National AI Strategy aimed at how the UK can become a world leader in AI Assurance, by specifying the government role and how the UK AI ecosystem should continue to work together to realise this ambition. Assurance, both as a market-based means of managing AI risks and as a complement to regulation, is pro-innovation as it empowers industry to ensure that AI systems meet their regulatory obligations. Yet at the same time it does not eschew regulation and other governance options.

The EU AI Act is a good piece of regulation but needs to address concerns raised by AlgorithmWatch including: consistent update mechanisms for all the risk categories; fundamental rights impact assessment before a high-risk system is put into use; a register of all high-risk AI systems since transparency is the first step towards accountability; a broad definition of AI to cover all systems that may have a significant impact on people's lives; a clear definition and scope of military and national security exemption; obligations to disclose the environmental impact of all AI including data systems to account for ecological impact; a comprehensive protection of people from systems that violate fundamental rights by closing loopholes for prohibitions.

How can we equip policymakers to better keep pace with the speed of AI and Machine Learning development?

Hybrid learning approaches that address skills gaps

There is a skills mismatch. Computer science graduates are often good at the technology bits, but poor on the ethical and regulatory issues and the reverse is true for policy makers. Therefore, while all AI undergraduates should take a course in the structure, governance, legal aspects, ethical aspects, political aspects of AI, policy makers should take a basic course on technology. The two groups should also have interactions in hybrid public policy and AI courses to ensure they all have the skills and ongoing interactions to ensure the safety, ethical characteristics of AI programs and their deployments, for example on ensuring privacy, security, provenance and how to detect the inherent biases in the data that will then get propagated into the AI technology.

Tutorials

Industrial bodies and other organizations working on AI should develop tutorials for policy makers. There are three types of tutorials: Hands-On; Translation; Implication.
Translation. I have previously worked on the IEEE Global Initiative Translation Tutorial Against Algorithmic Bias.

How can individuals be empowered to know and protect their rights in the face of potential threats to privacy and civil liberties linked to use of AI?

Increase society's knowledge of constitutional principles. In liberal democracies the triad of human rights, democracy and the rule of law are the core elements of western, liberal constitutions. The knowledge of constitutional rights and simplified rules of enforcing them for groups is the key first step to confront challenges brought by AI or any other future technology. This should be followed by an increased awareness of ethics and society's ability to know how ethics and the law interact. National laws and ethics should be situated within the international human rights standards due to the transnational nature of the companies deploying AI. Constitutional law and international human rights law should be included in high school and university curricular.

How can we ensure that the use of AI for "public safety" does not reinforce existing patterns of discrimination?

The companies and governments using AI for public safety should be held to international human rights standards. While governments have obligations under international human rights law, the private sector actors have responsibilities to proactively prevent discrimination in order to comply with existing human rights law and standards. When prevention is not sufficient or satisfactory, and discrimination arises, a system should be interrogated and harms addressed immediately. Public Safety technologies often affect some groups disproportionately, therefore extra measures should be taken.

At a more practical level, there is a need to devise the use of public safety technologies in a way that overcomes power asymmetries. This may include the depoliticization of these public safety technologies and whenever law enforcement are using these technologies, should do so in a neutral way. This helps to build trust and confidence in the use of technologies in historically marginalised communities. As an example, if crowd management AI technologies are used responsibly in marginalised communities, it is easy to get the community buy-in not just for that purpose but even when the authorities intend to repurpose the collected data for such other ends as public health.

“

I believe it is important to recognize that humans build AI-driven decision-making systems as proxies for the human brain. As our society builds AI, we are exploring who we are, how our brains work, and what drives our decisions.

”

Arthur McCallum

Founder

Neu.ro Inc

What are key factors in the way AI is developed/deployed that have the most impact on protecting human rights?

I believe that AI has the potential to meaningfully change the world for the better for all. That said, as AI practitioners we must ensure that this powerful technology is used in an ethically and environmentally conscious way. In order for AI to lead to positive change for humanity and the planet, we must ensure that it adheres to fundamental ethical values such as privacy, fairness, freedom of choice, and respect and care for life and the environment. It must also minimize the potential for any violation of rights, physical harm, injustice or environmental damage that could result from the operation of AI systems, whether due to bad data, poor design or deliberate or accidental misuse. Finally, AI development should adhere to a set of Machine Learning Operations (MLOps) best practices that ensures reproducibility, verifiability and monitoring of its effects on human society.

Neu.ro has been a pioneer in the growing community of AI technology companies focused on improving the efficiency of AI development, through MLOps best practices. These practices, when applied properly, have the potential to reduce the time and energy required to develop AI systems up to 3000x, according to recent research from UMass Amherst, Google AI and elsewhere. Further, Neu.ro is a leader in Green AI, providing a state of the art AI Cloud that runs on 100% renewable, zero-emission energy geothermal and hydro power.

How can we strengthen interdisciplinary engagement in technology development, deployment, governance, regulation and oversight?

As AI development is a new frontier, and is the subject of a tremendous amount of research interest internationally, it has been a fruitful area of cooperation and mutual learning between researchers, scientists and industry.

But like many other new technology waves before it, there is a tendency on the part of large tech players - including infrastructure providers, platform companies and tool creators to seek to stake out territory in the field by making their products non-interoperable and encouraging vendor lock-in.

There is a serious risk of AI infrastructure being monopolized by large centers of tech power - i.e. hyperscale AI castles that surround themselves with proprietary technical moats. Neu.ro believes that the values of interoperability and portability should be embraced as drivers of transparent cooperation in the AI space - and this is what Neu.ro seeks to build. AI is a fast moving and fast changing frontier - dividing the space into corporate fiefdoms is antithetical to the inter-disciplinary advances we have achieved to date.

What emerging regulatory frameworks are having the greatest impact on AI development at the present time?

Regulatory frameworks for AI are still in their infancy, with the EU currently leading the way in proposing a framework that would seek to prevent harm, preserve privacy and establish operational monitoring and data usage practices for high-risk systems that could potentially impact public safety, critical infrastructure or eligibility for services.

The risk, of course, is that regulatory overreach stifles innovation, but we are not opposed to the idea of such regulatory frameworks per se. It remains to be seen what the ideal balance between oversight and freedom is to promote progress in the field while ensuring it is used safely, fairly and for the benefit of all.

How does your team make decisions around integrating AI and Machine Learning into your product? How do you handle data collection, management, and model optimization? Who is at the table in these conversations? What are you optimizing for?

As an MLOps platform, Neu.ro facilitates AI development best practices and simplifies and automates resource utilization and tool interoperability for creating AI systems.

First, MLOps best practices support responsible AI by making the process of development:

- Continuous
- Reproducible
- Collaborative
- Accountable

MLOps best practices allow AI developers to follow a transparent process for the entire development lifecycle - from data collection, model design, model training, and monitoring and measurement of success criteria. MLOps create a repeatable workflow and a digital 'paper trail' that allows those outside the development process to audit data, methodology and results.

Furthermore, MLOps at Neu.ro allows for teams to monitor and optimize their CO2 footprint for the entire development and deployment process, allowing strict adherence to ESG policies by the AI team.

Finally, MLOps allows for the causes of failed results to be quickly identified, giving AI developers, product managers, AI ethicists and regulators actionable steps to correct issues and minimize AI risks.

How do you currently apply climate and sustainable development goal frameworks to design and development of AI?

Beginning in 2021, Neu.ro announced a major new initiative and strategic direction for the company: Green AI - to be supported by our new Zero-Emissions AI Cloud and our high-efficiency, waste reducing cloud and on-Orem MLOps software offerings.

Neu.ro's Green AI goals and values include:

- Provide a 100% renewable energy, zero emission option for AI development, training and deployment TODAY
- Support sustainable, repeatable and rapid AI development via implementation of MLOps best practices, improved efficiency and reduction of "digital waste"

- Help our AI development clients see their projects through an ethical AI lens and a climate lens—and enable them to act on what they find
- Assist our customers and partners in achieving net zero carbon emissions ahead of global regulatory requirements
- Engage openly and broadly on Green AI, Ethical AI and sustainability issues
- Allow for accelerated near-term ethical AI development and deployment while maintaining sustainability and carbon neutrality

Going forward, Neu.ro will track our progress towards these goals and will further seek to combine our efforts in this area with other interested parties via joint projects, meetings and commitments.

How can we ensure that the use of AI for "public safety" does not reinforce existing patterns of discrimination?

The promise of AI is the ability to make human-level decisions at a scale, speed and accuracy that surpasses human capacity. One risk of AI is that it fails to accurately reproduce human decision making systems. Another risk of AI is that it reinforces unfair patterns of existing human decision making systems. These are two distinct problems.

There are three fundamental pillars I would highlight to minimize bias in public safety AI. First, AI developers must deeply understand the limitations and biases of existing data sets, and diligently account for misrepresentation. Second, AI developers must be conscious of 'model' bias, as the design of AI itself can inadvertently create or reinforce discrimination. Finally, we must recognize that AI is still a nascent technology that will make mistakes. Therefore, we should establish fundamental areas where AI-driven systems should not be used, or should be used only with human oversight as a critical component.

Looking at your crystal ball of a positive tech future. Let's say we have human-centered AI. What would humane AI look like? What would its pillars be?

I believe it is important to recognize that humans build AI-driven decision-making systems as proxies for the human brain. As our society builds AI, we are exploring who we are, how our brains work, and what drives our decisions. This is a fascinating process where we place a mirror on ourselves. I believe that, over time, our AI journey will lead us to make more enlightened human decisions.

I see Humane AI of the future as a symbiotic human-machine system that enhances our ability to solve global problems. It will be based on increasingly accurate algorithms, capable of modeling certain cognitive aspects of the human brain. It will leverage the power of environmentally responsible computational machines. It will use a combination of ethically sourced natural data and large amounts of synthetic data. It will be infused with the same human ethics that we strive to instill in all human systems. It will include human-in-the-loop quality control mechanisms that will minimize, but will never end, the risks of AI failure.

“

Developing and using AI that aligns with human values takes a structured approach anchored in governance and compliance. To get there, organizations need a framework for identifying and addressing potential issues throughout the AI lifecycle, from design to operation

”

Beena Ammanath

Executive Director

Deloitte Global AI Institute

What are the areas of AI and Machine Learning development whose potential social implications are currently being overlooked in the general discourse? Which areas get comparatively too much attention relative to their potential impact?

When AI's potential downsides are discussed in media reporting and public debate, unfairness resulting from AI bias receives much of the attention. It is a valid issue, as biased AI could lead to discriminatory decisions, such as in job applications and access to loans. However, there are other AI qualities whose impact may be as important, or even more important, depending on the AI model and how it is used. Important qualities can include things like reliability in outputs over time, transparency to all stakeholders, data privacy, and the safety and security implications from using AI. These and other elements impact whether AI use is in line with human values and by extension, whether we can trust it.

This is not to say that transparency, privacy and other qualities are equally important in all circumstances. Take autonomous vehicles. Bias and unfairness are not relevant for a computer vision AI used to identify people and things while the car is moving. Reliability, however, is extremely important, as one erroneous output can have life and death consequences. The larger insight is that organizations deploying AI need to carefully examine every AI use case to identify which dimensions of trust are most important.

What models and frameworks do you use to examine the assets and/or predict the impact of an emerging application of Machine Learning and AI? How well have these models worked for you?

Developing and using AI that aligns with human values takes a structured approach anchored in governance and compliance. To get there, organizations need a framework for identifying and addressing potential issues throughout the AI lifecycle, from design to operation. At the Deloitte AI Institute, where I am executive director, we created the Deloitte Trustworthy AI™ framework, which identifies six dimensions of trust—transparent and explainable, fair and impartial, robust and reliable, respectful of privacy, safe and secure, and responsible and accountable.

The framework is valuable as organizations develop comprehensive governance to guide AI programs, which can include new professional roles, committees, technologies, strategies, and amendments to decision making processes. Together, these elements of AI governance lead an organization toward compliance with relevant laws and regulations, as well as compliance with the organization's own values and ethical priorities.

What technical safeguards, oversight, and best-practices are needed to ensure safety by design and protection of human rights while mitigating harms?

Aligning AI with human rights and values cannot be left to chance. Taking a proactive, structured approach, effective AI governance emerges from the stakeholders engaged in the AI lifecycle, the processes they follow, and the supporting technologies they use. There are leading practices in each area that can help mitigate harm and promote the best outcomes.

People working with AI may need new knowledge and skills to assess ethics and trust in AI, and organizations may consider creating new positions and advisory groups, such as a chief ethics officer or an AI audit committee. In terms of processes, an essential step is setting clear waypoints at every major stage in the AI lifecycle where stakeholders assess whether AI is trustworthy, risks are mitigated, and the project is meeting expectations. These waypoints allow issues to be addressed before a model is deployed.

To aid in this, stakeholders might use complementary technologies that can validate AI outputs and reveal how complex, opaque AI models are truly working. There may also be technical approaches that can guard against potential harms. For example, edge computing can help mitigate privacy concerns because less personal data is transferred to a remote server.

This approach of aligning people, processes, and technologies marshals all of an organization's assets to intentionally manage AI to its greatest potential value, as well as its most ethical application.

How do we ensure that new and emerging technologies and systems do not reproduce and augment existing inequalities?

One of the most important steps we can take to ensure AI promotes equal access to opportunity and value is bringing a more diverse group of stakeholders to the decision making table. History teaches us that when transformational technologies emerge, women and minority stakeholders are too often under-represented. This is already the case in AI. By some analysis, women and minorities account for just 5% of the world's AI workforce. This is a tremendous missed opportunity.

As AI matures and is deployed at scale, we have a responsibility to intentionally include people from all backgrounds who can use their diverse lived experiences to shape the AI lifecycle. Ethical issues may be more apparent to one person over another. More broadly, AI that is developed and used by a diverse workforce is better suited to provide value in a diverse society.

“

One of the changes to strengthen interdisciplinary skills in the field of AI starts with education. Although there has been a crescent movement in computer sciences and other technical degrees to include data and AI ethics in their curriculum, it is not enough.

”

Bruna de Castro e Silva

AI Ethics and Governance
Saidot

Doctoral Researcher
Tampere University

How can we strengthen interdisciplinary engagement in technology development, deployment, governance, regulation and oversight?

All actors involved in technology development, deployment, regulation, and oversight must acknowledge that AI systems are sociotechnical systems, therefore, they should be addressed together with the people, organizations and institutions using and governing them. We must look at the complexity and embeddedness of social structures in big data and AI to understand that just designing systems and applying them in the world will not solve our problems. Here is where interdisciplinary engagement is crucial. Trustworthy AI systems for the public good which will equally benefit everyone require expertise and capabilities from different disciplines (e.g., computer sciences, engineering, ethics, law, sociology, humanities) and from internal and external stakeholders.

One of the changes to strengthen interdisciplinary skills in the field of AI starts with education. Although there has been a crescent movement in computer sciences and other technical degrees to include data and AI ethics in their curriculum, it is not enough. AI Ethics should be a core discipline and not an adjacent subject as it usually is today. AI projects inside organizations and institutions must include cross-sectoral and interdisciplinary teams composed of product owners, engineers, lawyers, AI Ethics officers, and business managers, etc. This organizational change is also essential for developing responsible technology.

Tech policymakers and regulators should play a crucial role in steering regulatory frameworks and standards that support and foment responsible innovation only. Finally, increasing the access to interdisciplinary research and the collaboration with experts are efficient measures that allow for greater understanding of sociotechnical challenges present in big data and AI.

How can individuals be empowered to know and protect their rights in the face of potential threats to privacy and civil liberties linked to use of AI?

The first step is to critically analyze the positions, narratives, and values embedded in and constructed through ongoing discourses around AI systems and services. The dissemination of a holistic understanding of the complex and multifaceted dynamics of diverse roles and interests shaping the governance of AI technologies will particularly empower individuals in a less privileged position providing them with much needed tools and capabilities to actively engage in and directly impact the overall digitalization of society.

Participatory construction of best practices, co-design approaches for engaging citizens in critical deliberations and decision-making throughout the AI lifecycle, and frameworks for multistakeholder AI governance will contribute to a more inclusive AI ecosystem, leading to unparalleled benefits for our business, governments, and societies.

How do we ensure that new and emerging technologies and systems do not reproduce and augment existing inequalities?

First, we need to rethink the digital divide and acknowledge that as long as inequality exists in the production, access, ownership and quality of data that feeds AI models it will be necessarily reflected in the outcomes of AI and data-driven technologies.
Second, we cannot ignore that even if the digital divide is addressed, the current power

relations, inequalities and oppressions existing in society are echoed in the data. Big data paints a digital picture of reality and our societies and they are full of biases, discrimination, and injustices. We cannot address the potential harms of AI without tackling these societal issues regardless of the element of technology.

Third, we need to redefine transparency and explainability in a completely different level in order to open the black box of algorithms and understand how the data is being used to create predictions and inferences about people. It's not enough to improve fairness on datasets without a clear and accessible understanding of how algorithms function and how to challenge their outcomes.

Fourth, data/AI literacy has become an integral requirement for the exercise of democracy, to guarantee the implementation of human rights and ultimately to protect our human dignity. If people are not aware that AI and big data have been driving decisions in virtually all spheres of our lives, shaping the way we function as a society, our autonomy will be significantly harmed and so will be our dignity as human beings.

Looking at your crystal ball of a positive tech future. Let's say we have human-centered AI. What would humane AI look like? What would its pillars be?

Responsible innovation must equally provide benefits and opportunities for all, and this goal will be only achieved with inclusion. It is not enough that AI systems do good, technology should empower people who need it most and it should help improve everyone's quality of life and wellbeing. Big data and AI have incredible potential to help us tackle the most critical challenges of our modern societies such as global health, the energy crisis, urbanization, and climate change. AI has also the potential to drive progress across all the UN Sustainable Development Goals by reducing costs, managing risks, streamlining operations, accelerating growth, and fueling innovation. Humane AI will require that we consciously and responsibly leverage all this potential in an equitable, fair, and inclusive way.

Tell us about your role:

I work on the fascinating intersection of ethics, law, and technology. My expertise, interests, and research subjects focus on the risks and opportunities for human rights posed by AI. My mission is to propose solutions to ensure that AI and emerging technologies are designed, developed, deployed, and regulated in a way that not only respects but also promotes human rights.

How did you carve out your career in the Responsible Tech ecosystem?

Digitalization is one of the most impactful phenomena in our modern lives and it has been reshaping how our entire society functions and, consequently, how our laws and legal systems function as well. AI offers significant opportunities for education, work, social care, health, environmental protection, law enforcement, among other areas, but there are several issues that need to be considered as AI has also the potential to undermine or violate human rights. The legal discipline has always been my passion. As a legal expert, when observing all the urgent and evolving questions posed by AI, understanding and critically addressing the nexus between law and technology became my passion too.

“

Simply stated, we need to stop prioritizing profits over human well-being. On a societal level, this means rethinking our economic “growth at all costs” paradigm, which fails to account for externalities, and realigning AI to benefit society and uphold human rights.

”

Camille Carlton

Senior Policy and Communications Manager

Center for Humane Technology

What emerging regulatory frameworks are having the greatest impact on AI development at the present time?

Many countries and state alliances have established guiding strategies or principles for AI, but far fewer have regulatory frameworks. In fact, the best approach to regulating AI – a holistic, overarching scheme versus sector-driven governance – is an ongoing debate. For instance, in Europe, we've seen overarching regulatory frameworks proposed, such as the EU AI Act and the Digital Services Act. On the other hand, the U.S. has taken a governance of sectors approach with groups like the Federal Trade Commission (FTC) and the Department of Housing and Urban Development (HUD) regulating AI within their purviews.

The novelty of AI regulation prompts the question of what and how we should govern to ensure AI serves humanity. An impactful approach would be to govern AI as a system – this means overseeing not just algorithms but the data and the platforms or firms that develop models, mine data, and use that data to train the models. For firms, in particular, regulation should also disincentivize destructive operating models by ensuring that firms internalize the costs of AI-related harms. This systems approach also means governing AI development and AI deployment, since deploying an AI system for a different use than its original purpose, without contextual consideration, can have consequential effects. Finally, and most importantly, we need the active participation of a wide set of interdisciplinary stakeholders, including those impacted by AI systems, to inform every stage of design, deployment, and regulation.

How can we equip policymakers to better keep pace with the speed of AI and Machine Learning development?

Two important approaches come to mind.

The first would be to enhance the technical expertise of policymakers and their staff when it comes to AI and ML. While this isn't an exhaustive list, there are several great programs I can think of working to do this: Aspen Tech Policy Hub, Tech Congress, and Open Philanthropy Fellowship all work to train technologists and tech experts in policymaking. And other organizations, like the Partnership for Public Service, run programs that bring together tech experts and public officials for mutual learning.

The second would be to build the capacity of government bureaucracy to regulate AI from a systems perspective – either by enhancing the expertise of existing departments or agencies, or by establishing new ones. This can be done by directly training staff (as above) and by bringing in interdisciplinary experts to serve as advisors. Building the capacity of government bureaucracy would allow our elected policymakers to rely on complementary expertise – as they do with food, pharmaceuticals, and transportation – to inform regulation.

What specific structural changes to incentives and business models are needed in order to prioritize user well being and human rights? What kinds of incentive systems (companies, people, regulators, societal) need to be tapped and how?

Simply stated, we need to stop prioritizing profits over human well-being. On a societal level, this means rethinking our economic "growth at all costs" paradigm, which fails to account for externalities, and realigning AI to benefit society and uphold human rights.

What's exciting is that we're starting to see movement in the right direction. For instance, regulators are pushing to move the cost of tech platforms' harms onto their balance sheets with bills like CA 2408 which establishes a duty of care and allows for private right of action if this duty is not met. Additionally, the rise of benefit corporations and double bottom line models has opened the door to new incentive structures. Extending the fiduciary responsibility of corporations to their employees, their customers, and the communities they affect would significantly shift the impact of tech platforms on society.

Looking at your crystal ball of a positive tech future. Let's say we have human-centered AI. What would humane AI look like? What would its pillars be?

Currently (and rightfully so) many discussions take a negative approach to AI, asking how we can minimize its harms. But truly humane AI would extend beyond this, pushing for a positive approach in which AI is beneficial – enhancing human well-being, dignity, and collaborative problem solving – for everyone. Importantly, when assessing AI's impact, we need to look beyond the aggregate and also focus on the individual level. We need to evaluate aggregate and individual impact because historically, many extraordinary gains from technology have benefited society "on average" while increasing inequalities astronomically, particularly for marginalized communities. So, in considering both, we should push humane AI to enhance collective well-being while also distributing benefits more equitably and actively minimizing disparities.

We can use a broad range of frameworks – such as Sen's Capabilities Approach or the UN's human rights-based approach (HRBA) – to inform a holistic set of pillars for humane AI. But, as I mentioned above, an essential part of developing humane AI pillars should be taking a participatory approach by including a wide set of interdisciplinary stakeholders and ensuring that those most impacted by AI systems are directly involved.

Finally, it's worth remembering that achieving humane AI isn't about the technology itself – but the choices that we as a society make in the design, deployment, and governance of AI. Coming together across the private sector, academia, government, and civil society will be critical to realigning AI with society's best interests.

Tell us about your role:

As the Policy and Communications Lead at Center for Humane Technology (CHT), my role has two distinct parts.

First, I help the team advance high-leverage Solutions, particularly in the policy space. Specifically, I help create briefings, provide counsel on taking a “systems thinking” approach to regulating the tech ecosystem, and make meaningful connections between those designing technology and those advancing regulation in order to help close the understanding gap.

Second, I help shed light on how social media’s ad-driven, engagement maximizing business model drives systemic harms. We do this through our podcast, Your Undivided Attention, our newsletter, The Catalyst, informational videos, and other types of media.

How did you carve out your career in the Responsible Tech ecosystem?

Virtual coffee chats. Towards the end of my master’s degree, I started mapping out organizations I was interested in and “cold messaging” folks from those organizations on LinkedIn, asking if they’d chat with me. It was earlier into the pandemic, and people were incredibly gracious with their time, connecting me with others and sending me opportunities as they came by. At the same time, I joined as many webinars as I could to learn more about the issues that interested me. This approach opened up a handful of opportunities for me and played an important role in helping me grow my network in the responsible tech space.

Oh, and I was connected to my current role at Center for Humane Technology thanks to the All Tech is Human job board – so big shout out to the work David Ryan Polgar, Rebekah Tweed, and Cate Gwin are doing here!

“

AI is no exception to the fact that technological development and its impacts cannot be divorced from the environments and incentives which produce it.

”

Damini Satija

Head of Algorithmic Accountability Lab

Amnesty Tech

What are key factors in the way AI is developed/deployed that have the most impact on protecting human rights?

The power dynamics inherent in the way new technologies are designed, funded, sold and purchased - that is the key factor shaping AI's impact on people and society. AI-driven tools are being conceptualised and built in a world that is systemically unequal and chronically discriminatory towards marginalised individuals and communities, and this is reflected and baked into almost each decision point that sits behind a technology's deployment, whether that is who it serves, why it is built in the first place or the safeguards that do (or don't) sit around it.

AI is no exception to the fact that technological development and its impacts cannot be divorced from the environments and incentives which produce it. In the context of social media companies, this is a profit-making business model. In the context of government technologies, this is whatever social and political narratives are prevalent at the time. Virginia Eubanks, author of Automating Inequality, illustrates this really powerfully in her writing. As such, we end up with AI that replicates the status quo, including everything that is broken and punitive about it. We end up with systems that can automate, scale and entrench the oppression of marginalised communities.

How can we strengthen interdisciplinary engagement in technology development, deployment, governance, regulation and oversight?

There are so many answers to this and I don't think there will be one way. For everyone it will depend where they sit in the wider system and if it's at the point of development, deployment, governance, regulation, oversight or somewhere else. It's incumbent on everyone who works in this space to think hard and go out of their way to design interdisciplinary thinking in their work, within their teams and organizations and in the way they evaluate the impacts of technologies. For those who are able to hire new roles, it's essential that you do everything possible to bring people in from other disciplines. Professors and teachers can play a really important role here in developing interdisciplinary curriculums and collaborations between students who will then go on to work in this space. Those doing research and policy development should also look to participatory and community-led or centric methods which can (if done correctly) force you into a more interdisciplinary lens. If you work in tech development, then take an active interest in policy and regulation, and vice versa. I think this one is very much a collective effort and responsibility sits at the individual and organisational levels.

What are the areas of AI and Machine Learning development whose potential social implications are currently being overlooked in the general discourse? Which areas get comparatively too much attention relative to their potential impact?

Right now, I am most concerned with how AI technologies mediate relationships between state and citizen, and more crucially enable the state to exercise undue power over citizen life. AI is increasingly being adopted in the delivery of government services such as social security, housing, education, and more. We talk a lot about algorithms in online spaces but we need to make it known that we are interacting with algorithms in all areas and they increasingly dictate our life outcomes. I think the UK exams algorithm scandal somewhat shifted the dial on this and was almost the ‘Cambridge Analytica’ moment for algorithms (in the UK) but didn’t do enough, and obviously I wish we didn’t need that level of scandal for everyone to pay attention.

I also think we get stuck on algorithmic/AI bias, sometimes at the expense of other harms such as surveillance or digital exclusion. This is counter-productive because bias very quickly leads into a more technical or policy discussion, which is important, but again at the expense of wider rights-focussed interrogation of AI.

As a small appendage, I also think that military use and development of AI is not studied enough by the wider tech and human rights community and too confined to the specialist tech and security community. This needs to change as research into AI is often funded through defence channels and military AI tools will increasingly be used in the name of national security, for both non-violent and violent activities.

How can we equip policymakers to better keep pace with the speed of AI and Machine Learning development?

This might be contrarian, but I don’t think policymakers being behind is as much of the problem as it used to be (though I can’t speak for everywhere in the world). I think civil society, researchers, academia, journalists and others have done a great job of documenting tech’s harms and making sure their work is noticed. There is a huge body of work in this field to draw on and experts to consult. Being ‘ill-equipped’ is not an excuse for policymakers anymore. I think what remains the problem is the speed of policymaking relative to the speed of technological development and why we continue to see a discord between tech regulation and tech development (without going into detail on regional differences in regulation patterns).

Some governments are trying to better predict tech development e.g. through the use of future methodologies, so that their regulatory frameworks account for future changes in technology. One thing policymakers can and should do, in my opinion, is remain outcomes focussed rather than technology specific i.e. design regulation to prevent particular human rights harms so that even new technologies are governed, regulated and prohibited pre-deployment. This necessitates being clear on how technologies will be assessed, who will assess them and the evidence that is sufficient to prove harm - there should be no loopholes for the developers of new technology to exploit.

For those working outside of policy, we should keep striving to ensure decision-makers hear what we have to say and our policy recommendations.

How can individuals be empowered to know and protect their rights in the face of potential threats to privacy and civil liberties linked to use of AI?

This is a huge question and one that really stands in the way of seeking accountability and justice for algorithmic harms because it's not just about knowing your rights, it is about knowing when you've been harmed by an AI-driven tool. How can we advocate for our rights if we don't even know when we have been subject to algorithmic decision-making in the first place? And if we are aware we've been subject to one, do we know our rights in the face of these tools and their adverse impacts? I think the onus is on those of us already working in this space to get creative and launch bigger scale awareness raising projects. We should think about how to teach children about the impacts of AI, and parents too so they can talk to their kids about these issues at home. We should also lobby for effective redress and remedy mechanisms in AI regulation so impacted people have a way to seek accountability when they are harmed by AI technologies. We should demand governments and companies to disclose when they use algorithmic tools. We should provide legal support to those harmed by algorithmic technologies. Overall, I don't have the clearest or fullest answer on this one (and am working through it with others in this space) but want to point out that I think it's one of the biggest questions at this point in time.

How do we ensure that new and emerging technologies and systems do not reproduce and augment existing inequalities?

There are a lot of interventions to think about here at each stage of a technological system right from its very conceptualisation through to deployment, ongoing use and retirement. There is a lot of literature out there on this. But first and foremost, our policymakers must be bolder and more progressive in enforcing red lines where technologies are fundamentally flawed, intolerable and will always augment existing inequalities. This includes instituting bans and moratoriums - it shouldn't have taken George Floyd's murder to finally see a cascade of company moratoriums on law enforcement use of facial recognition technology.

Tell us about your role:

I am Head of the Algorithmic Accountability Lab at Amnesty Tech. We are a multidisciplinary team with a goal to hold governments and other powerful actors to account on their misuse of algorithmic technologies. We are focussed on exposing the human rights impacts of algorithmic systems used to deliver social welfare and essential public services such as housing, health, education, benefits and more. Through research, campaigns, advocacy and litigation, we aim to elevate and centre the stories of those affected by algorithmic harms, shed light on how marginalised communities are affected by welfare algorithms in all regions of the world, interrogate the broader socio-technical and political forces which drive adoption of algorithmic systems and seek justice where rights have been denied or violated.

“

If we don’t include the voices of those we’ve historically discriminated against as well as listen to their experiences to better understand how we can break the pattern, we risk continuing to repeat those same injustices.

”

Dr. Dédé Tetsubayashi

CEO + Founder
incluu

Responsible Innovation Manager
Meta (Facebook)

How can we strengthen interdisciplinary engagement in technology development, deployment, governance, regulation and oversight?

Ensuring all technology development, deployment, governance, regulation, and oversight is interdisciplinary requires intentionally designing with a consortium of diverse players - from technologists & designers to social justice activists, policy-makers, community-leaders, social scientists & behaviorists, advocates, and more. Currently, the de-facto approach to the development of technology is the belief that only extremely specialized engineers understand the short, medium and long-term impact on communities and peoples who use said technology. Unfortunately, that leaves out most of the population of folks who use technology in day-to-day life, and doesn't take into consideration that those engineers' expertise may be limited to acquired knowledge, without being tempered or enhanced by lived experience. In actuality, each step in the process needs to consist of at least one representative from intersectional social identities to comprise diversity in thought, access, and impact expected in the building of that technology. This requires developing and deploying technology more slowly and thoughtfully, integrating all levels of social and legal requirements such as governance, policies resulting in positive impact, security, privacy and accessibility by design, etc.. Although the time from design to deployment is slightly longer, it results in ensuring each representative understands the technology being built, is equipped with the tools to advocate for positive outcomes, and builds the foundation to ensure long-term sustainable growth tied to positive outcomes are considered the ultimate goal-posts to reach.

How does your team make decisions around integrating AI and Machine Learning into your product? How do you handle data collection, management, and model optimization? Who is at the table in these conversations? What are you optimizing for?

My team practices and teaches others how to build products for inclusion, equity, and accessibility - in other words, we design and build with intentionality for positive outcomes over intent, iteratively testing for these outcomes at each step of the development process, and optimizing for accessibility and positive outcomes. At the table in these conversations are the people who are experiencing the most disproportionate difficulties in either accessing or having positive experiences using our product or platform for what it was intended to do. Data collection and management is done at each stage of the development process - if we don't first collect the data, we can't see patterns, analyze connections, and make changes to improve the product. But access to the data gathered is managed through role-based-access-controls and neutral parties so as to prevent as much introduction of bias as possible. Our focus remains on centering stress and edge use-cases, the people at the margins, those excluded, because when we build with the people who have the most difficulties using the product or platform, forge relationships of trust with them to build collaboratively, we end up working with their entire adjacent network of support.

How can we ensure that the use of AI for "public safety" does not reinforce existing patterns of discrimination?

AI technology is currently in its infancy and just like any infant, it needs to be taught, guided and nudged to understand historical behavior or patterns in multiple contexts, before we can expect it to be able to break out of existing patterns of discrimination. So, as we teach a child society's morals, we also need to provide teaching guardrails to AI. In order to ensure that it doesn't simply repeat our discriminatory behavior, we could potentially consider using AI to better understand how those existing patterns or systems of discrimination were formed, how, where, and against who that discriminatory behavior is perpetrated. Next, we can use that "map of discrimination" to create targeted plans of action to reduce discrimination. Further, ensuring we're not building AI in an ivory tower or tech silo is critical - we have to build and test with adversely impacted members of the "public", representative members of society, or a consortium of leaders who can hold both us and the technology accountable to mitigating and dismantling those structures of harm. If we don't include the voices of those we've historically discriminated against as well as listen to their experiences to better understand how we can break the pattern, we risk continuing to repeat those same injustices. Real-life harm from AI's reinforcement of discrimination is already happening in court cases, applications for jobs, housing, credit, and more, because we aren't collaborating with people negatively impacted by badly-taught AI that's been left to make decisions without human input.

What technical safeguards, oversight, and best-practices are needed to ensure safety by design and protection of human rights while mitigating harms?

All teams should practice building products for equitable outcomes and non-discrimination as the product celing and product floor. This means they must invest in research, building trust-based relationships with the people using their products and invite feedback that is incorporated into the builds at regular cadences. Moving away from silos, and building cross-functionally with legal, security & governance teams enables anticipating, planning for and mitigating potential risks during every step. This also prevents the team from having to duplicate efforts or redesign an entire flow during the deployment phase because of a risk or harm that was identified too late. Bringing in external experts to provide reviews and guidance on the actual usage of the product also ensures that testing is conducted with intersectional social identities. Regular customer feedback from diverse intersectional social identities and an iterative integration process inviting all critical stakeholders' (both internal and external) input during each phase of the product life-cycle empowers teams to understand why it's necessary to design with the goal of and be accountable to building for positive or equitable outcomes and mitigating harms.

How do we ensure that new and emerging technologies and systems do not reproduce and augment existing inequalities?

We need to build AI (or other new and emerging technologies) with guardrails, and keep in mind that no such technology can be fully independent of human oversight until it becomes an adult. Simultanesouly, we as the human overseers and guiders need to be aware of the multitude of ways we repeat our own patterns of discrimination, so that we can interject new behavior and establish new patterns that are less discriminatory. This does not mean that we can remove discrimination completely because we are human and we are imperfect, but awareness of our systems of injustice and patterns of discriminatory behavior can help us learn and form new habits. Similarly, we can apply this technique to how we build AI and other emerging technologies by integrating transparency, inclusion, accessibility, security and privacy controls into the development process. We need to be able to interrogate the models used to ascertain how and why certain decisions are made. Without this understanding, we cannot mitigate harms because we won't know what harms are being perpetrated.

What specific structural changes to incentives and business models are needed in order to prioritize user well being and human rights? What kinds of incentive systems (companies, people, regulators, societal) need to be tapped and how?

Companies, people, regulators, and all of us must have a mind-shift regarding rapid growth, profitability, large margins and bottom lines. If we continue to prioritize maximizing profits through any means necessary as ultimate goals for products, we'll continue to create products that cause harm rather than mitigate it, because we won't be centering immediate and long-term sustainable growth, and we won't be centering user well being and human rights. It's possible to build ethically and responsibly with both people and profits in mind as long as we're not prioritizing moving fast at the cost of all else. It requires moving more slowly, with intentionality and with a focus on using ethical and equitable practices to design and develop products for a world where we prioritize human rights and well being. Getting those products to market may take slightly longer, but a slight delay to market is better than having to scrap your product or start from scratch because one didn't prioritize user safety, well being and human rights by design into product builds. Executive leadership must also hold these goals as critical in their mission, in their yearly goals and OKRs, and provide the coverage and tools necessary to enable each member of the organization to hold themselves accountable to upholding those values as they build products. Building for more than just profits should further be tied to incentive and payment structures so delivering against company goals is centered on positive social impact, thus enabling a cultural shift.

> “
> **It is imperative that citizens take an active role in the protection of their rights and ensure that the responsibilities of ensuring accountable, transparent and fair AI is not only left to AI developers.**
> ”

Diana Nyakundi

Tech Policy Fellow- Artificial Intelligence
Lawyers Hub

How can individuals be empowered to know and protect their rights in the face of potential threats to privacy and civil liberties linked to use of AI?

Governments need to foster AI literacy that will ultimately enable its citizenry to fully understand the risks that AI poses, its benefits and most importantly, its impacts. It is imperative that citizens take an active role in the protection of their rights and ensure that the responsibilities of ensuring accountable, transparent and fair AI is not only left to AI developers.

Secondly, public participation in the creation of AI policies by citizens should be encouraged to ensure that they have a voice with regards to the protection of their rights from the onset.

Lastly, I think it is vital to have data protection principles such as lawfulness and transparency embedded in the development of AI systems. This way, individuals can be fully aware of the collection of their personal data, what it is being used for and be provided with an opportunity to consent to the collection and use of this data. This will certainly empower individuals to protect their right to privacy.

What technical safeguards, oversight, and best-practices are needed to ensure safety by design and protection of human rights while mitigating harms?

The governance of AI systems should include AI Risk and Impact Assessment which would enable the identification and assessment of the risks that specific AI systems pose, differentiating these risks based on their level of risk and adopting measures proportionate to those risks that would act as prevention and mitigation measures.

How do we ensure that new and emerging technologies and systems do not reproduce and augment existing inequalities?

There is a tendency to think that AI driven inequalities are as a result of lack of data or lack of accurate data. However, what we fail to understand is that these inequalities are a reflection of our societies which only transcends to the data collected and even further into the building and design of AI systems. It is therefore important to ensure that there is a wide array of developers and stakeholders in the creation of these systems in order to integrate diversity of views, backgrounds and needs in the design of AI.

Looking at your crystal ball of a positive tech future. Let's say we have human-centered AI. What would humane AI look like? What would its pillars be?

The pillars of humane AI would be;

1. A collaborative approach between humans and AI to reduce blind spots and the potential risks that AI poses, to give customers a personalized experience and to make much more informed decisions.
2. Non discriminatory algorithms free from bias that would reflect representation of all humans regardless of sex, gender and race.
3. A human rights-based approach in the development of AI systems.

Tell us about your role:

I am a tech policy fellow at the Lawyers Hub with a focus on artificial intelligence. I conduct research on AI policies, development, news and innovations especially within the African context.

I also work in the same scope on privacy and data protection issues, internet governance, tech and democracy especially on content moderation and censorship.

How did you carve out your career in the Responsible Tech ecosystem?

Interestingly, All Tech is Human was a great foundation for me in this space. I was very green in the tech policy space and I was looking to expand my network, skills and find solid opportunities in the field. I heard about it from a friend and signed up as a volunteer. From that experience, I expanded my networks and I was exposed to a tonne of opportunities. That is how I learnt about the Center for AI and Digital Policy (CAIDP) and applied for a research fellow role in the Fall Policy Clinic 2021. My experience at CAIDP was intense, rigorous, eye opening and led me into the world of Artificial Intelligence. I thereafter got an opportunity at my current role as a tech policy fellow, AI with the Lawyers Hub.

“I’m concerned that approaches taken to human rights due diligence can be ad hoc rather than strategic—for example, that attention is focused on a somewhat random mix of products that for whatever reason are in the spotlight, rather than taking a systematic look across all products, all human rights, and all impacted populations to prioritize the most significant”

Dunstan Allison-Hope

Vice President

BSR [Business for Social Responsibility]

How can we strengthen interdisciplinary engagement in technology development, deployment, governance, regulation and oversight?

My work focuses on human rights due diligence at technology companies. Over the past five years there has been a transformational shift in the range of business functions engaged in human rights due diligence, and the quality of this work is enhanced every time that product, data science, engineering, sales, compliance, and government affairs (for example) are involved. My simple belief is that more of the same is needed here, and that increasing awareness of the building blocks of a human rights-based approach across different professional disciplines is essential.

At the same time, I'm concerned that approaches taken to human rights due diligence can be ad hoc rather than strategic—for example, that attention is focused on a somewhat random mix of products that for whatever reason are in the spotlight, rather than taking a systematic look across all products, all human rights, and all impacted populations to prioritize the most significant. I'm seeing encouraging signs that this is changing for the better, and my hope is that companies maintain "human rights risk registers" prioritizing risks to people, just as they maintain "enterprise risk registers" prioritizing risks to enterprise value creation—and seek to better understand the link between the two.

What are the areas of AI and Machine Learning development whose potential social implications are currently being overlooked in the general discourse? Which areas get comparatively too much attention relative to their potential impact?

I spend a lot of time assessing human rights impacts with technology companies that develop and sell AI and machine learning products into other industries, such as retail, financial services, and logistics. When doing this work it is striking how often the key variable in human rights impact are decisions made by the company deploying the technology, not decisions made by the company developing and selling the technology. For this reason, I think the biggest factor being overlooked in the general discourse is the role of non-technology companies deploying AI. The scrutiny of technology companies is entirely appropriate, but we risk taking our eye off the ball if that is all we do.

In addition, a key gap remains the role of systems integrators, distributors, and re-sellers of technology, who play a key role in determining what Solutions are deployed and in which sectors. They are often notably absent from the general discourse.

What emerging regulatory frameworks are having the greatest impact on AI development at the present time?

The AI Act in the European Union will have a big impact, not least by making it clear that companies deploying AI (i.e. not just the companies developing and selling AI) have a responsibility to assess and address adverse impacts, though these provisions can be strengthened. The impact of the AI Act would be greater if there was more alignment with the human rights due diligence expectations of the UN Guiding Principles on Business and Human Rights, and a more intentional effort towards interoperability with emerging regulatory frameworks for mandatory human rights due diligence by companies doing business in the European Union.

What models and frameworks do you use to examine the assets and/or predict the impact of an emerging application of Machine Learning and AI? How well have these models worked for you?

We use methodologies based on the UN Guiding Principles on Business and Human Rights to assess the potential adverse human rights impacts of AI. Several key features of this approach are extremely valuable: a thorough assessment against all international human rights instruments, not just a subset of them; paying particular attention to rights holders at heightened risk of becoming vulnerable or marginalized; prioritizing action by the company according to the most severe risks to people; emphasizing that all harms should be addressed, and that a simple “benefits outweigh costs” approach should not be taken.

There are two areas where these frameworks can be approved. First, I think we need to pay greater attention to roles and responsibilities across entire systems holistically, and not just focus on the single company doing the assessment. Second, we need to better understand how to apply the UN Guiding Principles to product assessments —as distinct from company operations, like factories, farms, and offices— and the UN has a project underway to accomplish this.

“Among the many social impacts of AI, I believe Cultural Erasure is the most overlooked. Take language models as an example. These models can write text that is impressively similar to human-written prose. The internet-scraped data used to train these models, however, is predominantly in English; undermining global linguistic diversity. As a result, models’ outputs are heavily influenced by western cultures that are dominant in the training data.”

Dr. Ellie Sakhaee

Senior Program Manager, Responsible AI

Microsoft

What are the areas of AI and Machine Learning development whose potential social implications are currently being overlooked in the general discourse? Which areas get comparatively too much attention relative to their potential impact?

Among the many social impacts of AI, I believe Cultural Erasure is the most overlooked. Take language models as an example. These models can write text that is impressively similar to human-written prose. The internet-scraped data used to train these models, however, is predominantly in English; undermining global linguistic diversity. As a result, models' outputs are heavily influenced by western cultures that are dominant in the training data. This trend is also evident in multimodal models. For instance, when DALL-E 2, a recent text to image AI model from OpenAI, is prompted with "a wedding" or "a restaurant," it generates images aligned with western norms, such as white bridal dress or dining tables with western decorations.

Language is a reflection of a culture. As these AI models get increasingly adopted across the world, their effect on erasing local cultures becomes more pronounced. Lack of representation of low-resource languages – those for which data can not be easily scraped from the internet – in training data not only leads to lower quality of service for those who speak such languages but more importantly, would result in erasure of local traditions and cultures over time.

How can we equip policymakers to better keep pace with the speed of AI and Machine Learning development?

Prior to joining Microsoft, I was a Technology Policy Fellow working on emerging-tech public policy. I saw first-hand why public policy remains lagging behind the rapidly growing field of AI and how communities like All Tech Is Human can effectively bring change. I believe raising awareness, offering Solutions, and building coalitions are key. Let's talk about them.

First, reports. Policymakers cannot solve a problem if they don't know what it is and how it is impacting real people. Reports that highlight a specific issue area or summarize AI trends and developments are helpful resources to raise awareness of how AI technology touches the lives of people in different ways.

Second, policy proposals. Policymakers may not be experts in AI, but they do listen to experts (well, at least the ones who care!). Reporting on an issue area is helpful, but what is more likely to make a difference is an accompanying pragmatic proposal. That's what policymakers are looking for to find the best solution quickly and effectively.

Third, building coalitions. Policymakers want Solutions that have public support. A solution that comes from a unified voice has far greater impact. In my experience, a successful coalition is one that has collected opposing viewpoints and has adjusted the proposed solution to reflect diverse voices and opinions.

Raising awareness about issue areas and building Solutions to address them pave the way to better AI policy and equip policymakers with channels to keep pace and update the policies as the technology evolves.

What can we do to help technologists and policymakers from multiple intersectional social identities contribute to human-centered AI without feeling like they're the token minority?

Having minority voices at the table surely sheds light on blind spots and helps reduce some of the unconscious biases that creep into AI systems during design and development. However, no two individuals have identical worldviews. Setting number-based minority hiring targets and expecting a handful of individuals to represent a whole demographic group is a flawed expectation.

Instead of tokenizing individuals, we should opt for creating channels and mechanisms through which the wider community can be engaged in AI systems design. Take the U.S. government rulemaking process as an example. Agencies follow a Notice and Comment process, which requires the agency to notify the public of the proposed new or changed rule, and to accept public comments. The feedback is later analyzed to be reflected in the rules.

Feedback channels to collect perspectives from those impacted by the AI systems widens the reach of developers to diverse voices. Such feedback mechanisms would enable individuals from multiple intersectional social identities to raise concerns and have a say in how systems should be improved, without a few individuals feeling tokenized.

Looking at your crystal ball of a positive tech future. Let's say we have human-centered AI. What would humane AI look like? What would its pillars be?

I like to think of Humane AI as not necessarily human-centered, but rather, nature-centered AI. AI systems should not harm the natural world, animals, climate, or the ecosystem. I believe we should rethink our definitions of Humane AI to care for our planet and for nature, which so much of our being depends on.

Currently, human-centered AI principles revolve around transparency, accountability, privacy, fairness, inclusion, safety and security. A missing pillar here is the effect of AI on the environment and living beings. The environmental pillar in nature-centered AI would go beyond carbon emissions of training large AI models and considers AI ecosystem more holistically, from mining rare earth for chips and batteries, to building energy-efficient data centers, all the way to sustainable e-waste decommissioning.

How did you carve out your career in the Responsible Tech ecosystem?

I am a Computer Scientist by training. Prior to pivoting my career to Responsible Tech, I was a Lead Machine Learning Scientist, leading teams on AI algorithms for self-driving cars. As I continued to develop models and algorithms, I couldn't help but notice the gap between the world of technologists and the downstream effects of AI systems on real people. Technologists are often much focused on improving accuracy, not so much on mitigating downstream impacts. Observing this gap, I started looking out for opportunities to pivot my career. At this point, I didn't know there was an interdisciplinary field called Responsible Tech! I just knew what I wanted to do. I was fortunate to be selected for a fellowship that allowed me to bring my tech expertise to the world of public policy and in return learn about technology policy. During this fellowship, my passion for responsible tech only grew, and it was clear that Responsible AI was a natural choice for the next step in my career.

I am thrilled to be part of this community now, and look forward to connecting with as many fellow responsible tech enthusiasts!

> “
> **I think we need to start realising that innovation is not only about tactical innovation; innovation only happens when a lot of different expertises gets together.**
> ”

Gemma Galdon-Clavell

Founder & CEO

Eticas Consulting

How can we strengthen interdisciplinary engagement in technology development, deployment, governance, regulation and oversight?

I think we need to start realising that innovation is not only about tactical innovation; innovation only happens when a lot of different expertises gets together. If you want to build the building, you not only use plumbers, but you also need electricians, painters, architects. You need lots of different trades to undertake a complex task. We need to realise that technological innovation is still a very complex issue, and that the data that is being used by AI systems is personal data, so it is data from complex societies. We cannot make the most of that data if we only rely on engineers – it is like relying only on plumbers to build buildings. We are bound to build very weak constructions if we don't take everyone into account. Understanding that technological innovation is not only about the technical is the first step to building multidisciplinary efforts around the development, deployment and oversight of technical innovation.

What are the areas of AI and Machine Learning development whose potential social implications are currently being overlooked in the general discourse? Which areas get comparatively too much attention relative to their potential impact?

I think there is quite a lot of awareness of the social implications of new technologies, but I think we still don't know how to mitigate those impacts. We do realise that AI impacts privacy; we do realise that AI systems are somewhat changing democracy and how it intervenes in the public dividers. There is a lot of talk about content moderation and polarisation, and it doubles in social media. There is also some understanding about mental health issues that come with addiction to new technical platforms. Maybe there is less awareness about the fact that new technologies are imposing memory on all of us. For the first time in history, forgetting is more expensive – almost impossible – than remembering, whereas through our history it was the other way around.

We are beginning to grasp the deep consequences and implications of the generalisation of data and AI systems, but maybe we don't know yet how to build those concerns into practical things that will make technology safer. Like cars, which didn't come with seatbelts or speed limits – we added both to make the most of those technical systems – we still need to develop the seatbelts, the speed limits, the emission limits, the crossroads and the red lights for AI and new technologies. So we need to turn those concerns into practical protection tools.

How can we equip policymakers to better keep pace with the speed of AI and Machine Learning development?

One of the big problems that we have at the moment is that when most people talk about technology, they are actually talking about science fiction, and, unfortunately, that is also the case with politicians. Oftentimes the public and political debate is very far away, very remote, from the actual technical dilemmas and challenges that we are facing. So I think that we need to raise awareness among policymakers of what all these technologies can actually do and stop fulfilling what is called the immoral law. In technology we tend to overestimate the impact of future technologies and underestimate the current impact of existing technologies. We need to start protecting people from the current harms of technology and stop talking about robotics, the substitution of humans and general consciousness, because it's nowhere near and that may never happen. We are seeing bias and discrimination in decision making and lots of other harms that are happening right now. We are having very specific, concrete victims, after vulnerable populations, so I think that we need to equip policymakers with a better understanding of technology as the first step to kind of recuperate political spaces for strategic anticipation of social dynamics, and for definitions of the protections of society. We need to make sure that innovation does not infringe on fundamental rights and values.

What models and frameworks do you use to examine the assets and/or predict the impact of an emerging application of Machine Learning and AI? How well have these models worked for you?

We actually use our methods. We are global pioneers in algorithmic auditing and in the creation of responsible data infrastructures and responsible data policies. We take what we learn from working with cutting edge practitioners in the public and private sectors and turn that into methodologies that we can share with other actors and also release to the public. Unfortunately, because we are at the very beginning of designing what responsible innovation and responsible AI actually looks like, we cannot rely on the work of others. We need to do it ourselves. I often say that we are at the same stage as medicine was when you could buy cocaine in pharmacies. It took us a long time to come up with an ecosystem of regulation that ensures that anything sold in a pharmacy is safe for human consumption. I think that today we have the same problem – we have the cocaine right before us, but we still don't know how to make sure that it doesn't harm humans. We need to develop an ecosystem of regulation just like we developed ecosystems of regulation about medical innovation, mobility, food innovation. We need to make sure that the technology sector is held to the same standards as any other area of innovation.

How can individuals be empowered to know and protect their rights in the face of potential threats to privacy and civil liberties linked to use of AI?

Actually, I think it is quite unfair to put the responsibility of protection on the individuals. I think we need to understand that privacy is a collective right – that is, we not only make decisions about our data, we also make decisions about other people's data. In the same way that we don't ask anyone to have a degree in nutrition to enter a supermarket and make sure that they buy something that is good for them, we shouldn't be asking citizens to have a degree in engineering to know what they should be consuming when it comes to technology. The best thing that people can do is demand that we have ecosystems of protection, that the public administrations around us are able to reimagine what protection means in the digital domain and make sure that they create the tools to ensure that this protection is actually effective. Right now we have lots of regulations that aim to protect us but there is lack of reinforcement, so what we need is to demand that those principles and laws that we have – these concerns for fundamental rights, discrimination and privacy – are turned into specific practices and tools that mean that we are effectively protected whenever we use any kind of software or technological innovation.

What technical safeguards, oversight, and best-practices are needed to ensure safety by design and protection of human rights while mitigating harms?

The best tool, and the one that we are investing in, is algorithmic audits. We have found that through auditing algorithms we can identify when bias, harms and inefficiencies make it into a system and impact vulnerable populations. Once we have identified it, we can actually correct it. We can build better AI systems; we can build AI and decision making tools that do not discriminate against specific and vulnerable populations. We can make systems that ensure that we all have the same rights when facing a technical solution. For us, algorithmic auditing is a crucial piece in the ecosystem of protection that AI urgently needs and we have a lot of faith in it to become a standard and a mainstream practice. Five years from now we'll look back and we will be alarmed about how in 2022 we were implementing AI and advanced technical systems without proper audits.

What specific structural changes to incentives and business models are needed in order to prioritize user well being and human rights? What kinds of incentive systems (companies, people, regulators, societal) need to be tapped and how?

In fact, one of the reasons why Eticas exists is because we realised that, in order to change what happens, you need to press different buttons to create a window of opportunity for change. At the moment the market has no incentives to do things well, to act responsibly and to develop a responsible innovation, and this is probably the thing that needs to change the most. And I think the incentives only come from the fear of a fine; that is ok, but will not seduce the market into doing things better. I think we need to convince the market that a biased system is a bad system also for their own objectives. If a bank is discriminating against women, as they all do, that is not only a discriminatory decision but also a bad business decision. I think that we need to do a lot more in raising awareness in the private sector and making sure that they have the right incentives to change, and that we make sure that it's understood that AI development has huge potential but also huge risks that need to be mitigated. The best AI will only come about once we are able to build responsible AI, just like with cars. Cars are much better from having seatbelts and speed limits; AI will be much better from being audited and having other protection mechanisms that ensure that its use on human beings is safe.

Tell us about your role:

Dr. Gemma Galdon-Clavell is a leading voice on technology ethics and algorithmic accountability. She is the founder and CEO of Eticas, where she is responsible for leading the management, strategic direction and execution of Eticas' vision. Her multidisciplinary background in the social, ethical, and legal impact of data-intensive technology allows her and her team to design and implement practical Solutions to data protection, ethics, explainability, and bias challenges in AI. She has conceived and architected the Algorithmic Audit Framework which now serves as the foundation for Eticas flagship product, the Algorithmic Audit.

Her academic work has been published in Science and Public Policy, Information Polity, Ethics and Information Technology, Citizen Science Theory and Practice or Urban Studies, and in leading academic publishers such as Routledge, Springer, and Sage.

“

People speak about bias all the time. While this is a fair attack on the issues with the system, correcting for bias is not easy.

”

Hessie Jones

Venture Partner
MATR Ventures

Co-founding Member
MyData Canada

What are key factors in the way AI is developed/deployed that have the most impact on protecting human rights?

Datasets are always the starting point in developing models that are fair.

1. Are they fairly representative from the outset?
2. Also, to what extent does the data hold personal information?
3. And can this personal information be expunged or anonymized to minimize the risk of model memorization or model leakage?

ML models are meant to be developed with fairness in mind, however by simply leveraging models that have worked in the past, we perpetuate the implicit bias. In many cases, many models have been created with confirmation bias in mind i.e. developing an intended outcome in line with original objectives. This limits access for individuals who have not been represented in the models. 2) With regard to personal information, we need to ask ourselves, do we really need this data to improve our models OR do we look for patterns outside of personal information that can help us improve model efficacy? The harms that we witness today are that organizations have not thought about the implications of their current practices until they happen. By limiting the use of personal information we can move towards developing fairer models. Just because we can, it doesn't mean we should.

How can we strengthen interdisciplinary engagement in technology development, deployment, governance, regulation and oversight?

In any organization, transformation means that people, process and technology need to be in sync. For AI to be useful, it must also meet the needs of the organization. The directives need to be top-down but also embedded within each stakeholders's responsibility. The customers want to verify claims about the level of privacy protection with respect to their sensitive information. Regulators want to minimize the risk to individuals that lead to harmful outcomes. Academics want to conduct impartial research on the impacts associated with large-scale AI. AI developers want to verify that competitors will also follow best practices rather than cut corners to gain advantage. The business wants to mitigate risk to customers as well as reputational and financial risk. In today's environment, we aim for reproducibility to ensure that results are correct. This ensures transparency and confidence in understanding exactly what has been done, and reduces the risk of error. We aim for use cases, best practices and publicize results, models and code to verify claims. We can leverage Open communities like OpenMined, MS Seal, TF-Federated, TF-Encrypted to contribute use cases and learnings when it comes to data privacy, for example. We also leverage independent third party assessments: privacy impact assessments, algorithmic impact assessments and security audits in advance of mandated legislation.

What are the areas of AI and Machine Learning development whose potential social implications are currently being overlooked in the general discourse? Which areas get comparatively too much attention relative to their potential impact?

People speak about bias all the time. While this is a fair attack on the issues with the system, correcting for bias is not easy. A model that already produces the right results for an organization e.g. profitability or ideal customer need to be open to sub-par business performance for a time in order to ensure model fairness moving forward. Bias is a natural tendency for humans. Eliminating bias in models means wiping out many years of known approaches that have driven the 'right' or 'intended' business results. There needs to be more written about how we can, in a short span of time, correct the biases that business has historically inflicted in order to be profitable. More importantly, are businesses willing to do this? For example, data proxies for race or income have been the mainstay for many industries and government organizations: for hiring, loan adjudication, insurance. By eliminating their use will businesses be willing to live with the results given they are beholden to their shareholders.

What emerging regulatory frameworks are having the greatest impact on AI development at the present time?

In Canada the Office of the Privacy Commissioner (OPC) called for revised legislation that allows Canada to ensure the digital economy, in Canada, delivers on the promise of digital innovation by creating a robust and balanced regulatory framework that protects privacy and enables business and fosters digital innovation and responsible data sharing.

As well, the European Commission is proposing a new framework on artificial intelligence that will address the risks created by AI applications; identify the high-risk applications and set out requirements and obligations for providers of these high risk applications.

The recent Belgian Protection Authority has determined the IAB Europe data collection system which is behind 80% of apps and websites in the EU is unlawful. This is significant as it begins to pave the way for transformation within an industry that has historically targeted user interests, activities and behaviour without transparency and without user consent.

The final area that will have considerable impact is the current 9th circuit ruling that data scraping is not unlawful when it comes to public websites. This is the fight that Linkedin has been fighting for a number of years against a computer analytics company, which has been allowed access to Linkedin's public member profiles. Under the 9th Circuit Court of Appeals, anyone accessing data from public websites does not constitute "access without authorization." This will have implications that may perpetuate data collection methods like Clearview, especially in cases of secondary use.

How can we equip policymakers to better keep pace with the speed of AI and Machine Learning development?

Policymakers need to become part of open ecosystems of best practices: within law, within open innovation communities which enable discourse, and new ways of tackling current issues. In advance of any legislation, there are many cross-functional professional groups that are developing standards through IEEE, Kantara Initiatives etc. Organizations and Slack groups like MyData, HuggingFace, Aggregate Intellect, Responsible AI Institute, Montreal AI Ethics Institute engage in programs, workgroups and provide resources that allow participants to stay up to date. I've currently engaged with a cross-functional disciplinary team across industry in collaborating towards developing the PII policy framework, and developing the NLP specifications for detection, decisioning and transformation.

How does your team make decisions around integrating AI and Machine Learning into your product? How do you handle data collection, management, and model optimization? Who is at the table in these conversations? What are you optimizing for?

Big Science's year-long research workshop was launched to address the impact that Artificial Intelligence and Natural Language Processing (NLP) have on a powerful new artificial intelligence technology called large language models, and ultimately, on society. Working groups were formed, one of which was one to address the protection of Personally Identifiable Information (PII) in large datasets for large models. I was part of this working group. We ran a few parallel working groups 1) Data detection/remediation 2) Policy/definition of personal data (PII) 3) Probing global legislations to understand data privacy from the perspectives of data as property, the individual rights when it comes to their personal information.

Policy people, researchers, business, developers – technical and non-technical — were at the table to collaborate on principles of Anonymity/Privacy, Transparency, Autonomy, Inclusion/Representation with approaches to detect and remediate personal information in large datasets.

We quickly realized that the goal of an open privacy framework outgrew Big Science, and needed an existence of its own, so we gathered privacy practitioners across industry to help our goal of making privacy tools more readily accessible to all.

Our quest to make a difference in data privacy protection on a global scale continues as we continue the work to create a standard that protects personal information through AI/ML.

Looking at your crystal ball of a positive tech future. Let's say we have human-centered AI. What would humane AI look like? What would its pillars be?

We need two sides:

1) Promotion of privacy preserving free-flow of data across borders i.e. harmonization of privacy and data protection standards across borders
2) Technology aware digital ecosystems to promote transformation supported through regulation. This will allow low cost adoption by business; low friction human-centric products that will empower individuals while enabling responsible data sharing.

We have to ensure that this new privacy regime is comprehensive. This means:

1) Addressing the issues of data justice, and bias in algorithms and how people are represented as a result of their production of data. This includes algorithmic impact assessments and establishment of clear guidelines for fair explanation while protecting trade secrets.
2) Prioritizing children's privacy as a means keeping children safe but ensuring young people can thrive.
3) Emphasizing citizen control of their information as well as explicit positive consent.
4) Also extending this to applicable third party providers across jurisdictions and strengthening enforcement
5) Make data portability a reality. This is happening now with the use of verifiable credentials. We need to continue to encourage investment in innovation for responsible data practices and Solutions.
6) Giving businesses the tools to meet their core objectives while respecting Consent by Design and supplying them with incentives to support privacy preserving best practices.
7) Educate and empower individuals through healthy digital choices, Privacy best practices, and digital literacy.

> “Policymakers need to become part of open ecosystems of best practices: within law, within open innovation communities which enable discourse, and new ways of tackling current issues.”

Ivana Bartoletti

Global Chief Privacy Officer
Wipro

Founder
Women Leading in AI Network

What emerging regulatory frameworks are having the greatest impact on AI development at the present time?

AI systems do not exist in isolation and laws do apply. Human rights, privacy and data protection, liability and tort law. Nevertheless, a lot of the existing legislation may need updating to deal with, for example, algorithmic harm. This is because issues, such as algorithmic discrimination are very complex and simply saying that automated systems have a degree of human control will not be - as too many think - the ultimate solution. Algorithmic discrimination is subtle and individuals may be harmed without them even knowing, especially if the discrimination happens by proxy.

At this present time, we have seen Data Protection Authorities playing a key part in safeguarding individual rights when they are breached by AI systems. Key cases such as Deliveroo, Foodino, Uber or the recent fraud case in the Netherlands have highlighted issues around data accuracy, privacy by design, fairness as crucial themes to be upheld.
The EU AI Act will be very important as it will interact with the General Data Protection Regulation while reaching out even further as it will regulate AI impacting not only individual rights but also, contravening the fundamental values of the EU.

The FTC in the US is also being very vocal and countries, such as China, have put clear governance around algorithms. Singapore relies on the FAIT principles and legislation governing privacy, such as the Brazilian one, also deals with automated decision making.
Also, worth adding that there is a plethora of initiatives around accountability, the push for Algo Impact Assessment to be deployed in public administration (like it happens in Canada).

How can individuals be empowered to know and protect their rights in the face of potential threats to privacy and civil liberties linked to use of AI?

First, privacy by design is important as we need to be able to harness the value of data while safeguarding people's rights. So it is important to push for that, as well as for privacy enhancing technologies. Consumers and individuals can make choices of consumption based on that, and companies combining tech with privacy will then have a competitive advantage. Second, individuals should be able to access redress in case of harm generated by AI. My view is that we need Ombudsman bodies where individuals can appeal. Third, we need law enforcers with teeth. Companies should be forced to delete algorithms, share them or receive fines should be not complying with fairness, privacy and human rights by design principles. I think it will be very difficult for individuals to be responsible for tackling the harms happening to them. The ubiquity of data extraction and the opacity of AI systems mean that the burden cannot be on individuals having to claim their right - but on companies having to do the right thing. This is why it is essential in my view that we add a public dimension of privacy, looking at privacy as a common good and also looking at privacy for groups and not just individuals.

How can we ensure that the use of AI for "public safety" does not reinforce existing patterns of discrimination?

Partly, I think we need to ban systems like predictive policing. These contravene the assumption of innocence principle and end up being like self fulfilling prophecies that lock people into existing prejudices by wrapping controls and measures around them. It is crucial that we understand what the automation of existing stereotypes mean in law enforcement, policing and safety. It is like enabling self fulfilling prophecies at large scale which is the opposite of what both public policy and human life should be all about. The bottom line is that there may be systems that regardless of technical fixes may produce bias due to their use - one of them is surveillance technologies, such as facial recognition which will end up being deployed in certain areas rather than others.

What can we do to help technologists and policymakers from multiple intersectional social identities contribute to human-centered AI without feeling like they're the token minority?

Human centered AI is not a job of a token minority. It must become a board strategy and the board's responsibility. Every single day shows us the potential and the opportunities of AI alongside the inherent risks. This shows how the work of advocates, civic society and organisations is never undone. I think there is one thing that must be done, and that is to show the inextricable link between the success of a company and its approach to privacy and responsible technology. We also need to be able to embrace technology, like privacy preserving methodologies, to ensure we can combine data harnessing with human rights. This requires new skills and the ability to master a new language - activists need to speak the language of tech, too.

> “
> **For AI recommendation engines to be truly humane, however, creative strategies will need to be developed to help people discover aspirations and satisfactions that they could not have foreseen, instead of confining them within ways of living already established as pleasurable.**
> ”

James Brusseau

Philosophy professor

Pace University - New York City

What are the areas of AI and Machine Learning development whose potential social implications are currently being overlooked in the general discourse? Which areas get comparatively too much attention relative to their potential impact?

Too much attention focuses on artificial general intelligence (AGI) because the concept is senseless. Here's what I mean. My premise is that knowledge gets produced in plural ways: humans create understandings through art, through analogy, through analytic reasoning, and each method remains incomprehensible to the others. (You can prove an analytic conclusion, but it's senseless to prove an artwork.) This kind of split, it seems to me, also comes between humans and AI. Their two ways of producing knowledge are foreign to each other, irreconcilable.

If that's right, then there's no such thing as "general intelligence" because there isn't a single and enveloping kind of intelligence. Instead, there are multiple, localized productions, each with indigenous strengths and weaknesses. This explains why AI outperforms humans when it comes to calculating the shortest route to visit all the spots on a pub crawl, and humans do better when choosing which beer to drink at each stop. Differences like this will never be erased and can never be bridged, and not only because AI doesn't drink.

Then, at the other extreme, if it's true that AI produces knowledge not better or worse than humans, but differently, it follows that AI will probably fail comically at some tasks, but there may also emerge pockets of experience where AI advances are literally inconceivable for us, analogous to the way calculators astonish us in math. Especially in healthcare, this potential impact may be worth more attention than it currently receives.

How can we equip policymakers to better keep pace with the speed of AI and Machine Learning development?

Working with a collective of philosophers, computer scientists, and medical doctors organized from the Frankfurt Big Data Lab, I have directly experienced this problem. Our group performed ethics evaluations of AI-first startup companies, and each time we went back and forth for months gathering information and debating between disciplines, which is great intellectually, but it rendered the process impracticable in the real world. We were lucky to review three companies in a year.

So, the speed problem is real.

One solution is to employ natural language processing to scrape public information — government filings, white papers, watchdog statements, news reports— and then to fold the data into measurements of company performance in terms of privacy, fairness, explainability and similar.

Along with the Computer Science Department at the University of Trento, I'm working on just this project now. First, we need to capture the sense of an ethical principle in quantitative terms, along with a company's specific rendition of it. For example, what does the principle of privacy mean, and how does the specific company Facebook commit to privacy. Then we need to scan the online linguistic world to rate the company's performance against its own claims and, more broadly, to estimate performance on a continuum defined ideally by the privacy principle.

The work is challenging, but using AI to apply AI ethics to AI-intensive companies may be the only way to keep pace with the technology for policymakers and, more significantly in my view, also for those investors who detect an alignment between ethical and financial performance.

How do we ensure that new and emerging technologies and systems do not reproduce and augment existing inequalities?

In this narrow range I adhere to accelerationism: the only way out is through. The answer is not to slow or limit AI, but more, faster.

For example, despite burdensome regulations, innovative fintechs are filtering data to discover alternative metrics for lending risk. From there, non-predatory loans can be profitably extended to underserved communities because credit-worthy borrowers are identified among those who had previously been frustrated by conventional credit scoring.

Looking at your crystal ball of a positive tech future. Let's say we have human-centered AI. What would humane AI look like? What would its pillars be?

The deepest humanistic threats emerge not when AI is less than perfect, but more than perfect. Already in our everyday lives we catch glimpses of this when satisfactions are received even before we want them. It happens watching Netflix when the next film starts just as we conclude the previous: before there is a chance to ask whether we want to watch that particular next film, or even continue watching at all, it has already begun and we are captured.

There is a hint of oppression here – we are trapped in front of the screen – but it is odd because the origin is our own personal information. It is only because Netflix knows so much about our viewing patterns that we keep viewing. Of course the stakes are low on a lazy binge night, but the experience does reveal this shift: oppression used to be something that someone did to us, now, we do it to ourselves.

Humane AI will be directed at just this paradox of suffocation by our own personal information and contentment. In the area of predictive analytics, work is being done to address this, and of course it is always possible to feed users random suggestions in a crude attempt to break them out of their ruts. For AI recommendation engines to be truly humane, however, creative strategies will need to be developed to help people discover aspirations and satisfactions that they could not have foreseen, instead of confining them within ways of living already established as pleasurable.

“

The aim should be what we could call a human-centered market economy whose profitability increases if human rights are respected, and even better, promoted. Only what protects human rights should thus be profitable, or more profitable. This would obviously increase the incentive to adapt AI based economy in the service of humanity. Such a human rights-based data economy could become a sector of great innovation. Unnecessary data production and decentralisation of data control is a key to this.

”

Jan Juhani Steinmann

Dr.phil. / Lecturer in Philosophy
University of Vienna

Postdoc Researcher in Philosophy
Institut Catholique de Paris

What are key factors in the way AI is developed/deployed that have the most impact on protecting human rights?

There are two ways of responding to this: one that is intrinsic to AI and one that is critical of AI. To the first: The decisive factor is probably that the algorithms on which the specific forms of AI are based are as publicly accessible as possible or at least allow to be viewed by, for example, governmental bodies or, even better, by international bodies such as the UN or politically neutral, ethical control authorities (consisting in particular of philosophers, artists, and theologians). On the second: Overall, however, the influence of AI should be reduced, i.e. it should be avoided as far as possible that AI merges more and more with humans, i.e. that AI occupies human rationality (which is hermeneutically, sensitively, intuitively, existentially disposed) by the automated, algorithmic rationality of AI. The greatest influence on protecting the human being is therefore, the reduction of unnecessary use and influence of AI. AI should remain a complementary instrument that humans can use in a sovereign manner. This is in particular a pedagogical task.

How can individuals be empowered to know and protect their rights in the face of potential threats to privacy and civil liberties linked to use of AI?

The easiest solution is to reduce the sphere of influence of AI on our lives, i.e. to mitigate its unnecessary use and thus the constant production of controllable data. Simple as that. This requires more clarity and education in the area of data protection, along with political interventions to increase data protection and prevent data monopolies. Ideally, a globalised data society can establish a legally anchored decentralisation of data access, data ownership and data use, in which neither a national or international security service, nor companies or private individuals (except the data subjects themselves) have complete access to the available data. Such a decentralisation of data would increase the balance of controllability of data and thus protect individual human rights. To protect and ensure such interests Individuals can become politically involved, despite the utopian character of these interests.

How do we ensure that new and emerging technologies and systems do not reproduce and augment existing inequalities?

Unfortunately, I consider this illusory. Digitalisation, and with it all new and emerging technologies and systems, ultimately only reproduce the human condition. This, in turn, has always been predisposed to inequalities for a variety of ontological and anthropological reasons. It thus tends, entropically as it were, to increase, or at least, maintain inequalities. Digitalisation will not change this, but will rather promote inequalities in its own matter.
This pessimistic interpretation does, however, not hinder us to reduce inequalities and injustice.

A structural attempt in this regard would be to implement human rights as binding in the further development of technologies. For this to happen, human rights would have to be consistently extended to the rapidly evolving digital sphere, or digital human rights would have to be made binding in all digital contexts. As it is well known, first steps in this direction have already been taken (see, for example, the UN's "Digital Human Rights").

Overall, however, people must protect themselves best and reduce inequalities for themselves and in their environment. There are no legal bodies that relieve us of this personal responsibility.

What can we do to help technologists and policymakers from multiple intersectional social identities contribute to human-centered AI without feeling like they're the token minority?

Whether human-centered AI is even possible is, for several reasons, already open to debate. But let's assume that it is possible, which in my understanding ultimately means that however AI develops in the future (still far beyond deep learning or DAOs): Humans must always remain sovereign over AI: legally, politically, psychologically as well as in their individual, existential self-determination. This includes, above all, the protection of individual freedom and the potentials of manipulation through AI.

The discourse between technologists and policy makers as well as philosophers, politicians and lawyers (etc.) must therefore be conducted in such a way that the dangers and advantages of AI are made equally transparent. Mere growth and particular power interests (political or economic) must be put in the background in the greater service of humanity. Society as a whole must therefore strengthen the incentives for the responsible technologists and policy makers to work not for particular interests, but for collective interests that protect human rights. The properly informed and educated society should thus get behind them on the basis of these ideals.

What specific structural changes to incentives and business models are needed in order to prioritize user well being and human rights? What kinds of incentive systems (companies, people, regulators, societal) need to be tapped and how?

This is a very relevant question: The aim should be what we could call a human-centered market economy whose profitability increases if human rights are respected, and even better, promoted. Only what protects human rights should thus be profitable, or more profitable. This would obviously increase the incentive to adapt AI based economy in the service of humanity. Such a human rights-based data economy could become a sector of great innovation. Unnecessary data production and decentralisation of data control is a key to this. These incentives can be created on the one hand by the states, but also by ethically conscious corporations and, above all, the users and consumers themselves. Once again, information, education and ethical awareness are of great importance in this adaptation of business models towards human-centered AI.

> “
> **To ensure emerging technologies do not reproduce and amplify existing inequalities, designers, policy-makers, scholars and activists need to listen to the experiences, perspectives and expectations of those who are often excluded from these conversations, and yet often have the most to lose though the deployment of the technologies. Mere ‘consultation’ when the decisions are already made is not enough.**
> ”

Jeannie Marie Paterson

Professor of Law

Co-Director
Centre for AI and Digital Ethics
(CAIDE)

What are key factors in the way AI is developed/deployed that have the most impact on protecting human rights?

Techno-utopianism and techno-solutionism risk leading to harmful deployments of AI. These positions (falsely) frame AI as a solution to most social problems and creates problems for AI to solve, without a genuine inquiry into whether it is needed or even an effective solution. One response to countering these trends is capacity building. The aim should be more people, in particular those impacted by algorithmic applications, and more people from more places, and especially other than the global North, participating in discussion and debate about the kind of uses of AI that are acceptable and those that are not. We should also aim to build fine grained, nuanced and applied understandings of what responsible, human-centred AI might look like 'on the ground' and the role for human rights in informing these uses.

What are the areas of AI and Machine Learning development whose potential social implications are currently being overlooked in the general discourse? Which areas get comparatively too much attention relative to their potential impact?

Government and business discourse about AI tends to promote the futuristic benefits of the technology. Sometimes this has no grounding in reality. Significantly, it minimises the risks associated with AI systems such as algorithmic decision-making or surveillance technologies. In other words the possibilities of AI get too much attention, and there is insufficient attention on the risks of AI technologies to privacy, autonomy, fairness, and equity. Additionally there is a tendency to refer to principles of AI ethics and ethical auditing as Solutions to those risks without any follow-up discussion of what those interventions might actually require.

At CAIDE, we would like to see more focus on the risks of AI, without that focus being dismissed as standing in the way of innovation. We would also like to see genuine engagement with the key principles of AI ethics that interrogates what those principles mean when applied to specific technologies and in particular contexts. We would like greater real focus on accountability and contestability. We would like greater engagement with human rights perspectives on AI, such as discrimination and disability rights, as well as issues of data sovereignty and access to technology. We would also like to see more focus on the environment and on non-human life in debates about responsible AI.

How can we equip policymakers to better keep pace with the speed of AI and Machine Learning development?

To equip policy makers to better keep pace with the speed of AI and ML development, first, we need policymakers to engage with, listen to and represent different cohorts of people, particularly otherwise marginalised cohorts, and skills, including people with a background in humanities. Secondly, we need to demystify AI so as to allow robust discussions about where, when, and how (if at all) AI should be deployed. Thirdly, we need to move beyond a focus on 'innovation' at all costs. Smaller aspirations about what technology can and should do might in many cases be better for humanity and for the environment.

What models and frameworks do you use to examine the assets and/or predict the impact of an emerging application of Machine Learning and AI? How well have these models worked for you?

The Centre for AI and Digital Ethics is a cross disciplinary research team. We try to model in our deliberations about research, policy and teaching on the ideals that we propound for responsible AI. We use practical ethics, philosophy and models of ethical AI, such as set out by the OECD. Significantly, we aim to apply those frameworks to particular contexts and uses of AI to make them practical and operational. We aspire to listen to the experiences, perspectives and expectations of people likely to be most affected by the deployment of algorithmic decision-making, surveillance technologies and other applications of AI. We collaborate with other groups working to understand the role of aspirations for responsible AI in society. We believe in the power or art and the humanities to provoke conversations about AI and human rights. We work with climate and sustainable development teams to think about how aspirations for a tech utopian world might be moderated by a respect for the planet, the environment and all living things.

How can individuals be empowered to know and protect their rights in the face of potential threats to privacy and civil liberties linked to use of AI?

This is a tricky question as many uses of technology are designed or have the effect of disempowering individuals and obscuring new hierarchies of power and influence. At CAIDE we think education is important in empowering individuals. We also believe that the ongoing work of activists, artists, scholars and policy makers is essential to ensure that the risks to privacy and civil liberties inherent in uses of AI, such as in algorithmic decision-making about rights or the widespread use of surveillance technologies, are acknowledged and addressed. Ideally, human rights activists and ethicists need to work with computer and data scientists to ensure that ‘safety by design’ and ‘human rights by design’ are not used as mere rhetoric to justify harmful technologies. Additionally, we think the geopolitical implications of AI and access to technologies should be included as part of the discussion about responsible AI.

How do we ensure that new and emerging technologies and systems do not reproduce and augment existing inequalities?

To ensure emerging technologies do not reproduce and amplify existing inequalities, designers, policy-makers, scholars and activists need to listen to the experiences, perspectives and expectations of those who are often excluded from these conversations, and yet often have the most to lose though the deployment of the technologies. Mere ‘consultation’ when the decisions are already made is not enough.

What specific structural changes to incentives and business models are needed in order to prioritise user well being and human rights? What kinds of incentive systems (companies, people, regulators, societal) need to be tapped and how?

To prioritise user well-being and human rights in AI, first, we need stronger law and law enforcement to centre the importance of responsible AI. Secondly, we also need more public focus on what responsible AI might look like, and the benefits of this approach. The preoccupation with innovation is harmful in the long term. Third, we need international collaboration on these important issues. Fourthly, we need to ensure that the benefits of AI are not sequestered in the global North. All countries should participate in the advantages of new and emerging technologies, and the discussions about what responsible AI should look like. Finally, we need to collaborate with climate activists - responsible AI must be climate respecting.

“

Embracing responsible principles—in design, in data, in algorithms—should be perceived not as an extra step before a product ships or goes live, but as a long-term strategy and fundamental shift in culture that builds trust across all stakeholders.

”

Jenna Hong

Product Manager

Microsoft AI Development and Acceleration

What are key factors in the way AI is developed/deployed that have the most impact on protecting human rights?

Every part of the AI product lifecycle has an impact. As such, Responsible AI is a necessary piece to the puzzle - to be addressed at each step and built into the core of development. Products are meant to be user-centered, so protecting human rights and designing for fairness are indisputable parts of ensuring that technology is built to enhance human experiences, address user needs, and empower people to achieve more. Embracing responsible principles - in design, in data, in algorithms - should be perceived not as an extra step before a product ships or goes live, but as a long-term strategy and fundamental shift in culture that builds trust across all stakeholders. It can also play a key part in boosting brand power and recruiting talent.

In today's Age of AI, we also need to keep pace with innovation; with new technology comes new application spaces and consequences we haven't faced before. It's important to ensure that we're tackling issues proactively, and continuously. This might mean acknowledging that we don't know what might happen - and that we can't fully guarantee, or "check off" Responsible AI - so it becomes crucial to design processes that help us better understand the implications of a piece of technology, whether that's ways to mitigate harm or track metrics that keep us better informed.

How can we strengthen interdisciplinary engagement in technology development, deployment, governance, regulation and oversight?

Oh man, I really believe in the power of multidisciplinary thought. Diversity breeds innovation, and even more so if AI is driven by thinkers from all different backgrounds and identities - not just those who are elbow-deep in optimization and experimentation, but also the social scientists, designers, linguists, philosophers, economists, educators, policymakers, artists, etc. that bring valuable perspectives to things like how AI is regulated and what human-AI interaction looks like. The table should be a wide one.

AI systems are also not neutral - they are sociotechnical in nature, and reflect the assumptions, priorities, and values of the people involved in their development. By leveraging a wide array of experiences and viewpoints, subject matter experts can co-create systems that behave in trustworthy, fair ways for all users.

This kind of engagement is strengthened through amplifying and multiplying - those who come together at the heart of Responsible AI development should be supported by others, who can then take insights back to their own communities and find meaningful parallels in their work.

I think it's also important that - to quote High School Musical - we're all in this together. From big corporations to smaller groups, it's inspiring to see people get together to tackle some of these complex sociotechnical problems - but it's also crucial that we're sharing notes. If one organization's principles are wildly different from another's, outcomes start to deviate. By sharing learnings and best practices across efforts, we enrich our understanding of what Responsible Tech means and becomes.

How can individuals be empowered to know and protect their rights in the face of potential threats to privacy and civil liberties linked to use of AI?

AI literacy is the new digital literacy. Artificial Intelligence is decidedly present in our lives and work, so there's a growing need to educate the public about its capabilities and implications. I think it starts with democratizing these larger ideas around Responsible AI, and lowering the barrier by employing principles of transparency and accessibility early on, so that people can feel empowered to learn and advocate for themselves. Just as Responsible AI can be baked into every step of the product lifecycle, it can be embedded within core curricula. Not just for students, but for all backgrounds and experience levels. And it doesn't really stop at a single point - we're constantly rethinking how to prepare crowds to be successful in an increasingly AI-powered society.

Looking at your crystal ball of a positive tech future. Let's say we have human-centered AI. What would humane AI look like? What would its pillars be?

My pillars revolve around storytelling, multidisciplinary thought, and inclusive design - a positive tech future starts with people that feel empowered to ideate and innovate in a way that benefits society and minimizes harm. What I really appreciate about those in the human-centered AI space is that they find their way here by caring, and by meandering. No two people have the same journey to intersections like these, but the passion is real and it is shared. I've also seen interest at the intersection of technology, policy, ethics, etc. blow up in the past few years, and I believe it will continue. Responsible Tech should be driven by the brilliant, dedicated individuals who have found themselves here, and carried forward by those who listen to their causes and promote their ideas. My hope is that more and more people will recognize and commit to Responsible Tech - not just the superstar visionaries that are leading the cause.

For the general public, I look towards a future where the perception of AI is not rooted in fear. The narratives about AI taking over our jobs or even the world can be fun, but it's a common image - and one we have to contextualize. People are creative, nuanced, and irreplaceable. Jobs might shift, but maybe so that we can tap into more of that unique human intellect.

@ students reading this - keep going! Ask hard questions, talk to people that inspire you, and form a perspective. You have more influence than you think.

Tell us about your role:

I'm a product manager at the Microsoft AI Development and Acceleration Program (MAIDAP). As part of the New England Research and Development center, our projects span across Microsoft's product groups at the heart of AI/ML innovation such as Office, Devices, Cloud, Gaming, and others. I lead product vision for our projects, which take on a bunch of different forms - incubation efforts, feature work, product development - to solve a diverse set of user problems. Our teams are highly collaborative and interdisciplinary, which I love, and I've been lucky to be able to learn about and advocate for embedding Responsible AI practices into all different points of the product development lifecycle.

How did you carve out your career in the Responsible Tech ecosystem?

As a double major in cognitive science + computer science, I was naturally drawn to, and inspired by, the ideas that fuel human-centered AI. ‘Human-centered’ in every sense of the word: I became fascinated in research around the science of human intelligence (and how it differs from machine intelligence), and it ended up stretching my mind to the other end of the spectrum, around applying AI for social good. It took me to gain interest in things like brain development, story understanding, computer vision… etc. I loved the way AI traced roots to psychology and behavioral science, and extended out to spaces like ethics and social science.

My research reflected this range, from neuroscience experiments in emotional contagion to AR spaces for perceptual memory. I then leaned into the idea of democratizing AI education for K-12 students by developing curricula with researchers at the MIT Media Lab.

I carried this passion to my job at MAIDAP; as I was rotating across products as a PM, I sought out ways to be engaged in the larger Responsible AI ecosystem. Microsoft has a vibrant mesh of Responsible AI practitioners and researchers, and it really energized me. I also looked to external thought leaders, which brought me to the ATIH community.

My mentors often tell me: create the job you want. So “carve out my career in the ecosystem” is actually pretty accurate - I took my dispersed interests and sparked stimulating conversations, and it’s led me to even more of them.

“

The legislative space has changed a lot over the last few years, and regulators are increasingly stepping up to take action. But we also need to ensure that this is reflected in international norms as well, and applied consistently throughout the world.

”

Jennifer Easterday

Executive Director and Co-Founder

JustPeace Labs

How can we strengthen interdisciplinary engagement in technology development, deployment, governance, regulation and oversight?

This is a question that is very top of mind for me. It is critical that we work across sectors and disciplines to tackle the challenges of AI. I think this requires systems-based thinking about the impact of AI in our world, horizontal and meaningful engagement with groups who are most vulnerable to the negative impacts of AI, and more opportunities for multi-stakeholder collaboration. This requires a lot of awareness raising across disciplines and the difficult work of translation of ideas and approaches. It also requires vulnerability, embracing a learning mentality, and allowing space for disagreement. And, speaking from the civil society perspective, institutional support and funding to support such future-oriented programming.

What are the areas of AI and Machine Learning development whose potential social implications are currently being overlooked in the general discourse? Which areas get comparatively too much attention relative to their potential impact?

Very specifically, the impact of AI and ML on the exacerbation of violent conflict is being overlooked. Many take a reactive perspective to conflict and crisis situations, and/or look at the more commonly discussed topics of mis/dis-information and content moderation on conflict or AI in tools of war. But we need to think beyond this to look at how AI is impacting conflict (during all stages—including fragility, pre-conflict, inter-communal violence, full-blown armed conflict, and post-conflict), mapping users of AI, data and training sets, algorithmic accountability, and conducting additional research on connections between AI/ML and conflict situations. Any of the “typical” risks posed by AI are proportionately more serious and non-remedial as they move into situations of conflict. They urgently demand more attention.

How do we ensure that new and emerging technologies and systems do not reproduce and augment existing inequalities?

I think these are the hardest questions of our time. “How” to do this is really a moving target, because we often create potential approaches to reducing these risks but technology outpaces their implementation. As a lawyer, I tend to think that impactful, smart, and well-implemented regulation will be a significant tool here. And we are getting there – the legislative space has changed a lot over the last few years, and regulators are increasingly stepping up to take action. But we also need to ensure that this is reflected in international norms as well, and applied consistently throughout the world. There needs to be real accountability for corporations with a global footprint.

At the same time, and while we wait for regulations to catch up, there is a significant need to work WITH corporations and technology developers to understand the risks of their products and shift thinking about prioritizing profit-motives above risks to users, communities, and the world. And this requires a systems-based approach, working with multiple stakeholders, educating users and consumers, developers, and civil society about how to do this in a meaningful way.

How did you carve out your career in the Responsible Tech ecosystem?

It was a long, meandering path! I started working in international criminal law and transitional justice, and then did doctoral research on challenges of peace-building. At the same time, I consulted on a variety of projects looking at the use of technology for atrocity crimes accountability. It all helped me realize that there were a lot of emerging risks when transferring tech tools to fragile contexts that would require a multi-sector, inclusive approach to mitigating the risks of technology while taking advantage of the many benefits that technology offers. Getting here involved a lot of feelings of imposter syndrome, pursuing fascinating but ultimately unproductive projects and lines of research, and meeting a lot of inspiring and brilliant people who motivated me to keep breaking ground.

“

Inclusion requires, at least, proactivity, resource allocation, and continuous dialogue. I would say that, in the context of inequalities and underfunded activities outside of the private sector, compensating people for their time and expertise is a fundamental step toward building more inclusive spaces in human-centred AI.

”

Jess Reia (they/them)

Assistant Professor of Data Science
University of Virginia

What are key factors in the way AI is developed/deployed that have the most impact on protecting human rights?

There are several factors to guide us in the pursuit of the protection of human rights in AI and ML. Digital rights, civic engagement mechanisms, bottom-up policy design, interdisciplinarity, and multistakeholderism are critical factors in this equation. Another key factor is to ask ourselves whose voices are (or are not) heard. Are we considering the impact on historically marginalized communities when developing and deploying AI?

One of my roles as an educator is to ask difficult questions and try to figure out answers to them with my students. Data science is a trendy career right now, and I see a lot of potential in how we shape the future of the field towards responsibility and a human rights-centred approach. Given my experience working with the deployment of data-centric systems in Latin America and Canada, I try to emphasize the impacts of AI beyond borders–the transnational nature of certain technologies has a significant impact on society, especially when they are developed in the Global North and deployed in the Global South. Given the differences in context, regulatory frameworks, democratic or authoritarian governments, the robustness of civil society, and local history, AI can exacerbate inequalities in unexpected ways. Hopefully, the transnational aspect can also help us to share best practices, think about the problems, and collectively create Solutions to advance a public interest approach to AI while protecting humans and fundamental rights.

How can we strengthen interdisciplinary engagement in technology development, deployment, governance, regulation and oversight?

As someone with an unusual background, I see interdisciplinarity as a crucial approach to public interest technology. Being part of various departments (Public Policy, Media Studies, Law, and Data Science) allowed me to learn different perspectives, contest narrow views, and build bridges between people with unique trajectories. I was lucky to navigate spaces of policy and lawmaking and to collaborate with advocates, scholars and activists in several countries. However, interdisciplinary engagement with technology is often hard to achieve, requiring resources and continuous effort. It requires a long-term commitment, as well as openness and strong teamwork.

Data practitioners must consider ethical principles and the implications of their work on society; and regulators should understand the technical and non-technical, computational and non-computational aspects of the matter that needs regulation. Moreover, proper governance mechanisms should consider numerous variables and interests in a broad spectrum; responsible oversight needs to understand what is going wrong, what is causing harm and if people and organizations can be held accountable. The complexity of this challenge should not be impeditive to further develop mechanisms to foster interdisciplinary engagement. Collaboration and multiple perspectives help us tackle the ever-changing issues that developing and deploying AI might cause. I also believe in the role of multistakeholderism in this process, given that we are dealing with fast-paced innovation that affects people and communities differently–ask who does not have a seat or a say at the decision-making table and work to change skewed power dynamics.

How can we equip policymakers to better keep pace with the speed of AI and Machine Learning development?

Some mechanisms and approaches can be useful in equipping policymakers (and civil servants) to keep pace with AI and ML development. First, capacity-building programs that build upon their existing knowledge and expertise while filling in gaps and providing up-to-date, critical tools to deal with the ever-changing AI and ML development. Second, partnerships and spaces to exchange experiences with local and regional civil society organizations, universities, think tanks, and the private sector–without losing sight of public interest. Third, these initiatives will require financial sustainability. It is important to consolidate consistent, inclusive, and available programs for more than one election cycle. Finding ways to sustain such efforts beyond industry funding, without strings attached and focused on public interest technology is challenging.

Doing fieldwork in Brazil, my team and I discovered a worrisome trend of industry-led policymaking for data-centric initiatives in cities that rarely engage with residents and disregard human rights and current debates around data protection. This is the kind of development we should fight against. I believe in a multistakeholder approach, allied to leveraging resources and knowledge already available, as well as relying on international networks of collaboration. Most policymakers and civil servants are eager to learn and tackle issues related to AI and ML but still struggle to find the time, resources and information needed. We need to share our best practices and our failures openly. As scholars and advocates of public interest technology, we can help change this narrative.

What can we do to help technologists and policymakers from multiple intersectional social identities contribute to human-centered AI without feeling like they're the token minority?

Inclusion requires, at least, proactivity, resource allocation, and continuous dialogue. I would say that, in the context of inequalities and underfunded activities outside of the private sector, compensating people for their time and expertise is a fundamental step toward building more inclusive spaces in human-centred AI. Let's pay people for their contributions. Inclusion also implies humility and a commitment to improvement. We will make mistakes and must learn from them, mitigate harm, and actively work to do better. We have so much to gain from learning together. Another key aspect of this process is a focus on equity. We must strive to create spaces where multiple intersectional social identities can feel welcome, heard, and valued. Respect for what people are bringing to the conversation is as important as a willingness to listen and make their voices heard. Human-centred AI needs to be built on diverse and meaningful participation from various communities.

Tell us about your role:

I am an Assistant Professor of Data Science at the University of Virginia. My job involves teaching current and future data scientists, conducting research in Responsible Data Science and Data Justice and working to build bridges between various communities and organizations. I am a free culture activist, having openness, public interest tech, and social justice at the core of my everyday work.

Prior to joining the School of Data Science at UVA, I worked at McGill University as a Mellon Fellow in Canada, an AI Fellow at New Cities, and Professor at the Center for Technology and Society at FGV Law School in Rio de Janeiro. My work focused on the power imbalances between people, technologies and spaces, especially in the Americas.

As a non-binary, Latinx person in Data Science, living and working in the Global North, a significant part of my work needs to be focused on holding space for many voices that are still invisible in this space and amplifying the amazing work being done, relentlessly, to create a better future for us.

“

In the domain of machine learning, ML models are a reflection of the data that is used to train them. If we can improve our data problem, we can improve our ML problem.

”

Jessie J. Smith

PhD Student

University Colorado - Boulder

How can we strengthen interdisciplinary engagement in technology development, deployment, governance, regulation and oversight?

Stop using buzzwords and jargon to gatekeep the tech space! Knowledge is power and it is unfortunately wielded in the hands of few individuals in the AI discipline. Anyone can code, anyone can learn math, and anyone can become a data scientist. If we all begin to work together to educate ourselves and each other about how technology works, we can gain a collective power to independently audit the technological systems that we interact with every day. With increased awareness about how the technology that we use works, we have an increased power to create positive change.

What models and frameworks do you use to examine the assets and/or predict the impact of an emerging application of Machine Learning and AI? How well have these models worked for you?

In the past I have worked in several industry settings with teams to create a generalizable framework that can be used to assess if a machine learning application and its outputs are aligned with the values and principles of an organization. I've also adopted an audit framework that was originally introduced in the peer-reviewed publication: "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing." In my work, I have emphasized the importance of working with stakeholders (data scientists who implement the technology and real users who are impacted by the technology) when evaluating an ML model's potential for harm and impact. After assessing qualitatively what the potential impact of a model could be, quantitative statistical tests can be run to evaluate how well the model is actually performing in those areas. If any red flags are raised, then discussions with the data scientists who created the model can begin, with mitigation plans to improve the model's performance in key areas to alleviate potential harms that were discovered.

Finally, I also led the team that created REAL ML, a tool to help machine learning researchers appropriately engage with limitations in their work and to describe those limitations publicly in their research papers. This helps uncover potential areas of impact before novel machine learning approaches are adapted and implemented in an industry setting: https://github.com/jesmith14/REAL-ML.

How do we ensure that new and emerging technologies and systems do not reproduce and augment existing inequalities?

In the domain of machine learning, ML models are a reflection of the data that is used to train them. If we can improve our data problem, we can improve our ML problem. Some may assume that improving datasets means "collecting more data," however, there are many instances where collecting more data could be more harmful than helpful. The key is to collect quality data that is more representative of all users who will interact with the system – not just the "average" user. Data augmentation techniques can be helpful for balancing datasets across demographics such as ethnicity and gender, but they don't always improve the accuracy of a system for users whose identities aren't captured by the data distribution. Additionally, data annotation tasks are often written by few people with their own biases and lived experiences.

Even if we could build better annotation guidelines to create more robust datasets, our ML systems might still perpetuate inequalities. Data taken from the real world reflects the inequality of our history and present. Methods of evaluating bias in datasets and ML model outputs can help us measure and understand when these kinds of biases exist in our systems, such as recognizing if a natural language processing model associates women with stereotypically feminine tasks, or vice versa. The interventions that we can take after recognizing bias are context dependent, and no solution will ever be the perfect solution – but awareness and intentionality is key to fostering responsibility for the impacts of these systems.

Looking at your crystal ball of a positive tech future. Let's say we have human-centered AI. What would humane AI look like? What would its pillars be?

Human-centered AI has previously been defined as a system that understands its users while also helping users understand the system. I believe that one of the core pillars of human-centered AI is transparency – an ability for users to understand how the technology works, how it impacts them, how it was designed, how their actions influence its inputs and outputs, and how they can use this knowledge to gain better agency in their digital lives. As I've said before, there are no perfect Solutions, but there are informed, responsible, and transparent Solutions.

How did you carve out your career in the Responsible Tech ecosystem?

In the last year of my Software Engineering degree, I began to notice the disparity between the tools I was being taught to build and the impacts that those tools might have on real people. I started to do independent research on the topic of technology ethics, and then I approached several professors on campus to start working on machine learning ethics research projects. In my final year of my undergrad, I decided to apply to grad school to pursue these topics further, and the rest is history!

“

We must not take for granted the existence of one open, global internet but should look for opportunities to reinforce its existence and universality, and of course to promote global access. We cannot afford to see the rise of a splinternet, with a second internet that offers cost-efficiency in return for mass surveillance and lower human rights standards.

”

Kate Jones

Associate Fellow

Chatham House

How can we strengthen interdisciplinary engagement in technology development, deployment, governance, regulation and oversight?

The imbalance of scale, budget, speed and technical knowledge between the private and public sectors means that it's currently really difficult for the public sector to exercise governance over the private sector. Ideally, there would be multi-stakeholder initiatives, starting in the US and extending internationally, to strengthen governance and oversight. Such initiatives would need to be funded by, and perhaps led by, big tech, and should have the genuine aim of achieving governance, regulation and oversight that facilitates pursuit of profit while also taking full account of a broad set of public interest considerations. It would be fantastic to see big tech pick up the baton of establishing multi-stakeholder governance forums or bodies, with the endorsement of government.

What are the areas of AI and Machine Learning development whose potential social implications are currently being overlooked in the general discourse? Which areas get comparatively too much attention relative to their potential impact?

We need much more attention to the potential impact of AI on our minds, our decision-making processes, our values and ethics. And we need to be vigilant to ensure that AI benefits everyone - for example, by contributing to implementation of the Sustainable Development Goals - rather than simply benefiting the world's well-off.

Identification of social implications is only the first step: for example, discrimination in AI has now been widely identified, but there is a very long journey to remedy it. Beyond discrimination, we should be more aware of the risk of deploying tech that essentially relies on comparison and stereotyping, in a society where the better approach has been seen to be assessment of each individual's needs and potential according to his or her own characteristics and merits.

And finally, for all the talk of privacy protection, in my view we haven't yet found the right model that will allow for the large-scale data sharing that enables AI, while increasing personal protection and control over individual data. Individuals ought to have much more clarity about the data held and shared about them, and arguably a stake in its value.

What emerging regulatory frameworks are having the greatest impact on AI development at the present time?

At present, there are many competing standards of AI ethics, each slightly different and employing concepts (like “transparency”, “accountability”) in different ways. Usually it is difficult to tell whether companies have complied with their own standards, and there’s no remedy for breach. The introduction of algorithmic impact assessment (AIA) and certification of high-risk AI through the EU AI Act are positive steps, as is the UK’s work on development of an AI assurance industry. But it is key to ensure that the underlying technical standards are robust, and made with regard to human rights, so as to avoid the risk of ‘ethics-washing’.

International human rights law would help us to adopt consistent standards. There is a lack of understanding of the subtlety of human rights law, and its ability to weigh competing rights and interests using concepts of necessity and proportionality.

International human rights law is a hard-won consensus on the protection of human life, dignity and equality that has been in place for many years, is widely understood around the world, and already benefits from impact assessment and accountability processes both national and international. It should play a far greater role in the development of both AIA and technical standards.

Those championing human rights in AI ought to focus on raising awareness of human rights among both the tech industry and policy makers. Human rights ought to be mainstream. Too often it’s being discussed by specialists, in a different room from the decision-making on tech development and AI ethics.

How can individuals be empowered to know and protect their rights in the face of potential threats to privacy and civil liberties linked to use of AI?

This is a crucial point. Individuals can only “know and protect their rights” if they do in fact have enforceable rights as a matter of law or contract. Vague commitments to AI ethics often do not amount to legally enforceable contractual rights.

All companies have a responsibility to respect rights in their activities, and governments have a duty to protect their residents from corporate abuses of rights.In Europe, all individuals have legally enforceable rights against their government, under the ECHR. If an individual’s rights - to privacy, equality, freedom of speech, freedom of thought etc. - are not adequately protected, she can bring a claim against her government. We are right at the beginning of AI litigation and the parameters of rights in the AI domain are still being worked out (see for example the Bridges case in the UK, on use of facial recognition technology by a police service).

We need much more clarity for individuals as to what rights they have, and whether those rights have been met. This entails more discussion and litigation as to the parameters of existing rights in AI; and much clearer commitments from companies (including through algorithmic impact assessment and audit) on how they are giving effect to human rights in their policies.

What specific structural changes to incentives and business models are needed in order to prioritize user well being and human rights? What kinds of incentive systems (companies, people, regulators, societal) need to be tapped and how?

First, we must not take for granted the existence of one open, global internet but should look for opportunities to reinforce its existence and universality, and of course to promote global access. We cannot afford to see the rise of a splinternet, with a second internet that offers cost-efficiency in return for mass surveillance and lower human rights standards.

Second, we ought to get our house in order as regards surveillance, in particular by taking a long hard look at the monetisation of data. Data monetisation has already been shown to incentivise poor practices, from attention-maximising design to recommender system prioritisation of emotive and divisive social media posts. A new model is needed that allows for data collection and sharing where needed to facilitate AI, yet protects the rights of the individual not to live a life surveilled, i.e. protects personal and sensitive data. This may mean changing the cost/profit base for some tech companies.

Looking at your crystal ball of a positive tech future. Let's say we have human-centered AI. What would humane AI look like? What would its pillars be?

Humane AI would 'level up' society (to coin a current British phrase). It would enable delivery of the Sustainable Development Goals. It would enable the delivery of aid, goods and services to those who need them most. It would make life more efficient and safer for the benefit of every individual.

It is difficult to see how humane AI will reach these goals while it is very largely being developed and implemented by the for-profit sector. Getting there will require a new approach, just as a new approach to social welfare was developed in the mid-20th century.

How did you carve out your career in the Responsible Tech ecosystem?

My background is as a diplomat, lawyer and academic. I began my career as a British solicitor with a City law firm. For many years I was a legal adviser and diplomat with the UK's Foreign, Commonwealth and Development Office, with a particular focus on international human rights law, including postings at the British representations to the United Nations in Geneva and the Council of Europe in Strasbourg. I also spent several years as the Director of the Diplomatic Studies Programme and member of the Law Faculty at the University of Oxford, where I taught both diplomatic practice and public international law while pursuing my research interest in responsible tech and human rights.

“

Emerging technologies are virtually guaranteed to reproduce and augment existing inequalities. Our response should be to push for rules that mandate redistributive justice when autonomous systems inevitably afflict the afflicted and comfort the comfortable.

”

Kevin Klyman

Lead Technology Researcher, Avoiding Great Power War Project

Harvard’s Belfer Center for Science and International Affairs

How can we equip policymakers to better keep pace with the speed of AI and Machine Learning development?

First, policymakers need to be empowered to exercise their democratically-mandated regulatory functions. This entails hiring staff who are experts in AI, having resources to commission third party investigations, and carving out time to ask questions about confusing technical concepts. Policymakers who spend much of their time fundraising or managing their bureaucratic fiefdoms will never understand homomorphic encryption.

Second, policymakers need to be trained to distrust talking points from large technology companies. In the United States, Facebook and Amazon spend more than any other company on lobbying; in Europe, Apple, Google, and Microsoft are also leading spenders. This lobbying is not benign, it serves to confuse policymakers by convincing them that AI systems are unlikely to have unintended consequences.

Third, experts should collaborate and speak with a common voice to the extent possible. For example, AI ethics researchers have constructed over 100 different frameworks for ethical AI. If experts instead worked together to push lawmakers in a single province or city to adopt an ethics-based data governance framework it just might work.

How can individuals be empowered to know and protect their rights in the face of potential threats to privacy and civil liberties linked to use of AI?

A positive step is the proliferation of privacy preserving tools and platforms. Encrypted password managers browsers, search engines like DuckDuckGo, and the Brave browser are now widely available. However, these tools are time consuming and often deliver worse performance than surveillance-oriented alternatives. While public education is valuable (and data privacy day on January 28th is loads of fun), a more effective approach is pushing for regulations that require AI systems to be built with privacy by design. Supporting organizations that aggregate and act on user complaints related to algorithmic harm is another necessary tactic.

Media outlets that receive funding from large corporations should also be held accountable for their unwillingness to question how AI systems might restrict civil liberties. For instance, newspapers gave countless glowing accounts of the ways in which algorithmic tools were used to mitigate the pandemic, but paid little attention to the fact that government spending on such tools traded off with investment in more effective low-tech pandemic Solutions.

How do we ensure that new and emerging technologies and systems do not reproduce and augment existing inequalities?

Disruptive new technologies should be governed by democratic institutions. Well funded and independent data protection offices, surveillance oversight boards, and competition authorities are the best institutions we currently have to prevent multinational technology companies from using their gargantuan market power to manipulate governments as though they are vassal states.

That being said, emerging technologies are virtually guaranteed to reproduce and augment existing inequalities. Our response should be to push for rules that mandate redistributive justice when autonomous systems inevitably afflict the afflicted and comfort the comfortable.

Tell us about your role:

I am a researcher at Harvard's Belfer Center for Science and International Affairs, where I primarily focus on the technology competition between the United States and China. I am currently working on projects related to U.S. government investment in the semiconductor industry, applying risk assessments to foundation models, constraining the use of AI systems in decision making processes related to nuclear weapons, the decision by Big Tech to take part in the war over Ukraine, and the ways in which public-private partnerships contribute to data colonialism. In my spare time I am a freelance writer—this year my writing has been published by TechCrunch, The Diplomat, South China Morning Post, and InkStick.

How did you carve out your career in the Responsible Tech ecosystem?

I began my career in Responsible Tech when I realized that my peers in my computer science classes at UC Berkeley had no awareness of the negative externalities that might come from the tools they would build. After graduating, I received the John Gardner Public Service Fellowship, which provided me with funding to work wherever I wanted. I chose to work at the artificial intelligence lab of the UN Secretary-General, where I had the opportunity to work with some of the world's leading experts on privacy, responsible data, and digital development. While at the UN, I wrote a new UN-wide privacy policy related to government-backed collection of health data that was adopted by the World Health Organization and the World Bank.

“

I think it’s important that we stop anthropomorphizing AI. Intelligence might emerge from complex systems but these systems are not that complex and most of the time we’re giving them too much credit.

”

Kolja Verhage

Manager Digital Ethics

Deloitte

How can we strengthen interdisciplinary engagement in technology development, deployment, governance, regulation and oversight?

One of several good practices I've learned from doing digital ethics projects for organizations is that the person who coordinates the implementation and execution of ethics-based policies and governance frameworks should have an interdisciplinary profile. There is a good reason for this. Doing proper ethics requires an understanding of all stages of the lifecycle of a digital solution (and how they are being governed) from a broad range of perspectives. As an example: What I've seen when building and managing AI ethics boards is that their value really comes from the interdisciplinary engagement from board members that bring a rich variety of experiences and perspectives to the table.

To strengthen interdisciplinary engagement we need to actively gather good practices around digital ethics implementations in organizations. Because there are currently no widely accepted governance standards on how to make digital ethics work in corporate or public policy, far too many organizations are facing a steep learning curve and are therefore hesitant to start. To strengthen the necessary interdisciplinary engagement more work should be done on collecting good practices and creating standards based on them.

What emerging regulatory frameworks are having the greatest impact on AI development at the present time?

In my view, the regulatory framework with the greatest impact is the upcoming EU AI Act. An incredible amount of effort has gone into it that makes me proud to be a European. What's interesting is that the AI Act takes a unique route by regulating the development of a technology instead of the outcome of its use. From a safety perspective I think generally it's a great strategy. But its important to understand there are also potential pitfalls.

One of the consequences of this approach is that it does not offer individuals the right or avenue for redress when they are negatively affected by an AI system. For example, in the financial sector banks are developing AI models to combat money laundering. These models have good intentions but are, by nature, highly discriminatory. Without any avenues for redress those that are disproportionately impacted by these models have no way to get their voices heard. This will become especially important if the AI Act follows the precedent set by the GDPR and becomes a template for AI legislation around the world.

Another wonderful development I'd like to mention is the law that was recently passed in New York City prohibiting employers from using AI systems for HR purposes without those tools first being audited for bias. If successful, I hope this law becomes a template for many other AI applications. There are some incredible people in the AI auditing community and I would love for them to get their hands dirty.

What technical safeguards, oversight, and best-practices are needed to ensure safety by design and protection of human rights while mitigating harms?

From my perspective the answer to ensuring the protection of human rights lies in the organizational processes embedded around the decision-making and design of a digital solution. In very basic terms it starts with asking the question “should we do this?” or “is this a good idea?” instead of simply asking “is this legal?” In fact, in many organizations the legal department is kept out of digital ethics implementations. The reason for this is that doing ethical analyses from a compliance perspective (which is more natural for people with a legal background) is counter-productive. To answer a question like “should we do this?” you need to think in values rather than laws.

There exist a wide variety of “digital ethics-by-design” implementations, from doing value-sensitive design labs or dilemma-thinking trainings with developers to building advisory AI ethics boards or setting up internal digital ethics ambassadors. For all these implementations we should embrace the “learning-by-doing” approach. It’s important that we collect and share good practices across the industry and collaborate internationally to connect the dots. I feel that’s the only way we’re going to see broad adoption.

How do we ensure that new and emerging technologies and systems do not reproduce and augment existing inequalities?

Depending on the type of inequality we are concerned with, “digital ethics-by-design” as a methodology for creating public and corporate policies could act as a safeguard to these inequalities, as it looks to the broader impacts of technology and how it’s being deployed.

On a global scale the impact of the distribution of resources, either technology itself or the expertise needed to build it, must be considered in light of the post-colonial disparities that still plague developing countries. Understanding the conflicting values by engaging with different stakeholders in multi-stakeholder forums can help identify and potentially mitigate some of the inequalities rather than exacerbate them. One clear example of this are the composition of working groups in which technical standards for emerging technologies are created. Developing countries hardly have a say in these procedures but the impact on their economies is very real.

Looking at your crystal ball of a positive tech future. Let's say we have human-centered AI. What would humane AI look like? What would its pillars be?

First of all, I think it's important that we stop anthropomorphizing AI. Intelligence might emerge from complex systems but these systems are not that complex and most of the time we're giving them too much credit. They're a great technology, in the same way that highways are a great technology. Like AI, highways guide our behavior and are essential to the functioning of our economy.

So what would a humane highway look like? Of course, this completely depends on how people interact with it (its socio-technical context) and a range of cultural, economic and environmental factors. Because this context is complex and constantly changing, its impossible to give one definitive answer.

Nonetheless, creating "humane" technology starts by taking hyper specific uses of technology and their social context into account. So at least there should be a development methodology in place that understands this, offers avenues for redress and adjusts accordingly when it's found to be causing harm. At the same time we should be encouraging people to be creative: dream bold, take calculated risks and let our ingenuity drive us to improve our shared human condition. That's what Humane AI should be all about: Creating the guidelines that empower our human creativity to build responsible technologies that alleviate suffering.

Tell us about your role:

My role as manager of the Digital Ethics team at Deloitte is two-fold. Most of our work involves helping a wide range of multinational corporations and government agencies to build responsible digital Solutions that pursue their organizational goals in ways that build trust in their use of technology. Generally speaking, we do this by operationalizing value-based (or ethical) principles in the governance of digital technologies. Trust is a two-way street and the way I see it there is a lack of trust is a misalignment between an organization's values and the impact of its technology on its stakeholders. Building trust means getting alignment between the two.

The second part of my job is what I call "spreading the gospel of digital ethics." We're a relatively new team and many organizations still don't fully understand the opportunities they are missing out on by not getting the governance of technology right. So my team and I write blogs, give presentations, organize roundtables, and many more things to get our message across. We've been getting incredible responses and I truly feel blessed to be managing such an expert and passionate group of people in this growing field.

“

If we are discussing ‘public safety’ in the framework of policing, then I believe that there is no full way to ensure that AI will not reinforce existing patterns of discrimination. The policing system at its core is racist — we cannot add systems to an already discriminatory one to make it ‘safer’ for our communities, especially those that are marginalized.

”

Lama Mohammed

Associate
Glen Echo Group

Fellow
Internet Law and Policy Foundry

What are key factors in the way AI is developed/deployed that have the most impact on protecting human rights?

There is a myriad of factors that go into the development of AI that greatly impacts society on a large scale. From a sociological perspective, these advanced technologies are being developed at the largest technology companies that greatly employ cis, upper-middle-class white men. When designing AI, these experiences are shaping the technology that is being used by people of all backgrounds and ethnicities, and it is, therefore, critical that those who are designing the technology can represent its users. This idea of diversity is also essential in the testing phase. If technologists are only testing their technology on a specific population, then how is it possible for others to enjoy and use the technology? The lack of representation in designing and testing is essential to creating inclusive AI and equitable technology.

What are the areas of AI and Machine Learning development whose potential social implications are currently being overlooked in the general discourse? Which areas get comparatively too much attention relative to their potential impact?

An AI and ML development area that is overlooked is climate change. While there are several technological efforts to reduce carbon emissions in the atmosphere, I think it's important to understand how AI is also contributing to the climate crisis through e-waste and how we can sustainably create these technologies.

We also focus heavily on creating diverse datasets so we can design facial recognition systems that recognize the faces of people from different ethnicities and backgrounds. However, what happens when we reach that threshold? How do we make sure that when we can recognize more faces of individuals we humanely use this technology? I can see facial recognition technology already being weaponized to surveil marginalized communities all over the world, and I worry that this use will become even greater once we build an AI system that can equally detect faces.

How can we equip policymakers to better keep pace with the speed of AI and Machine Learning development?

Over the years, different private companies have partnered with government agencies to support their technology efforts, whether it is in increasing an agency's cybersecurity or using technology to develop secure voting machines. This effort is commonly known as "public-private" partnerships, and we can continue to expand this effort to help educate policymakers about the different uses of technology and its effects, as well as work together to create policy Solutions with different voices at the table that prioritize the needs of technologists and policymakers.

I would also encourage policymakers to do more open town hall events on issues related to socially responsible technology (cybersecurity, privacy, trust and safety, etc.) to encourage consumers' voices on the issue. Allowing consumers and the general public to showcase their voices allows policymakers to bring their priorities to technology companies, whether it is their privacy or more about what technology companies can do better to protect children. Town halls also allow for diverse voices a seat in this conversation. While public-private partnerships are effective, they also still carry inequalities in representation since government and technology are largely comprised of white men, enhancing a greater need for public participation on this issue.

How can individuals be empowered to know and protect their rights in the face of potential threats to privacy and civil liberties linked to use of AI?

I still think those who are actively working in this space care about it so much because they have access to resources through their schools and offices about the subject matter. For a lot of people, anything related to the field of technology, policy, or both is relatively difficult to understand. Individuals like us in the space who understand the potential privacy threats and civil liberties need to do more to bring these issues to light in a way that connects to broader issues that our greater society cares about, such as climate change, voting and democracy, freedom of speech, and other global values.

In my conversations with people who do not know much about responsible tech will say, “oh, I don’t really care for privacy because they already have so much information on me anyway.” This learned helplessness has been a personal uphill battle for me and what I have used to tackle it is explaining how the lack of privacy for a single individual impacts everyone else they know or even do not know on such a large scale. Yes, you may not care that a wide array of your facial recognition data is on the web, but did you know that it was likely used to develop a surveillance system by an authoritarian government or our prison system? Making individuals aware of how this is a collective societal problem is one of the best and most effective ways to get people involved in the space.

How can we ensure that the use of AI for “public safety” does not reinforce existing patterns of discrimination?

This answer mostly relies on what we define “public safety” as. If we are discussing “public safety” in the framework of policing, then I believe that there is no full way to ensure that AI will not reinforce existing patterns of discrimination. The policing system at its core is racist — we cannot add systems to an already discriminatory one to make it “safer” for our communities, especially those that are marginalized.

We need to completely restructure our understandings of public safety that is entirely grounded in human value and opportunity, and abolish this practice of cruel punishment that the prison system does. There is more opportunity for us to invest in human-centered Solutions to public safety, such as greater investment in public education, public health (both physical and mental), and equal economic opportunities.

Tell us about your role:

I am currently an Associate at the Glen Echo Group in Washington, DC — but I work remotely from the New York City neighboring area. The Glen Echo Group is a communications and public relations firm assisting technology companies, university centers, think tanks, and nonprofits with their general and political communications.

In my role, I work on policy and communications with clients whose areas focus on artificial intelligence, augmented and virtual reality, cybersecurity, the digital divide, and privacy. Many of our clients work to advance socially responsible technology, whether it is creating equitable tech, researching and bringing issues within tech to light, or working with policymakers and other influencers on accountable policy.

How did you carve out your career in the Responsible Tech ecosystem?

I received my undergraduate degree from American University in Washington, D.C. During my first year, Mark Zuckerberg testified for the first time in Congress on Facebook's involvement with Cambrdige Analytica during the 2020 U.S. presidential election.

Following this hearing closely, I noticed two things: the communications gap between technologists and policymakers in attempts to understand one another and the impact technology has on our democracy as it relates to individuals' privacy and their civil liberties.

From here, I realized that I wanted to do work that involved this intersection between policy, law, and technology that focused on human value and society at large, thereby leading me to a career housed in responsible technology.

> “
> There is too much power in the hands of people who stand to suffer the least from AI’s flaws, and not enough understanding or clout in the institutions we might ask to regulate it, to level this playing field or call a halt to this race into dystopia
> ”

Laura Walker McDonald

Senior Advisor,
Digital Technologies and Data Protection
International Committee of the Red Cross

What are the areas of AI and Machine Learning development whose potential social implications are currently being overlooked in the general discourse? Which areas get comparatively too much attention relative to their potential impact?

I think we are suffering massively from the distortion in our global thinking caused by the concentration of research and experimentation next to wealth. Whether it's in big tech companies or research centers, AI and Machine Learning skill, computing power, and funding is deeply entwined with the companies and capital enclaves of the Global North. This means that the principles and assumptions governing AI reflect only the reality for a minority of people – the most privileged in the most privileged places. Even the data is distorted, either misrepresenting or simply ignoring most people's life experiences. I find that very concerning, particularly when, even as we acknowledge how far we have to go to address concerns around bias in AI and Machine Learning, we're still moving forward to employ these technologies in transport, weapons systems, the very systems that govern decisions made about us and for us. And it is not clear that there is any accountability for this, or any way to stop it. There is too much power in the hands of people who stand to suffer the least from AI's flaws, and not enough understanding or clout in the institutions we might ask to regulate it, to level this playing field or call a halt to this race into dystopia.

How can we equip policymakers to better keep pace with the speed of AI and Machine Learning development?

Lack of knowledge about technology has always been such a huge issue. Really basic things – in the US, we've seen legislators unsure of how Facebook makes its money ('Senator, we run ads'.) This is not about everyone learning to code – I think the interplay between society and technology is far more important. Alongside training on leadership skills and cyber security, basic human-centered design principles like the Principles for Digital Development would be a good start; along with the same kind of essential information about cables, cloud infrastructure and databases that we are all taught about the weather as children. You need to understand how the nuts and bolts hang together in order to decide how they should be managed. And this kind of learning would also help with critical thinking about technologies that maybe won't have so much utility for solving global problems – maybe like cryptocurrency and non-fungible tokens.

But I don't want to oversimplify this. It's a very hard problem, particularly with the role technology now plays in global competition. The incentive is to keep experimenting and going as far as fast as you can before another country out-competes you, not to slow down and think about the implications of what is being developed. And investing in organizational knowledge about digital, say in governments and international organizations, would make a huge difference but would be a momentous commitment of time and money. I would love to see that commitment made – but I don't know that I expect it.

Tell us about your role:

The ICRC is a global, independent humanitarian organization helping people affected by conflict and armed violence, and promoting the rules that protect victims of war. We have invested significantly in recent years to understand and respond to the ways that technology, including AI, is changing our world. For example, we have developed new data protection guidance that governs how we collect, use and manage data in our work; we are developing new digital humanitarian services, such as RedSafe, which provides information and tools for people on the move; and we talk to governments, the private sector and other actors about how they are using, building and regulating technologies like biometrics and – yes – AI.

My role is one of several across the world which is part of an ICRC delegation, in my case, the regional delegation to the US and Canada. That means that I have a focus on the partners and programs we're working most closely with here in North America. It's fascinating work, looking at how technology influences geopolitics, and vice versa; how new technologies are changing the way wars are fought; and helping colleagues to use digital in their humanitarian work. But most of all my job is to help us think about and use digital technologies in alignment with the fundamental humanitarian principles of neutrality, impartiality, independence and most of all, humanity.

How did you carve out your career in the Responsible Tech ecosystem?

My career is about community and learning. My first aid job was with the British Red Cross. I learned a lot, and got interested in tech for aid and development. In 2010 I joined FrontlineSMS. We made software that let you send, receive and manage SMS without using the Internet. By 2012 I was co-CEO.

Our user community taught me how hard it is to tie technology to the outcomes you're trying to achieve. When we spun out FrontlineSMS as a for profit in 2014, I set up SIMLab as a think tank and consultancy focused on impactful, responsible and inclusive technology projects. We closed, inelegantly, in 2017, because I hadn't been able to build a sustainable business model. I wrote about what I'd learned from the experience, and so many people have told me it helped them. And people still cite our work, which we left behind as openly licensed resources.

I consulted on various projects for a while with baby in tow, before joining the Digital Impact Alliance in a leadership role in 2019. I grew their policy influencing team, and worked on the Digital Principles. Last summer, the ICRC opened up this position and I jumped at the chance to come 'home' to the Red Cross and Red Crescent Movement.

I would not be here without my community, particularly my techladies and techpersons who are less well served by the patriarchy. We have to have one another's back! Hit me up any time for a chat.

“Effective governance is critical for the "ethics-by-design" framework. This starts with the organization's leadership establishing as its North Star, AI principles consistent with those adopted by the OECD. It also entails creating documentation, decision making, training, accountability, compliance, and other mechanisms to ensure that the organization follows this North Star and factors in diverse viewpoints.”

Lee J. Tiedrich

Distinguished Faculty Fellow in Ethical Technology

Duke University

What are key factors in the way AI is developed/deployed that have the most impact on protecting human rights?

To protect humans, it's critical to establish frameworks consisting of laws, standards, policies, and other tools that collectively foster AI innovation in a manner that's safe, sustainable, ethical, protects human rights, and otherwise operationalizes the OECD AI principles. These frameworks must align with sound scientific principles and be implementable. They can increase trust by helping society identify trustworthy AI applications.

In addition, to protect humans, AI developers should adopt an inter-disciplinary "ethics-by-design" approach that fosters collaboration among experts in human rights, law, ethics, policy, business, sustainability, computer science, and other disciplines, throughout the AI product life cycle. This holistic approach will help ensure that human rights, ethics, sustainability, and legal compliance are addressed upfront in AI product design and development, as well as throughout deployment and wind-down.

Effective governance is critical for the "ethics-by-design" framework. This starts with the organization's leadership establishing as its North Star, AI principles consistent with those adopted by the OECD. It also entails creating documentation, decision making, training, accountability, compliance, and other mechanisms to ensure that the organization follows this North Star and factors in diverse viewpoints.

Finally, others in the AI ecosystem have important roles in protecting humans too. AI deployers should familiarize themselves with applicable laws as well as explanations provided by AI developers, including on how to use AI tools in ways that protect individuals. Individuals and organizations also should take advantage of opportunities to provide feedback to AI developers and deployers as well as policymakers, enforcement agencies, and other relevant organizations.

How can we strengthen interdisciplinary engagement in technology development, deployment, governance, regulation and oversight?

As discussed above, AI developers can strengthen inter-disciplinary engagement by implementing an "ethics-by-design" framework supported by an appropriate governance structure. This inter-disciplinary model can be adapted by others in the AI ecosystem. For example, AI deployers can draw upon inter-disciplinary experts to help ensure that their activities are sustainable and comply with applicable legal and contractual obligations as well as directions provided by AI developers in relevant explanations.

As also discussed above, society needs appropriate AI frameworks consisting of laws, standards, policies, and other tools that are grounded in science. Some mechanisms exist to support inter-disciplinary collaboration in policymaking. For example, at Congress' direction, the National Institute of Standards and Technology is working on an AI risk management framework with input from a broad range of stakeholders. Similarly, the draft EU AI Act contemplates pre-market conformity assessment requirements for high-risk AI as well as the development of technical tools. The EU-US Trade and Technology

Council establishes structures for cross-border collaboration on technical tools that also should promote interoperability. Policymakers should continue to build on this good inter-disciplinary foundation.

The need for inter-disciplinary collaboration also extends to enforcement. For instance, the U.S. Department of Justice announced an initiative to combat redlining together with the U.S. Attorney's Offices. It also has partnered with the EEOC to warn against potential discrimination stemming from AI hiring tools. Other agencies, such as the FTC, have relevant enforcement authority too. As agencies exercise this enforcement authority, it's important that they have access to relevant technical expertise.

What emerging regulatory frameworks are having the greatest impact on AI development at the present time?

Among emerging AI regulatory frameworks, the proposed EU AI Act is presently having the greatest impact on AI development. It would apply to any AI deployed within the EU, even if the developer resides outside the EU. Embracing a risk-based approach, the proposed Act would create four categories of AI applications and calibrate the regulatory requirements for each category based on the associated risk. For instance, the proposed EU AI Act would ban those AI applications that are deemed to pose an unacceptable level of risk. It also would establish several regulatory requirements for "high risk" AI applications, including pre-market conformity assessment requirements and post-market surveillance obligations. The proposed EU AI Act also addresses enforcement and provides for hefty penalties for non-compliance.

In the United States, while Congress may not pass in the near term the proposed Algorithmic Accountability Act or any other similar AI specific legislation, the growing number of state and local AI laws and ordinances, such as in New York City and California, are having increased impact. It is noteworthy that the expanding patchwork of differing AI laws and regulations within and outside the United States can increase the cost and difficulty of compliance for AI developers and deployers, which in turn, can pose barriers to entry, particularly for small and medium sized businesses. To reduce these barriers and promote competition, policymakers should continue to foster more cross-jurisdictional harmonization. In the United States, Congress could achieve more harmonization through federal legislation that preempts inconsistent state and local requirements. Bi-lateral and multi-lateral efforts can help foster more international harmonization.

Tell us about your role:

After practicing law for three decades at a global law firm, I recently joined Duke University as a Distinguished Faculty Fellow in Ethical Technology, with a dual appointment at Duke Law School. In addition to teaching AI Law and Policy, I have developed and am teaching an Ethical Tech Practicum where students from different Duke programs work in inter-disciplinary teams to help real world clients address ethical tech matters. The course combines teaching relevant legal, policy, and ethics doctrine with hands-on experiential learning.

In addition to teaching, I continue to engage in relevant policy developments. For example, I currently am a member of the Global Partnership on AI (GPAI) Multi-stakeholder expert group, and co-chair the GPAI IP Subcommittee and serve on the GPAI AI and Climate Steering Committee. Consistent with my commitment to education, I also am co-authoring the first AI law school case book to help expand AI education more broadly.

How did you carve out your career in the Responsible Tech ecosystem?

Having studied both electrical engineering and law, I always have gravitated to helping emerging technology pioneers navigate the evolving legal and policy landscape so they can make responsible new technologies available for society. At the beginning of my career, this involved working with pioneers of wireless and internet technologies, well before many people had cell phones or high speed internet access. Drawing upon my engineering studies, I recognized early on the importance of inter-disciplinary collaboration and the need for our legal and policy frameworks to align with science. Throughout my career, while the technologies have changed, my goal of fostering inter-disciplinary cooperation to enable society to benefit from responsible technology has remained constant.

“

Most tech companies are not consulting with an appropriate number of stakeholders before developing and deploying AI. We need to create platforms that bring together not only human rights experts, but also diverse representatives from civil society and experts in the field of peacebuilding who bring experience applying a conflict sensitivity lens to business and technology.

”

Lisa Schirch

Starmann Chair and Professor of the Practice in Peacebuilding
University of Notre Dame

What are key factors in the way AI is developed/deployed that have the most impact on protecting human rights?

Most tech companies are not consulting with an appropriate number of stakeholders before developing and deploying AI. We need to create platforms that bring together not only human rights experts, but also diverse representatives from civil society and experts in the field of peacebuilding who bring experience applying a conflict sensitivity lens to business and technology.

How can we strengthen interdisciplinary engagement in technology development, deployment, governance, regulation and oversight?

Right now, the human rights field is involved and invited to such oversight boards. But there is a lack of expertise in peacebuilding. I am currently working with a broad network of peacebuilding organizations engaged in applying new forms of "peace tech" or analyzing the negative impacts of technology on conflict dynamics. These groups have distinct insights and concerns that are not represented by the human rights community. In addition, the field of human security is missing from these conversations. Human security platforms have been formed at the local level in countries like the Philippines to provide a way for companies to hear from local community stakeholders.

What are the areas of AI and Machine Learning development whose potential social implications are currently being overlooked in the general discourse? Which areas get comparatively too much attention relative to their potential impact?

There is not enough analysis or attention to the impact of AI and machine learning on social movements, peace processes and elections as well as state surveillance that is putting democracy and human rights activists, as well as mediators and peacebuilders, at risk. For example, smart city AI and health apps deployed during the pandemic are being used by states to repress civil society and undermine democratic movements. While current forms of AI are undermining peace and democracy, there are also ways of using AI and ML to benefit peace and democracy - and that is not getting enough attention.

What specific structural changes to incentives and business models are needed in order to prioritize user well being and human rights? What kinds of incentive systems (companies, people, regulators, societal) need to be tapped and how?

AI is a weapon. It can be a weapon of mass destruction. The oversight of AI needs to treat it with this seriousness. The profit model of all companies needs to support human rights and democracy, flat out. If the AI rewards data fostering hate, polarization and disinformation for the sake of profit, it should be sanctioned and ended.

Incentives to create and use AI that aids peace and democracy should be rewarded at every level of a company, and also with a prize for most innovative positive use of AI.

“The most important factors to keep in mind are truly respecting human dignity and fundamental rights so that your design or deployment does not infringe upon them or consider humans as data points to exploit, oppress, manipulate for your own (corporation or government’s benefit)”

Merve Hickok

Founder
Alethicist.org

Research Director
Center for AI & Digital Policy

What are key factors in the way AI is developed/deployed that have the most impact on protecting human rights?

The most important factors to keep in mind are truly respecting human dignity and fundamental rights so that your design or deployment does not infringe upon them or consider humans as data points to exploit, oppress, manipulate for your own (corporation or government's benefit). If we truly respect the dignity of a human with all its complexities, diversity, aspirations and dreams, we would not be subjecting them to algorithmic systems which compare them against certain norms that we subjectively created and expect everyone to adhere - whether that is race, gender, ability, sex, culture, income, etc.. We would not be using systems that try to predict whether someone is trustworthy, creditworthy, criminal and such. We would be trying hard to narrow equity gaps, solve problems at their root (instead of focusing on a part which we think we can turn into code), expand access to resources to those unprivileged, rather than treating everyone as possible fraudster and rank order them in our systems.

What emerging regulatory frameworks are having the greatest impact on AI development at the present time?

There are two emerging regulatory frameworks that will have a lot of impact on AI development and deployment, and one voluntary framework, which currently has the widest global endorsement.

First one is that the draft EU AI Act tries to provide a harmonized regulation regarding AI systems across, so that member countries do not have more/less restrictive regulations. The Act also tries to protect fundamental rights within the same framework, using a risk-based approach. Since its publication in April 2021, it has been the focus of many debates. It is the most comprehensive legislation attempt to regulate these systems, and has the potential to have wider global impact as a blueprint (just like GDPR did) adopted by other nations. The heated debates continue in European Commission committees.

The other framework is Council of Europe's work on a legally binding international instrument on AI systems. Different from the EU AI Act, the focus of this instrument is how to protect fundamental rights, rule of law and democracy. CoE has successfully developed international conventions adopted widely such as Convention 108+ and Budapest Convention. Since any country can possibly adopt and ratify this possible future convention without being a member state to CoE, the impact might be wider than the EU AI Act eventually.

The last framework is UNESCO'S Recommendations on AI Ethics. 193 member countries have endorsed it in November 2021. Although this is a voluntary scheme, there are some reporting mandates which can help move the agenda forward and provide motivation for different countries to act more responsibly.

How can individuals be empowered to know and protect their rights in the face of potential threats to privacy and civil liberties linked to use of AI?

We need to be careful about not putting the responsibility and onus of protection of rights on the individuals. Yes, the individual citizens and consumers can enhance their digital literacy but there is such a power imbalance between the individuals vs corporations and individuals vs state that it is impossible for individuals to protect their privacy and liberties by themselves. The systems have to be designed, deployed and monitored in ways that do not jeopardize a human's dignity and rights.

What can we do to help technologists and policymakers from multiple intersectional social identities contribute to human-centered AI without feeling like they're the token minority?

First of all, we need to expand this framing from technologists and policy makers to a multi-disciplinary one. You might be a non-tech, non-policy related employee/consumer/advocate with multiple intersectional social identities and would like to contribute, or would like to raise concerns. We need mechanisms and space within organizations to 1)be able to flag the issues we see, 2) trust that we will not be punished for voicing our concerns, and 3)trust the system that our concerns will be objectively reviewed and decided upon.

Another necessary point is to move responsible AI/tech conversations from 'corporate social responsibility' side to embedded business practices. As long as these conversations sit separate from the actual business interactions, product development, incentives, risk management etc, the attention and resources they will receive will be limited. When you accept that your product is better and more innovative because of responsible practices, and that your consumers and employees fare better because of it, you will put the necessary resources behind it rather than tokenizing minorities or exploiting their experiences and knowledge.

“

Policymakers need to grasp that AI is not a ‘natural’ process, but always the result of a complex social construction (of questions, data, categories, rules). And they need to have a sense of where AI is being applied and why, and a sensitivity to the potential gaps between ‘technical’ justifications for what AI is asked to do and the social questions which AI is being asked to address.

“

Nick Couldry

Professor of Media, Communications, and Social Theory

London School of Economics

What are the areas of AI and Machine Learning development whose potential social implications are currently being overlooked in the general discourse? Which areas get comparatively too much attention relative to their potential impact?

Most overlooked are the inherent LIMITS of AI and ML. Well-designed ML can perform tasks of processing that are not possible for humans, but always within parameters that humans can and should set, even when this means the parameters within which AI is allowed itself to learn from the patterns it detects in data. But there may be areas of knowledge that are not susceptible to capture in forms that can be readily quantified, and patterns of interaction whose analysis humans would not want to delegate to machines (love, grief etc). There are also social limits to the effectiveness of processes for managing AI and ML: AL and ML will not win social acceptance unless they operate within processes that are transparent, credibly ethical, and open to inclusive human intervention.

How can we equip policymakers to better keep pace with the speed of AI and Machine Learning development?

Policymakers need to grasp that AI is not a 'natural' process, but always the result of a complex social construction (of questions, data, categories, rules). And they need to have a sense of where AI is being applied and why, and a sensitivity to the potential gaps between 'technical' justifications for what AI is asked to do and the social questions which AI is being asked to address. The key is the ability to translate between the 'technical' and the social : where that is not possible, or the means to do so are under-resourced, problems will occur, and effectiveness will break down.

How can individuals be empowered to know and protect their rights in the face of potential threats to privacy and civil liberties linked to use of AI?

This is difficult. Education about what AI is and how it works needs to be widespread, but it is not realistic to expect people to give up their scarce time to 'learn' this, so media must play a part in explanations. But a basic knowledge of what AI does needs to be supplemented by empowering citizens to feel able to ask the SOCIAL questions about AI's goals, means and results, that they should be asking. Here the myths about AI as automatically objective or neutral or unquestionable need to be exploded publicly and replaced with an awareness of the contextual value of AI as a way of thinking about and discovering the world.

How do we ensure that new and emerging technologies and systems do not reproduce and augment existing inequalities?

The people who design AI, implement it, and then apply it in multiple contexts must be as diverse as possible, but that in itself is not enough. They need to be empowered, by their institutional contexts, to question, to challenge, to redesign, and sometimes to halt those aspects of current AI practice with which they disagree or about which they have reservations. And they need to feel empowered and resourced to reach out beyond the 'AI experts' to wider citizens to consult about what they are doing and be confident about having and learning from that inclusive debate. AI needs to be embedded in a civic process, if it is to work, because it is a form of social knowledge, not just something that can be safely black-boxed from social life.

Looking at your crystal ball of a positive tech future. Let's say we have human-centered AI. What would humane AI look like? What would its pillars be?

It would operate according to goals and through methods that were chosen by human beings, on terms and under conditions monitored and accountable to human beings, and in general terms to serve the goal of enabling a sustainable and good life for human beings and the planet. But 'human beings', I don't mean just any human beings, but specific social and civic groups of human beings who have the resources and are empowered to think together about what society needs from AI, what risks it might involved, and what would it mean to address those opportunities and risks in an ethical, inclusive and democratically accountable way.

Tell us about your role:

I teach and research about media and social theory in the Department of Media and Communications at LSE. Until a decade ago, my work was on the power of traditional media, especially television. In the past decade I have increasingly focussed on issues of platform power, data extraction and AI, and the issues of power and ethics they raise. My latest books were Media: Why It Matters (Polity 2019) and The Costs of Connection (with Ulises Mejias, 2019, Stanford University Press).

"

I like to stress the concept of intentional inclusion, by which I mean ensuring there is a process in place that recognises both relevant systemic inequalities and inequities as well as the biases arising from those developing technology, and provides effective tools to confront them. Practicing intentional inclusion is particularly important in the context of many new and emerging technologies and systems as many of these have the ability to reproduce and entrench existing inequalities at scale.

"

Nora Lindström

Digital Development Leader

What are key factors in the way AI is developed/deployed that have the most impact on protecting human rights?

We continue to live in societies where “computer says no” is often taken at face value. What I mean by that is that it continues to be common for people to accept and not question answers or Solutions provided by technology. AI has the potential to make how technology behaves even more obscure, making it difficult for people to critically assess the predictions, recommendations, or decisions (!) made by AI systems.

There needs to be increased recognition that humans design and develop technological systems and Solutions, including AI, and that these systems are as flawed and have as many shortcomings as we have – and potentially more, and at larger scale.

To minimise the potential harms caused by AI systems, it’s important we keep a human in the loop and ensure machines do not make autonomous decisions that impact people’s lives. We also need to ensure AI systems are transparent and that we understand how an AI makes predictions, as well as ensure AI-based decision-making systems can be readily held accountable.

Finally, I would like to stress the need to have diverse teams when developing AI and seek diverse inputs and feedback in particular from people directly and indirectly affected by the AI, as no team can ever represent “all”.

How does your team make decisions around integrating AI and Machine Learning into your product? How do you handle data collection, management, and model optimization? Who is at the table in these conversations? What are you optimizing for?

At CRS, we tend to use AI and machine learning for relatively straightforward use cases. For example, we use a combination of satellite imagery and image recognition (deep learning) to estimate the number of people living in a specific area. This eliminates the need to go house to house to count the number of people living there. We know the results aren’t perfect, but for our purposes – such as providing suggestions for where water, sanitation, and health (WASH) infrastructure should be located - a rough estimate is enough. The final locations must obviously be validated and agreed upon with the relevant communities.
One of our more complex uses of machine learning is in a project in Malawi we have devised a data collection and analysis scheme to measure and predict resilience among households prone to food insecurity one to two months out. Data prediction is achieved with machine learning. The goal is to be more proactive in identifying which households are the most food insecure and respond in advance. Crucially, it is community members themselves that collect the data, and we close the data loop by providing the insights gained from data analysis back to them, so that they can validate the results and use their own agency to take action on its basis.
Across our work, we optimise for improving programme quality, impact, and/or reach.

How do we ensure that new and emerging technologies and systems do not reproduce and augment existing inequalities?

This is a really tough question as it requires actions on multiple fronts. It also needs to be acknowledged that it is not only new and emerging technologies that have the ability to reproduce and augment existing inequalities – much of the technology that we already use does so too. An example of this could be how the English alphabet has become the standard digital alphabet. As a result, I have to change my name every time I fly internationally – I can't be Nora Lindström because the system does not support the letter "ö", I have to be Nora Lindstroem. Once, I forgot to be Nora Lindstroem when booking my flights, and was subsequently denied boarding. For many, one's name is fundamental to one's identity, and having to change the spelling of one's name when flying is not an experience unique to me, so this is great example of widespread technology-facilitated discrimination that we already live with.

So how do we prevent it? I like to stress the concept of intentional inclusion, by which I mean ensuring there is a process in place that recognises both relevant systemic inequalities and inequities as well as the biases arising from those developing technology, and provides effective tools to confront them. Practicing intentional inclusion is particularly important in the context of many new and emerging technologies and systems as many of these have the ability to reproduce and entrench existing inequalities at scale.

Looking at your crystal ball of a positive tech future. Let's say we have human-centered AI. What would humane AI look like? What would its pillars be?

A couple of years ago, when working for Plan International, I came up with the idea of Equality Tech. Equality Tech is defined as technology that in itself advances equality. We know that technology can discriminate and entrench inequalities, so why couldn't it also do the opposite? The idea is that by embracing the inherent bias in technology, we can develop digital products that help us challenge harmful norms and stereotypes, and that nudge us towards more inclusive behaviours. Examples of this type of technology already exists – including in digital Solutions many of us use every day: Microsoft Office 365's Editor has a setting that allows you to spell-check your language for inclusiveness and gives suggestions of more inclusive language, like using the term "firefighter" instead of the gendered "fireman". For me, a positive tech future is one where technology helps us create a more inclusive and equitable world where everyone's human rights are respected.

Tell us about your role:

I'm a digital development leader currently serving as Senior Director of ICT4D at Catholic Relief Services. My team's role is to enable the agency to leverage technology for increased programmatic quality, impact and reach. One of the tools we use for this is machine learning and AI. We use data collected by our programmes around the world and leverage advance analytics to gain further insights from the data, with which we can then improve programme outcomes.

I am also the Chair of the Digital Principles Advisory Council, the role of which is to advice the Digital Impact Alliance (DIAL) on their stewardship of the Principles for Digital Development.

How did you carve out your career in the Responsible Tech ecosystem?

I never set out to have a career in tech, let alone Responsible Tech! I'm a social scientist by training and "fell" into the tech space by working on projects where we used technology for increased impact. Over time - and failures - I came to recognise the risks associated with leveraging technology particularly in the global development and humanitarian sector where we work with many vulnerable groups, and started educating myself on how we can use technology more responsibly, such as by adhering to the Principles for Digital Development. Through this journey, I also realised that responsible use of technology and data ultimately results in having a more sustainable (positive) impact.

“

To avoid replicating and exacerbating existing biases, machine learning models need to be regularly updated, flexible and, where feasible, supervised and reviewed by human moderators. Feeding the right training data —large, diverse and accurate— is essential.

”

Patrick Grady

Project Lead
Internet Commission

How can we strengthen interdisciplinary engagement in technology development, deployment, governance, regulation and oversight?

Interdisciplinary engagement should first be encouraged within tech companies. Throughout product development and deployment, cross-functional collaboration can often bridge gaps between the intentions of developers and the expectations of external stakeholders (users, regulators, etc.) Opening internal communication channels can encourage staff from different departments and functions, often carrying different disciplinary backgrounds, to capitalise on each other's expertise. When deploying AI, in particular, technical teams should engage closely with those from legal and privacy. Formalising regular cross-functional interdisciplinary engagements, such as annual safety summits, can assure opportunities for colleagues to share insights and learnings.

It is now standard practice for tech companies to engage with external subject matter experts and researchers when developing and deploying products – this marks commendable progress. When using AI to shape products or enforce policy, companies should not neglect to consult users themselves, who are often overlooked in policy and product development. Some tech companies already engage with users, through community consultations or by including representatives (such as high-profile 'creators') at high-level advisory forums. Fostering feedback from users potentiates a more accountable enforcement process and may inform the development of useful features that give users choices over the product they use.

What emerging regulatory frameworks are having the greatest impact on AI development at the present time?

The European Union's Artificial Intelligence Act captures a vast scope of AI applications and could have a GDPR-esk effect well beyond European borders. However, I'm most interested in the upcoming Digital Services Act in Europe – for companies dealing with user-generated content, and for millions of users, it will have the largest immediate impact.

The latest proposals reveal potential provisions that require larger platforms to explain how, and for which purposes, automated tools are deployed. This includes, for instance, shedding light on the use of recommender systems and explaining why certain parameters are in place. Qualitative insights such as these will not only provide a useful resource for regulators, researchers and all those hoping to better understand companies' decision-making but will impel companies themselves to consider the intentions and impact of their use of AI.

What technical safeguards, oversight, and best-practices are needed to ensure safety by design and protection of human rights while mitigating harms?

AI plays a critical role in keeping online environments safe and free from fake engagement. Its deployment improves the speed and accuracy of the removal of harmful content and can protect human reviewers from the most egregious cases. When deploying automated tools, however, organisations should ensure the right checks and balances are in place – and doing so at the earliest stage in design is a smarter approach that saves expensive retrofitting.

To avoid replicating and exacerbating existing biases, machine learning models need to be regularly updated, flexible and, where feasible, supervised and reviewed by human moderators. Feeding the right training data – large, diverse and accurate – is essential. Models should be built with better explainability, and ideally designed with safeguards that protect journalistic use of content, to best uphold human rights whilst mitigating harm.

What specific structural changes to incentives and business models are needed in order to prioritize user well being and human rights? What kinds of incentive systems (companies, people, regulators, societal) need to be tapped and how?

Greater transparency can incentivise companies to prioritise user well being and human rights. Transparency reports can be a useful tool for companies to exhibit the machinery of their use of automated tools, and their use of AI, and can serve to inform the public and build trust. Together with emergent regulation, public pressure has demanded more sophisticated reporting from companies. Reports are now more granular, segregated and useful for those looking to understand the impact of internet services on users' well being. As a result of public scrutiny, many companies are developing innovative, 'contextual' metrics to better measure the harm occurring on their platforms. Where greater transparency becomes the standard or required, companies will be incentivised to develop products that are safer by design.

Tell us about your role:

I lead the Internet Commission's accountability reporting project: a year-long study of content moderation practices. We focus on accountability: exploring the consistency between a company's stated purpose and actual practice.

Using our evaluation framework, a questionnaire of 122 qualitative and quantitative indicators, we collect several companies' data to try and understand their approach to moderating content. After several rounds of supplementary questioning and hours of interviews with staff across different verticals, we build confidential case studies. These form the basis of a knowledge exchange between companies – a unique inter-industry opportunity for trust and safety specialists to share experiences, successes and failures in content moderation. We then amalgamate our findings in a yearly report.

How did you carve out your career in the Responsible Tech ecosystem?

The most important nudge toward this field happened when specialising in digital policy through postgraduate political science and philosophy studies. Considering the deluge of competing interests and ethical dilemmas that furnish this field, I became convinced of the need to open dialogues between siloed groups (researchers, technologists, ethicists, regulators, etc.) Happily, there are now plenty of organisations working to do just this!

“

As long as there is an ‘us and a them’, as long as there is a ‘we’ who are ‘helping’ a ‘them’, tokenisation or worse has already begun. I’d suggest you could think deeply about your own identity, about how you are yourself marginalised in some ways and privileged in other ways.

”

Dr. Peaks Krafft

Senior Lecturer
University of the Arts London

How do we ensure that new and emerging technologies and systems do not reproduce and augment existing inequalities?

Stop building new technologies and instead redistribute those financial and labour resources to directly intervene on inequality. As long as a small group of owners and producers have the power to develop technology, technology will reproduce inequality by nature of that unequal distribution of power. As long as carbon-intensive technologies are being manufactured, rare earth minerals are being mined, and global trade networks are being maintained, the planet will continue to decay.

What can we do to help technologists and policymakers from multiple intersectional social identities contribute to human-centered AI without feeling like they're the token minority?

Who is the "we" meant to be in this question? As long as there is an "us and a them", as long as there is a "we" who are "helping" a "them", tokenisation or worse has already begun. I'd suggest you could think deeply about your own identity, about how you are yourself marginalised in some ways and privileged in other ways. In what situations would you want solidarity for a struggle you are facing or want to feel more welcome? Why is it that there are not already meaningfully diverse and inclusive spaces in the tech industry?

What specific structural changes to incentives and business models are needed in order to prioritize user well being and human rights? What kinds of incentive systems (companies, people, regulators, societal) need to be tapped and how?

As long as the conversation is about AI, the right people for positive social change will not be part of the conversation. AI is a mysterious and poorly defined term that functionally creates barriers to entry for anyone who doesn't think they have a solid understanding of what AI is. Since AI isn't actually anything in particular, and since AI means different things to different people, predominantly those individuals who have existed in some privileged sphere that grants a sense of entitlement to important political conversations will feel welcomed into a conversation about AI. The conversation must be about substantial social issues that create common ground with civil rights activists, non-government organisations, community and religious leaders, and civil society groups. And since the conversation is then not about AI and about topics that others have already been well-trodden for decades or centuries, the question is how can we learn about join in their conversations without causing problems due to our positions of power and privilege, not about how they can join ours.

Looking at your crystal ball of a positive tech future. Let's say we have human-centered AI. What would humane AI look like? What would its pillars be?

Ria Kalluri brilliantly poses the question as to "not ask if artificial intelligence is good or fair, ask how it shifts power". A truly humane and human-centred approach to AI in the United States would be for tech companies to dissolve themselves, forfeiting their resources towards reparations, community development, and prison abolition in line with the Black Panthers' still-relevant Ten Point Program. As Arvind Narayanan and others have pointed out, "AI" is not a technology but a rhetorical device best characterised as "snake oil". To the extent that AI automates certain kinds of labour, it does so in a way that disproportionately benefits tech companies and imperialist states, and primarily not even through any gains in efficiency but rather by providing liability shields. AI, insofar as it is essentially a rhetorical device intertwined with concentration of resources towards corporate and state power, cannot be made humane except by redirecting the substantial investments in it towards socialist programmes.

Tell us about your role:

My current role is as Senior Lecturer and Course Leader of the MA Internet Equalities degree programme at the University of the Arts London's Creative Computing Institute. MA Internet Equalities explores how power relations are organised, embedded and perpetuated in internet technologies, and how they can be re-organised or challenged through critical, creative and activist practice. In my role I recruit and look after students, manage the teaching team for the degree, and teach on the degree a bit myself. I also have responsibilities for conducting research in my role as Senior Lecturer. My own practice involves critically-oriented computer science research, academic organising, and community organising, especially recently on four issues in higher education and tech: social impacts of technology; personal and institutional accountability; White supremacy in organisations; and conflicts of interest from tech funding.

How did you carve out your career in the Responsible Tech ecosystem?

I got my PhD in computer science, thinking at first when I began in 2012 that I would become a machine learning researcher working on applications to social science. Instead I slowly became aware of the many substantially concerning sociopolitical issues with any applications of machine learning, and particularly applications to social problems. As I explored these questions more, and benefiting from many fantastic mentors and collaborators, I discovered that I felt much better about my work and felt my work was much better aligned with my values when I was orienting myself towards understanding and challenging power in the tech industry rather than supporting it, whether intentionally or not, whether directly or indirectly. My background as a practitioner in computer science gave me relatively unique perspectives and networks compared to many others in the Responsible Tech ecosystem.

“

The creation of AI Advisory Councils in many countries are another mechanism that have served governments well, enabling them to have timely access to industry and academic expertise as well as broad perspectives from across civil society. Governments should invest greater resources into these mechanisms – to solicit advice more frequently, conduct research on emerging topics of policy relevance, and to promote public engagement and awareness.

”

Philip Dawson

Policy Lead
Schwartz Reisman Institute
for Technology and Society

What emerging regulatory frameworks are having the greatest impact on AI development at the present time?

Emerging regulatory frameworks such as the European Union's Artificial Intelligence Act combined with standardization initiatives being led by ISO, the IEEE and national standards bodies like NIST and CEN-CENELEC are poised to have the greatest impact on AI development right now. On a sectoral level, AI-based medical devices and financial services products have already attracted significant attention from regulators, as has the use of AI in hiring. Governments have also adopted specific directive, policies and rules to operationalize AI in government, often targeting public procurement of AI systems or the use of AI in administrative decision-making.

How can we equip policymakers to better keep pace with the speed of AI and Machine Learning development?

Partnerships with leading AI labs and interdisciplinary academic institutes that bring together legal, regulatory, computer science and public policy expertise can play a significant role in helping government keep pace. Certain governments have also worked to develop curricula on AI for the public service. The Schwartz Reisman Institute for Technology & Society recently partnered with the Canadian School of Public Service on a lengthy series of lectures on AI, which has been a great success. Other institutes such as the Brookings Institute, the Centre for European Policy Studies, institutes at Stanford University and many others have begun offering courses for policy makers. The creation of AI Advisory Councils in many countries are another mechanism that have served governments well, enabling them to have timely access to industry and academic expertise as well as broad perspectives from across civil society. Governments should invest greater resources into these mechanisms — to solicit advice more frequently, conduct research on emerging topics of policy relevance, and to promote public engagement and awareness.

Tell us about your role:

I lead the Schwartz Reisman Institute for Technology and Society's work on regulatory innovation for artificial intelligence (AI), which we believe will be critical to controlling AI harms — including to human rights – and unlocking the social and economic benefits of AI. Specifically, we believe that soft law instruments, such as standards and conformity assessments, and assurance tools, such as audits, impact assessments and certifications, will play increasingly important roles in the design of agile and resilient regulatory systems that are capable of responding to the unique challenges of the age of AI. In order to address the speed, complexity and scale of AI, we believe that a combination of automation and AI-assisted technologies will be necessary for organizations to operationalize these standards and tools at scale, or for government to manage increasing the regulatory burden. My work involves leading policy research projects, identifying strategic partnerships and initiatives that, hopefully, will help policy makers, industry, researchers and civil society actors understand how to work together to create regulatory models that are fit for purpose in the age of AI.

How did you carve out your career in the Responsible Tech ecosystem?

A lawyer by trade, I first entered a career in "Responsible Tech" as part of a team that developed standards and guidance for the operation of unmanned aircraft systems (more commonly referred to as "drones") at the International Civil Aviation Organization, the United Nations specialized agency that sets standards for aviation. I was struck by the challenge faced by our membership of 192 countries, which hoped to realize the significant benefits from drone delivery, mapping or urban air mobility, but also had a mandate to protect important aviation safety objectives. Discussions with large aviation companies, startups, international organizations, governments and trade associations centered around the need for harmonized regulatory frameworks, based on standards and certifications, as well as new technologies that would enable the safe integration of drones into national airspace. Many of these discussions are relevant to artificial intelligence and the regulation of emerging technologies more broadly, as we try to realize the benefits of innovation while minimizing harms. The expertise I developed at ICAO, as well as the international networks and experience in consensus-based multi-stakeholder policy and standards processes, has served me well in developing a career in AI policy and regulatory development — and I had a lot of help from good colleagues, friends, and mentors along the way!

“

When we think about abuse and threats to individual rights brought by abusive data and geolocation collection practices, the data brokerage market, ad-tracking systems, and surveillance technologies, for instance, our best bet is to apply collective pressure on policymakers, from whom we can demand better public policies, and on companies, in our capacities as users and consumers.

”

Talita Pessoa

Tech Policy Consultant

What are the areas of AI and Machine Learning development whose potential social implications are currently being overlooked in the general discourse? Which areas get comparatively too much attention relative to their potential impact?

Whenever AI is mentioned in the general debate, it's often a depiction of AI focused on superintelligent and conscious machines, usually followed by a moral framing of "is AI good or bad?" that taps into our very human fears of being replaced or rendered obsolete. Other times, we get the reverse picture, with AI-powered technologies being presented as infallible Solutions to complex and multifaceted difficult and multifaceted problems, such as those posed by economic development challenges. The problem with framing AI as this intangible and sci-fi" thing is that it takes public attention away from the actual risks and implications brought by the AI systems that already exist and are currently being deployed. If we dive into specific problems currently brought by the use of AI on automated decision making, for example, trying to address them would take our focus to much more worldly topics such as privacy and data collection practices, the quality of datasets used, the underrepresentation of cultures and languages, and how humans working behind the AI curtain are treated. While discussing these issues may not feel as exciting as singularity or superintelligent computers, they are far more consequential to shaping the world and economies we'll be living in.

How do you currently apply climate and sustainable development goal frameworks to design and development of AI?

What I love so much about the UN 2030 Agenda for Sustainable Development is that it ambitiously recognizes the most pressing world issues are interconnected and cannot be fully tackled in silos. The Agenda broke with a historical tradition of separating civil and individual liberties from economic rights - the 17 SDGs and 169 targets are integrated, indivisible, and reflect the economic, social, and environmental dimensions of sustainable development. And then we have two SDGs truly cross-cutting in terms of their impacts and relevance to all the others: SDG#5, which is about gender equality, and SDG#13, about climate action.

Whether nation states are making enough progress to achieve all goals by 2030 is a story still being written, but the Sustainable Development Agenda establishes an important and timely roadmap for the responsible development and design of AI systems that may affect our societies. If we start thinking about potential harms and unintended consequences, the SDGs and their targets provide a holistic checklist of societal aspects to be considered by AI developers.

Another cool aspect of the Agenda is it also bases an important ongoing conversation promoted by the UN around the Global Digital Compact, a document expected to outline shared principles for an open, free, and secure digital future for all. While multilateral organizations such as the UN are far from being perfect, they have unparalleled potential to foster international cooperation and gather the voice of truly diverse groups. The process is open to all in civil society, academia, governments, and tech companies.

How can individuals be empowered to know and protect their rights in the face of potential threats to privacy and civil liberties linked to use of AI?

On an individual level, we can benefit from the work of so many talented people and organizations putting their time and energy into developing more responsible technology and AI, as well as bridging the knowledge gap for non-technical audiences. Just to name two favorites of mine: the comic book series called “We are AI” by Prof. Julia Stoyanovich and Falaah Arif Khan from R/AI - Center for Responsible AI, and the privacy enhancing tools made available by the Electronic Frontier Foundation.

However, it is important to recognize that any individual impact we can have on protecting our rights in the face of threats to privacy and civil liberties will inevitably be limited. When we think about abuse and threats to individual rights brought by abusive data and geolocation collection practices, the data brokerage market, ad-tracking systems, and surveillance technologies, for instance, our best bet is to apply collective pressure on policymakers, from whom we can demand better public policies, and on companies, in our capacities as users and consumers.

What technical safeguards, oversight, and best-practices are needed to ensure safety by design and protection of human rights while mitigating harms?

In the decades before terms like ESG, B Corps and sustainable brands became popular, companies dealt with their impacts on human rights, the environment, and society through what was called Corporate Social Responsibility. One of the most telling signs on whether a company takes their social responsibility seriously is to identify where the “CSR” department is located within their organizational chart. If the CSR team has direct engagement with the activities that make up the company’s core business, for instance, they are more likely to influence daily operations and advocate that business decisions consider issues beyond economic considerations like profit or costs. A company can also be very engaged in philanthropic activities, however, its potential for creating the most meaningful positive societal impact will almost always lie within its core business and business model.

This is also true for the development and design of AI technology. I’m a big defender of embedding safety checks in the design and engineering steps of a roadmap, using a risk-based approach.There are amazing resources out there in terms of safety-by-design and privacy-as-code. When it comes to AI, the Solutions to the risks and potential harms created by the technology will require creative Solutions that go beyond procedural compliance and inform the technical way in which products are built. This of course adds friction to the process, and, as we know, the industry famously likes to move fast and break things. But there is also an argument to be made that building sustainable products makes for better business decisions.

How do we ensure that new and emerging technologies and systems do not reproduce and augment existing inequalities?

From the business side, companies can avoid their technology systems from reproducing and amplifying societal and regional inequalities by ensuring that local context and societal power dynamics are taken into account throughout their design, production and deploying steps. This means building processes that focus on the messy parts of that technology being deployed in the different parts of the real world, a consideration often overlooked by optimistic engineers and product managers. It also means committing to hiring and maintaining a disciplinary, culturally, gender, and regional diverse workforce at all hierarchical levels. Finally, it also means fostering a culture of transparency and multi stakeholder engagement with civil society, academic researchers, and advocacy groups.

Tell us about your role:

I work as an independent consultant helping tech companies, international organizations, and governments in projects related to tech policy, political risk, ESG and human rights, particularly in Latin America. My role involves analytical and research work, advocating for policy positions, and engaging with different stakeholders in multicultural and politically-sensitive environments.

I studied Law at the University of São Paulo and I hold a master's degree in Global Politics and Security from Georgetown University's School of Foreign Service. Some of my previous work experiences include the United Nations, the World Bank, the Public Defense of the State of São Paulo, and a Washington, DC-based political risk consulting firm.

How did you carve out your career in the Responsible Tech ecosystem?

My interests in privacy and human rights go far back in my career. I first started to seriously follow the tech policy debate while working in New York with Brazil's Mission to the United Nations, where I dealt with topics of Sustainable Development and the UN development agencies.

While pursuing my masters at Georgetown, where I took all of the tech and development classes that I could, a job working directly with tech policy did not feel like a possibility for non-STEM backgrounds. In hindsight, this may have been the case because tech seemed like an uncommon career path for traditional IR students, and companies in the tech space did not seem to heavily recruit within our academic community.

Staying true to my passion for this industry, however, I kept studying the subject and engaging with experts in the ecosystem on my own. After working with a political consulting firm for a while, I managed to build relationships with partners in the field and slowly pivot to where I am now.

“

Trustworthy AI requires participation from a wide range of stakeholders so that policies and design practices reflect perspectives historically excluded from technology innovation.

”

Valérie Morignat, PhD

Founder & CEO

Intelligent Story

What are key factors in the way AI is developed/deployed that have the most impact on protecting human rights?

The development of trustworthy AI implies transformations that are endogenous and exogenous to the AI lifecycle. A foundational step is to educate leaders on AI's transformative effects on society and sensitize them to the importance of ethical risk oversight and mitigation. In addition, ethical AI governance across the organization is essential to mitigate risk and maximize the value of AI responsibly. In establishing an AI governance body, organizations can foster ethical practices at scale by setting forth the principles, missions, policies, processes, controls, programs, and incentives necessary to operationalize ethical AI. Organizational AI governance also supports mutually beneficial engagements with external stakeholders, including academics, regulators, and advocacy groups. Trustworthy AI requires participation from a wide range of stakeholders so that policies and design practices reflect perspectives historically excluded from technology innovation. The multidisciplinary and diverse makeup of AI teams is another critical factor. We know that teams with members with diverse backgrounds and skills are more likely to identify complex and ambiguous systemic risks, fallacies, flaws, and unsupported assumptions that may adversely impact minoritized communities and society. In addition, organizations should implement the principles of Responsible AI throughout the AI lifecycle and in their key performance indicators. These principles include fairness, inclusiveness, privacy, safety, transparency, explainability, auditability, accountability, and the environmental footprint of AI models.

In summary, successfully deploying trustworthy AI relies on AI ethics-ready leadership and workforce; transparent governance; proactive engagements with advocacy groups; diverse, inclusive, and multidisciplinary AI teams; responsible innovation frameworks; transparent and auditable systems; incentivizing ethical practices; and ethical risk oversight and mitigation.

How can we strengthen interdisciplinary engagement in technology development, deployment, governance, regulation and oversight?

Before becoming a computer science field, the quest for autonomous, intelligent machines implied the cross-pollination of many disciplines, including the creative arts, law, theology, philosophy, and many others. During the Middle Ages, philosophers Roger Bacon and Albertus Magnus created a talking head and an android. Leonardo Da Vinci built several automata and a self-propelled cart. A logician, mathematician, and philosopher of the Enlightenment, Leibniz anticipated computational AI in his pioneering use of the binary system. The origins of autonomous, intelligent machines teach us a lot about the role of interdisciplinary thinking in breakthrough innovation. AI's cross-functional development today confirms its significance.

One way to promote interdisciplinary engagement is integrating social, legal, political, epistemological, and ethical issues into AI education. Diverse stakeholder participation in AI design and governance is also key to ensuring AI standards and practices reflect viewpoints historically excluded from AI development. In leading AI companies, cross-functional teams already use ideation tools and documentation to foster shared mental models. They also engage multidisciplinary stakeholders in setting policies and standards.

However, interdisciplinary AI design thinking is at an early stage. Organizational silos, legacy systems, communication challenges between disciplines, divergent KPIs, and mindsets impede interdisciplinary innovation. A key to success is cultivating interdisciplinary engagements with diverse external stakeholders and hiring leaders adept at analogical reasoning and developing shared mental models across teams. The ability to execute at the intersection of several fields requires interdisciplinarity and cross-functional reasoning, which are pivotal to building teams that are ethical, culturally responsive, and network-driven.

What emerging regulatory frameworks are having the greatest impact on AI development at the present time?

I consider the Artificial Intelligence Act (AIA) proposed by the European Commission in April 2021 the most influential regulatory framework. The AIA will impact AI ventures and policymaking in European and non-European countries alike.

The proposal involves principled compliance requirements and a proportionate risk-based regulatory approach tailored to application domains. The AIA deems AI that impacts human rights or user safety as high-risk when applied to critical infrastructure, education, training, employment, workforce management, essential public and private services, law enforcement, border control, migrations, and the administration of justice. Such systems must be non-discriminatory, traceable, transparent, auditable, robust, accurate, secure, and subject to human oversight. The AI operators of high-risk systems will be required to maintain detailed technical documentation, perform compliance assessments, register in the EU database, monitor performance, and enhance coordination with market surveillance agencies.

Through reinforcing the central role of ethics, the EU's landmark legal framework accelerates the true potential of AI. In addition, the act addresses the risks to users' psychological integrity and existential autonomy. Specifically, the AIA forbids subliminal behavior distortion, commercial exploitation of vulnerable users, and social scoring that produces decontextualized effects. The proposal stresses the importance of responsible competitiveness frameworks that mitigate risks through long-term strategies focused on people. It offers policymakers and regulators a foundation to build an interoperable approach to AI standards and governance. Businesses should view it as a unique opportunity to lead the way in responsible AI governance.

Tell us about your role:

I am the founder and CEO of Intelligent Story, a San Francisco-based AI consultancy firm specializing in ethical risk, AI strategy and design, and workforce AI-readiness. I also serve as AI Ethics and Policy Advisor with The Cantellus Group (San Francisco) and Professor of AI Ethics, Strategy and Responsible Design at aivancity, School for Business, Technology & Society (Paris).

As an ethical AI expert, responsible design strategist, and executive education partner to the public and private sectors, my work is interdisciplinary. In addition to serving as a subject matter expert to consortia and government agencies on AI ethics and policy issues, I help corporations and higher education institutions succeed in their AI transformation. My missions include devising AI strategies, performing AI risk and opportunity analysis, building ethical AI governance, forging multistakeholder partnerships, advising AI product teams, and pioneering cross-functional executive education programs that foster responsible AI competitiveness.

“

As AI becomes ubiquitous, it is imperative that we center the design, development, and deployment of AI on people and outcomes, not on privilege and profits.

”

Vilas Dhar

President

Patrick J. McGovern
Foundation

What are key factors in the way AI is developed/deployed that have the most impact on protecting human rights?

As AI becomes ubiquitous, it is imperative that we center the design, development, and deployment of AI on people and outcomes, not on privilege and profits. Decisions about AI are not purely technology questions: the consequences influence the fundamental ways we interact as humans and must incorporate dignity and agency as key inputs. Technology for its own sake is exciting, but it rarely delivers long-term sustainable returns; including social welfare and equitable distribution in conversations about AI will lead to better long-term gains. A holistic and ethically-grounded approach to AI technology can improve the human condition in the digital age. We have the opportunity to build tools to address our most pressing global challenges - malnutrition, climate change, and equitable economic systems. At the same time, we can create new forms of economic participation by ensuring that menial tasks are automated to create new skilling opportunities for workers.

How can we equip policymakers to better keep pace with the speed of AI and Machine Learning development?

A basic level of digital literacy is something we can all benefit from – we don't need to be technologists to be informed as to how emergent technologies might affect our lives - now or in the future. It's important that policymakers (and all of us!) have access to non-partisan tools and training to better understand AI and other technologies - and we have a responsibility to create resources and shared narratives that promote general digital literacy.

We're partnering with Stanford University's Human-Centered Artificial Intelligence (HAI) Institute to support the implementation of rigorous education programs to prepare policymakers with an understanding of AI technologies and trends, as well as with Oxford University's Institute for Ethics in AI and the Harvard Belfer Center's Technology and Public Purpose Project to support cross-sector educational initiatives and collaboration on AI design and implementation.

How do we ensure that new and emerging technologies and systems do not reproduce and augment existing inequalities?

Civil society must continue to demand transparency and accountability about the changes and their impacts on society. For decades, we've been told that participating in a connected digital world requires giving up our personal data, our privacy, and even our agency. It's true that social media platforms have revolutionized our ability to stay connected, participate in civil society, build community, nurture relationships across distance and borders, and benefit from e-commerce. But simultaneously, we've seen the emergence of disturbing practices surrounding data collection, use, surveillance, and manipulation. Institutions that derive value from our data should be challenged to reassess their fundamental assumptions about the use of our assets to promote their commercial interests and realign their practices to shared social benefit.

What specific structural changes to incentives and business models are needed in order to prioritize user well being and human rights? What kinds of incentive systems (companies, people, regulators, societal) need to be tapped and how?

As part of their core business strategy, many companies are already heavily investing in technology to transform capacity and reach. Such investments also present unexplored - and often unexpected - opportunities to leverage learnings to accelerate social outcomes, which present the potential to create stakeholder value and trust in the company, sustain new markets and add enterprise value, as well as advance social progress.

Looking at your crystal ball of a positive tech future. Let's say we have human-centered AI. What would humane AI look like? What would its pillars be?

Technology can extend and equip our ability to express our rights, build social movements, find economic agency and identity, and build communities of purpose around the world. I'm optimistic and energized about human-centered AI - a powerful tool that, when coupled with thoughtful implementation, can help to close the distance between problem and solution - ultimately delivering new insights and actions for a more sustainable, just, and equitable world.

> “
> **We shouldn’t automate for the sake of automation itself – not when we risk that the said automation would have a detrimental effect on people, social interactions or the environment. At the end of the day, automation should aim at making people’s lives easier rather than more complicated, unfair and anxious.**
> ”

Vyara Savova
Human Rights & Tech Lawyer
vyara.io
Public Policy Expert
The European Crypto Initiative

What are key factors in the way AI is developed/deployed that have the most impact on protecting human rights?

We shouldn’t automate for the sake of automation itself – not when we risk that the said automation would have a detrimental effect on people, social interactions or the environment. At the end of the day, automation should aim at making people’s lives easier rather than more complicated, unfair and anxious.

The key factors should include:

- Doing more good than harm – if the harm is potentially too much, then the use of AI should be limited.
- Sharing easily digestible information with the people and communities that will be affected and involving them in the discussion as early as possible.
- Being well aware and open about the ethical limitations and risks that come with the desired automation - non-representable datasets, blackboxes, etc.

How can we strengthen interdisciplinary engagement in technology development, deployment, governance, regulation and oversight?

Nothing beats direct communication – it would be best to get experts from all disciplines to come together and express their worries. There might be a need for “interpreters”, such as people with both technical and legal backgrounds, business and regulatory, that would help merge the issues and point discussions in the right direction.

What are the areas of AI and Machine Learning development whose potential social implications are currently being overlooked in the general discourse? Which areas get comparatively too much attention relative to their potential impact?

In my opinion we aren’t talking about the risks of social scoring enough – apart from the direct social scoring that we might not really get exposed to (hopefully), there’s also the risk of indirect social scoring that is industry-specific, but important nonetheless. For example, if we allow for AI to evaluate an aspect of our healthcare status that would later be included in our digital health dossier, this will serve as a very niche, but still important score that will influence aspects of our future medical treatment.

What emerging regulatory frameworks are having the greatest impact on AI development at the present time?

The European Union will most definitely set the scene for the broader AI regulations worldwide. Another influential one is the Chinese take on AI, which might not get widely copied, but will likely be as effective when it comes to regulating the day-to-day applications of AI and machine learning.

How can we equip policymakers to better keep pace with the speed of AI and Machine Learning development?

Speaking from an European perspective, it is crucial that we start by differentiating between what policymakers suggest as politicians from their expert suggestions. I believe that we should focus on influencing the experts by equipping them with the best possible arguments when they go into discussions and negotiations with other core decision makers.

How can individuals be empowered to know and protect their rights in the face of potential threats to privacy and civil liberties linked to use of AI?

First, we need to make it clearer for them that there are rights to be protected to begin with. And by “we” I mean all people involved in the discussion around responsible AI. It is necessary as, for example, not all people might recognise the unfair treatment they receive from an automated system as a form of discrimination.

Resources & Organizations

Suggested resources from our profile interviews

Our interviewers' answer to:

Do you have any resources we should know about (Related to the seven key areas)?

"Afua Bruce's and Amy Sample Ward's framework (in "The Tech that comes next") is a great resource to understand the different touchpoints that a technology like AI can have in our society, and what questions we need to ask ourselves to make it more equitable and inclusive.

- Alberto Rodriguez, Senior Program Manager, Public Interest Technology

---

"Principles and values included in the Responsible Artificial Intelligence: Recommendations to Guide the University of California's Artificial Intelligence Strategy report.

- Alexa Koenig, Executive Director, Human Rights Center, UC Berkeley School of Law

---

"While AI has reached a level of maturity, we are still only at the beginning of the AI era. As such, many of the ideas, priorities, and considerations around our trust in this powerful technology are still being defined. To provide some clarity on just what constitutes trustworthy AI and how to achieve it, I wrote Trustworthy AI: A Business Guide for Navigating Trust and Ethics in AI. The book is primarily written for a professional audience, and it digs into the dimensions of trust, the philosophical questions they raise, and how organizations can manage and govern AI to its greatest potential value."

- Beena Ammanath, Executive Director, Deloitte Global AI Institute

"I'd point people to the following:
A Human Rights-Based Approach to Content Governance
A Human Rights Impact Assessment of End-to-End Encryption
Human Rights Due Diligence of Products and Services
Applying the UNGPs to Technology
Human Rights and the Right to Science

- Dunstan Allison-Hope, Vice President, BSR

"I am the co-host and co-founder of The Radical AI Podcast, where we discuss topics surrounding responsible technology design with world-renowned scholars and experts on these topics. You can find more information and listen to previous episodes at radicalai.org, or join our Twitter community by following us @radicalaipod."

- Jessie J. Smith, PhD Student - University of Colorado - Boulder

---

"My research paper on AI and human rights will be published by Chatham House in September 2022. I have written widely on other aspects of technology and human rights, with a particular focus on disinformation and political discourse, technical standards, and the right to freedom of thought. My website and LinkedIn."

- Kate Jones, Associate Fellow, Chatham House

"I cannot recommend the work of the Stop Killer Robots coalition (formerly the Campaign to Stop Killer Robots) highly enough. Their tireless efforts to organize opposition to the use and development of fully autonomous weapons have pushed over 30 countries to endorse a ban on such weapons, paving the way for a binding treaty outlawing new categories of extremely dangerous weapons. In terms of a specific resource, I would point to Human Rights Watch's report, "Stopping Killer Robots," which was authored by Mary Wareham, the former coordinator of the coalition. The report is the first to outline the position taken by every UN member state that has voiced an opinion relating to lethal autonomous weapons systems. It is a helpful guide that contains clues for how individuals around the world might pressure their own governments to take a bolder stance against killer robots."

- Kevin Klyman, Lead Technology Researcher at the Avoiding Great Power War Project, Harvard's Belfer Center for Science and International Affairs

---

"My book: "Social Media Impacts on Conflict and Democracy: The Techtonic Shift"

- Lisa Schirch, Starmann Chair and Professor of the Practice in Peacebuilding - Notre Dame University

“At CAIDP, we provide semester long AI policy training (Research Group) to advocates, human right defenders and professionals from around the world (currently 40 countries). Our policy work aims to promote development and use of AI in ways that respect fundamental rights, rule of law and democratic values. Our applications for Fall 2022 semester are open now until July 1st.

We have just published the 2nd edition of our AI & Democratic Values Index which reviews national AI strategies of 50 countries against the public system use of AI on the ground and holds governments accountable.

Council of Europe has now taken over from CAHAI’s work and moving forward with a possible convention / legally binding instrument that will ensure AI systems take into account human rights impact assessments and protect rights and rule of law. This is a global level work that requires close attention.”

- Merve Hickok, Founder, AIethicist.org ; Research Director, Center for AI & Digital Policy

---

“Websites for the book The Costs of Connection on colonizedbydata.com and sup.org. Tierra Comun network of activists and academics centered in Latin America.”

- Nick Couldry, Professor of Media, Communications and Social Theory, London School of Economics

---

“MA Internet Equalities; Algorithmic Equity Toolkit; Defining AI in Policy versus Practice; UK Tech Workers Know Your Rights; United Tech and Allied Workers; Devin Guillory, Combatting Anti-Blackness in the AI Community (The Ethics of AI in Context); Malcolm X Interview At UC Berkeley (1963)”

- Dr. Peaks Krafft, Senior Lecturer, University of the Arts London

---

“Eticas Consulting Resources”

- Gemma Galdon-Clavell, Founder & CEO, Eticas Consulting

“A bunch! Microsoft has a number of open source tools embedded within the Responsible AI dashboard, which was introduced in 2021 as a comprehensive experience with capabilities for data exploration, fairness assessment, model interpretability, error analysis, and more. “On the design front, the Microsoft HAX toolkit is a great set of resources that helps prioritize human-AI interaction principles when building new experiences. There’s also a lot of meaningful efforts that have come out of MIT, such as RAISE and Joy Buolamwini’s Coded Bias.”

- Jenna Hong, Product Manager, Microsoft AI Development and Acceleration

Organizations

5Rights Foundation (@5RightsFound)
"5Rights Foundation exists to make systemic changes to the digital world that will ensure it caters for children and young people, by design and default, so that they can thrive. 5 Rights work with, and on behalf of, children and young people to reshape the norms of the digital world in four priority areas: design of service, child online protection, children and young people's rights and data literacy." 5rightsfoundation.com

Resource: Data Literacy

Access Now (@accessnow)
"Access Now defends and extends the digital rights of users at risk around the world. By combining direct technical support, comprehensive policy engagement, global advocacy, grassroots grantmaking, legal interventions, and convenings such as RightsCon, we fight for human rights in the digital age." accessnow.org

Resources: RightsCon, Digital Security Helpline

Access Partnership (@accessalerts) *"We focus on policy as it affects technology, and our government relations expertise allows us to predict the future of regulation and manage the outcomes of policy trends globally. To do this, we deploy effective, proprietary and stress-tested processes that ensure our clients' government affairs goals are met every time."* AccessPartnership.com

Resource: Reports

Accountable Tech (@accountabletech)
"We are facing a crisis of truth. Accountable Tech advocates for the social media companies at the center of today's information ecosystem to strengthen the integrity of their platforms and our democracy." Accountabletech.org

Resource: The Tech Transparency Project (TTP) is a research initiative of Accountable Tech that seeks to hold large technology companies accountable.

Africa Digital Rights' Hub (@hub_adr)
"The Africa Digital Rights' Hub is a not-for-profit think tank registered in Ghana that advances and promotes research and advocacy on digital rights across the African continent. Interested in the impact of digital technology on people living in the Continent, the Hub brings together academic researchers, stakeholders, policy makers, regional and international bodies to address digital rights issues in Africa." africadigitalrightshub.org

Resource: Blog

AIandYou (@AIandYou2)
"AIandYou is a community-facing platform intended to foster a more inclusive and empathetic artificial intelligence ecosystem. AIandYou is creating a global dialogue between AI leaders and marginalized communities in order to prepare for AI's impact at the local level, strengthen our local communities through the use of AI and identify solutions that minimize bias." aiandyou.org

Resources: Why does AI matter to my community?, AIandYou Podcast

---

AI Ethics Lab (@aiethicslab)
"AI Ethics Lab brings together researchers and practitioners from various disciplines to detect and solve issues related to ethical design in AI. Through collaboration between computer scientists, practicing lawyers and legal scholars, and philosophers, the Lab offers a comprehensive approach to ethical design of AI-related technology. Our goal is to enhance technology development by integrating ethics from the earliest stages of design and development for the mutual benefit of industry and communities. Our work aims to provide ethics guidance to researchers, developers, and legislators." aiethicslab.com

Resources: AI Ethics Analysis, AI Ethics Roadmap, AI Ethics Strategy, AI Ethics Training

---

AI Infrastructure Alliance (@AiInfra)
"The AI Infrastructure Alliance is helping make the canonical stack in AI/ML a reality, by bringing together the essential building blocks for the Artificial Intelligence applications of today and tomorrow. We represent some of the top venture backed companies, projects and service providers in the rapidly developing AI/ML space."
https://ai-infrastructure.org

Resource: AI Infrastructure Landscape

AI Now (@ainowinstitute)
"The AI Now Institute aims to produce interdisciplinary research and public engagement to help ensure that AI systems are accountable to the communities and contexts in which they're applied." ainowinstitute.org

Resources: Research, Policy

AI for Peace (@AI4Peace)
"At AI for Peace, we believe AI has the potential to change our lives substantially in the next decade. It can lead to rapid improvement of our lives and welfare, but it can also lead to negative consequences, even if that is not the intention. We believe that with a technology as powerful and complex as AI, constructive dialogue and engagement between academia, industry and civil society is critical to maximizing the benefits and minimizing the risks to human rights, democracy and human security. We want to make sure that peace-builders, humanitarians and human rights activists are well informed, and their critical voices heard in this process." aiforpeace.org

Resource: Newsletter

AI Policy Exchange (@aipolex)
"AI Policy Exchange is an international cooperative association of individuals and institutions working at the intersection of AI and public policy. The association operates independently and on a non-profit basis with its day-to-day affairs managed by the AI Policy Exchange Secretariat, based out of New Delhi, India anchoring its Editorial and Research Networks to produce deliverables that can create an AI-literate society and inform better AI policies." aipolicyexchange.org/

Resource: Why you should be at the center of machine learning policies, ethics & responsible development

---

Alan Turing Institute (@turinginst)
"The Alan Turing Institute is the UK's national institute for data science and artificial intelligence...The Institute's mission is to: undertake data science research at the intersection of computer science, mathematics, statistics and systems engineering; provide technically informed advice to policy makers on the wider implications of algorithms; enable researchers from industry and academia to work together to undertake research with practical applications; and act as a magnet for leaders in academia and industry from around the world to engage with the UK in data science and its applications." turing.ac.uk/

Resource: Turing Institute Data Ethics

AlgorithmWatch (@algorithmwatch)
"AlgorithmWatch is a non-profit research and advocacy organization that is committed to watch, unpack and analyze automated decision-making (ADM) systems and their impact on society." algorithmwatch.org/en

Resource: Policy & Advocacy

Algorithmic Justice League (@AJLUnited)
"The Algorithmic Justice League is an organization that combines art and research to illuminate the social implications and harms of artificial intelligence. AJL's mission is to raise public awareness about the impacts of AI, equip advocates with empirical research to bolster campaigns, build the voice and choice of most impacted communities, and galvanize researchers, policymakers, and industry practitioners to mitigate AI bias and harms." ajl.org/

Resources: Actionable Auditing: Investing the Impact of Biased Performance Results of Commercial AI Products; Coded Bias Documentary, Facial Recognition Technologies: A Primer

---

Alliance for Peacebuilding
The group is an open community of practice to advance "digital peacebuilding", defined as the analysis of and response to online conflict dynamics and the harnessing of digital tools to amplify peacebuilding outcomes." allianceforpeacebuilding.org/digital-peacebuilding-cop

---

All Tech Is Human (@AllTechIsHuman)
"All Tech Is Human is a non-profit organization committed to building the Responsible Tech pipeline; making it more diverse, multidisciplinary, and aligned with the public interest. We believe that we can build a better tech future by changing those involved in it. By uniting a diverse range of participants across civil society, government, and industry we are uniquely positioned in helping co-create a tech future that is aligned with the public interest." AllTechIsHuman.org

Resources: Responsible Tech Guide, HX Report: Aligning Our Tech Future With Our Human Experience, Job Board, University Ambassadors, Mentorship Program

American Civil Liberties Union (ACLU) (@ACLU)
"The ACLU is an non-profit, non-partisan organization of people who believe in the power of action. We are united by the quest – "We the people dare to create a more perfect union." Whether in the courts, statehouses, Congress or communities, we fight to defend the rights that the Constitution guarantees to all of us —regardless of who we are, where we come from, whom we love, or what we believe. Together, we take up the toughest civil rights and liberties challenges of our time. We seek to inspire those who want change to become the ones who make change."

Resource: Privacy & Technology

Amnesty International, Technology & Algorithmic Accountability Lab (@amnestytech)
"Amnesty Tech is a global collective of advocates, hackers, researchers and technologists... We aim to: Bolster social movements in an age of surveillance, Challenge the systemic threat to our rights posed by the surveillance-based business model of the big tech companies, Ensure accountability in the design and use of new and frontier technologies, Encourage innovative uses of technology to help support our fundamental rights." amnesty.org/en/tech

Resources: Disrupting Surveillance, Big Data and AI, Censorship

---

ARTICLE 19 (@article19org)
"ARTICLE 19 works for a world where all people everywhere can freely express themselves and actively engage in public life without fear of discrimination. We do this by working on two interlocking freedoms: the Freedom to Speak, and the Freedom to Know. When either of these freedoms come under threat, ARTICLE 19 speaks with one voice." article19.org

Resources: Global Expression Report, Digital Rights, Privacy and surveillance

---

Asian Americans Advancing Justice - Telecommunications and Technology (@AAAJ_AAJC)
"As access to technology transitions from being a benefit to an absolute necessity in our everyday lives, understanding the telecommunications- and technology-related needs of our diverse community becomes increasingly important. While digital innovation continues to be a boon for Asian Americans, it also represents a potential source of mischief and real-world harm. Our work focuses on ensuring that the principles of opportunity, fairness, and equity are protected online." advancingjustice-aajc.org/telecommunications-and-technology

Resources: Community Resource Hub: Privacy, Facial recognition technology, Algorithmic bias, Social media surveillance

Aspen Tech Policy Hub (@AspenPolicyHub)
"The Aspen Tech Policy Hub is a West Coast policy incubator, training a new generation of tech policy entrepreneurs. Modeled after tech incubators like Y Combinator, we take tech experts, teach them the policy process through fellowship and executive education programs in the Bay Area, and encourage them to develop outside-the-box solutions to society's problems." AspenTechPolicyHub.org

Resource: Aspen Tech Policy Hub Projects

Black in AI (@black_in_ai)
"Black in AI increases the presence and inclusion of Black people in the field of AI by creating space for sharing ideas, fostering collaborations, mentorship and advocacy."
blackinai.github.io/#

Resource: Programs

Berkman Klein Center (@BKCHarvard)
"The Berkman Klein Center's mission is to explore and understand cyberspace; to study its development, dynamics, norms, and standards; and to assess the need or lack thereof for laws and sanctions...We are a research center, premised on the observation that what we seek to learn is not already recorded. Our method is to build out into cyberspace, record data as we go, self-study, and share. Our mode is entrepreneurial nonprofit." cyber.harvard.edu

Resource: Projects and Tools

Business & Human Rights Resource Centre (@BHRRC)
"We work with everyone to advance human rights in business. We track over 9000 companies, and help the vulnerable eradicate abuse. We empower advocates. We amplify the voices of the vulnerable, and human rights advocates in civil society, media, companies, and governments. We strengthen corporate accountability. We help communities and NGOs get companies to address human rights concerns, and provide companies an opportunity to present their response.We *build corporate transparency. We track the human rights policy and performance of over 9000 companies in over 180 countries, making information publicly available."* business-humanrights.org/en/

Resource: Technology & Human Rights

Business for Social Responsibility (BSR) (@BSRnews)
"BSR™ is an impact-driven sustainability organization that works with its global network of leading companies to create a world in which all people can thrive on a healthy planet. With offices in Asia, Europe, and North America, BSR™ provides insight, advice, and collaborative initiatives to help business see a changing world more clearly, create long-term value, and scale impact."

Resource: One area of expertise: Human Rights

Carnegie Endowment for International Peace (@CarnegieEndow)
"In a complex, changing, and increasingly contested world, the Carnegie Endowment helps countries take on the most difficult global problems and safeguard peace and security through independent analysis, strategic ideas, support for diplomacy, and training the next generation of international scholar-practitioners." carnegieendowment.org/

Resources: Technology and International Affairs, Cyber Policy Initiative

---

Center for AI and Digital Policy (@theCAIDP)
"The Center for AI and Digital Policy aims to ensure that artificial intelligence and digital policies promote a better society, more fair, more just, and more accountable – a world where technology promotes broad social inclusion based on fundamental rights, democratic institutions, and the rule of law." caidp.org

Resource: Artificial Intelligence and Democratic Values Index, 2021 report (2022)

The Center for Data Ethics and Justice (@uvadatascience)
"Centering ethics and justice at the core of data science The Center for Data Ethics and Justice equips researchers, faculty and students at the University of Virginia to address relevant ethical, social and political issues that intersect with data science." datascience.virginia.edu/center-data-ethics-and-justice

Resource: Projects

Center for Data Innovation (@DataInnovation)
"The Center for Data Innovation conducts independent research and formulates public policies to enable data-driven innovation in the public and private sector." datainnovation.org

Resources: More Than Meets The AI: The Hidden Costs of a European Software Law, Principles to Promote Responsible Use of AI for Workforce Decisions, The Artificial Intelligence Act: A Quick Explainer

Center for Democracy & Technology (@CenDemTech)
"The Center for Democracy & Technology. Shaping tech policy & architecture, with a focus on the rights of the individual...Our team of experts includes lawyers, technologists, academics, and analysts, bringing diverse perspectives to all of our efforts." Cdt.org

Resources: Making Transparency Meaningful: A Framework for Policymakers, CDT and AAPD Report – Centering Disability in Technology Policy: Issue Landscape and Potential Opportunities for Action, CDTs collection of reports and insights

---

Change the Terms (@changeterms)
"To ensure that companies are doing their part to help combat hateful conduct on their platforms, organizations in this campaign will track the progress of major tech companies —especially social media platforms— to adopt and implement these model corporate policies and give report cards to these same companies on both their policies and their execution of those policies the following year." changetheterms.org

Resource: Adopt the Terms

Citizen Digital Foundation (@Citizendigital1)
"Citizen Digital Foundation aims to promote safe and responsible navigation and innovation in the digital ecosystem...We address issues of digital distraction, cybersecurity, attention-harvesting, subconscious behaviour manipulation, disinformation, polarisation and teenage mental health through interventions, consultancy, and advisory for students, educators, parents, media professionals, corporates, technologists, entrepreneurs and policy makers."
https://citizendigitalfoundation.org/

Resource: Global Conversations

Center for Inclusive Change *"At The Center for Inclusive Change, we guide organizations to help them bring together the three critical business practices that are essential in ensuring AI is human-centered, sustainable, fair, and equitable: 1) sound DEI&B practices (the foundation); 2) governance-centered AI policy / procedure / structure (the framework), and 3) inclusive change management (where the rubber meets the road). Our approach is purposefully practical and instructive, because failure is not an option when it comes to safe AI."* inclusivechange.org/

Resource: AI Governance Solutions

Common Sense Media (@CommonSense)
"Common Sense is dedicated to helping kids thrive in a world of media and technology. We empower parents, teachers, and policymakers by providing unbiased information, trusted advice, and innovative tools to help them harness the power of media and technology as a positive force in all kids' lives." Commonsensemedia.org

Resources: Resource for Parents, Resource of Educators, Resource for Advocates, Tweens, Teens, Tech, and Mental Health: Coming of Age in an Increasingly Digital, Uncertain, and Unequal World 2020

Center for Internet and Technology Policy "*The Center for Information Technology Policy (CITP) is a nexus of expertise in technology, engineering, public policy, and the social sciences. In keeping with the strong University tradition of service, the Center's research, teaching, and events address digital technologies as they interact with society."* citp.princeton.edu

Resources: Our Work: Privacy & Security, National Security & Surveillance

---

Center for Long-Term Cybersecurity (@CLTCBerkeley)
"The Center for Long-Term Cybersecurity was established in 2015 as a research and collaboration hub in the School of Information at the University of California, Berkeley. The mission is to help individuals and organizations address tomorrow's information security challenges to amplify the upside of the digital revolution." cltc.berkeley.edu

Resources: AI Policy Hub, AI Security Initiative, Decision Points in AI Governance, Guidance for the Development of AI Risk and Impact Assessments, The Flight to Safety-Critical AI: Lessons in AI Safety from the Aviation Industry, Toward AI Security: Global Aspirations for a More Resilient Future, "What, So What, Now What?": Adversarial Machine Learning

---

Center for Responsible AI at NYU (@AIResponsibly)
"The Center for Responsible AI is a comprehensive laboratory for accelerating responsible AI practices. We are building a future in which responsible AI is the only kind accepted by society." AIResponsibly.com

Resources: Comparing Apples and Oranges: Fairness and Diversity in Ranking, Disaggregated Interventions to Reduce Inequality, Innovating Public Procurement to Mitigate the Risks of Artificial Intelligence Systems, Taming Technical Bias in Machine Learning Pipelines, Transparency, Fairness, Data Protection, Neutrality: Data Management Challenges in the Face of New Regulation

Center on Privacy and Technology (@GeorgetownCPT)
"*The Center on Privacy and Technology is a think tank based at the [Georgetown University] Law Center . It aims to bridge the gap between the policy and academic worlds on privacy, and train students to be leaders in privacy practice, policymaking, and advocacy.*" law.georgetown.edu/privacy-technology-center

Resource: Publications

Center for Security and Emerging Technology - CyberAI Project (@CSETGeorgetown)
"The CyberAI project focuses on the intersection of cybersecurity and artificial intelligence (AI). We seek to understand how advances in AI may alter the current state of cybersecurity and, conversely, how the cybersecurity of AI systems affects their safe and trusted development, fielding and operations"

Resource: AI and the Future of Disinformation Campaigns, Deepfakes: A Grounded Threat Assessment, Exploring Clusters of Research in Three Areas of AI Safety, Key Concepts in AI Safety: Interpretability in Machine Learning, Machines, Bureaucracies, and Markets as Artificial Intelligences, Trends in AI Research for the Visual Surveillance of Populations

---

Center for Strategic and International Studies (@CSIS)
"The Center for Strategic and International Studies (CSIS) is a bipartisan, nonprofit policy research organization dedicated to advancing practical ideas to address the world's greatest challenges." csis.org

Resource: Cybersecurity and Technology

Center for Technology Innovation at Brookings (@BrookingInst)
"Founded in 2010 and led by Director Nicol Turner Lee*, the Center for Technology Innovation (CTI) at Brookings focuses on delivering research that affects public debate and policymaking in the arena of U.S. and global technology innovation. Our research centers on identifying and analyzing key developments to increase innovation; developing and publicizing best practices to relevant stakeholders; briefing policymakers about actions needed to improve innovation; and enhancing the public and media's understanding of technology innovation."*
brookings.edu/about-the-center-for-technology-innovation/

Resource: TechTank

Confederation of Laboratories for Artificial Intelligence Research in Europe (CLAIRE) (@vision_claire)

"CLAIRE seeks to strengthen European excellence in AI research and innovation. The network forms a pan-European Confederation of Laboratories for Artificial Intelligence Research in Europe. Its member groups and organisations are committed to working together towards realising the vision of CLAIRE: European excellence across all of AI, for all of Europe, with a human-centred focus." claire-ai.org/

Resource: Important Documents

Data for Black Lives (@Data4BlackLives)
"Data for Black Lives is a movement of activists, organizers, and mathematicians committed to the mission of using data science to create concrete and measurable change in the lives of Black people. Since the advent of computing, big data and algorithms have penetrated virtually every aspect of our social and economic lives." d4bl.org

Resource: Consentful Tech Curriculum

Data & Society (@DataSociety)
"*Data & Society is an independent nonprofit research organization. We believe that empirical evidence should directly inform the development and governance of new technology. We study the social implications of data-centric technologies and automation, producing original research that can ground informed, evidence-based public debate. We combine academic rigor with creative outreach to connect, convene, and sustain expert and practitioner networks. Since 2014, Data & Society has defined the field with original research and programming to break down disciplinary silos and connect provocative thinkers across sectors."* datasociety.net

Resource: Media Manipulation & Disinformation

DataKind (@DataKind)
"We are living inside a data revolution that is transforming the way we understand and interact with each other and the world - and it has only just begun. Every field is now having its "data moment," giving mission-driven organizations brand new opportunities to harness data to advance their work. From poverty alleviation to healthcare access to improved education, data science has the potential to move the needle on seemingly insurmountable issues, but only if there is close collaboration between data science and social sector experts." datakind.org

Resources: DataKind Playbook, Chapters: Bengaluru, San Francisco Bay Area, Singapore, United Kingdom, Washington DC

Deloitte Global AI Institute
"The Deloitte AI Institute applies research and eminence to help drive transformation and leadership in the Age of With™" deloitte.com/us/en/pages/deloitte-analytics/articles/advancing-human-ai-collaboration.html

---

Digital Grassroots (@digigrassroots)
"Digital Grassroots is a youth and female non-profit working to increase digital citizenship on Internet governance and digital rights among youth from underrepresented communities globally. Through open leadership, community engagement programs and mentorships, we promote youth activism in shaping the Internet ecosystem." digitalgrassroots.org/index.html

Resource: Community Tools

Digital Freedom Fund (DFF) (@df_fund)
"The Digital Freedom Fund supports strategic litigation to advance digital rights in Europe. DFF provides financial support and seeks to catalyse collaboration between digital rights activists to enable people to exercise their human rights in digital and networked spaces." digitalfreedomfund.org

Resources: Projects: Decolonising Digital Rights, Digital Rights For All

---

Digital Impact Alliance (DIAL) (@DIAL_community)
"The Digital Impact Alliance (DIAL) is a "think, do, replicate" tank housed at the United Nations Foundation. Our vision is of a world where services can safely reach everyone, everywhere using the power of digital technology and data. Our mission is to overcome the systemic barriers preventing digital solutions from going to scale." digitalimpactalliance.org

Resource: Principles for Digital Development

Distributed AI Research Institute (@DAIRInstitute)
"We are an interdisciplinary and globally distributed AI research institute rooted in the belief that AI is not inevitable, its harms are preventable, and when its production and deployment include diverse perspectives and deliberate processes it can be beneficial. Our research reflects our lived experiences and centers our communities." dair-institute.org/

Resource: Research

Emotional AI Lab
"This international research group examines the social and cultural impact of artificial intelligence technologies that function in relation to data about human emotion, moods and affective states." emotionalai.org

---

Encode Justice (@encodejustice)
"We're a coalition of youth activists and changemakers fighting for human rights, accountability, and justice under AI. Harnessing a global network of volunteers from all over the United States and World, we champion informed AI policy and encourage youth to confront the challenges of the age of automation through political advocacy, community organizing, educational programming, and content creation." medium.com/encode-justice

Resource: Blog

The Ethics and Governance of Artificial Intelligence Initiative
"*Launched in 2017, the Ethics and Governance of AI Initiative is a hybrid research effort and philanthropic fund that seeks to ensure that technologies of automation and machine learning are researched, developed, and deployed in a way which vindicate social values of fairness, human autonomy, and justice."* aiethicsinitiative.org

Resource: Ethics and Governance of AI Initiative

Electronic Frontier Foundation (@EFF)
"The Electronic Frontier Foundation is the leading nonprofit organization defending civil liberties in the digital world. Founded in 1990, EFF champions user privacy, free expression, and innovation through impact litigation, policy analysis, grassroots activism, and technology development. EFF's mission is to ensure that technology supports freedom, justice, and innovation for all people of the world." eff.org/

Resources: Whitepapers, Legal Cases

Electronic Privacy Information Center (@EPICprivacy)
"*Electronic Privacy Information Center is an independent nonprofit research center in Washington, D.C. EPIC's mission is to focus public attention on emerging privacy and related human rights issues. EPIC works to protect privacy, freedom of expression, and democratic values, and to promote the Public Voice in decisions concerning the future of the Internet."*
epic.org/

Resource: AI & Human Rights

The Engine Room (@EngnRoom)
"We are an international organisation that helps activists, organisations, and other social change agents make the most of data and technology to increase their impact." theengineroom.org

Resource: Research

EthicsNet (@KinderMachines)
"The EthicsNet team is working to empower people to infuse their personal sense of values into Artificial Intelligence, teaching it through examples, like a very young child. We want to make it easy to collect examples of behaviour, to create datasets of behavioral norms which best describe the creeds of specific demographics, as well as discovering those universal values that we all share." Ethicsnet.org

---

ForHumanity (@ForHumanity_Org)
*"To examine and analyze the downside risks associated with the ubiquitous advance of AI & Automation, to engage in risk mitigation and ensure the optimal outcome... ForHumanity***."**
forhumanity.center

Resource: Biometric: Humanity's Most Precious Data (Feb 2022)

Freedom House (@freedomhouse)
"Freedom House is founded on the core conviction that freedom flourishes in democratic nations where governments are accountable to their people; the rule of law prevails; and freedoms of expression, association, and belief, as well as respect for the rights of women, minority communities, and historically marginalized groups, are guaranteed." freedomhouse.org/

Resource: Technology & Democracy

Future of Privacy Forum (@futureofprivacy)
"A nonprofit organization serving as a catalyst for privacy leadership & scholarship, advancing principled data practices in support of emerging technologies" fpf.org

Resources: Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models, Data Protection Principles in Machine Learning, The Privacy Expert's Guide to AI And Machine Learning, The Spectrum of Artificial Intelligence – An Infographic Tool, Unfairness By Algorithm: Distilling the Harms of Automated Decision-Making, Warning Signs: Identifying Privacy and Security Risks to Machine Learning Systems

The Future Society (@thefuturesoc)
“*AI will change many aspects of our lives. These changes are not up to chance: everyone and every community should decide how AI is used. Citizens, politicians, programmers, scientists, businesspeople, artists, ... everyone should understand these changes and decide what to do about them. We at The Future Society believe intelligent collective discussions and actions on this topic are needed and our mission is to advance the responsible adoption of AI for the benefit of humanity.*” thefuturesociety.org

---

Giga Initiative (@Gigaglobal)
“UNICEF and the International Telecommunication Union (ITU) have ... joined forces to create Giga: a global initiative to connect every school to the Internet by 2030.” Giga is a big supporter and early adopter of crypto-based financing mechanisms.” giga.global/

Resource: How NFTs can give more children the chance to learn online

---

Global Internet Forum to Counter Terrorism (GIFT) (@GIFCT_official)
“The Global Internet Forum to Counter Terrorism (GIFCT) is an NGO designed to prevent terrorists and violent extremists from exploiting digital platforms. Founded by Facebook, Microsoft, Twitter, and YouTube in 2017, the Forum was established to foster technical collaboration among member companies, advance relevant research, and share knowledge with smaller platforms. Since 2017, GIFCT’s membership has expanded beyond the founding companies to include over a dozen diverse platforms committed to cross-industry efforts to counter the spread of terrorist and violent extremist content online.”

Resources: Campaign Toolkit, Content Incidental Protocol

Global Partners Digital (@GlobalPartnersD)
“Global Partners Digital (GPD) is a social purpose company working to enable a digital environment underpinned by human rights. We do this by making policy spaces and processes more open, inclusive and transparent, and by supporting public interest actors to participate strategically in them.” gp-digital.org/who-we-are/

Human Rights Watch "Technology and Rights" Section (@hrw)
"The internet and other technologies are critical tools to defend rights and hold powerful actors to account. But technology can also be used in ways that curtail rights and deepen inequality. We defend human rights in the digital age. We document how governments and companies restrict online speech and access to information. We investigate how digital surveillance tools, from hacking to facial recognition, are used to target activists, racial and ethnic minorities and workers. We expose the impact of AI and other data-driven technologies on the rights of workers and people living with poverty. We advocate for laws and policies that promote privacy, digital inclusion, and respect for human rights by social media platforms." hrw.org/topic/technology-and-rights

Resources: Automated Hardship, Stopping Killer Robots

Human Rights Data Analysis Group (@hrdag)
"We Are Statisticians For Human Rights
When we partner with human rights defenders, from truth commissions to UN missions to local activists, we help them understand how data science can be used to answer questions about human rights violations." hrdag.org

Resource: Projects

Human Technology Foundation (@HumanTechF)
"The Human Technology Foundation network has several thousand members and operates in Paris, Montreal and Geneva. Indeed, if most technologies are neither good nor bad in themselves, they are not neutral either: they carry an intentionality and a vision of the human being that must be questioned. From this perspective, the Human Technology Foundation is striving to put technology back at the heart of social debates." human-technology-foundation.org

Resource: Report "Artificial Intelligence, Insurance & Solidarity" - January 2020

HURIDOCS (Human Rights Information and Documentation Systems) (@HURIDOCS)
"HURIDOCS (Human Rights Information and Documentation Systems) is an NGO that helps human rights groups gather, organise and use information to create positive change in the world. Since 1982, we have developed methodologies and tools that make it easier not only to manage collections of evidence, law and research, but also to analyse them for insights. In all that time, the technology has evolved but our passion for the cause remains unchanged. Information is a force for good—as long as human rights defenders are able to safely and efficiently make sense of it."

Resources: Machine learning for human rights information, HURIDOCS is a Google AI Impact Grantee, Topic: Machine Learning

Information Technology and Innovation Foundation (@ITIF_dc)
"ITIF is an independent, nonpartisan research and educational institute focusing on the intersection of technological innovation and public policy. Recognized as one of the world's leading science and technology think tanks, ITIF's mission is to formulate and promote policy solutions that accelerate innovation and boost productivity to spur growth, opportunity, and progress." itif.org

Resources: Creating an AI Bill of Rights Is a Distraction, How Can AI Improve Educational Outcomes in the United States?, How Congress and the Biden Administration Could Jumpstart Smart Cities With AI, Using AI to Fight Disinformation in European Elections

---

The Institute for Ethical AI & Machine Learning (@EthicalML)
"The Institute for Ethical AI & Machine Learning is a research centre based in the UK. We carry out highly technical research to answer some of the most challenging questions present in the intersection between Machine Learning and industry." ethical.institute

Resources: The AI-RFX Procurement Framework, The Responsible Machine Learning Principles, XAI - eXplainableAI Framework

---

Institute for Ethics in Artificial Intelligence (@IEAITUM)
"*The IEAI conducts inter-, multi-, and transdisciplinary research that promotes active collaboration between the technical, engineering and social sciences, while also actively courting interaction with a wide group of international stakeholders from academia, industry and civil society. This exhaustive approach enables the IEAI to truly and comprehensively address a growing group of ethical challenges arising at the interface of technology and human values. It also aids in the development of thoroughly operational ethical frameworks in the field of AI.***"** ieai.sot.tum.de

Resources: Annual Report 2021

Institute for Security and Technology (@IST_org)
"*The Institute for Security and Technology builds solutions to enhance the security of the global commons. Our goal is to provide tools and insights for companies and governments to outpace emerging global security threats. Our non-traditional approach is biased towards action, as we build trust across domains, provide unprecedented access, and deliver and implement solutions."* securityandtechnology.org

Resources: Countering Strategic and Nuclear Risks, Strengthening and Defending the Information Environment to Support Democracy, Digital Cognition & Democracy Initiative

International Committee of the Red Cross (ICRC) - Virtual Reality & Innovation (@ICRC)
"Using new and accelerating technologies, the ICRC continues to develop virtual environments as one of the many tools used to teach, motivate and maintain universal respect for IHL." icrc.org/en/what-we-do/virtual-reality

Resource: Contact virtualreality@icrc.org

Internet Society (@internetsociety)
"The Internet Society supports and promotes the development of the Internet as a global technical infrastructure, a resource to enrich people's lives, and a force for good in society. Our work aligns with our goals for the Internet to be open, globally connected, secure, and trustworthy. We seek collaboration with all who share these goals. Together, we focus on: Building and supporting the communities that make the Internet work; Advancing the development and application of Internet infrastructure, technologies, and open standards; and Advocating for policy that is consistent with our view of the Internet." internetsociety.org

Resource: Impact Report 2021

Internet Society Foundation (@ISOC_Foundation)
"At the Internet Society Foundation, we focus on funding initiatives that strengthen the Internet in function and reach so that it can effectively serve all people. Our work advances the vision of the Internet Society (ISOC): The Internet is For Everyone. To this end, we support efforts to ensure that the Internet is open, globally-connected, secure, and trustworthy. We champion the use of the network as a critical technical infrastructure that can bring communities better education, healthcare and economic opportunity among other important areas of focus. We believe that by working together, we can use the Internet to shape a better future for us all and positively impact humanity worldwide." isocfoundation.org

Resource: Funding Areas

JustPeace Labs (@justpeacelabs)
"JustPeace Labs (JPL), a women-founded and lead 501(c)(3) organization, advocates for and supports the responsible use and deployment of emerging technologies in high-risk settings—communities experiencing conflict, transitioning from conflict or enduring systematic human rights abuses." justpeacelabs.org/

Resources: Technology in Conflict: Conflict Sensitivity for the Tech Industry report, Ethical Guidelines for PeaceTech

Lawyers Hub (@AfricaLawTech)
"The Lawyers Hub is a Legal-Tech organisation that works on Digital policy and Justice innovation. Headquartered in Kenya and serving the global south, The Lawyers hub runs the Africa Digital Policy Institute, Africa Law Tech association, The Africa Startup Law Accelerator and convenes the annual Africa Law Tech Festival and the Africa Legal Innovation week on Justice innovation. The organisation is the publisher of the Africa Journal on Law & Tech. The Lawyers Hub work is focused on Privacy and Data Protection, Artificial intelligence, Intellectual Property, Digital Identity, Internet Governance, Digital Tax and lending, Tech and Democracy." lawyershub.org/

Resources: Lawyers Hub Kenya, Lawyers Hub Nigeria, Lawyers Hub South Africa

---

Lips (@lips_zine on IG)
"*Lips is a feminist technology organization building products designed to unlock opportunities for previously underserved and intersectionally marginalized communities. Our first product is a social media platform designed for women and LGBTQ individuals seeking a space to express themselves through art without the unhealthy aspects of mainstream internet culture such as online harassment, censorship and plagiarism. We are building more inclusive Machine Learning and Contextual AI technologies that can be used across industries to improve the online experience of traditionally marginalized communities.*" lipsdistro.co

Resource: lips.social

Milton Keynes Artificial Intelligence (MKAI) (@MKAI)
"*MKAI is an inclusive community of diverse thinkers that, together, are shaping the future of Ethical Artificial Intelligence - Our diversity is our strength. Our vision is to connect diverse minds and deliver impactful community-led projects that make artificial intelligence (AI) more inclusive, accessible and rooted in sustainable human values."* mkai.org

---

Montreal AI Ethics Institute (@mtlaiethics)
"The Montreal AI Ethics Institute is an international non-profit organization democratizing AI ethics literacy. We equip citizens concerned about artificial intelligence to take action because we believe that civic competence is the foundation of change. You are our best shot at a future where humans and algorithms bring out the best in each other."
montrealethics.ai/

Montreal Institute for Genocide and Human Rights Studies (@MIGSinstitute)
"The Montreal Institute for Genocide and Human Rights Studies (MIGS) is Canada's leading think tank working at the intersection of human rights, conflict and emerging technologies. The institute serves as a leadership and ideas incubator that convenes stakeholders with the goal of developing better policies to protect human rights." concordia.ca/research/migs.html

Resources: Digital Mass Atrocity Prevention Lab

My Data Rights Africa (@MyDataRightsAf1)
"Through the eyes of a feminist, the intersections of Artificial intelligence, privacy and data protection are explored in the context of South Africa.The context of gendered marginalisation of women, gender diverse people and sexual minorities forms the basis of understanding the data concerns. AI and gender; feminist methodology and policy are assessed from a feminist perspective to develop recommendations for gender responsive policy and regulations and action from the civil society for engagement on data rights." mydatarights.africa/

Resource: Feminist Methodology

OECD.AI (@OECDinnovation)
"The OECD AI Policy Observatory is a tool at the disposal of governments and businesses that they can use to implement the first intergovernmental standard. OECD.AI *combines resources from across the OECD, its partners and all stakeholder groups.* OECD.AI *facilitates dialogue between stakeholders while providing multidisciplinary, evidence-based policy analysis in the areas where AI has the most impact."* oecd.ai

Resources: AI Wonk Blog, Live Data on AI Research, OECD Network of Experts on AI (ONE AI), The Global Partnership on AI (GPAI), OECD AI Principles Overview

---

Online Hate Prevention Institute (@OnlineHate)
"The Online Hate Prevention Institute (OHPI) is an Australian Harm Prevention Charity. We aim to reduce the risk of suicide, self harm, substance abuse, physical abuse and emotional abuse that can result from online hate. Our focus ranges from cyber-racism, online religious vilification and other group-based forms of online hate, through to the cyber-bullying of individuals...OHPI conducts research, runs campaigns and provides public education, recommends policy changes and law reform, and seeks ways of changing online systems to make them more effective in reducing the risks posed by online hate. We aim to find ways to create systemic changes that reduce the risk of harm both now and into the future." https://ohpi.org.au

Resource: Online Hate

Open Voice Network (@openvoicenet)
"The Open Voice Network (OVON), a Linux Directed Fund, seeks to make human voice to machine technology worthy of user trust - a task of critical importance as voice emerges as a primary, multi-device, multi-lingual portal to the digital, IoT, and metaverse worlds, and as independent, specialist voice assistants take their place in people's pockets, wearables, vehicles, homes, and workplace." openvoicenetwork.org/about

Resources: OVON Ethical Use Guidelines for Voice Experiences v1.0, OVON Privacy Principles Unique to Voice White Paper v2.0, OVON Voice Registry System Initiative OVON State of Voice - CES 2022 Future of Voice - Voice Metaverse

---

Oxford Internet Institute, University of Oxford (@oiioxford)
"The Oxford Internet Institute's (OII) mission is to understand this transformation. Our research draws on many different disciplines, essential to tackling the major challenges of the 21st century. From digital politics to the ethics of artificial intelligence, we aim to address the societal implications of life online to inform public policy, advise industry and enhance daily life for people around the world.

Our research excellence strategy involves a comprehensive program for conducting original research, recruiting the world's top researchers, and advancing both the scientific and humanistic understanding of the impact of the internet, data, and information technologies on society." https://www.oii.ox.ac.uk

---

Partnership on AI (@PartnershipAI)
"A non-profit partnership of academic, civil society, industry, and media organizations creating solutions so that AI advances positive outcomes for people and society. By convening diverse, international stakeholders, we seek to pool collective wisdom to make change" partnershiponai.org

Resource: Resources

Patrick J. McGovern Foundation (@PJMFnd)
"A global philanthropy advancing AI and Data for Good by serving as global changemakers and optimists advancing AI and data solutions to create a thriving, equitable, and sustainable future for all. The Foundation is the legacy of visionary IDG founder and CEO Patrick J. McGovern who believed in the potential for technology to be a force for good." mcgovern.org

Resources: Accelerating the use of data for equality, Data4Change Accelerator

Pew Research Center, Internet & Technology (@pewresearch) (@pewinternet)
"Pew Research Center is a nonpartisan fact tank that informs the public about the issues, attitudes and trends shaping the world. We conduct public opinion polling, demographic research, content analysis and other data-driven social science research. We do not take policy positions." pewresearch.org/topic/internet-technology

Resources: Technology Policy Issues, The Future of Digital Spaces and Their Role in Democracy

---

PeaceTech Lab (PLT), U.S. Institute of Peace (@PeaceTechLab)
"PeaceTech Lab (PTL) drives action-oriented solutions by bringing a diversity of experts together: data scientists, social scientists, engineers, MBAs, global influencers, media ambassadors, and creatives. It is through these alliances that we can collectively develop effective peacebuilding solutions." peacetechlab.org/

Resources: Programs: Hate Speech, Misinformation, Election Violence Prevention, and more!

Pollicy (@PollicyOrg)
"Pollicy is a feminist collective of technologists, data scientists, creatives and academics working at the intersection of data, design and technology to craft better life experiences by harnessing improved data." pollicy.org

Resources: Engendering AI: A Gender and Ethics Perspective on Artificial Intelligence in Africa, Engendering Artificial Intelligence, Disabled but not disqualified, Afro Feminist Data Futures Report

---

Privacy International (@privacyint)
"Governments and corporations are using technology to exploit us. Their abuses of power threaten our freedoms and the very things that make us human...That's why PI is here: to protect democracy, defend people's dignity, and demand accountability from institutions who breach public trust." Privacyinternational.org

Resource: Advertisers on Facebook: who the heck are you and how did you get my data?

Queer in AI (@QueerInAI)
"Queer in AI's mission is to raise awareness of queer issues in AI/ML, foster a community of queer researchers and celebrate the work of queer scientists." queerinai.com

Resource: Resources

Ranking Digital Rights (@rankingrights)
"Ranking Digital Rights (RDR) works to promote freedom of expression and privacy on the internet by creating global standards and incentives for companies to respect and protect users' rights...We do this by ranking the world's most powerful digital platforms and telecommunications companies on relevant commitments and policies, based on international human rights standards. We work with companies as well as advocates, researchers, investors, and policymakers to establish and advance global standards for corporate accountability." RankingDigitalRights.org

Resources: Theory of Change, Recommendations for governments and policymakers

---

Responsible AI Network - Africa (RAIN Africa) (@AIAfricaNetwork)
"The transnational nature of AI technology platforms makes the ethical and application of AI a global concern. Because AI inherently interacts with its surroundings, cultural, political and environmental differences in context may produce different sets of ramifications (positive or negative) resulting from the use of AI-based technology. There is therefore a growing need to understand how AI may impact or be accepted by society in various regions around the world." rainafrica.org

Resource: Artificial Intelligence Needs Assessment Survey in Africa (March 2021)

Superrr (@superrr)
"Superrr is a community and a lab. We develop visions and projects with the goal to create more equitable futures. We do so by researching technologies, building networks and shaping new narratives. Superrr is playful, visionary and feminist. The mission of Superrr Lab is to explore and develop the potentials of new technologies for society and diversity. We challenge existing paradigms by bringing new perspectives and stakeholders to the discussion. We work with civil society, political decision makers and the tech industry." superrr.net

Resource: Feminist Tech Policy Card Deck

S.T.O.P (Surveillance Technology Oversight Project) (@STOPSpyingNY)
"S.T.O.P. litigates and advocates for privacy, working to abolish local governments' systems of mass surveillance. Our work highlights the discriminatory impact of surveillance on Muslim Americans, immigrants, the LGBTQ+ community, indigenous peoples, and communities of color, particularly the unique trauma of anti-Black policing." stopspying.org

Resource: NYC Internet Remastered: A Privacy and Equity Analysis of the New York Internet Master Plan

Stanford Institute for Human-Centered Artificial Intelligence (@StanfordHAI)
"Stanford HAI advances AI research, education, policy, and practice to improve humanity. Stanford HAI leverages the university's strength across all disciplines, including: business, economics, genomics, law, literature, medicine, neuroscience, philosophy and more. These complement Stanford's tradition of leadership in AI, computer science, engineering and robotics." hai.stanford.edu

Resources: Building a National AI Research Resource, Evaluating Facial Recognition Technology: A Protocol for Performance Assessment in New Domains, Recommendations on Updating the National Artificial Intelligence Research and Development Strategic Plan, Stanford HAI Artificial Intelligence Bill of Rights

---

Stop LAPD Spying Coalition
"The Stop LAPD Spying Coalition is a community organization founded in 2011. We work to build community power toward abolishing police surveillance. We are rooted in the Skid Row neighborhood of downtown Los Angeles. Their Data Driven Policing initiative strives to stop the usd of data to collect and mine madd data to determine which people and places will be policed. They work to stop the police from drawing from a vast web of surveillance sources, data brokers, state agencies, and open-source information, putting this data at police fingertips." stoplapdspying.org/

Resource: Reports and Resources

Tony Blaire Institute (@InstituteGC)
"Helping leaders & governments to deliver for their people. We equip political leaders and governments to build open, inclusive and prosperous societies in an interconnected global world." institute.global/tony-blair

Resource: Government Advisory Projects

Urban AI (@Urban__AI)
"Urban AI is a Think Tank which federates an international ecosystem and a multidisciplinary community. Together, we propose ethical modes of governance and sustainable uses of urban AI." urbanai.fr

---

Virtual Activism
"We have been working on the intersection of technology and human rights and development since 1998. We have been providing training workshops on technology to human rights organizations mainly based in the Middle East but also elsewhere. We are in consultative status with ECOSOC. We are currently working on Ethics and AI" virtualactivism.org

Resource: Emerging Technologies

WeProtect Global Alliance (@WeProtect)
"WeProtect Global Alliance brings together governments, the private sector, civil society and international organisations to develop policies and solutions to protect children from sexual exploitation and abuse online." weprotect.org

Resources: Global Threat Assessment 2021, Implementing the Global Strategic Response framework, Implementing the Model National Response framework

---

Women in AI Ethics (@WomeninAIEthics)
"The Women in AI Ethics™ *(WAIE) is a global initiative with a mission to increase recognition, representation, and empowerment of women in AI Ethics. This initiative started with the first* 100 Brilliant Women in AI Ethics™ *list in 2018, which is now published annually to recognize rising stars as well as pioneers in this space. To increase representation of women at AI/tech conferences and companies, we have also launched an open* online directory of Women in AI Ethics™ *to make it easier for conference organizers and recruiters to recruit talented women working hard to make AI ethical, inclusive, and accessible for all."* womeninaiethics.org

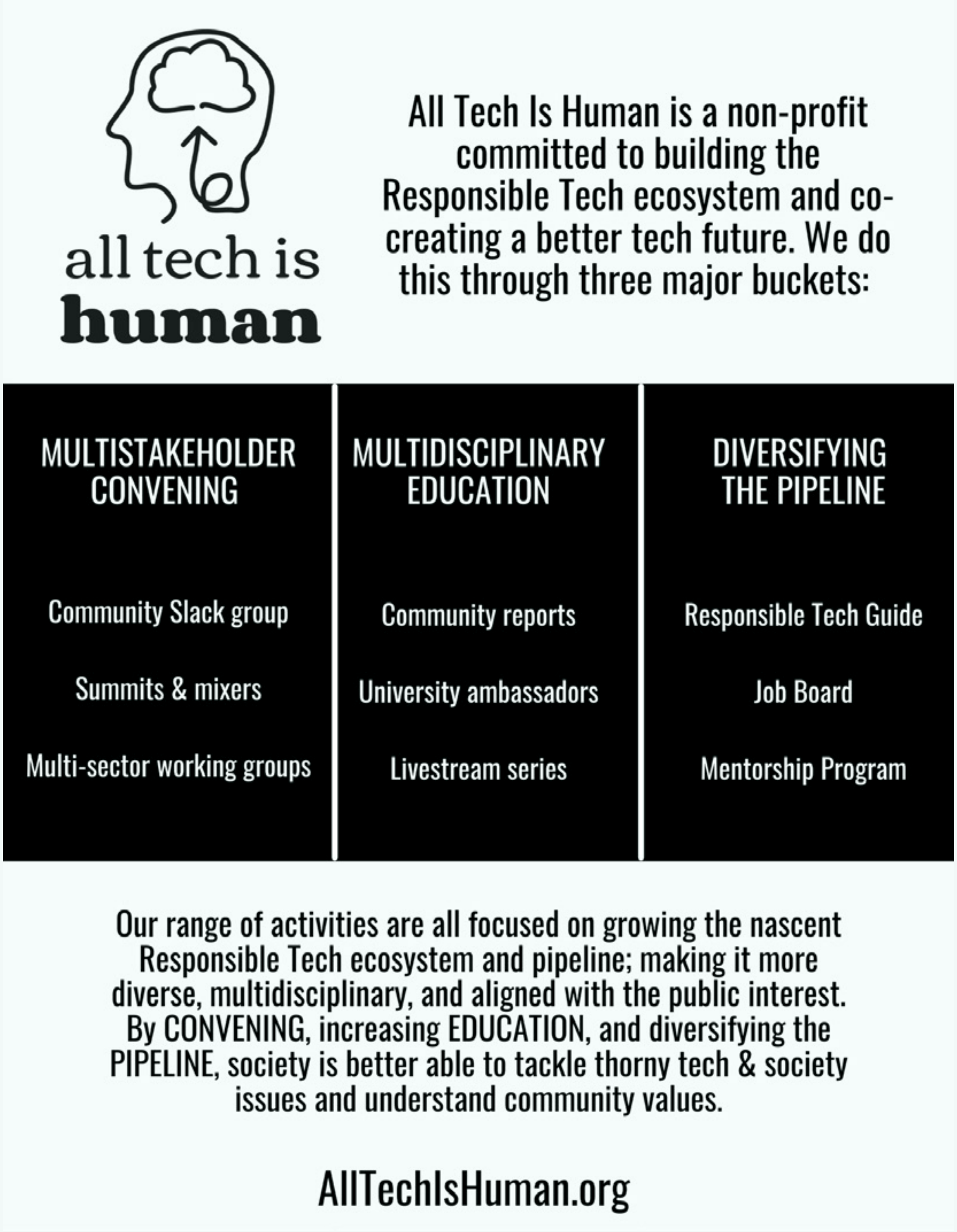
Resource: Directory

Women of Color Advancing Peace, Security, and Conflict Transformation (WCAPS)
"*At WCAPS, we believe global issues demand a variety of perspectives. That's why in 2017, we created a platform devoted to women of color that cultivates a strong voice and network for its members while encouraging dialogue and strategies for engaging in policy discussions on an international scale."* wcaps.org

Resources: Working Groups - Cyber Security and Emerging Technology, Defense and Intelligence; Chemical, Biological, Radiological & Nuclear Security Policy

World Economic Forum AI and Machine Learning (@wef)
"The World Economic Forum's AI and Machine Learning team brings together key stakeholders from the public and private sectors to co-design and test policy frameworks that accelerate the benefits and mitigate the risks of AI and ML. Project areas include standards for protecting children, creating an AI regulator for the 21st century, and addressing the unique challenges of facial recognition technology." weforum.org

Resources: AI Ethics Framework, Empowering AI Leadership, Ethics by Design: An organizational approach to responsible use of technology, Generation AI: Developing Artificial Intelligence Standards for Children, Human-Centered Artificial Intelligence for Human Resources, Responsible Limits on Facial Recognition Technology, Responsible Limits on Facial Recognition Use Case: Flow Management



all tech is human

All Tech Is Human is a non-profit committed to building the Responsible Tech ecosystem and co-creating a better tech future. We do this through three major buckets:

MULTISTAKEHOLDER CONVENING	MULTIDISCIPLINARY EDUCATION	DIVERSIFYING THE PIPELINE
Community Slack group	Community reports	Responsible Tech Guide
Summits & mixers	University ambassadors	Job Board
Multi-sector working groups	Livestream series	Mentorship Program

Our range of activities are all focused on growing the nascent Responsible Tech ecosystem and pipeline; making it more diverse, multidisciplinary, and aligned with the public interest. By CONVENING, increasing EDUCATION, and diversifying the PIPELINE, society is better able to tackle thorny tech & society issues and understand community values.

AllTechIsHuman.org

all tech is
human

AI and Human Rights:
Building a Tech Future Aligned With the Public Interest

June 2022

alltechishuman.org