



Ethics in Data Science

CS316 Data Science

Adapted from:

http://ai.stanford.edu/blog/ethical_best_practices/

<https://www.datascience-pm.com/10-data-science-ethics-questions/>

Data Science Ethics: Key Questions

- ← How does a Data Science project affect society?
- ← Impact of the project on disadvantaged or vulnerable populations?
- ← Were any tests written to determine if the datasets were biased?
- ← Did any team discussions center around transparency of the trained model?
- ← Any time spent considering other ethical hypotheticals?

Growing attention to ethics...?

← Fairness, Accountability, and Transparency

← FAcCT:

<https://facctconference.org/index.html>

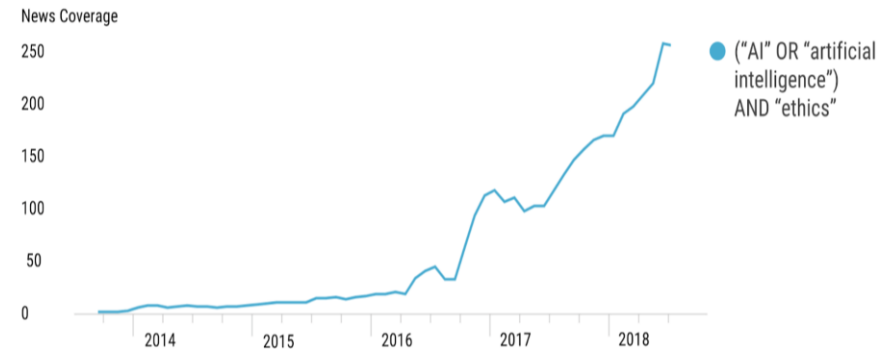
← [Ethical AI in Action World Tour](#)

← More informal talks (following some high-profile mistakes)

← General scientific ethics courses

Talk of AI and ethics is on the rise

Quarterly news mentions of ("AI OR artificial intelligence") AND "ethics" 2014 - Q3 2018



Source: cbinsights.com

 CBINSIGHTS

This lecture: some best practices to know

A Framework for Ethical AI and Data Science

- ← With AI's growing influence, ethical considerations ensure that technologies are developed responsibly.
- ← The primary goal of ethical thinking in data science (and everywhere, really) is to **avoid unintended consequences of your work**
 - Of course, this assumes the actor is intentionally good... we don't have time to cover how to handle intentionally bad actors)
- ← How?
 - ← **Education**
 - ← **Communication**
 - ← **Distribution**
 - ← **Advocacy**

Education

- ← Basics of AI Ethics
- ← Legal and policy communities have thought about ethics in AI at least as much as AI researchers have thought about its development
- ← Automated systems are not inherently neutral. They reflect the priorities, preferences, and prejudices of those who have the power to mold artificial intelligence.
- ← Consider: opacity of machine learning algorithms
 - ← Opacity of **secrecy**
(corporate, government)
 - ← Opacity of **technical illiteracy**
(black box algorithms)
 - ← Opacity of **scale**
(unavoidable algorithmic complexity)

Education

← Consider: potential harms of fully-automated decision-making (1)

Potential Harms from Automated Decision-Making

| Individual Harms | | Collective / Societal Harms |
|---|---|--|
| Illegal | Unfair | |
| Loss of Opportunity | | |
| Employment Discrimination E.g. Filtering job candidates by race or genetic/health information | Unfair E.g. Filtering candidates by work proximity leads to excluding minorities | Differential Access to Job Opportunities |
| Insurance & Social Benefit Discrimination E.g. Higher termination rate for benefit eligibility by religious group | Unfair E.g. Increasing auto insurance prices for night-shift workers | Differential Access to Insurance & Benefits |
| Housing Discrimination E.g. Landlord relies on search results suggesting criminal history by race | Unfair E.g. Matching algorithm less likely to provide suitable housing for minorities | Differential Access to Housing |
| Education Discrimination E.g. Denial of opportunity for a student in a certain ability category | Unfair E.g. Presenting only ads on for-profit colleges to low-income individuals | Differential Access to Education |

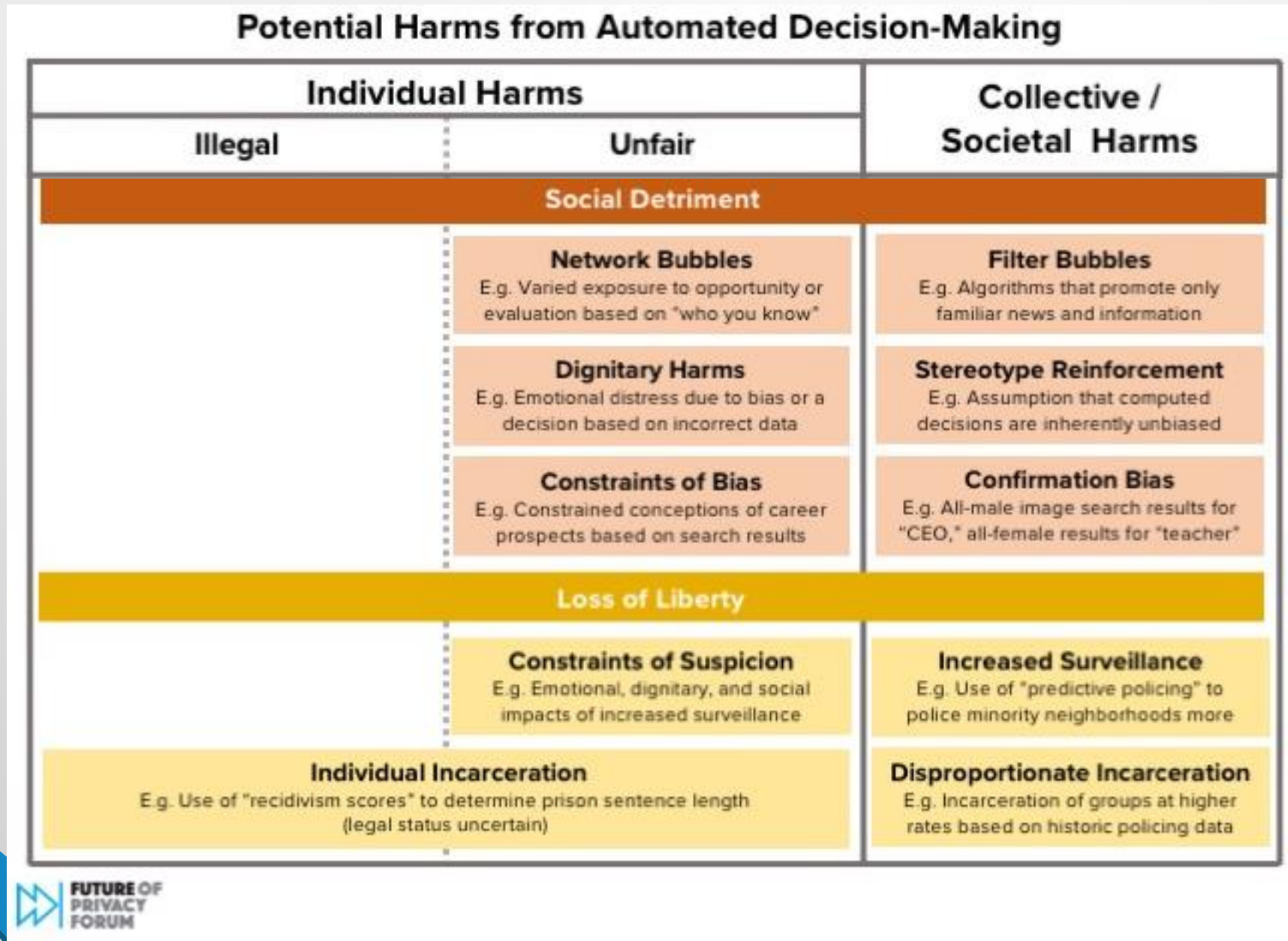
Education

← Consider: potential harms of fully-automated decision-making (2)

| Potential Harms from Automated Decision-Making | | |
|--|--------|--|
| Individual Harms | | Collective / Societal Harms |
| Illegal | Unfair | |
| Economic Loss | | |
| Credit Discrimination E.g. Denying credit to all residents in specified neighborhoods ("redlining") | | Differential Access to Credit |
| E.g. Not presenting certain credit offers to members of certain groups | | |
| Differential Pricing of Goods and Services E.g. Raising online prices based on membership in a protected class | | Differential Access to Goods and Services |
| E.g. Presenting product discounts based on "ethnic affinity" | | |
| Narrowing of Choice E.g. Presenting ads based solely on past "clicks" | | Narrowing of Choice for Groups |

Education

← Consider: potential harms of fully-automated decision-making (3)



Education

← Potential mitigation strategies (1)

| Harms | Description | Mitigation Tools | |
|---|---|--|-----------------|
| Individual Harms – Illegal | | | |
| <ul style="list-style-type: none"> Employment Discrimination Insurance & Social Benefit Discrimination Housing Discrimination Education Discrimination Credit Discrimination Differential Pricing Individual Incarceration | Existing law defines impermissible outcomes, often specifically for protected classes | <ul style="list-style-type: none"> • Data methods to ensure proxies are not used for protected classes & data does not amplify historical bias • Algorithmic design to carefully consider whether to use protected status inputs & trigger manual reviews • Laws & policies that use data to identify discrimination | |
| Individual Harms – Unfair (with illegal analog) | | | |
| <ul style="list-style-type: none"> Employment Discrimination Insurance & Social Benefit Discrimination Housing Discrimination Education Discrimination Credit Discrimination Differential Pricing Individual Incarceration | Individual harms that could be considered illegal if they involved protected classes, but do not in this case | <ul style="list-style-type: none"> • Business processes to index concerns; ethical frameworks & best practices to monitor & evaluate outcomes • Laws & policies include tools like DPIAs to measure impact or enable rights to explanation | |
| Key | | | |
| Loss of Opportunity | Economic Loss | Social Stigmatization | Loss of Liberty |

Education

← Potential mitigation strategies (2)

| Harms | Description | Mitigation Tools |
|---|--|--|
| Collective/Societal Harms (with illegal analog) | | |
| <ul style="list-style-type: none"> Differential Access to Job Opportunities Differential Access to Insurance Benefits Differential Access to Housing Differential Access to Education Differential Access to Credit Differential Access to Goods & Services Disproportionate Incarceration | Group level impacts that are not legally prohibited, though related individual impacts could be illegal | <ul style="list-style-type: none"> • Same as above section • Laws & policies should consider offline analogies & whether it is appropriate for industry to identify & mitigate |
| Individual Harms – Unfair (without illegal analog) | | |
| <ul style="list-style-type: none"> Narrowing of Choice Network Bubbles Dignitary Harms Constraints of Bias Constraints of Suspicion | <p>Individual impacts for which we do not have legal rules.</p> <p>Mitigation may be difficult or undesirable absent a defined set of societal norms</p> | <ul style="list-style-type: none"> • Business processes to index concerns, ethical frameworks & best practices to monitor & evaluate outcomes • Laws & policies should consider whether it is appropriate to expect industry to identify & enforce norms |
| Collective/Societal Harms (without illegal analog) | | |
| <ul style="list-style-type: none"> Narrowing of Choice for Groups Filter Bubbles Stereotype Reinforcement Confirmation Bias Increased Surveillance of Groups | Group level impacts for which we do not have legal rules or societal agreement as to what constitutes a harm | <ul style="list-style-type: none"> • Same as above section |
| Key | | |
| Loss of Opportunity | Economic Loss | Social Stigmatization |
| | | Loss of Liberty |

Education

- ← Consider: facial recognition in public places
 - ← City centers, airports
 - ← Concerns of error, function creep, and privacy
 - ← Emotional privacy
 - ← Masking emotions
 - ← Social cohesion
 - ← <http://blog.practicaethics.ox.ac.uk/2014/03/computer-vision-and-emotional-privacy/>

Education


- **Exclusion and demographic bias:** AI systems may inadvertently exclude certain groups or reinforce stereotypes if the training data does not represent diverse demographics.
- **Overgeneralization and confirmation bias:** AI models might incorrectly generalize patterns from biased data, leading to conclusions that reinforce pre-existing biases.
- **Topic overexposure (availability heuristic) and underexposure:** Overemphasis on popular topics in training data can lead to disproportionate representation, while less common topics may be underrepresented, skewing the AI's understanding.

← <http://aclweb.org/anthology/P16-2096>

← NLP ethical best practices <http://aclweb.org/anthology/W17-1604.pdf>

Education

Remedies: Pyramid of Possible Responses to Unethical Behavior.



| | |
|-----------------|--|
| Disclosure | to document/to reveal injustice to regulators, the police, investigative journalists ("Look what they do!", "Stop what they do!") |
| Resignation | to distance oneself III ("I should not/cannot be part of this.") |
| Persuasion | to influence in order to halt non-ethical activity ("Our organization should not do this.") |
| Rejection | to distance oneself II; to deny participation; conscientious objection ("I can't do this.") |
| Escalation | raise with senior management/ethics boards ("You may not know what is going on here.") |
| Voicing dissent | to distance oneself I ("This project is wrong.") |
| Documentation | ensure all the facts, plans and potential and actual issues are preserved. |

Education

- ← Consider: “dual-use” technologies
 - ← Technologies designed for civilian use but which may have **military applications**
 - ← Google’s Project Maven, software for automated drone surveillance for the Pentagon

Sign this letter

Dear Sundar,

We believe that Google should not be in the business of war. Therefore we ask that Project Maven be cancelled, and that Google draft, publicize and enforce a clear policy stating that neither Google nor its contractors will ever build warfare technology.

- ← Microsoft employees have protested the company’s involvement in the same Department of Defense program
- ← Amazon’s collaboration with the US Immigration and Customs Enforcement (ICE)

Education

- ← Codes of Ethics
- ← Some have advocated for a “Data Science Hippocratic Oath”
- ← IEEE and ACM organizations have explicit codes of ethics
 - ← <https://www.ieee.org/about/corporate/governance/p7-8.html>
 - ← <https://www.acm.org/code-of-ethics>
- ← AI research is arguably unique, but NeurIPS publishes each year a Code of Conduct
 - ← <https://nips.cc/public/CodeOfConduct>
- ← Many other institutions and “influencing” organizations have begun making their own



IEEE Code of Ethics

[Related information](#) >

The following is from the IEEE Policies, Section 7 - Professional Activities (Part A - IEEE Policies).

> [Download a copy of the IEEE Code of Ethics](#)

7.8 IEEE Code of Ethics

We, the members of the IEEE, in recognition of the importance of our technologies in affecting the quality of life throughout the world, and in accepting a personal obligation to our profession, its members and the communities we serve, do hereby commit ourselves to the highest ethical and professional conduct and agree:

- I. To uphold the highest standards of integrity, responsible behavior, and ethical conduct in professional activities.
 1. to hold paramount the safety, health, and welfare of the public, to strive to comply with ethical design and sustainable development practices, to protect the privacy of others, and to disclose promptly factors that might endanger the public or the environment;
 2. to improve the understanding by individuals and society of the capabilities and societal implications of conventional and emerging technologies, including intelligent systems;
 3. to avoid real or perceived conflicts of interest whenever possible, and to disclose them to affected parties when they do exist;
 4. to avoid unlawful conduct in professional activities, and to reject bribery in all its forms;
 5. to seek, accept, and offer honest criticism of technical work, to acknowledge and correct errors, to be honest and realistic in stating claims or estimates based on available data, and to credit properly the contributions of others;
 6. to maintain and improve our technical competence and to undertake technological tasks for others only if qualified by training or experience, or after full disclosure of pertinent limitations;
- II. To treat all persons fairly and with respect, to not engage in harassment or discrimination, and to avoid injuring others.
 7. to treat all persons fairly and with respect, and to not engage in discrimination based on characteristics such as race, religion, gender, disability, age, national origin, sexual orientation, gender identity, or gender expression;
 8. to not engage in harassment of any kind, including sexual harassment or bullying behavior;
 9. to avoid injuring others, their property, reputation, or employment by false or malicious actions, rumors or any other verbal or physical abuses;
- III. To strive to ensure this code is upheld by colleagues and co-workers.
 10. to support colleagues and co-workers in following this code of ethics, to strive to ensure the code is upheld, and to not



ACM Code of Ethics and Professional Conduct

ACM Code of Ethics and Professional Conduct

Preamble

Computing professionals' actions change the world. To act responsibly, they should reflect upon the wider impacts of their work, consistently supporting the public good. The ACM Code of Ethics and Professional Conduct ("the Code") expresses the conscience of the profession.

The Code is designed to inspire and guide the ethical conduct of all computing professionals, including current and aspiring practitioners, instructors, students, influencers, and anyone who uses computing technology in an impactful way. Additionally, the Code serves as a basis for remediation when violations occur. The Code includes principles formulated as statements of responsibility, based on the understanding that the public good is always the primary consideration. Each principle is supplemented by guidelines, which provide explanations to assist computing professionals in understanding and applying the principle.

Section 1 outlines fundamental ethical principles that form the basis for the remainder of the Code. Section 2 addresses additional, more specific considerations of professional responsibility. Section 3 guides individuals who have a leadership role, whether in the workplace or in a volunteer professional capacity. Commitment to ethical conduct is required of every ACM member, ACM SIG member, ACM award recipient, and ACM SIG award recipient. Principles involving compliance with the Code are given in Section 4.

The Code as a whole is concerned with how fundamental ethical principles apply to a computing professional's conduct. The Code is not an algorithm for solving ethical problems; rather it serves as a basis for ethical decision-making. When thinking through a particular issue, a computing professional may find that multiple principles should be taken into account, and that different principles will have different relevance to the issue. Questions related to these kinds of issues can best be answered by thoughtful consideration of the fundamental ethical principles, understanding that the public good is the paramount consideration. The entire computing profession benefits when the ethical decision-making process is accountable to and transparent to all stakeholders. Open discussions about ethical issues promote this accountability and transparency.

1. GENERAL ETHICAL PRINCIPLES.

A computing professional should...

1.1 Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing.

This principle, which concerns the quality of life of all people, affirms an obligation of computing professionals, both individually and collectively, to use their skills for the benefit of society, its members, and the environment surrounding them. This obligation includes promoting fundamental human rights and protecting each individual's right to autonomy. An essential aim of computing professionals is to minimize negative consequences of computing, including threats to health, safety, personal security, and privacy. When the interests of multiple groups conflict, the needs of those less advantaged should be given increased attention and priority.

On This Page

Preamble

1. GENERAL ETHICAL PRINCIPLES.

1.1 Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing.

1.2 Avoid harm.

1.3 Be honest and trustworthy.

1.4 Be fair and take action not to discriminate.

1.5 Respect the work required to produce new ideas, inventions, creative works, and computing artifacts.

1.6 Respect privacy.

1.7 Honor confidentiality.

2. PROFESSIONAL RESPONSIBILITIES.

2.1 Strive to achieve high quality in both the processes and products of professional work.

2.2 Maintain high standards of professional competence, conduct, and ethical practice.

2.3 Know and respect existing rules pertaining to professional work.

2.4 Accept and provide appropriate professional review.

2.5 Give comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks.

2.6 Perform work only in areas of competence.

2.7 Foster public awareness and understanding of computing, related technologies, and their consequences.

2.8 Access computing and communication resources only when authorized or when compelled by the public good.

2.9 Design and implement systems that



Education

NEURAL INFORMATION PROCESSING SYSTEMS FOUNDATION CODE OF CONDUCT

We the participants, employees, and other individuals involved with Neural Information Processing Systems, come together for the open exchange of ideas, the freedom of thought and expression, and for respectful scientific debate which is central to the goals of this Conference. This requires a community and an environment that recognizes and respects the inherent worth of every person.

RESPONSIBILITY

All participants, organizers, reviewers, speakers, media, sponsors, and volunteers (referred to as "Participants" collectively throughout this document) at our Conference, workshops, and Conference-sponsored social events are required to agree with this Code of Conduct both during an event and on official communication channels, including social media.

Sponsors are equally subject to this Code of Conduct. In particular, sponsors should not use images, activities, or other materials that are of a sexual, racial, or otherwise offensive nature. This code applies both to official sponsors as well as any organization that uses the Conference name as branding as part of its activities at or around the Conference.

Organizers will enforce this Code, and it is expected that all Participants will cooperate to help ensure a safe and inclusive environment for everyone.

POLICY

The conference commits itself to providing an experience for all Participants that is free from the following [1]:

- **Harassment, bullying, and discrimination** which includes but is not limited to:
 - Offensive comments related to age, race, religion, creed, color, gender (including transgender/gender identity/gender expression), sexual orientation, medical condition, physical or intellectual disability, pregnancy, or medical conditions, national origin or ancestry.
 - intimidation, personal attacks, harassment, unnecessary disruption of talks or other conference events.
- **Inappropriate or unprofessional** behavior that interferes with another's full participation including:
 - sexual harassment, stalking, following, harassing photography or recording, inappropriate physical contact, unwelcome attention, public vulgar exchanges, derogatory name-calling, and diminutive characterizations.
 - Use of images, activities, or other materials that are of a sexual, racial, or otherwise offensive nature that may create an inappropriate or toxic environment.
 - Disorderly, boisterous, or disruptive conduct including fighting, coercion, theft, damage to property, or any mistreatment or non-businesslike behavior towards participants.
 - Zoom bombing or any virtual activity that is not related to the topic of discussion which detracts from the topic or the purpose of the program. This includes inappropriate remarks in chat areas as deemed inappropriate by presenters/monitors/event leaders.
 - Individuals and organizations that make false claims or accusations related to Neural Information Processing Systems' business or inappropriately make comments online as if they represent Neural Information Processing Systems without advanced approval.
- **Scientific misconduct** including fabrication, falsification, or plagiarism of paper submissions or research presentations, including demos, exhibits or posters.

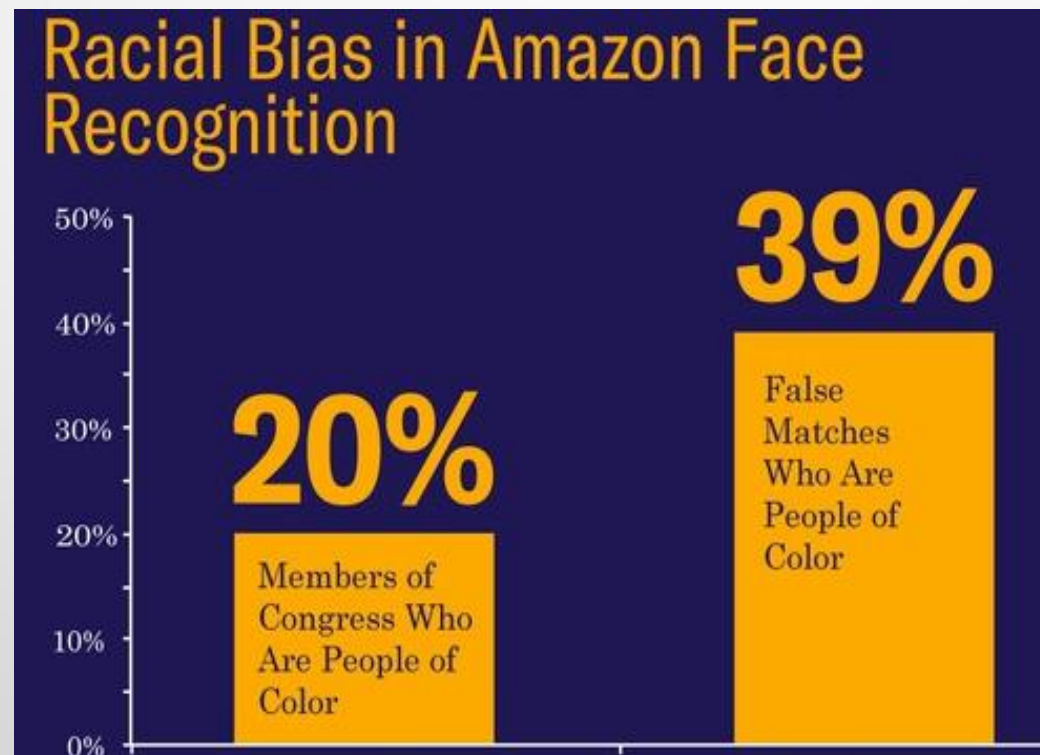
This Code of Conduct applies to the actual meeting sites and conference venues where Neural Information Processing Systems business is being conducted, including physical, virtual venues, and official virtual engagement platforms, including video, virtual streaming, and chat-based interactions. Neural Information Processing Systems is not responsible for non-sponsored activity or behavior that may occur at non-sponsored locations such as hotels, restaurants, physical, virtual, or other locations not otherwise deemed a sanctioned space for Neural Information Processing Systems sponsored events. Nonetheless, any issues brought to the Hotline Relations Counselors will be considered. However, it is also the case that Neural Information Processing Systems cannot actively monitor voluntary social media platforms and cannot follow-up on every transaction occurring between individuals who voluntarily engage in argument and altercation outside the Neural Information Processing Systems sponsored events virtual or otherwise.

ACTION

If a Participant engages in any inappropriate behavior as defined herein, the Conference organizers may take action as deemed appropriate, including: a formal or informal warning to the offender, expulsion from the conference with no refund, barring from participation in future conferences or their organization, reporting the incident to the offender's local institution or funding agencies, or reporting the incident

Communication

- ← Potential **misuses** and ethical considerations of new AI and data science algorithms / packages are rarely identified and pointed out, either in documentation or in academic papers
- ← Amazon's Rekognition product for facial recognition did not warn about the high false positive rate associated with its default parameters



Communication

- ← New “Ethical Considerations” section in published works (academic, code documentation, blog posts, etc)
- ← Margaret Mitchell, Senior Research Scientist at Google, Tech Lead for Google’s ML fairness effort
 - ← 2017 paper flagging patient suicide risk in clinical settings given their writings as input
 - ← Point out clear cases of potential ethical misuse and how their study mitigated these concerns

2 Ethical Considerations

As with any author-attribute detection, there is the danger of abusing the model to single out people (*overgeneralization*, see Hovy and Spruit (2016)). We are aware of this danger, and sought to minimize the risk. For this reason, we don’t provide a selection of features or representative examples. The experiments in this paper were performed with a clinical application in mind, and use carefully matched (but anonymized) data, so the distribution is not representative of the population as a whole. The results of this paper should therefore *not* be interpreted as a means to assess mental health conditions in social media in general, but as a test for the applicability of MTL in a well-defined clinical setting.

Communication

- ← Standardizing means of communicating aspects of new datasets and AI services
- ← Datasheets for datasets
- ← Data statements for NLP
- ← Policy certificates for RL
- ← Declarations of AI service conformity

Dataset Fact Sheet

Metadata



Title COMPAS Recidivism Risk Score Data

Author Broward County Clerk's Office, Broward County Sheriff's Office, Florida

Email browardcounty@florida.usa

Description Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

DOI 10.5281/zenodo.1164791

Time Feb 2013 - Dec 2014

Keywords risk assessment, parole, jail, recidivism, law

Records 7214

Variables 25

priors_count: *Ut enim ad minim veniam, quis nostrud exercitation* **numerical**

two_year_recid: *Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.* **nominal**

Missing Units 15452 (8%)

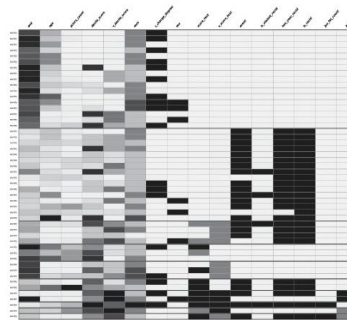


This dataset contains variables named "age," "race," and "sex."

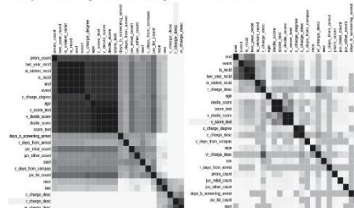
Probabilistic Modeling

Analysis

12



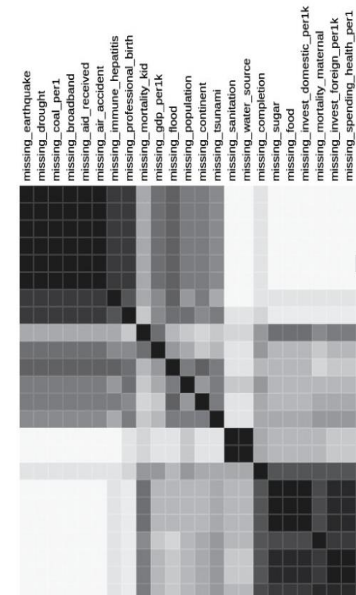
Dependency Probability Pearson R



Missing Units

Clustering Variable Missing Variable

race r_days_from_arrest



Distribution

- ← Approval and Terms of Access for datasets, code, and models

- ← ImageNet

- ← One of the most important computer vision datasets of the decade (1000 object classes, 1,281,167 training images, 50,000 validation images and 100,000 test images)

- ← Downloading it requires agreeing to terms of access!

- ← Admittedly increases overhead for host lab or organization, but helps mitigate the dual-use problem

- ← A “Responsible AI License” for code and pre-trained models

Distribution



14,197,122 images, 21841 synsets indexed

[Home](#) [Download](#) [Challenges](#) [About](#)

Not logged in. [Login](#) | [Signup](#)

Download

Download ImageNet Data

The most highly-used subset of ImageNet is the [ImageNet Large Scale Visual Recognition Challenge \(ILSVRC\)](#) 2012-2017 image classification and localization dataset. This dataset spans 1000 object classes and contains 1,281,167 training images, 50,000 validation images and 100,000 test images. This subset is available on [Kaggle](#).

For access to the full ImageNet dataset and other commonly used subsets, please login or request access. In doing so, you will need to agree to our terms of access.

Terms of access:

[RESEARCHER_FULLNAME] (the "Researcher") has requested permission to use the ImageNet database (the "Database") at Princeton University and Stanford University. In exchange for such permission, Researcher hereby agrees to the following terms and conditions:

1. Researcher shall use the Database only for non-commercial research and educational purposes.
2. Princeton University and Stanford University make no representations or warranties regarding the Database, including but not limited to warranties of non-infringement or fitness for a particular purpose.
3. Researcher accepts full responsibility for his or her use of the Database and shall defend and indemnify the ImageNet team, Princeton University, and Stanford University, including their employees, Trustees, officers and agents, against any and all claims arising from Researcher's use of the Database, including but not limited to Researcher's use of any copies of copyrighted images that he or she may create from the Database.
4. Researcher may provide research associates and colleagues with access to the Database provided that they first agree to be bound by these terms and conditions.
5. Princeton University and Stanford University reserve the right to terminate Researcher's access to the Database at any time.
6. If Researcher is employed by a for-profit, commercial entity, Researcher's employer shall also be bound by these terms and conditions, and Researcher hereby represents that he or she is fully authorized to enter into this agreement on behalf of such employer.
7. The law of the State of New Jersey shall apply to all disputes under this agreement.

Distribution

BOX 9

Ethical considerations in deciding whether to share Google AI advances

We generally seek to share Google research to contribute to growing the wider AI ecosystem. However we do not make it available without first reviewing the potential risks for abuse. Although each review is content-specific, key factors that we consider in making this judgment include:

- **Risk and scale of benefit vs downside** – What is the primary purpose and likely use of a technology and application, and how beneficial is this? Conversely, how adaptable is it to a harmful use, and how likely is it that there are bad actors with the skills and motivation to deploy it? Overall, what is the magnitude of potential impact likely to be?
- **Nature and uniqueness** – Is it a significant breakthrough or something that many people outside Google are also working on and close to achieving? Is sharing going to boost the capabilities of bad actors, or might it instead help to shift the playing field, so good actors are more able to offset the bad? What is the nature of Google’s involvement — are we openly publishing a research paper that anyone can learn from, or are we directly developing a custom solution for a contentious third-party application?
- **Mitigation options** – Are there ways to detect and protect against bad actors deploying new techniques in bad ways? (If not, it might be necessary to hold back until a ‘fix’ has been found.) Would guidance on responsible use be likely to help, or more likely to alert bad actors?

Distribution

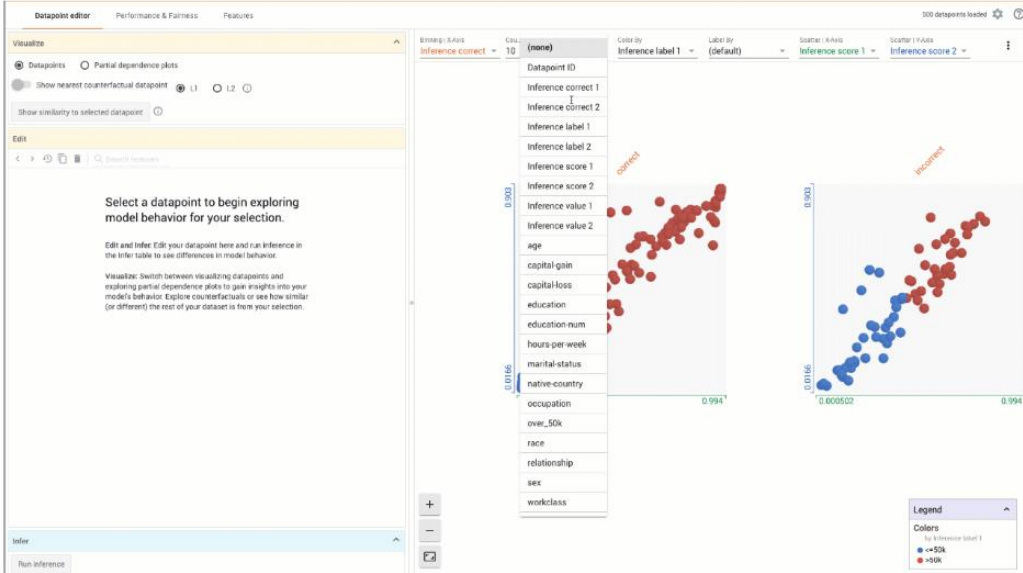
- ← Use, share, and create emerging tools to detect bias and explore datasets for ethical considerations
- ← IBM's AI Fairness 360
 - ← An extensible open source toolkit that helps you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle
 - ← <https://aif360.res.ibm.com/>

The screenshot shows the homepage of the IBM AI Fairness 360 project. The browser address bar displays <https://aif360.res.ibm.com>. The navigation bar includes links for Home, Demo, Resources, Events, Videos, and Community. The main heading is "AI Fairness 360", followed by a descriptive paragraph: "This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. We invite you to use and improve it." Below this, there are three buttons: "Python API Docs", "Get Python Code", and "Get R Code". A section titled "Not sure what to do first? Start here!" contains seven cards: "Read More", "Try a Web Demo", "Watch Videos", "Read a paper", "Use Tutorials", "Ask a Question", and "View Notebooks". Each card provides a brief description and a right-pointing arrow. A "Contribute" section is also visible at the bottom left, encouraging users to add metrics and algorithms in GitHub and share Jupyter notebooks.

Distribution

← Google's What-If

← <https://pair-code.github.io/what-if-tool>









The screenshot displays the 'What-If Tool' web interface. At the top, the title 'What-If Tool' is on the left, and navigation links for 'GET STARTED', 'TUTORIALS', 'DEMOS', 'FAQs', 'GET INVOLVED', and 'GITHUB' are on the right. The main content area features a heading: 'Visually probe the behavior of trained machine learning models, with minimal coding.' Below this is a red 'GET STARTED' button. The interface is divided into several panels. On the left, there's a 'Datapoint editor' with options for 'Visualize' and 'Partial dependence plots'. The central panel shows a list of features: 'age', 'capital gain', 'capital loss', 'education', 'education num', 'hours per week', 'marital status', 'native country', 'occupation', 'over_50k', 'race', 'relationship', 'sex', and 'workclass'. To the right of this list are two scatter plots. The first plot is labeled 'correct' and shows a positive correlation between 'age' and 'Inference score 1'. The second plot is labeled 'incorrect' and shows a similar positive correlation. A legend at the bottom right indicates that blue dots represent '≤50k' and red dots represent '>50k'.



















Advocacy

- ← This is where we ALL come in
- ← Bring up concerns in talks and classrooms (like this one!)
- ← Dedicate part of the syllabus (like this one)
- ← Take an entire class on AI and Ethics

Advocacy

- ← Obtain and promote more diverse research perspectives
- ← In 2017, Joy Buolamwini (PhD student, MIT media lab) found facial recognition platforms at Microsoft, IBM, and Face++ did very poorly when identifying women and minorities

| Gender Classifier | Overall Accuracy on all Subjects in Pilot Parliaments Benchmark (2017) |
|---|---|
|  Microsoft | 93.7%  |
|  FACE++ | 90.0%  |
|  IBM | 87.9%  |

| Gender Classifier | Darker Male | Darker Female | Lighter Male | Lighter Female | Largest Gap |
|---|--|--|--|--|--|
|  Microsoft | 94.0%  | 79.2%  | 100%  | 98.3%  | 20.8%  |
|  FACE++ | 99.3%  | 65.5%  | 99.2%  | 94.0%  | 33.8%  |
|  IBM | 88.0%  | 65.3%  | 99.7%  | 92.9%  | 34.4%  |

Advocacy

- ← While each service touted an excellent overall accuracy, certain subgroups performed very poorly
- ← Error analysis reveals that:
 - ← 93.6% of faces misgendered by Microsoft were those of darker subjects.
 - ← 95.9% of the faces misgendered by Face++ were those of female subjects.
- ← Buolamwini created <http://gendershades.org/> and contacted each company regarding their inclusion and diversity practices during development

Advocacy

← Small and large-scale initiatives

← AI4ALL

← Women in AI

← Black in AI

← AI Now and NYU

← Human-Centered AI Institute at Stanford

← Ada Lovelace Institute

← AAAI/ACM Conference on AI, Ethics, and Society

Advocacy

- ← Timnit Gebru was hired as Google's director of AI ethics in 2018
- ← She authored a paper on the risks of large language models (LLMs) acting as stochastic parrots
- ← Google management requested that Gebru either withdraw the paper or remove the names of all the authors employed by Google
- ← In December 2020, she was fired.
- ← Hundreds of Google employees co-signed a letter condemning Google's actions
- ← Advocacy can be uncomfortable but it is necessary to drive change

Advocacy

Sign this letter

Dear Sundar,

We believe that Google should not be in the business of war. Therefore we ask that Project Maven be cancelled, and that Google draft, publicize and enforce a clear policy stating that neither Google nor its contractors will ever build warfare technology.

- ← Employees of Amazon and Microsoft have likewise worked to withdraw their respective companies from DoD contracts

10 Data Science Ethics Questions

- **Q1: Which laws and regulations might be applicable to our project?**
 - Which laws and regulations might be relevant?
 - What these laws are designed to protect or accomplish?
 - What the impact may be of not taking them into account?
 - This includes considering recent regulations such as [GDPR](#) (the European General Data Protection Regulation).

10 Data Science Ethics Questions

- **Q2: How are we achieving ethical accountability?**
 - It should be clear who will be accountable to minimize the harm that could be done by the project.
 - Accountability includes ensuring the project team proactively identifies potential stakeholders and evaluates harms such as possible disproportionate effects that may arise from the application of a model.

10 Data Science Ethics Questions

- **Q3: How might the legal rights of an individual be impinged by our use of data?**
 - For the project to be ethical, the organization must have the right to use the data for their specific purpose.
 - For example, privacy issues should not only focus on who owns the collected data, but also the rights that need to be applied to downstream users of that data.

10 Data Science Ethics Questions

- **Q4: How might individuals' privacy and anonymity be affected by our aggregation and linking of data?**
 - Consideration should be given to how privacy will be maintained through the transmission, storage and merging of the data.

10 Data Science Ethics Questions

- **Q4: How might individuals' privacy and anonymity be affected by our aggregation and linking of data?**
 - Consideration should be given to how privacy will be maintained through the transmission, storage and merging of the data.

10 Data Science Ethics Questions

- Q5: How do we know that the data is **ethically available** for its intended use?
 - Being able to access and collect data does not mean that it is ethical to use that data.
 - Care must be taken to understand who owns the data, what are their rights and expectations, and is the data being used the way that the contributors intended?

10 Data Science Ethics Questions

- **Q6: How do we know that the data is **valid** for its intended use?**
 - A data science project should ensure that the data that is used for the project is suitable for the intended use within the project.
 - One aspect of data validity is data accuracy. For example, imputing missing values or excluding records with missing values could have a significant impact on the downstream analytical results (which might amplify bias).
 - Another data validity concern is related to 'fitness of purpose' with respect to how specific data will be used.

10 Data Science Ethics Questions

- **Q7: How have we identified and minimized any bias in the data or in the model?**
 - Data science machine learning models can be built using data that has a bias, and thus, the model might also learn this bias (for example, the use of machine learning algorithms has shown the capability of inheriting racial and social biases).
 - Bias might come from the fact that the data used to build the model was biased.
 - The data science team needs to be aware that the choices with respect to training data might have profound impacts on others.

10 Data Science Ethics Questions

- **Q8: How was any potential modeler bias identified, and if appropriate, mitigated?**
 - There could be subjectivity within the model building process, in that model building involves subjective decisions, and that these decisions can result in biases and prejudices.
 - There can be subjectivity when decisions must be made within the project, such as with respect to what metric one should optimize, which algorithm to use, which data sources to use, or if one data point should be used as a proxy for a missing fact.

10 Data Science Ethics Questions

- **Q9: How transparent does the model need to be and how is that transparency achieved?**
 - An explanation in understandable terms as to why a specific decision is recommended often cannot be supplied — even by the team that builds the model.
 - This makes explainability and comprehensibility very difficult.
 - Many models are effectively a black box.
 - Model transparency is particularly important when model output might disadvantage a certain subgroup, or in situations where there is a high degree of regulation or a right of challenge (e.g, lending money).

10 Data Science Ethics Questions

- **Q10: What are likely **misinterpretations** of the results and what can be done to prevent those misinterpretations?**
 - Most predictive models are statistical in nature. They provide no guarantees.
 - With this in mind, the data science project manager should ensure that the analytical decisions are made as a result of a data science project that reflects the scale, accuracy and precision of the data that was used in creating the model.

Conclusions

- ← Data science and artificial intelligence are only going to become more intertwined with our daily lives (self-driving cars, smart homes, internet-of- things)
- ← Automated decision-making has the potential to shape our civilization on a large scale
- ← Understanding this technology and the strengths and limitations of its abilities is critical as we integrate it ever more deeply into our everyday routines
- ← Being able to interface not only with researchers, but with policymakers, legislators, and the public is going to be essential
- ← Can no longer afford to hide behind the ivory tower and ignore the implications of our work, and its unintended consequences

References

← AI Ethics Resources

<https://www.fast.ai/2018/09/24/ai-ethics-resources/>