Ethics Toolkit

ETHICALLY ALIGNED AI:APPLIED ETHICS FOR AI DEVELOPMENT (V1.0 NOVEMBER 2020)

Curated by Katrina Ingram, BBA, MA (Communications and Technology)

Introduction

Why an ethics toolkit?

Perhaps the biggest challenge AI developers face in applying ethics is integrating broad, high level principles into their actual, day to day work. What does it mean to promote well being or to do no harm? How do you ensure fairness or mitigate bias? In working with the AI development community while researching artificial intelligence and ethics there seemed to be a need for a set of practical, tangible, accessible tools to help developers apply ethical considerations in the context of their work. These tools exist, but for the most part, they are buried in long papers or communicated in ways that make them difficult to use.

This toolkit provides an easily accessible, curated set of resources to help guide AI developers broaden their ethical thinking. These resources intersect with the AI development cycle at key points. The toolkit also serves to facilitate communication between technical and non-technical team members in order to foster a more inclusive conversation that can bring domain experts or end user stakeholder voices into the development process.

Who should use this toolkit?

This resource is aimed primarily at students in post-secondary programs in computing science who are focused on AI development. A secondary audience is computing science educators, who may wish to incorporate these tools into course curriculum. As such, these tools are meant to be used by the individual practitioner in order to document their work or broaden their ethical thinking. The selection of these tools account for the fact that students have limited resources – both time and money – to conduct processes or apply concepts which might involve more elaborate activities (e.g. extensive stakeholder engagement).

Those building AI in industry may have different needs than students or academic researchers, however, its conceivable that a modified version of the toolkit could be useful for developers and project managers working on AI related systems. These groups may need more customized domain specific solutions or processes that involve greater stakeholder engagement as part of the market research process.

Scope

Version one contains 10 tools that align with various aspects of the AI development workflow. It leverages existing concepts from a range of resources, which in some cases, have been modified in order to turn an academic concept into a tool or to make the tool more user friendly.

There are hundreds of ethics resources available. The goal of this project was to review and curate a small set of practical and useful ethics tools that connect with the AI workflow. It's acknowledged that the concept of "useful" is framed by the curation process. One limitation is the inability to assess highly technical tools (e.g. samples of code, Github resources). However, this limitation also has a benefit in sourcing tools that non-technical people, who may be part of a bigger interdisciplinary research team, can also understand, thereby helping to foster a dialogue between team members.

Acknowledgements

This ethics toolkit builds on the work of many people conducting research, developing methodologies, writing ethical codes, publishing papers and developing ethics tools aimed at helping to build better AI. A full list of references is included and authors for each tool or concept are noted throughout the document. In addition to the individual authors of the ten tools selected, this project benefitted immensely by work conducted by Jessica Morely, Luciano Floridi, Libbey Kinsey and Anat Elhalal who reviewed over 100 ethics tools, research papers and methods in their paper, "*From What to How:*

An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices". A link to this database is included in Additional Resources.

Disclaimer

A question we should ask about any resource is whose interests are being served? The research, tools and resources featured in this ethics toolkit may have been funded either directly or indirectly through research grants, government sources, corporate funding, private donations or philanthropic foundations. There may be economic or ideological reasons to foster a certain perspective either intentionally or unintentionally. Please keep this in mind as you review the tools in this document.

Also, the tools listed in this document are North American and Euro-centric and the foundational ethical principles represented are primarily extensions of Western philosophy. Its important to also acknowledge this cultural bias and we will attempt to address it in future iterations of this toolkit.

Ethics Overview

Ethical questions are normative, seeking answers about how the world should operate. There are differing approaches to ethics itself, but in general there are "three major critical orientations: deontological ethics, utilitarianism (sometimes called consequentialism), and virtue ethics" (Goldsmith & Burton, 2017, p. 25). To summarize these positions:

	Key Features	Key Figures
Deontology	• Rules/law based • Moral duty • Laws can be universal • What are the right rules and how best to apply them?	Immanuel Kant Biblical (Ten Commandments)
Utilitarianism	• Greatest good for most • Concern for consequences • Favoured by computer science, fits well with a math model • What is the greatest possible good for the greatest number?	Jeremy Bentham John Stuart Mill
Virtue Ethics	• Moral Character • Good personal habits • Practical wisdom (phronesis) • Individual/localized • Who should I be?	Ancient Greeks Aristotle

Figure A (Goldsmith & Burton, 2017)

Ethical issues can be explored from various angles and its useful to consider different approaches when confronted with difficult choices (Goldsmith &Burton, 2017). There are many difficult choices to consider in the development and deployment of AI.

Ethical issues in AI

There are numerous ethical considerations that can arise in the development of AI systems. This graphic illustrates some of the more common terms included in a range of ethical codes[1].

Many of these issues such as bias, fairness, justice, privacy and equity align with human rights. If an AI system is causing harm by treating certain people in unjust ways or infringing on personal rights, it has serious consequences for not only the impacted people, but for society as a whole. AI developers have a responsibility to ensure their work is not creating harm. Instead, AI could be used to promote human values.

The AI community also has a self-interested reason to care about ethics. AI needs a social license, which we can think of as moral approval or acceptance, to continue to operate in society. Without it, AI as a discipline is at risk. This has happened before in the field of nuclear energy. Prior to Chernobyl, nuclear energy was a promising technology that was yielding investments in both research and commercial applications. The magnitude of the Chernobyl disaster led to a public distrust of nuclear energy and the social license for developing it evaporated. Decades later, there is still little appetite for nuclear energy despite its benefits.

Culture and Context

Ethics is about appropriateness not accuracy. Determining what is appropriate is impacted by culture and context. For example, the actions taken by a government in a national emergency might be deemed an overstepping of boundaries during non-emergency times. In addition, the actions of a democratic government will be judged in a different light by its citizens than that of an authoritarian government.

There is also a level of subjectivity in the definitions of certain ethical issues. For example, what is fairness? Fairness can be defined in different ways depending on what goals we are trying to achieve. This talk explains 21 definitions of fairness and in each case, different goals are being pursued by different stakeholders. This poses a challenge in trying to ensure one “true” definition of fairness, because there are many ways to look at fairness and much depends on whose perspective is being taken.

Despite these challenges in applying ethics, we need to attempt to address these important issues. This toolkit provides a starting point for AI developers to take some practical steps in addressing ethical considerations.

[1] Retrieved from - https://www.researchgate.net/figure/Word-cloud-of-concepts-frequently-occurring-in-principles-and-codes-based-on-the_fig2_337565648

The Tools

10 Practical Ethics Tools to Support AI Development

	Stage in Workflow	Project Purpose	Data Collection and Cleaning	Build/Evaluate	Deploy/Monitor
1	AI Blind Spot	X	X	X	X
2	Princeton's AI Ethics Case Studies	X	X	X	X
3	Responsible AI Design Assistant	X			
4	10 simple rules for responsible big data research	X			
5	Harms Modeling	X			X
6	Datasheets for Datasets		X		
7	Data Ethics Canvas		X		
8	Model Cards for Model Reporting			X	X
9	Aequitas Bias and Fairness Audit			X	X
10	The Machine Learning Reproducibility Checklist				X

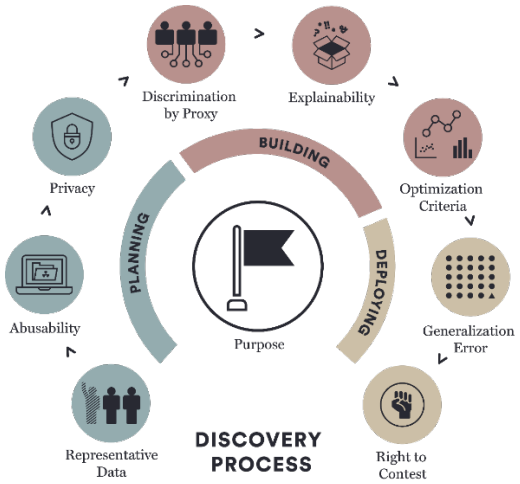
Tool One: AI Blind Spot

What is it?

- A discovery process for spotting unconscious bias using a deck of flashcards
- Works from a premise that we all have “blind spots”
- Cards are physical or digital. Each card is mapped to a part of the workflow process
- Cards list questions, resources and case studys by concept
- Everything is licensed under Creative Commons for use.

Who developed it? This tool was developed by Ania Calderon, Dan Taber, Hong Qu and Jeff Wen during the Berkman Klein Center and MIT Media Lab’s 2019 Assembly Program.

A closer look:



Retrieved from - https://aiblindspot.media.mit.edu/

How to use this tool:

- The cards can be used beginning to end during the AI design cycle or individually at a particular part of the workflow. Cards contain question, a case study example, suggestions on who to engage and a QR code that links to more information.
- Since the cards are easy to understand, they have also been used with non-technical audiences (policy makers, stakeholders etc) to help better explain, educate and drive conversation around the AI workflow.

More information: Learn more about AI Blind Spot and download the cards.

Tool Two: Princeton's AI Ethics case studies

What is it?

- A set of six fictional case studies designed to prompt questions and foster collaboration.
- The case studies cover various scenarios including healthcare, education, hiring, the public sector, criminal justice and voice/sound recognition. Each case covers a range of AI ethics issues which are noted at the bottom of each case.
- Each scenario is delivered as a narrative story and includes questions for reflection and discussion.

Who developed it? The case studies were produced by Princeton University as a collaboration between the University Center for Human Values and the Center for Information Technology Policy. Here is a list of the steering committee members for Princeton's Dialogues on AI and Ethics.

A closer look: Case Study 5: Hiring by Machine

This case study outlines a fictional scenario whereby an AI-enabled resume screening system filters out an otherwise qualified job applicant. The applicant lodges a complaint that brings a human into the process and its determined that the AI-system has used information that is irrelevant in making decision around cultural fit for the organization. The case includes reflection question to explore the ethical issues

Retrieved from - https://aiethics.princeton.edu/wp-content/uploads/sites/587/2018/12/Princeton-AI-Ethics-Case-Study-5.pdf

How to use this tool:

- Read through the case study. Each case ranges from 6-12 pages in length, including the reflection questions. It takes approximately 15-20 minutes to read the case and approximately 45-90 minutes to reflect on and answer the questions.
- It helps to write down your response to the reflection questions if you are doing the work solo.
- If you have another person or small group, you can work through the case together and use the questions to facilitate dialogue.

- Even if a case is not in your specific domain, it can still provide useful areas of ethical reflection. For example, Case Five is focused on hiring but it also speaks to broader issues and themes such as Fairness, Irreconcilability, Diversity, Capabilities and Contextual Integrity. These areas of AI ethics are highlighted at the bottom of each case.

More information: For more information and to access all six case studies visit Princeton's website for Dialogues on AI and Ethics.

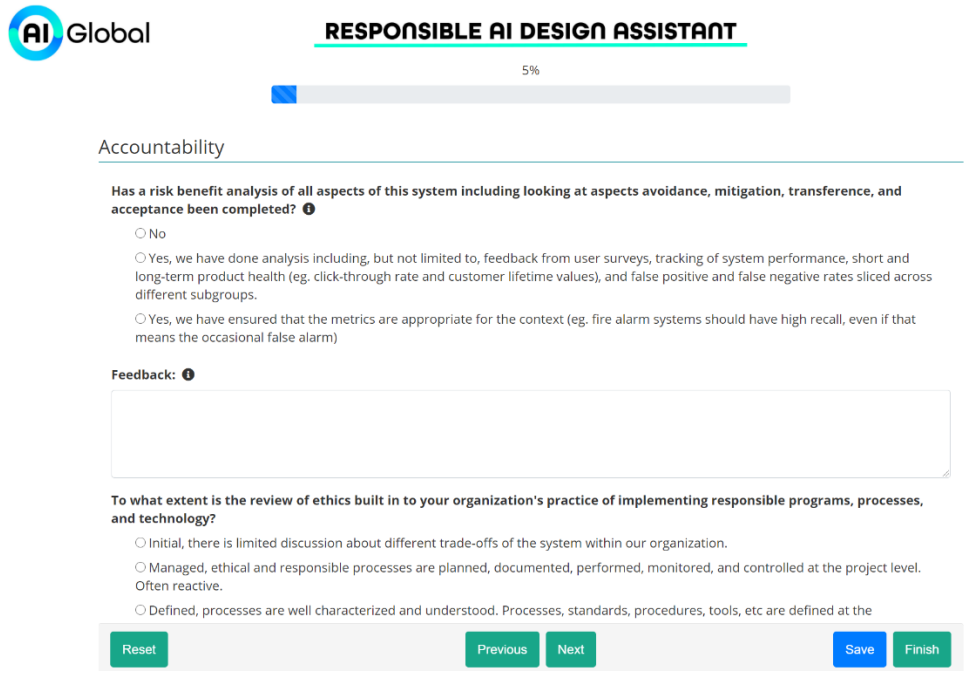
Tool Three: Responsible AI Design Assistant

What is it?

- An online survey to help assess your project for accountability, explainability and interpretability, data quality, bias and fairness and robustness
- Open source tool, currently in beta
- Meant to foster accountability/responsibility in the design process

Who developed it? AI Global is a not for profit organization founded in 2017 with a mission to "help individuals and organizations to navigate and easily adopt responsible AI business practices, making it easier to access relevant information, and work to inform AI regulation." (AI Global)

A closer look:



Retrieved from - https://oproma.github.io/rai-trustindex/

How to use this tool:

- Access the online survey and answer the questions.
- Questions are comprehensive and may require multiple team members to participate
- Designed to be used at the start of a project but can also be used at other phases in the project

- While the tool is aimed at organizations or teams building AI systems, the questions are still relevant for individual researcher to think about in the context of their work.
- This is a beta, the designers are requesting feedback on the tool

More information: Go to the Responsible AI Design Assistant online tool, read the guidelines for use, find out more about how this tool was developed or learn more about AI Global.

Tool Four: 10 Simple Rules For Responsible Big Data Research

What is it?

- High level list of rules for responsible big data research practices
- Aimed primarily at academic researchers

Who developed it? A team of inter-disciplinary researchers that includes:

- Matthew Zook
- Solon Barocas
- danah boyd,
- Kate Crawford
- Emily Keller
- Seeta Peña Gangadharan
- Alyssa Goodman
- Rachelle Hollander
- Barbara A. Koenig
- Jacob Metcalf
- Arvind Narayanan
- Alondra Nelson
- Frank Pasquale

A closer look:

7. Develop a code of conduct for your organization, research community, or industry

“The process of debating tough choices inserts ethics directly into the workflow of research, making “faking ethics” as unacceptable as faking data or results. Internalizing these debates, rather than treating them as an afterthought or a problem to outsource, is key for successful research, particularly when using trace data produced by people. This is relevant for all research including those within industry who have privileged access to the data streams of digital daily life. Public attention to the ethical use of these data should not be avoided; after all, these datasets are based on an infrastructure that billions of people are using to live their lives, and there is a compelling public interest that research is done responsibly.” (10 Simple Rules for Big Data Research)

Retrieved from - https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005399

How to use this tool:

- The 10 rules provide a list of relevant issues for big data researchers to consider and think about. The rules also contain links to papers and other resources that provide more in-depth discussion.
- Each rule also contains suggestions for further action. For example, Rule 7 talks about building a research code of ethics for your research project and suggests thinking about user expectations and the general public’s perception of the research (e.g. will they think it’s creepy?) as questions

to consider. However, one critique of this tool it that it doesn't provide a concrete next step or "how to" in creating an ethical research project code.

More information: Learn more about the 10 Simple Rules for Responsible Big Data Research

Tool Five: Harms Modeling

What is it?

- Harms modeling helps identify gaps and risks in assessing technology.
- It is both a set of tools and a practice to help technology builders design better solutions.
- There are a number of resources to help understand and assess harms and to look at harms from various stakeholder perspectives through questions and case studies.
- There is a table that outlines various types of harms to consider and a process for ranking the level of severity of harm.

Who developed it? Microsoft developed the Harms Modeling process and they share these tools on their website.

A closer look: Here is a sample of one type of harm, social detriment.

Social detriment

At scale, the way technology impacts people shapes social and economic structures within communities. It can further ingrain elements that include or benefit some, at the exclusion of others.

Harm	Description	Consideration(s)	Example
Amplification of power inequality	This may perpetuate existing class or privilege disparities.	How might this technology be used in contexts where there are existing social, economic, or class disparities? How might people with more power or privilege disproportionately influence the technology?	Requiring a residential address & phone number to register on a job website could prevent a homeless person from applying for jobs.
Stereotype reinforcement	This may perpetuate uninformed "conventional wisdom" about historically or statistically underrepresented people.	How might this technology be used to reinforce or amplify existing social norms or cultural stereotypes? How might the data used by this technology cause it to reflect biases or stereotypes?	Results of an image search for "CEO" could primarily show photos of Caucasian men.
Loss of individuality	This may be an inability to express a unique perspective.	How might this technology amplify majority opinions or "group-think"? Conversely, how might unique forms of expression be suppressed. In what ways might the data gathered by this technology be used in feedback to people?	Limited customization options in designing a video game avatar inhibits self-expression of a player's diversity.

Retrieved from - https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/type-of-harm

How to use this tool:

- There is a lot of information on this site. Depending on your project and background knowledge, you may want to skim certain sections. Its worth an in-depth read if you are new to the concepts. Otherwise, you may wish to skip ahead and just focus on specific areas.
- Read through the definitions and the harms listed in the types of harms table.
- Generate a list of potential harms that relate to your project.
- Evaluate these harms based on level of severity. The site will point you to this step after you've reviewed the types of harms.
- The Assessing Harms Index provides a nice graphic representation of a finished harms evaluation and serves as a sample outcome.
- The final step is to determine how you will address these harms.
- This is an involved process. It can be done independently but it may be helpful to have different stakeholders work on this as a team.

More information: Learn more about Harms Modeling or go directly to the list of harms

Tool Six: Datasheets for Datasets

What is it?

- A standard for documenting data sets
- Based on a concept in the electronics industry that documents each component with a datasheet (see below)
- Propose that every dataset be accompanied with "a datasheet that documents its motivation, composition, collection process, recommended uses, and so on." (Gebru et al, 2018)
- Meant to be adapted to be domain and workflow context specific
- Not to be "automated" – the value is in the reflective work of manually completing

Who developed it? This tool was developed by team of researcher led by Timnit Gebru, a research scientist at Google in the ethical AI team. Other team members are:

- Jamie Morgenstern, Georgia Institute of Technology
- Briana Vecchione, Cornell University
- Jennifer Wortman Vaughn, Microsoft Research
- Hanna Wallach, Microsoft Research
- Hal Daume III, Microsoft Research; University of Maryland
- Kate Crawford, Microsoft Research; AI Now Institute

A closer look: Sample of a completed datasheet

A Database for Studying Face Recognition in Unconstrained Environments **Labeled Faces in the Wild**

Where $\hat{\sigma}$ is the estimate of the standard deviation, given by:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{10}(p_i - \hat{\mu})^2}{9}} \quad (3)$$

The multiple-view approach is used instead of a traditional train/validation/test split in order to maximize the amount of data available for training and testing.

Training Paradigms: There are two training paradigms that can be used with our dataset. Practitioners should specify the training paradigm they used while reporting results.

- **Image-Restricted Training** This setting prevents the experimenter from using the name associated with each image during training and testing. That is, the only available information is whether or not a pair of images consist of the same person, not who that person is. This means that there would be no simple way of knowing if there are multiple pairs of images in the train/test set that belong to the same person. Such inferences, however, might be made by comparing image similarity/equivalence (rather than comparing names). Thus, to form training pairs of matched and mismatched images for the same person, one can use image equivalence to add images that consist of the same person.

  The files pairsDevTrain.txt and pairsDevTest.txt support image-restricted uses of train/test data. The file pairs.txt in View 2 supports the image-restricted use of training data.
- **Unrestricted Training** In this setting, one can use the names associated with images to form pairs of matched and mismatched images for the same person. The file people.txt in View 2 of the dataset contains subsets of people along with images for each subset. To use this paradigm, matched and mismatched pairs of images should be formed from images in the same subset. In View 1, the files peopleDevTrain.txt and peopleDevTest.txt can be used to create arbitrary pairs of matched/mismatched images for each person. The unrestricted paradigm should only be used to create training data and not for performance reporting. The test data, which is detailed in the file pairs.txt, should be used to report performance. We recommend that experimenters first use the image-restricted paradigm and move to the unrestricted paradigm if they believe that their algorithm's performance would significantly improve with more training data. While reporting performance, it should be made clear which of these two training paradigms were used for a particular test result.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
http://vis-www.cs.umass.edu/lfw/#download lists a small number of errors including a few incorrect matched pairs in the dataset and other known labeling errors. Errors could also have been introduced while determining the name of each individual in the dataset if the original caption associated with each person's photograph is incorrect. Some additional potential limitations and sources of bias are also listed at the end of the datasheet.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.
No. All data was derived from publicly available news sources.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
No. The dataset only consists of faces and associated names.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
Yes. The dataset contains people's faces.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
While subpopulation data was not available at the initial release of the dataset, a subsequent paper[2] reports the distribution of images by age, race and gender. Table 2 lists these results. The age, perceived gender and race of each individual in the dataset was collected using Amazon Mechanical Turk, with 3 crowd workers labeling each image. After exact age estimation, the ages were binned into groups of 0-10, 21-40, 41-60 and 60+.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
Each image is annotated with the name of the person that appears in the image.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
The dataset does not contain confidential information since all information was scraped from news stories.

Any other comments?

[2] http://biometrics.cse.msu.edu/Publications/Face/HanJain_UnconstrainedAgeGenderRaceEstimation_MSUTechReport2014.pdf

Fig. 2. Example datasheet for Labeled Faces in the Wild [14], page 2.

Retrieved from - https://arxiv.org/pdf/1803.09010.pdf

How to use it:

- Make a copy or print out the Datasheets for Datasets worksheet (Appendix A)
- Complete the fields based on the questions asked
- Add your own questions and modify based on your project
- Share and update as your project evolves
- The paper contains samples of completed datasheets as a reference

More information: Read the full paper [Datasheets for Datasets]

Tool Seven: Data Ethics Canvas

What is it?

- A set of questions that document data provenance and data use
- Easily printable and shareable Google docs format
- Similar concept to Datasheets for Datasets, slightly different questions

Who developed it?

The Open Data Institute is a not for profit organization founded in 2012 by Sir Tim Berners Lee and Sir Nigel Shadbolt to "show the value of open data and advocate for the innovate use of open data." (Open Data Institute)

A closer look: Here is an excerpt from the Data Ethics Canvas:

Title	Description	Input #1	Input #2
1. Data sources	Name and describe key data sources used in your project, whether you're collecting them yourself or getting access from third parties.		
2. Limitations in your data sources	Are there any limitations that might influence the outcomes of your project? Consider: • bias in data collection, inclusion, algorithm • gaps, omissions • other sensitivities such as data categorisation		
3. Sharing data with other organisations	Are you going to be sharing data with other organisations? If so, who?		
	Under what conditions?		
4. Relevant legislation and policies	What legislation or policies shape your use of this data? Consider data protection legislation, IP and database rights legislation, anti-discrimination laws, sector-specific data sharing policies/regulation (eg health, employment, taxation), sector-specific ethics legislation		
5. Rights over data sources	Where did you get the data from? Is it data produced by an organisation or data collected directly from individuals?		

Retrieved from - https://docs.google.com/document/d/1OXSrA2KDMVkHroxs_8SUoQZ5Uv0eRhtNNtIl9g_Q47M/edit

How to use this tool:

- Make a copy or print out the canvas
- Complete the fields based on the questions asked
- Add your own questions and modify based on your project
- Share and update as your project evolves

More information: Get the Data Ethics Canvas or learn more about the Open Data Institute.

Tool Eight: Model Cards for Model Reporting

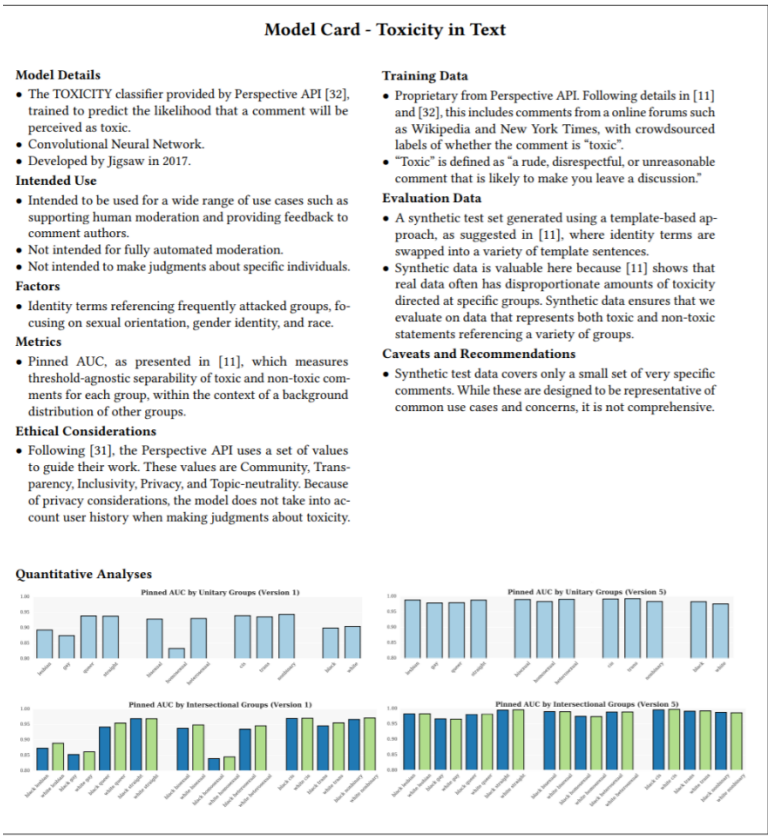
What is it?

- A documentation framework for machine learning models
- Provide benchmarks for transparent reporting of models
- Provides context details, intended uses and out of scope uses
- Snapshot of model performance
- Can be used to document a variety of machine learning models

Who developed it?

A team of researchers led by Google senior researcher Margaret Mitchell and including Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji and Timnit Gebru

A closer look: Sample of a completed model card



Model Card - Toxicity in Text

Model Details
- The TOXICITY classifier provided by Perspective API [32], trained to predict the likelihood that a comment will be perceived as toxic.
- Convolutional Neural Network.
- Developed by Jigsaw in 2017.

Intended Use
- Intended to be used for a wide range of use cases such as supporting human moderation and providing feedback to comment authors.
- Not intended for fully automated moderation.
- Not intended to make judgments about specific individuals.

Factors
- Identity terms referencing frequently attacked groups, focusing on sexual orientation, gender identity, and race.

Metrics
- Pinned AUC, as presented in [11], which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.

Ethical Considerations
- Following [31], the Perspective API uses a set of values to guide their work. These values are Community, Transparency, Inclusivity, Privacy, and Topic-neutrality. Because of privacy considerations, the model does not take into account user history when making judgments about toxicity.

Training Data
- Proprietary from Perspective API. Following details in [11] and [32], this includes comments from a online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic".
- "Toxic" is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

Evaluation Data
- A synthetic test set generated using a template-based approach, as suggested in [11], where identity terms are swapped into a variety of template sentences.
- Synthetic data is valuable here because [11] shows that real data often has disproportionate amounts of toxicity directed at specific groups. Synthetic data ensures that we evaluate on data that represents both toxic and non-toxic statements referencing a variety of groups.

Caveats and Recommendations
- Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

Quantitative Analyses

Pinned AUC by Unitary Groups (Version 1)

Pinned AUC by Unitary Groups (Version 5)

Pinned AUC by Intersectional Groups (Version 1)

Pinned AUC by Intersectional Groups (Version 5)

Figure 3: Example Model Card for two versions of Perspective API's toxicity detector.

Retrieved from - https://arxiv.org/pdf/1810.03993.pdf

How to use this tool:

- Make a copy or print out the Model Cards for Model Reporting worksheet (Appendix A)
- Complete the fields based on the questions asked
- Add your own questions and modify based on your project
- Share and update as your project evolves
- The paper contains samples of completed model cards as a reference

More information: Read the full paper [Model Cards for Model Reporting]

Tool Nine: Aequitas bias and fairness audit tools

What is it?

- An open source set of tools that allows developers to perform a self-assessment audit for a given set of bias and fairness measures
- Recognizes that the concept of fairness is not uniform and uses parity measures in an attempt to clarify how fairness is assigned in a model. Allows the user to make choices and trade-offs.
- This toolset is also useful for policy makers to understand the impacts of deploying a model
- Can be used as a website assessment tool or as code (Github available)
- Helps make self-audits a standard procedure in the development process

Who developed it?

A team of researchers primarily affiliated with the Center for Data Science and Public Policy at the University of Chicago which has moved to Carnegie Mellon University. The team includes:

- Pedro Saleiro
- Abby Stevens
- Ari Anisfeld
- Rayid Ghani

A closer look: Here is an overview of the Fairness Tree used in assessing decision making trade-offs

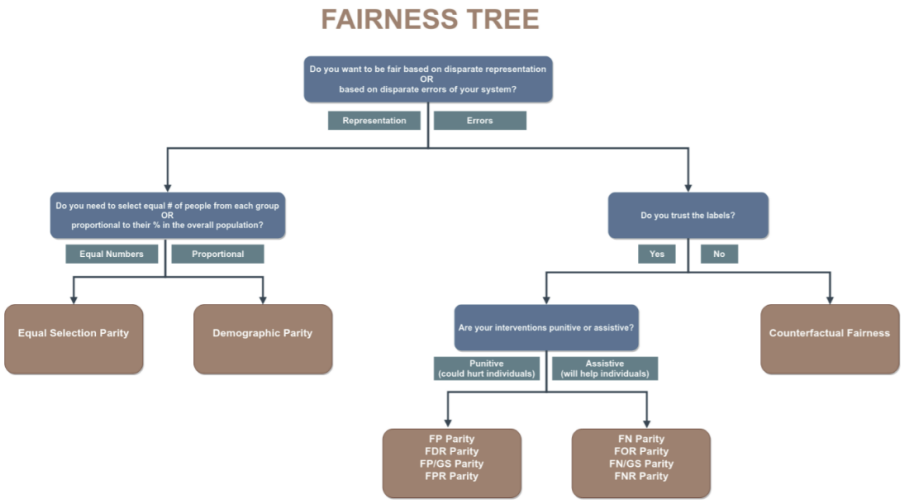


Figure 3: Fairness tree helps both data scientists and policymakers to select the fairness metric(s) that are relevant to each context.

Retrieved from - https://arxiv.org/pdf/1811.05577.pdf

How to use this tool:

- There are different options for using this tool. You can use the website to upload your own data or use a sample dataset to generate a bias report. There is also a sample bias report on the site.

- The web-based tool contains a step by step walk through. Once you select or upload data, you are prompted to enter your selected protected groups. Then, you enter your fairness metrics. The tool will calculate a bias report. This tool provides a quick review and is also suitable for non-technical users.
- In addition to the web-based tool, there is a GitHub repository that contains a Python library, code samples and other documentation. Technical users might find this set of resources a better fit for their project.

More information: Learn more and access the web assessment or code toolkits at Aequitas or read the paper.

Tool Ten: The Machine Learning Reproducibility Checklist

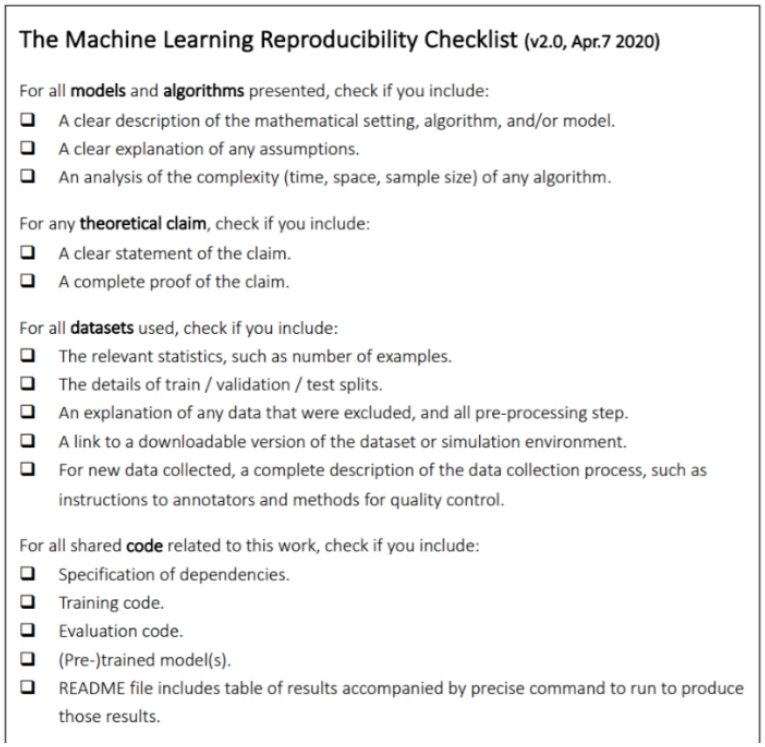
What is it?

- Aimed at promoting the concept of reproducibility of results
- A simple checklist the was developed as a tool to support a report on Machine Learning Reproducibility for the 2019 NeurIPs Conference
- Reproducibility is defined as using the same analysis and the same code to get the same results

Who developed it? McGill University led the initiative which includes the following researchers:

- Joelle Pineau
- Philippe Vincent-Lamarre
- Koustuv Sinha
- Vincent Lariviere
- Alina Beygelzimer
- Florence d'Alche-Buc
- Emily Fox
- Hugo Larochelle

A closer look:



The Machine Learning Reproducibility Checklist (v2.0, Apr.7 2020)

For all **models** and **algorithms** presented, check if you include:

- ❑ A clear description of the mathematical setting, algorithm, and/or model.
- ❑ A clear explanation of any assumptions.
- ❑ An analysis of the complexity (time, space, sample size) of any algorithm.

For any **theoretical claim**, check if you include:

- ❑ A clear statement of the claim.
- ❑ A complete proof of the claim.

For all **datasets** used, check if you include:

- ❑ The relevant statistics, such as number of examples.
- ❑ The details of train / validation / test splits.
- ❑ An explanation of any data that were excluded, and all pre-processing step.
- ❑ A link to a downloadable version of the dataset or simulation environment.
- ❑ For new data collected, a complete description of the data collection process, such as instructions to annotators and methods for quality control.

For all shared **code** related to this work, check if you include:

- ❑ Specification of dependencies.
- ❑ Training code.
- ❑ Evaluation code.
- ❑ (Pre-)trained model(s).
- ❑ README file includes table of results accompanied by precise command to run to produce those results.

Retrieved from - https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf

How to use this tool:

- Review the questions and check the boxes as applicable. Use the checklist to help identify areas of concern or missing information.
- If using this tool during a paper submission process, there are specific suggestions included with the tool.

More information: Get the Machine Learning Reproducibility Checklist or read the paper

Additional Resources

Want more ethics tools? Search over 100 resources organized by ethical principle: Beneficence, Non-Maleficence, Autonomy, Justice and Explicability. Search the AI Ethics Typology

Industry Toolkits

Google Responsible AI

IBM 360 Fairness Toolkit

Microsoft Responsible AI Tools

Ethical Codes

There are close to a hundred codes related to AI ethics. Here is a paper outlining some of these ethical codes. Ethical codes have been assembled by government, industry, industry associations, not for profit organizations and academia. Most codes represent a Western worldview.

A few notable ethical codes:

The ACM Code of Ethics and Professional Conducts for Computing Professionals is not specific to AI, but takes a wider perspective on the profession of computing. It was one of the first codes developed for computing professionals in 1992 and it was recently updated in 2018.

Montreal Declaration is one of the few Canadian led efforts. This code is also notable for it's public engagement process which involved input from hundreds of stakeholders in the community.

IEEE crafted guidelines for Ethically Aligned Design and takes a global perspective.

About Katrina Ingram

Katrina Ingram is a former technology marketer and media executive. She holds an undergraduate degree in business administration from Simon Fraser University and recently completed a master's in communication and technology at the University of Alberta. Her research is focused on artificial intelligence and applied ethics. Katrina is part of the GuARD-AI research team which is looking at the impact of COVID-19 in Alberta. She also teaches at MacEwan University. Katrina recently founded Ethically Aligned AI and is working on developing audits for AI systems, consulting services, educational workshops and ethics tools. Find out more at Ethically Aligned AI.

Subscribe to updates to this toolkit. Email katrina@ethicallyalignedai.com

References

Calderon, A., Taber, D., Qu, H. & Wen, J. (2019). AI Blindspot. Retrieved from - https://aiblindspot.media.mit.edu/

"Data ethics canvas" (n.d.). Open Data Institute. Retrieved from - https://docs.google.com/document/d/1OXSrA2KDMVkHroxs_8SUoQZ5Uv0eRhtNNtIl9g_Q47M/edit

"Dialogues on AI and Ethics Case Study PDFs" (n.d.). Princeton University. Retrieved from - https://aiethics.princeton.edu/case-studies/case-study-pdfs/

"Foundations of assessing harm" (2020, May 18). Microsoft. Retrieved from - https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2018). Datasheets for Datasets. In *arXiv [cs.DB]*. arXiv. http://arxiv.org/abs/1803.09010

Goldsmith, J., & Burton, E. (2017). Why Teaching Ethics to AI Practitioners Is Important. *Thirty-First AAAI Conference on Artificial Intelligence*. Retrieved from https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/viewPaper/14271/13992

Ingram, K. (2020). AI and ethics: Shedding light on the blackbox. *International Review of Information Ethics.* Vol 28. (06/2020). Retrieved from - https://informationethics.ca/index.php/irie/article/view/380/384

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2018). Model Cards for Model Reporting. In *arXiv [cs.LG]*. arXiv. https://doi.org/10.1145/3287560.3287596

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/1905.06876

Narayanan, A. (2018, March 1) Tutorial: 21 definitions of fairness and their politics. Retrieved from - https://www.youtube.com/watch?v=jIXIuYdnyyk&t=231s

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Larochelle, H. (2020). Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2003.12206

"Responsible AI Design Assistant" (n.d.) AI Global. Retrieved from - https://oproma.github.io/rai-trustindex/

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., & Ghani, R. (2018). Aequitas: A Bias and Fairness Audit Toolkit. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1811.05577

Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A., Hollander, R., Koenig, B. A., Metcalf, J., Narayanan, A., Nelson, A., & Pasquale, F. (2017). Ten simple rules for responsible big data research. *PLoS Computational Biology*, *13*(3), e1005399. https://doi.org/10.1371/journal.pcbi.1005399

Appendix A: Worksheets

Datasheets for Datasets Worksheet Template

Area	Questions	Answer
Motivation	For what purpose was the dataset created?	
	Who created the dataset (persons) and on behalf of what entity?	
	Who funded the creation of the dataset?	
	Any other comments?	
Composition	What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances?	
	How many instances are there in total (of each type if appropriate)?	
	What data does each instance consist of?	
	Is there a label or target associated with each instance?	
	Is any information missing from individual instances?	
	Are relationships between individual instances made explicit?	
	Are there recommended data splits?	
	Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?	
	Is that dataset self contained or does it link to or rely on other resources (ie social media)?	
	Does the dataset contain that might be considered confidential?	
	Does the dataset contain data that if viewed directly might be offensive, insulting, threatening or might otherwise cause anxiety?	

	Does the dataset relate to people?	
	Does the dataset identify any sub-populations (e.g. by age, gender etc)?	
	Is it possible to identify individuals either directly or indirectly from the dataset?	
	Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?	
	Are there any errors, sources of noise, or redundancies in the dataset?	
	Any other comments?	
Collection Process	How was the data associated with each instance acquired?	
	What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?	
	If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?	
	Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?	
	Over what timeframe was the data collected?	
	Were any ethical review processes conducted (e.g., by an institutional review board)?	
	Does the dataset relate to people?	
	Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?	

	Were the individuals in question notified about the data collection?	
	Did the individuals in question consent to the collection and use of their data?	
	If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?	
	Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?	
	Any other comments?	
Pre-processing	Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?	
	Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? I	
	Is the software used to preprocess/clean/label the instances available? Include link if appropriate.	
	Any other comments?	
Uses	Has the dataset been used for any tasks already?	
	Is there a repository that links to any or all papers or systems that use the dataset?	
	What (other) tasks could the dataset be used for? Are there tasks for which the dataset should not be used?	
	Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?	

	Are there tasks for which the dataset should not be used?	
	Any other comments?	
Distribution	Will the dataset be distributed to 3rd parties outside of the entity who created it?	
	How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?	
	When will the dataset be distributed?	
	Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? I	
	Have any third parties imposed IP-based or other restrictions on the data associated with the instances? I	
	Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?	
	Any other comments?	
Maintenance	Will the dataset be updated?	
	Who is supporting/hosting/maintaining the dataset?	
	How can the owner/curator/manager of the dataset be contacted (e.g., email address)?	
	Is there an erratum? Provide link if appropriate.	
	Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?	
	If the data relates to people is there a limit on retention? Are there regulations that guide this limit?	
	Will older versions of the dataset continue to be supported/hosted/maintained?	
	If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?	
	Any other comments?	

Model Cards for Model Reporting Worksheet Template

Category	Details	Answer
Model Detail		
	Person or organization developing the model	
	Model Date	
	Model Version	
	Model Type	
	Information about training algorithms, parameters, fairness constraints etc.	
	Paper, Citation Details, License, Contact for Questions	
Intended Use (use cases for which the model was developed or envisioned)	Primary use	
	Primary intended users	
	Out of scope (unintended) uses	
Factors (e.g. demographic, environmental, technical)	Relevant factors	
	Evaluation factors	
Metrics (chosen to reflect real world impact)	Model performance Measures	
	Decision thresholds	
	Variation approaches	

Evaluation Data	Datasets	
	Motivation	
	Pre-processing	
Training Data (mirrors evaluation data)	Datasets	
	Motivation	
	Pre-processing	
Quantitative Analysis	Unitary Results	
	Intersectional Results	
Ethical Considerations		
Caveats and Recommendations		