Northwestern | CENTER FOR ADVANCING SAFETY OF MACHINE INTELLIGENCE

Toward a Safety Science in Artificial Intelligence

The Center for Advancing Safety of Machine Intelligence

Alexander Einarsson, Andrea Lynn Azzo, Kristian Hammond

aeinarsson@u.northwestern.edu, {andrea.azzo, kristian.hammond}@northwestern.edu

Abstract

Artificial intelligence's (AI) unprecedented growth and breadth of application areas has resulted in a pressing need for a concrete safety field in AI. However, because safety in AI is a nebulous concept, the first challenge in developing a safety science in the field must be to define what it means for an AI system to be safe. Inspired by discussions and thoughts shared by a diverse group of researchers and stakeholders in a workshop on the topic, this work presents a series of steps for the AI community to define safety based on potential harms different AI systems may cause. The steps of 1) identification 2) mapping 3) quantification 4) remediation and 5) prevention will dampen the harms caused by existing AI systems in the short term, while also serving to inform all stakeholders in AI safety on how to minimize harm caused by novel systems in the long term. This work will be a resource for various groups, from developers of AI systems, to legislative groups and bodies, to anyone who has a vested interest in minimizing harm AI systems cause to the public.

1 Introduction

Artificial intelligence (AI) has found unprecedented growth over the past decade [26]. It has helped advance fields where it affects a large portion of humanity, including medicine, commerce, and business, and a majority of Americans will soon have their work influenced by AI [17, 24, 27, 40]. With this growth, and the lowered bar of entry in applying models to create AI systems, the risk of AI systems harming persons, groups of people, or society has risen considerably [12]. As that risk increases, the need for a safety field in artificial intelligence grows with it.

But while safety is relatively well-defined and measurable in fields such as aerospace engineering and medicine, safety in artificial intelligence remains abstract. This is likely largely caused by domain differences; it's clear if an aircraft is safe – on different levels of safety it 1) reaches its destination, 2) keeps its passengers unharmed, and 3) doesn't hurt the environment – and similarly so for a drug or method in medicine, that is not always the case in AI systems. An AI system can be considered safe if it doesn't cause harm to humans, or does more good than it does harm, or causes less harm than the system it is replacing, or even has the potential to do less harm than the replaced system in the future. This is the first challenge that must be overcome in establishing AI safety as a field: defining what it means for an AI system to be "safe."

To this end, the Center for Advancing Safety of Machine Intelligence (CASMI) held a workshop in January 2023 with the express purpose of identifying what prevents the community from being able to define what safety in AI means. Attendees included a wide array of stakeholders, from safety engineers in other fields, to AI researchers, to lawyers. The post-talk discussions were held according to Chatham House rules, and as such, some of the points that were brought up will be paraphrased in this paper without being credited to a particular attendee.[1] The opinions expressed in this work reflects the consensus opinion of the workshop attendees, unless otherwise stated.

2 Challenges

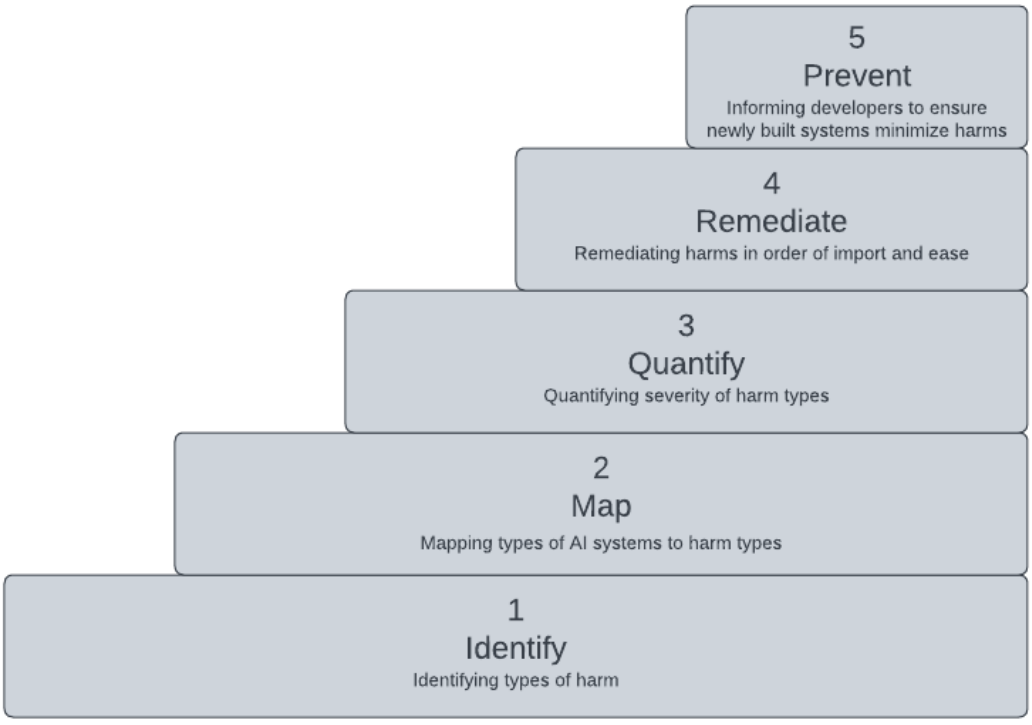
The primary challenge to defining a safe AI system lies in the difficulty of predicting malfunction or longer-term consequences. The range of how potential mishaps in AI systems can harm humans is too vast, and although there often are commonalities in harm caused by similar AI systems, the lack of historic documentation and oversight means developers and researchers seldom have frameworks for safety to follow to help them avoid common types of harms. Even if such frameworks did exist, they would be limited to helping developers avoid repeating mistakes but wouldn't prevent development of novel harmful AI systems.

Instead of trying to solve the issue of harmful AI systems on the system level, the focus should be on identifying, remediating, and preventing specific types of harms from AI systems. This effort would produce an extensive and concrete set of harms that AI systems may cause to either individuals, groups, or society. The resulting product would become a crucial resource for

[1] https://www.chathamhouse.org/about-us/chatham-house-rule

developers to use in their building of AI systems but would also importantly provide regulating bodies and auditing firms a product that they could use to determine whether AI systems are safe.

3 Steps



3.1 Step 1 (Identifying types of harms)

To create clear and concise definitions of the different types of harms, without awaiting documentation from too much harm that has already occurred, the consensus in the workshop was a proposed two-pronged approach.

First, as many other older fields have already encountered and remedied a wide variety of harms, and subsequently established positions specifically to predict and avoid harms, the AI community should collaborate with, and learn from, safety workers from these fields.

Second, a thorough analysis of incidents that have occurred is necessary. Efforts like the AI Incident Database and tesladeaths.com have been recording incidents and issues in AI systems for years as they have been reported on, and these efforts will be imperative in inductively

defining different types of harms.[2,3] Well-organized reporting databases allow researchers to establish different harm groups while providing developers and other stakeholders with an invaluable resource in finding potential harms for the AI systems they are developing.

While reporting and documenting all negative incidents in AI as harms is a necessary first step, it is not sufficient to identify different kinds of harms. We suggest that the most helpful labels for harms would be by who is harmed and how. The granularity of this division could be debated, but for the purposes of this paper, there will be three types:

1) The first harm type is the harm of individual persons. Incidents and issues for these systems have negative and unexpected impacts on single persons, without the issue being systemic discrimination of an entire group of people. Specifically, an automated system that mistakenly fires a person would be part of this group, as would a robot puncturing a can of bear spray, sending a group of people to the hospital, while a hiring system that learns to discriminate against women would not [7, 35, 38]. One of the more prominent examples of this type of harm would be autonomous car crashes. They affect individuals or small groups of people (in cases where a crash involves more than one person) but do not carry an inherent bias that negatively affects certain groups of people. Other examples of systems causing individuals harm include: Navigation apps routing people into wildfire territory [13]; Grading systems providing inaccurate student scores [39]; And unwanted pregnancies as a result of using a “natural cycle” birth control app [28].

2) The second harm type is the harm of groups of people. Incidents and issues for these systems include any type of systemic bias against a group of people and tend to be born out of historic bias in the data, rather than just poor development practices. Debiasing these systems would require collecting new data, strategically chosen to not carry an inherent bias, or correcting the existing data and then retraining the system. Examples of systems that caused harm against groups of people are wide ranging and include: Harm against racial groups, with facial recognition systems mislabeling Black people as gorillas, voice recognition systems primarily being trained to recognize white male voices, and predictive policing systems disproportionately targeting low-income, Latino, and Black neighborhoods [2, 25, 37]; Harm against the elderly, with ageist systems being reinforced by ageist big data approaches [32]; Harm against groups with special needs, such as hiring discrimination against people with disabilities [1]; And harm against gender groups, including gender bias in hiring systems [7].

3) The third harm type is harm to society. Incidents and issues of this type have a negative impact for a large portion of society, without being biased against a certain group of people. A racially biased AI system would not be in this harm group, but a system that radicalizes a portion of society would. The increasing popularity of large language models is a significant risk factor for these types of harm, as they can persuade groups

[2] https://incidentdatabase.ai/
[3] https://www.tesladeaths.com/

of people at scale, without any assurance that the text they are generating has any factual basis. In the CASMI workshop on AI Safety, a theme found for these types of harms was an initial small harm to an individual scaling up as a type of “death by a thousand cuts.” In particular, social media’s driving young people to phone addiction through algorithms that push them to spend hours each day scrolling on their phone leads to depression because they see happier people than themselves on social media. Each of these harms is an individual harm, but because the AI systems reach so many people, it turns into a societal harm over time. Other societal harms include: recommendation systems generating online echo chambers, causing radicalization [41]; language models producing large scale misinformation [15]; AI systems causing financial harm by replacing workers [10]; and potential ramifications of language model misuse in education [16].

In addition to the laid out differences between these harm types, harms regularly have different causes, and ordered by scale are also likely more difficult to correct.

The individual harms can usually be ascribed to human error in development, whether from bugs in the system or from developer oversight. The examples of individual harms above can all be traced back to bugs in the code or poorly planned and developed systems. These harms may be easy to fix but can also be very difficult to discover before they materialize. As such, it is imperative to document these harms and map them to AI systems where they tend to occur, so new systems don’t repeat the mistakes of their predecessors. However, completely preventing novel individual harms may prove impossible.

Harm of groups of people are often caused by systems trained on biased datasets, rather than errors in the code. Consequently, even supposedly well-coded systems can cause harm of this type. As with individual harms, documentation and labeling of known harms, and mapping those harms to systems where they have occurred, is instrumental in avoiding repeating past mistakes. In addition, and unlike the individual harms, issues with the AI systems can be spotted before the harms appear, as the harms are caused by the systems’ training on flawed datasets. Collaborative inspection of the data by people with domain knowledge can help spot potential biases before development. It is noteworthy that, while harms can be prevented before development in a way individual harms cannot, it is harder to remediate existing harms, as that would require collecting “good” data or debiasing the existing data.

Societal harms are likely the most difficult to both identify and prevent before they occur. Inherently, they have low initial harm until the scale grows to the point where they harm society, and as such it could prove challenging to detect and correct the causes of the harm before the large-scale harm becomes evident. While coding errors and biased data may be partial contributors to societal harms, they are more commonly caused by systems’ functioning correctly, where unforeseen consequences present themselves with scale either because of neglected minor harms or malicious misuse. As such, documentation, labeling, and mapping is crucial for this type of harm as well. Additionally, as full-fledged societal harms can be so

impactful, collaboration between the AI community and other communities who may provide insight about potential harm vectors will be critical to prevent the harms before they appear.

3.2 Step 2 (Mapping systems types to harm types)

The second necessary step to help people identify how their systems may fail is to map which types of systems are more liable to cause which types of harm. This should be done both via conversations within the larger AI community, and by looking at work like the AI Incident Database to gain insight about which types of systems tend to be problematic in which ways. Initially, these may be fairly straightforward observations, such as social media recommender systems tending to generate echo chambers, but with time these mappings should be more granular.

While the final goal of this step should be decided on through conversations within the AI community, it would be desirable to create frameworks for different types of AI systems that clearly lay out potential or common harms for the systems. This would ensure that new developers, practitioners, or stakeholders have resources to turn to for advice in the future.

3.3 Step 3 (Quantifying Harms)

While the initial harm division is necessary for the early talks of establishing a Safety in AI field, and creates a foundation from which to further define the harms, it is clear that such broad division will not suffice in the long term. AI systems that may accidentally fire people are not as harmful as ones that may kill. Systems that may make it difficult for groups to be hired by a single company are not as harmful as ones that cause unfair persecution of groups. And systems that cause phone addiction are unlikely to be as harmful in the long term as ones that drive radicalization. As such, it will be imperative that the harm of different instances within the laid out harm types is quantified.

The specifics of this harm quantification should be discussed and agreed upon by the broader AI Safety community, and will be an ongoing process as new instances of harms are discovered, but likely it will come down to a combination of the level of harm and the scale of the system. In the Safety in AI workshop, scale was agreed upon to be a significant contributor to how harmful a system was, as much as the direct impact to individual persons harmed by the systems. The prime example brought up was how a single social media user scrolling through their feed and getting body image issues may not be seen as particularly harmful, but if the system reaches millions of users every day, the harm goes from relatively insignificant to catastrophic.

Fig. 1: Two-dimensional harm diagram.

The diagram above is only an example diagram to showcase the strength of observing harms as a two-dimensional issue, as it shows how less impactful harms can become equally as harmful as more impactful harms with higher usage. In addition, observing the top left example harm, the autonomous vehicle accidents, makes it clear that by scaling the problem up, bringing in more autonomous vehicles that have an equal accident rate to the current fleet of vehicles, this problem may veer into the high danger quadrant in the top right. This holds for all types of incidents in this diagram, that unless the probability of an incident type goes down as the scale increases, the incident type veers toward the right edge of the diagram, meaning it becomes more harmful.

3.4 Step 4 (Remediating existing harms)

Once the harms have been identified, located, and quantified, they can and should be remediated. Owners of systems that have been identified as potentially harmful must be persuaded to lessen the potential harm through whatever means the community has. At this point, these means are largely nonexistent, and as such steps 4 and 5 will be largely

speculative until the broader AI Safety community finds scalable solutions. In the workshop, a recurring proposal was that of AI Safety audits, with the suggestion of legislation and liability as persuasion techniques also being brought up more than once.

The audits would fill a similar need that Underwriters Laboratories (UL) was born from, when it became the de facto determinator of whether an electronic product was safe for use by the community.[4] Although the safety organization is not a regulating body, and as such can't stop people from using products that are unsafe, people tend to want UL's stamp of approval on products because it means they meet scientific safety, quality, or security standards.[5] Similarly, with a rigorous auditing process for AI systems, the developers could prove that their system is free from known harms, and as such could be used without having to worry about any potential liability.

Legislation from regulatory bodies was suggested, and while it was agreed that it would likely be necessary, the issue of time and knowledge must be considered. Until an AI Safety community exists that can agree on what laws may be necessary and has identified the existing harms that the laws would target, it's unrealistic to believe that governing bodies will create laws that are loose enough to allow for creativity, while being strict enough to rein in developers who would otherwise create harmful systems.

3.4.1 On regulations of AI created by governing bodies

Governing bodies such as the European Union have shown an interest in creating legislation to stop harm with the Artificial Intelligence Act,[6] but numerous groups have voiced concern about the act's vague language and insufficient coverage [31].[7] Several European countries have launched investigations into ChatGPT, but the investigations have tended to focus on whether OpenAI has gone against the General Data Protection Regulation (GDPR) [21, 22, 23]. GDPR is a privacy focused piece of legislation, and while it is broad enough to cover some harms that aren't strictly privacy based (as data published about persons must be accurate, GDPR can be used to target misinformation), far from all potential harms caused by language models are within its legislative reach.

Canada's proposed Artificial Intelligence and Data Act has received similar feedback from policy analysts and researchers,[8] who have written that the bill provides invalid examples of harmful or discriminatory outcomes and that it also lacks an approach to independent oversight [5, 36]. Brazil's Senate has followed suit and drafted an AI law focused on protecting the public from

[4] https://www.ul.com/about/history
[5] https://www.ul.com/look-ul-safety-mark-you-buy
[6] https://artificialintelligenceact.eu/the-act/
[7] https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/feedback_en?p_id=8242911
[8] https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document#s9

harm, but even if implemented, the law risks quickly becoming obsolete because of the number of newly implemented AI systems throughout Brazil [4, 8]. Japan published governance guidelines, but the country does not have legislative or regulatory provisions that govern AI [29]. China's approach to regulate generative AI, meanwhile, has drawn condemnation from Western nations because of the Chinese Communist Party's requirements that AI technologies must be in line with the country's "socialist core values" [6].

According to the Organisation for Economic Co-operation and Development, the United States has been a global leader in AI policies and strategies, with a total of 77 initiatives listed on the OECD's repository.[9] So far, the U.S. government's approach to regulate AI has been to introduce voluntary guidance, such as the National Institute for Standards and Technology's AI Risk Management Framework and the White House's Blueprint for an AI Bill of Rights.[10,11] Lawmakers in the House of Representatives and the Senate have introduced legislation, but nothing has been signed into law [20, 34]. Four federal agencies – the Consumer Financial Protection Bureau, the Department of Justice, the Equal Employment Opportunity Commission, and the Federal Trade Commission – say their current approach is to use existing laws to take action against companies for their use of AI [9].

At the state level, general artificial intelligence bills or resolutions were introduced in at least 17 states in 2022.[12] Illinois became the first state in the country to regulate employers' use of AI when making hiring decisions [19]. In Colorado, insurance companies are banned from using any algorithms or predictive models to discriminate against people [30]. Several states have passed bills creating a commission, task force, or oversight position to evaluate the use of AI [14]. However, rules differ at the state level and can be focused on specific issues [11].

3.5 Step 5 (Preventing future harms)

Finally, the ultimate goal of an AI Safety community must be that new AI systems don't cause harm to humans. Preventing these harms follows naturally from remediation, and harm prevention in future novel systems will be unattainable until harm in existing systems is remediated. In addition to building on the earlier steps, the prevention step presents a new issue: in the development of new AI systems, there are not currently systems in place that allow for persons outside of the development team and their immediate stakeholders to influence prevention efforts.

Consequently, developing audits and assessments that can measure both existing and potential harms in novel AI systems, which was mentioned in the remediation step, is imperative for prevention to be feasible long-term. It is likely that the major legislative bodies around the world will require AI systems to go through rigorous testing via assessments and audits in the future, but as seen in the Regulations section, those actors may lack both the necessary knowledge

[9] https://oecd.ai/en/dashboards/overview
[10] https://www.nist.gov/itl/ai-risk-management-framework
[11] https://www.whitehouse.gov/ostp/ai-bill-of-rights/
[12] https://www.ncsl.org/technology-and-communication/legislation-related-to-artificial-intelligence

and legislative reach to be entirely relied upon. Even if they have the knowledge and power necessary, this is a time sensitive issue. It would behoove the AI research community and stakeholders to push the early development of these audits and assessments, even if support from legislators will be necessary to require novel AI systems to pass through these tests in the future.

4 Establishing the field of Artificial Intelligence Safety

The end goal of this endeavor is to establish the field of AI Safety, and this field must grow out from a well-informed and diverse group of AI stakeholders, including but not limited to:

1) AI researchers and developers with deep knowledge of AI systems, how they may cause harm, and how harm can be avoided or mitigated.
2) Human rights and social justice organizations who communicate with different groups of people who may be harmed by AI systems.
3) Lawyers and legal scholars with insight on the strengths and weaknesses of the legal system and its capability to protect the public from harm from AI systems.
4) Safety engineers from other fields who have experience with ensuring that systems or components don't cause undue harm.
5) Any other stakeholders with unique or unusual insight in potential issues AI systems may cause.

Two recurring themes from comments in the January workshop, showcased in the list of steps above, are that this endeavor must be a community effort (a community that extends beyond artificial intelligence researchers and developers), and legislative bodies and auditing companies must incentivize developers of AI systems to avoid harm. Of those, building out the AI safety community must come first, and that community must be there to inform and lead legislators and assessment creators.

5 Future steps

As part of the push to build out the field of AI Safety, we will promote and support the field in academic circles. We want to continue to build the community of researchers and other stakeholders in academia, while also connecting to persons with influence in the industry and various legislative bodies. Simultaneously, CASMI will continue to set up workshops both targeted at specific applications, and broadly on defining the field and building a community. Further, we will work to introduce AI safety tracks at established conferences, with the long term goal of setting up a new conference or a journal focused on AI safety.

Further work is needed to explore methods to measure and to validate AI safety. To address this, CASMI is hosting another workshop on July 18-19 entitled, "Sociotechnical Approaches to Measurement and Validation for Safety in AI." The event will convene interdisciplinary thought

leaders from academia, industry, and government to holistically consider how to measure AI harms or near harms.

One approach to assess AI safety is through measurement modeling, which identifies the hidden ways that ideas about the world get encoded as data and models [18]. Machine learning models are built using a number of assumptions that guide their behavior and predictions [3]. The complex decisions they make are not well understood [33]. Measurement modeling makes the assumptions that are baked into models explicit [18]. The next CASMI workshop will explore this technique from a sociotechnical point of view. The goal will be to uncover hidden assumptions to better understand how systems work and how harms occur.

6 Acknowledgements

The Center for Advancing Safety of Machine Intelligence is a collaboration with the UL Research Institutes' Digital Safety Research Institute. UL Research Institutes provided funding for this work as well as the workshop on which it was based. The findings and opinions in this work may not represent the views of UL Research Institutes. This work was made possible by the attendees of the "Toward a Safety Science of AI" workshop, their thoughtful responses, and their diverse perspectives.

7 References

[1] Associated Press. "U.S. Warns of Discrimination in Using Artificial Intelligence to Screen Job Candidates." NPR, May 12, 2022. https://www.npr.org/2022/05/12/1098601458/artificial-intelligence-job-discrimination-disabilities.

[2] Bajorek, Joan Palmiter. "Voice Recognition Still Has Significant Race and Gender Biases." Harvard Business Review, May 10, 2019. https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases.

[3] Barocas, Solon, Andrew D. Selbst, and Manish Raghavan. "The Hidden Assumptions behind Counterfactual Explanations and Principal Reasons." arXiv.org, December 10, 2019. https://arxiv.org/abs/1912.04930.

[4] Belli, Luca, Yasmin Curzi, and Walter B. Gaspar. "AI Regulation in Brazil: Advancements, Flows, and Need to Learn from the Data Protection Experience." *Computer Law & Security Review* 48 (2023): 105767. https://doi.org/10.1016/j.clsr.2022.105767.

[5] Castro, Daniel. "Canada's Reasons for an AI Law Do Not Stand up to Scrutiny." Center for Data Innovation, March 27, 2023. https://datainnovation.org/2023/03/canadas-reasons-for-an-ai-law-do-not-stand-up-to-scrutiny/.

[6] Che, Chang. "China Says Chatbots Must Toe the Party Line." The New York Times, April 24, 2023. https://www.nytimes.com/2023/04/24/world/asia/china-chatbots-ai.html.

[7] Dastin, Jeffrey. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women." Reuters, October 10, 2018.

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[8] Evangelos Sakiotis, Anna Oberschelp de Meneses. “Brazil’s Senate Committee Publishes AI Report and Draft AI Law.” Inside Privacy, January 27, 2023. https://www.insideprivacy.com/emerging-technologies/brazils-senate-committee-publishes-ai-report-and-draft-ai-law/.

[9] Feiner, Laure. “U.S. Regulators Warn They Already Have the Power to Go after A.I. Bias - and They’re Ready to Use It.” CNBC, April 25, 2023. https://www.cnbc.com/2023/04/25/us-regulators-warn-they-already-have-the-power-to-go-after-ai-bias.html.

[10] Ford, Brody. “IBM to Pause Hiring for ‘back-Office’ Jobs That AI Could Kill.” Bloomberg.com, May 1, 2023. https://www.bloomberg.com/news/articles/2023-05-01/ibm-to-pause-hiring-for-back-office-jobs-that-ai-could-kill#xj4y7vzkg.

[11] Friedler, Sorelle, Suresh Venkatasubramanian, and Alex Engler. “How California and Other States Are Tackling AI Legislation.” Brookings, March 22, 2023. https://www.brookings.edu/blog/techtank/2023/03/22/how-california-and-other-states-are-tackling-ai-legislation/.

[12] Gardner, Kevin. “6 Ways AI Improves Daily Living.” Medium, December 30, 2019. https://becominghuman.ai/6-ways-ai-improves-daily-living-72e53a02145.

[13] Godlewski, Nina. “Navigation Apps Sent People in California to Roads That Were on Fire.” International Business Times, December 7, 2017. https://www.ibtimes.com/waze-google-maps-send-california-residents-straight-wildfires-2625610.

[14] Goldman, Sharon. “Ai Regulation: A State-by-State Roundup of Ai Bills.” VentureBeat, August 8, 2022. https://venturebeat.com/ai/ai-regulation-a-state-by-state-roundup-of-ai-bills/.

[15] Goldstein, Josh A., Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. “Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations.” arXiv.org, January 10, 2023. https://arxiv.org/abs/2301.04246.

[16] Halaweh, Mohanad. “Chatgpt in Education: Strategies for Responsible Implementation.” AAU Digital Repository, January 1, 1970. https://digitallibrary.aau.ac.ae/handle/123456789/980.

[17] Heethuis, Nick. “Council Post: Four Ways Artificial Intelligence Is Transforming e-Commerce.” Forbes, February 16, 2022. https://www.forbes.com/sites/theyec/2022/02/15/four-ways-artificial-intelligence-is-transforming-e-commerce/?sh=6649ec80797b.

[18] Jacobs, Abigail Z, and Deirdre K Mulligan. “The Hidden Governance in AI.” The Hidden Governance in AI, July 6, 2022. https://www.theregreview.org/2022/07/07/jacobs-mulligan-the-hidden-governance-in-ai/.

[19] Jedreski, Matt, Jeffrey S Bosley, and K.C. Halm. Illinois Becomes First State to Regulate Employers’ Use of Artificial Intelligence to Evaluate Video Interviews, September 3, 2019.

https://www.dwt.com/blogs/artificial-intelligence-law-advisor/2019/09/illinois-becomes-first-state-to-regulate-employers.

[20] Kelley, Alexandra. “Lawmakers Introduce Bill to Keep AI from Going Nuclear.” Nextgov.com, April 27, 2023. https://www.nextgov.com/emerging-tech/2023/04/lawmakers-initiate-several-efforts-put-guardrails-ai-use/385711/.

[21] Laing, Aislinn, Elvira Pollina, and Silvia Aloisi. “Spain Asks EU Data Protection Board to Discuss OpenAI’s Chatgpt.” Reuters, April 11, 2023. https://www.reuters.com/technology/spains-data-regulator-asks-eu-data-protection-committee-evaluate-chatgpt-issues-2023-04-11/.

[22] MacRedmond, David. “Germany Launches Data Protection Inquiry over Chatgpt.” TheJournal.ie, April 24, 2023. https://www.thejournal.ie/germany-investigates-chatgpt-over-data-privacy-6052226-Apr2023/.

[23] McCallum, Shiona. “CHATGPT Banned in Italy over Privacy Concerns.” BBC News, April 1, 2023. https://www.bbc.com/news/technology-65139406.

[24] Mehra, Asheesh. “Council Post: How Ai Is Transforming Healthcare.” Forbes, March 16, 2020. https://www.forbes.com/sites/forbestechcouncil/2020/03/16/how-ai-is-transforming-healthcare/?sh=45f875bf4ba0.

[25] Mehrotra, Dhruv, Surya Matt, Annie Gilbertson, and Aaron Sankin. “How We Determined Predictive Policing Software Disproportionately Targeted Low-Income, Black, and Latino Neighborhoods.” Gizmodo, December 2, 2021. https://gizmodo.com/how-we-determined-predictive-policing-software-dispropo-1848139456.

[26] Metz, Rachel. “How Ai Came to Rule Our Lives over the Last Decade | CNN Business.” CNN, December 23, 2019. https://www.cnn.com/2019/12/21/tech/artificial-intelligence-decade/index.html.

[27] Mueller, Julia. “Ai Could Affect up to 80 Percent of US Workforce: Research.” The Hill, March 27, 2023. https://thehill.com/policy/technology/3919941-ai-could-affect-up-to-80-percent-of-us-workforce-research/.

[28] Muir, Ellie. “Would You Trust an App to Tell You If You’re Fertile? Because It Might Be the Future.” The Independent, January 31, 2023. https://www.independent.co.uk/life-style/natural-cycles-birth-control-app-b2270679.html.

[29] Nakazaki, Takashi. Artificial Intelligence Comparative Guide - - Japan, November 23, 2022. https://www.mondaq.com/technology/1059766/artificial-intelligence-comparative-guide.

[30] Nieberg, Patty. “Colorado Bill Prohibits Insurer Use of ‘discriminatory’ Data.” AP NEWS, May 4, 2021. https://apnews.com/article/colorado-race-and-ethnicity-bills-business-government-and-politics-2821191ff2a9fb259a592b5c972eabc6.

[31] Pouget, Hadrien. “The EU’s AI Act Is Barreling toward AI Standards That Do Not Exist.” Lawfare, January 12, 2023. https://www.lawfareblog.com/eus-ai-act-barreling-toward-ai-standards-do-not-exist.
[32] Rosales, Andrea, and Mireia Fernández-Ardèvol. “Structural Ageism in Big Data Approaches.” Nordicom Review, June 1, 2019. https://sciendo.com/article/10.2478/nor-2019-0013.
[33] Räz, Tim, and Claus Beisbart. “The Importance of Understanding Deep Learning.” *Erkenntnis*, 2022. https://doi.org/10.1007/s10670-022-00605-y.
[34] *Schumer Launches Major Effort To Get Ahead Of Artificial Intelligence*, April 13, 2023. United States Congress. https://www.democrats.senate.gov/newsroom/press-releases/schumer-launches-major-effort-to-get-ahead-of-artificial-intelligence.
[35] Staff, NBC10. “Robot Punctures Bear Spray Can, More than 50 People Sickened at New Jersey Amazon Warehouse, Town Says.” NBC10 Philadelphia, December 5, 2018. https://www.nbcphiladelphia.com/news/local/Amazon-Warehouse-Robbinsville-Sickened-Workers-501976671.html.
[36] Tessono, Christelle, and Sonja Solomun. “How to Fix Canada’s Proposed Artificial Intelligence Act.” Tech Policy Press, December 6, 2022. https://techpolicy.press/how-to-fix-canadas-proposed-artificial-intelligence-act/.
[37] Vincent, James. “Google ‘fixed’ Its Racist Algorithm by Removing Gorillas from Its Image-Labeling Tech.” The Verge, January 12, 2018. https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai.
[38] Wakefield, Jane. “The Man Who Was Fired by a Machine.” BBC News, June 21, 2018. https://www.bbc.com/news/technology-44561838.
[39] Weckler, Adrian. “Explainer: Why Has One Line of Computer Code Caused Such Disruption to the Leaving CERT Grades?” Independent.ie, October 1, 2020. https://www.independent.ie/irish-news/explainer-why-has-one-line-of-computer-code-caused-such-disruption-to-the-leaving-cert-grades/39580586.html.
[40] Weitzman, Tyler. “Council Post: The Top Five Ways Ai Is Transforming Business.” Forbes, November 22, 2022. https://www.forbes.com/sites/forbesbusinesscouncil/2022/11/21/the-top-five-ways-ai-is-transforming-business/?sh=93bc55d8e7f8.
[41] Whittaker, Joe, Seán Looney, Alastair Reed, and Fabio Votta. “Recommender Systems and the Amplification of Extremist Content.” Internet Policy Review, June 30, 2021. https://policyreview.info/articles/analysis/recommender-systems-and-amplification-extremist-content.