AI-ML Ethics Modules for ESES - Version 1 with line numbers-December 2022

Shelley Stall[1] and AGU AI/ML Ethics Steering Committee and Workshop Participants[1]

[1]Affiliation not available

December 12, 2022

Posted on 12 Dec 2022 — CC-BY 4.0 — https://doi.org/10.22541/essoar.167080673.37483484/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

AI/ML Ethics in the Earth, Space, and Environmental Sciences

Modules for Considerations and Capability

Steering Committee

- Ayris A Narock, NASA / Adnet, 0000-0001-6746-7455
- Micaela Parker, Academic Data Science Alliance, 0000-0003-1007-4612
- Yuhan “Douglas” Rao, NOAA / North Carolina Institute for Climate Studies, 0000-0001-6850-3403
- Thomas Donaldson, The Wharton School
- Guido Cervone, Pennsylvania State University, 0000-0002-6509-0735
- Lance Waller, Emory University, Life Sciences/ NASEM, 0000-0001-5002-8886

Editorial team

- Guido Cervone, Penn State University, https://orcid.org/00-0002-6509-0735
- Caroline Coward, NASA Jet Propulsion Laboratory/Caltech, https://orcid.org/0000-0001-9848-5912
- Joel Cutcher-Gershenfeld, Brandeis University, https://orcid.org/0000-0001-7659-7024
- Christopher Erdmann, American Geophysical Union, https://orcid.org/0000-0003-2554-180X

Draft

- Brooks Hanson, American Geophysical Union, https://orcid.org/0000-0001-6230-7145
- Jeanne Holm, City of Los Angeles, UCLA, https://orcid.org/0000-0001-9759-5140
- John Leslie King, University of Michigan, https://orcid.org/0000-0002-8069-262X
- Laura Lyon, American Geophysical Union, https://orcid.org/0000-0003-0585-9853
- Ryan McGranaghan, Orion Space Solutions | NASA Goddard Space Flight Center, https://orcid.org/0000-0002-9605-0007
- Micaela Parker, Academic Data Science Alliance, https://orcid.org/0000-0003-1007-4612
- Delia Pembrey MacNamara, International Society for the Systems Sciences, https://orcid.org/0000-0003-3680-2323
- Ge Peng, UA Huntsville/MSFC IMPACT, 0000-0002-1986-9115
- Yuhan "Douglas" Rao, NCSU, https://orcid.org/0000-0001-6850-3403
- Erin Ryan, Booz Allen Hamilton, https://orcid.org/0000-0001-5981-9537
- Brian Sedora, American Geophysical Union, https://orcid.org/0000-0003-0825-5967
- Shashi Shekhar, UMN,
- Shelley Stall, American Geophysical Union, https://orcid.org/0000-0003-2926-8353
- Kristina Vrouwenvelder, American Geophysical Union, https://orcid.org/0000-0002-5862-2502
- Christopher D. Wirz, National Center for Atmospheric Research (NCAR), https://orcid.org/0000-0002-8990-5505

Workshop Participants

- Abby Azari, Space Sciences Lab, UC Berkeley, 0000-0002-8665-5459
- Abhinav Sharma
- Abhishek Gupta, Montreal AI Ethics Institute
- Alejandro Coca-Castro, The Alan Turing Institute, 0000-0002-9264-1539
- Alexa J. Halford, NASA Goddard Space Flight Center, 0000-0002-5383-4602
- Amanda Hoffman-Hall, Eckerd College, 0000-0002-8153-7664
- Amy McGovern, University of Oklahoma, 0000-0001-6675-7119
- Ann McCartney, NHGRI, 0000-0003-3191-3200
- Anna-Louise Ellis, Met Office, UK
- Ayris Narock, NASA Goddard Space Flight Center, ADNET Systems, Inc., 0000-0001-6746-7455
- Barbara J. Thompson, NASA Goddard Space Flight Center, 0000-0001-6952-7343
- Billy Williams, American Geophysical Union
- Brant Robertson, UC Santa Cruz, 0000-0002-4271-0364
- Brooks Hanson, American Geophysical Union, 0000-0001-6230-7145
- Caroline Coward, NASA Jet Propulsion Laboratory, 0000-0001-9848-5912
- Charlton David Lewis, II, DARPA Defense Sciences Office, 0000-0003-2112-5921
- Chris Bard, NASA Goddard Space Flight Center, 0000-0002-5926-0566
- Chris Erdmann, Michael J. Fox Foundation, 0000-0003-2554-180X
- Chris Slocum, NOAA, 0000-0001-6293-7323
- Christian Reyes, NASA Headquarters
- Christine Custis, Shenondoah University, 0000-0003-4985-4376
- Christine Kirkpatrick, NCSA, 0000-0002-4451-8042

Draft

- Christopher Luwanga, NYU Singapore, 0000-0002-6723-5563
- Christopher Wirz, NCAR, 0000-0002-8990-5505
- Daisuke Nagai, Yale University, 0000-0002-6766-5942
- Dan Crichton
- Daniel Duffy, NASA Goddard Space Flight Center, 0000-0003-0155-5019
- David John Gagne, NCAR, 0000-0002-0469-2740
- Delia Pembrey MacNamara, University of Hull, 0000-0003-3680-2323
- Edward L. McLarney, NASA Headquarters
- Emily Hirsh, 0000-0001-6340-3040
- Enrico Camporeale, University of Colorado, 0000-0002-7862-6383
- Erin Ryan, Kennesaw State Universtiy, 0000-0002-5825-9491
- Frank Soboczenski, King's College London, 0000-0003-2023-9601
- Ge Peng, University of Alabama Huntsville, 0000-0002-1986-9115
- Geeta Chauhan, Indian Veterinary Research Institute, 0000-0001-6517-6187
- Guido Cervone, Penn State, 0000-0002-6509-0735
- Jeanne Holm, City of Los Angeles
- Jeffrey S. Evans, The Nature Conservancy and University of Wyoming, 0000-0002-5533-7044
- Joel Gershenfeld, Brandeis University, 0000-0001-7659-7024
- John Leslie King, University of Michigan
- John Moisan, NASA, 0000-0002-8078-8939
- Joses Omojola, Louisiana State University, 0000-0001-5807-2953
- K. Adem Ali, College of Charlston
- Katie Creel, Northeastern University, 0000-0001-7371-2680
- Kevin Coakley
- Lance Waller, Emory University, 0000-0001-5002-8886
- Laura Carriere, NASA Goddard Space Flight Center, 0000-0001-9639-9594
- Laura Lyon, American Geophysical Union, 0000-0003-0585-9853
- Lauren M. Sanders, Blue Marble Space Institute for Science, NASA Ames Research Center, 0000-0001-9393-0861
- Lekha Patel, Sandia National Laboratories
- Louis Barbier, NASA, 0000-0003-0378-6830
- Luis Vega
- Lyara Villanova, The University of Tokyo
- Madhulika Guhathakurta, NASA, 0000-0001-5357-4452
- Malvika Sharan (she/her), The Alan Turing Institute, 0000-0001-6619-7369
- Manil Maskey, NASA, 0000-0002-5087-6903
- Maria Molina
- Matthew Argall, University of New Hampshire
- Melanie Sharif, University of Colorado Boulder
- Micaela Parker, Academic Data Science Alliance (ADSA), 0000-0003-1007-4612
- Michael M. Little, NASA
- Mike Little, WordPress
- Rajesh Sampath, Brandeis University, 0000-0003-0782-7687

Draft

- Richard Tran Mills, Argonne National Laboratory, 0000-0003-0683-6899
- Robert Morris
- Ryan McGranaghan, Orion Space Solutions, 0000-0002-9605-0007
- Ryan T. Scott, KBR/Space Biosciences Division, NASA Ames Research Center, 0000-0003-0654-5661
- Sandra Gesing, University of Illinois Chicago, 0000-0002-6051-0673
- Sarah Paik
- Shashi Shekhar, University of Minnesota, 0000-0002-9294-4855
- Shelley Stall, American Geophysical Union, 0000-0003-2926-8353
- Siddha Ganju, NVIDIA, 0000-0002-9462-4898
- Srija Chakraborty, USRA
- Steven Crawford, NASA
- Susan J Winter, University of Maryland, 0000-0002-4524-0927
- Sylvain Costes, NASA Ames, 0000-0002-8542-2389
- Tae Wan Kim, Carnegie Mellon University
- Thomas Donaldson, Wharton School of the University of Pennsylvania
- Victoria Da Poian, NASA, 0000-0003-1175-3078
- Yuhan (Douglas) Rao, NCICS, 0000-0001-6850-3403

Draft

Vision

The overarching goal of these Artificial Intelligence and Machine Learning (AI/ML) Ethics Modules is to facilitate the development of equitable and just AI/ML that maximizes potential benefits while minimizing the potential risks. AI/ML are increasingly central to understanding, monitoring, and modeling the Earth and its environments at all scales and in diverse public uses of Earth and space science. Ethical AI/ML are essential for high-quality geoscience and planetary science and for addressing and responding to climate change, severe weather, managing natural resources, and many other matters.

AI/ML can deliver results and provide information that can not be achieved by other methods. These technologies also bring the risk of bias and harm. Ethical standards, principles, and practices associated with AI/ML in geoscience research represent essential considerations for researchers and the broader community so that the observation, modeling, and forecasting of geo-phenomena (broadly defined) happens in appropriately open and inclusive ways that consider and mitigate potential adverse impacts on historically marginalized communities and society at large.

> ***"Every new technology has affordances and tendencies that tilt toward . . . benefit and harm, but how these techs play out in the public space has more to do with social institutions and humanistic education than with the technologies themselves."***
>
> – Richard Powers, novelist, professor, and winner of the 2006 National Book Award for "The Echo Maker" (quoted in the Champaign News-Gazette, January 26, 2014, discussing his novel, "Orfeo")

Draft

Introduction and Overview

AI and ML are seeing rapidly increasing applications across the Earth, environmental, and space sciences. This is thanks to increasingly large and diverse environmental data (real and synthetic) and new methodologies being developed and used by an increasingly connected global community. These and related techniques are particularly powerful in probing datasets, including in combining diverse datasets at different scales. AI/ML can be used to reveal new information, find signals in noisy data, and develop actionable predictions and forecasts. Various types of bias and harm may ensue from the source data, mismatches from data used in model development and in model operation, as well as the algorithms, and when uncertainties are not well understood or characterized.

The use of any technology or technique should be understandable, and provided with documentation on data and tools that allow for the validation and replication of any scientific results. The entire method should be explained and accessible. The use of any techniques should address potential biases, risks, and harms, especially as related to the promotion of justice and fairness. Research questions should avoid unfairness (e.g., in application of models and algorithms).

This document provides an ethical AI/ML framework and set of leading practices for AI/ML. This framework was developed through community input and facilitated discussion in the latter part of 2022, and led by a steering committee (see Appendix C). The work was guided by the American Geophysical Union (AGU), through a grant from NASA (Grant 80NSSC22K0734). The AGU is committed to leading in the ethical use of AI/ML in geoscience research.

The ethical framework is organized around seven modules, each of which is structured to provide description and considerations, support training and development, and achieve needed compliance. The seven modules are:

Module 1: Transparency, Documentating, and Reporting
Module 2: Intentionality, Interpretability, Explainability, Reproducibility, and Replicability
Module 3: Risk, Bias, and Impacts
Module 4: Trust and AI/ML
Module 5: Participatory Methods and Domain Expertise
Module 6: Outreach, Training, and Leading Practices

Draft

Module 7: Considerations for Organizations and Institutions, Publishers, Societies, and Funders

The seven modules can each be used separately, or they can be used together as a full set (with the order flexible). The first three modules are focused on core skills and practices (Transparency and Reporting; Intentionality, Interpretability, Explainability, Reproducibility, and Replicability; Risk/Bias/Impacts). The remaining four modules involve broader principles (Trust and AI/ML; Participatory Methods and Domain Expertise; Outreach, Training, and Leading Practices; Organizational, Society, and Community Considerations). A principal investigator (PI) might cover a series of these modules as part of the agenda in research team meetings. They can also be consulted on a “just-in-time” basis.

The executive summary collects the key points from all the modules and is repeated in each module. Each module is organized with the following elements:

Module Focus
Module Key Points
Module Learning Objectives
Module Vision
Module Definitions
Module Principles
Module Responsibilities and Leading Practices
Module Use Cases and Illustrative Examples
Module FAQs

This is meant to be a living framework, and the principles, responsibilities, and other elements will be regularly reviewed and updated as the technologies, applications, and institutions evolve.

Draft

Executive Summary

A set of two workshops, over two days each, brought together approximately 90 geoscience researchers utilizing AI/ML, along with ethics and social science professionals. The agenda included:

- An overview of current AGU research ethics policies
- A review of the current state of AI/ML ethics in research
- A review selected case examples of AI/ML research with ethical implications
- Establishing AI/ML ethics working groups
- Conducting a “pre-mortem” to anticipate what could possibly go wrong with AI/ML ethics
- Reviewing and discussing recommendations by Working Groups
- Ensuring language is interoperable and extensible
- Considering future trajectories of AI/ML and ethical implications
- Presenting the results to AGU, NASA, and other key leaders

Some of the highlights from these group discussions included:

- Ethics should be integrated across the AI/ML research life cycle.
- A “one size fits all” approach should be avoided with AI/ML ethics.
- The AI/ML ethics effort should be community driven. A top-down approach, especially if authoritarian, seldom works.
- Advances are needed so that human subjects review can play appropriate roles with respect to AI/ML research (e.g., Institutional Review Boards that govern human subjects research in universities and other settings)
- Appreciation that AI/ML ethics can be controversial and that ethical standards will evolve, particularly as the technology evolves.
- A leadership individual or group in AGU and other professional societies providing consultation and advice for researchers utilizing AI/ML, with the AGU Ethics Committee as a further resource.

A principle contained in the phrase from the disability movement, “nothing about us without us,” was embraced for this work and suggests a pluralistic effort backed up by core principles.

Draft

Key Points in Modules 1-7

Module 1 Key Points: Transparency, Documentating, and Reporting

Transparency in AI/ML modeling and analysis is both essential and hard to achieve. AI/ML models involve algorithms that are a product of training data and other inputs that operate in ways that are not entirely visible or knowable. At the same time, there are aspects of AI/ML models that can be described in documentation in ways that indicate intent. Further, models can have “what if” capabilities that enable users to assess how they operate with some measure of transparency.

Note that transparency and documentation primarily bolster trust, but they can also reveal cause for concern or mistrust. Transparency and documentation are a necessary (but not always sufficient) precursor to replicability, reproducibility, and explainability. Further, transparency and documentation must be weighed against other factors, such as proprietary rights and privacy. Note that not all data can or should be open for issues of privacy, proprietary and sovereign data, and related matters.

Available/accessible documentation and disclosure are central to transparency. For example, code attribution and other contributions made by those outside the circle of the project are required to facilitate transparency. Transparency needs to be considered throughout the whole lifecycle of AI/ML applications from conceptual development for applications. Note that all parts of the research cycle can’t be fully transparent – such as internal ideation on research design, but there should be transparency on early research design decisions that have implications for stakeholders, particularly vulnerable populations.

Module 2 Key Points: Intentionality, Interpretability, Explainability, Reproducibility, and Replicability

First, it is important to specify and justify the method chosen, and when possible, include alternatives considered. Model specification and documentation are needed, along with evidence that the model is operating as intended, and it is applied to the data and to solve problems it was developed to.

For a model to be used, it should be both reproducible and replicable. In general, this implies that results can be obtained again by the group who first developed the model, or by independent researchers that adopted it. Setting aside a verification dataset along with the expected output, can be used to ensure the replicability of results.

Draft

Documentation of steps in model development and testing is important both for replicability and explainability.

In some cases, pre-registration of hypotheses is helpful as an indication of explainability. However, many AI/ML applications involve exploratory, discovery science in which pre-registration of hypotheses is not possible. Even in these cases, some specification and documentation of research intent are important so that unexpected or negative findings are recognized as such, and further analysis can be conducted to determine the degree to which the findings are indeed robust and trustworthy.

Module 3 Key Points: Risk, Bias, and Impacts

Mitigating AI/ML bias, risk, and harm will enable AI geoscientists to promote impactful, transformative, beneficial research. This involves a responsibility for researchers to anticipate potential disparities in the application of models and algorithms, as well as the assessment of early and continuing results for negative impacts. The mitigation work is both proactive and reactive.

The responsibility for mitigating bias, risk and harm lies with researchers, users of the models, and funders of the research. Typically, the harm is unintentional but deeply embedded in the data, such as disparities among communities with robust weather data and others with less warning of weather events due to gaps in sensors and tracking systems that correlate with low income communities. Training data that doesn't reflect the diversity of society possess particular risks in AI/ML applications. Mechanisms to hear the voices of vulnerable populations who might be impacted by the application of AI/ML in research are especially important and these then need to be reflected in the AI/ML research (see module 5 on participatory methods). This can happen through advisory committees, community forums, and ongoing multi-stakeholder consortia associated with research initiatives. Funders are encouraged to build voice and mitigation mechanisms into the budgets for funded AI/ML research.

Investments in tools and methods to identify bias in geoscience data are encouraged. Examples of this include: 1) Society leadership can be embodied in the appointment of a chief AI/ML risk officer serving on a broader ethics committee or in the form of other resources that can provide the needed consultation and advice to society members and others as appropriate; 2) A consortium of relevant professional societies may provide the needed set of shared resources in a specific domain.

Draft

Module 4 Key Points: Trust and AI/ML

Trust in AI/ML is not something we can prescribe or guarantee, yet trust building with respect to AI/ML research is essential. For AI/ML models, systems, and developers to be seen as trustworthy there is a need for engagement throughout the research life-cycle, with adjustments they are responsive to inputs along the way. Trust in AI/ML is context-dependent and we need to consider trust from the research questions we ask, the data we are using, and the models we develop to how the output is communicated, interpreted, and used.

Trust in AI/ML requires open and transparent research (to the extent feasible). We need to communicate and quantify uncertainty, be able to explain what models do and do not do, and communicate successes and failures. Evidence of taking into account multiple perspectives in AI/ML research enhances trust. There are broader dimensions of trust in technology and trust in science that underlie trust-development with AI/ML.

Module 5 Key Points: Outreach, Training, and Leading Practices

Ethical AI/ML practices are essential for high-quality science and positive public impact. Increasing awareness of ethical AI/ML and advocating for its inclusion in all AI/ML work, must be a central tenet of any work by the data science community.

Adoption of ethical AI/ML practices requires a deliberate action on behalf of the researchers and others relevant to the research. Training and access to resources enables the development of these essential skills. Professional societies must commit to providing access to resources and training, and advocating for researchers' time to learn these practices and develop curricula to train the next generation.
Resources are not "one size fits all;" a broad, inclusive community with a wide variety of activities requires a commensurate breadth of training and educational materials. A modular approach to training materials is recommended so that materials can be combined in multiple ways. The training needs vary across early-career, mid-career and more senior researchers, with the time to participate in training and development being a key factor. A "leader as teacher" model is recommended where Principal Investigators (PIs) and mentors can bring modular material to research teams on a timely basis. "Pre-mortems" and post-mortems are recommended to anticipate what might go wrong in the planning of research involving AI/ML and subsequently to learn from outcomes.

Module 6 Key Points: Participatory Methods and Domain Expertise

Draft

A key guiding principle comes from the disability movement: "Nothing about us without us." No research should be conducted that impacts individuals and groups in society without their consent. This requires the formation of advisory groups, the utilization of stakeholder and rightsholder mapping surveys, the democratic selection of community representatives, and other mechanisms for input.

A key practice to ensure impacted community perspectives are included is the co-production of knowledge. This is valuable with stakeholders and essential with what are termed "rights holders" such as First Nations, Indigenous Canadian peoples who are neither Inuit nor Métis. This input is important in the planning and conduct of research, as well as on a continuing basis after the research is complete to address continuing implications of the research.

Open science principles are key, even if not all data can or should be open (e.g., asking researchers to publish data, NASA Information Policy NASA SPD-41a). The FAIR and CARE principles (data that is Findable, Accessible, Interoperable, and Reusable or FAIR and, with respect to indigenous and other vulnerable populations, approaches that advance Collective benefit, Authority to control, Responsibility, and Ethics or CARE) are relevant here. Note, however, that not all aspects of CARE or FAIR principles can be fully applied with AI/ML in research.

Extra resources are needed for participatory practices. Institutional Review Boards (IRBs) need to be informed about participatory methods, which may involve a balancing of benefits and risks associated with the use of AI/ML (not just the elimination of risk). Note that participatory methods vary with scale, from AI/ML applications that are local, regional, national, and international.

Module 7 Key Points: Considerations for Organizations, Institutions, Publishers, Societies, and Funders

Professional societies, universities, federal labs, industry labs, and other organizations and institutional actors have a leadership role when it comes to AI/ML ethics. Because the technologies are developing at rapid rates this calls for agile and adaptive approaches by these organizations and institutions.

Community-driven standards require funding for forums, town halls, and other mechanisms to surface and consider current practices. Tensions will surface, such as the tensions between transparency and privacy.

Draft

Professional societies and other publishers have a particular responsibility to promulgate standards relevant to the publication of research involving AI/ML models and algorithms. Funding agencies in the United States, European Union, and other settings operate under directives to ensure the ethical use of AI/ML, which can be a model for others. While industry typically treats aspects of AI/ML as proprietary, there are community liability issues that point to the carving out of “pre-competitive” spaces in which AI/ML practices, applications, and risks are shared.

Stakeholder “Pulse” Survey

A stakeholder “pulse” survey of a cross section of geoscientists (n=118; with additional details in Appendix A) was used to inform the working group sessions. The survey confirmed that there is wide support for 1) having clear ethical standards and guidelines for the use of AI/ML in research (95%), as well as for 2) ensuring explainability/interpretability (93%) and for 3) ensuring replicability when AI/ML is used in research (90%). These are 3 of the 16 indicator issues that were included in this survey, covering many aspects of AI/ML ethics. Most of these indicator issues are major “pain points” – rated both as very important and also as very difficult to do by more than half of the respondents. Importantly, a large majority (82%) did not support researchers using AI/ML in any way they chose – without attention to ethical standards or guidelines.

Draft

Module 1: Transparency, Documentating, and Reporting

Module 1 Focus

Transparency, documentating, and reporting on uncertainties with AI/ML ethics in research are essential. This module sets a key ethical framework for many of the following modules, which rely on transparency and full documentation of the work – not just availability of data and code, but of who participated in the work, and how issues were addressed, including uncertainty and bias.

Module 1 Key Points

Transparency in AI/ML modeling and analysis is both essential and hard to achieve. AI/ML models involve algorithms that are a product of training data and other inputs that operate in ways that are not entirely visible or knowable. At the same time, there are aspects of AI/ML models that can be described in documentation in ways that indicate intent. Further, models can have "what if" capabilities that enable users to assess how they operate with some measure of transparency.

Transparency and documentation primarily bolster trust. Transparency and documentation are a necessary (but not always sufficient) precursor to replicability, reproducibility, and explainability. Transparency and documentation can also be a cause for concern or mistrust: they must be weighed against other factors, such as proprietary rights and privacy. Not all data can or should be open for issues of privacy, proprietary and sovereign data, and related matters.

Available and accessible documentation and disclosure are central to transparency in AI/ML work, including the data, training data, models, model validation, protocol and methods, and uncertainties. In addition, code attribution and other contributions made by those outside the circle of the project (see for example, Module 6 on outreach) are required to facilitate transparency and trust. Including or consulting additional experts on the data or code or other stakeholders can improve understanding, and their roles and contributions should be disclosed. This is part of the broader principle in research ethics of giving credit to those giving input. Transparency needs to be considered throughout the whole lifecycle of AI/ML applications from conceptual development for applications.

Draft

Module 1 Learning Objectives

- Knowing how to achieve transparency when using AI/ML in research.
- Considerations in the documentation needed with AI/ML models.

Module 1 Vision

Transparent and accessible documentation of research design and uncertainties (following FAIR, CARE, OCAP, TRUST, etc. principles on data, report key design decisions, etc.), including data and model biases, are needed at every step of a bio-geo-physical AI/ML project. Reasons for not being transparent should be provided. Guidelines are established for reporting on data collection, data preprocessing, model construction and training (parameter values, etc), model validation, results reporting, explainability, and leading practices for using these data/pretrained models in downstream applications. Recommendations will include the importance of subject matter (e.g., bio-geo-physical science) experts at all steps of pipeline development (Module 6), preference for explainable bio-geo-physical science informed AI/ML models (Module 2), providing post-hoc explanation of blackbox models, providing sensitivity analysis for key design decisions, etc.

Module 1 Definitions

- Transparency: State of making information available for others to see what has been done (National Academies Press, 2019).
- What is it and what does it mean and what are the parameters?
    - Documentation and reporting as a part of research methods
    - Convenient access to relevant information about a research project for those having a legitimate interest in that project.

Module 1 Principles

Transparency

- ❖ **Indicate how leading AI/ML practices are followed** in your research or where departures from leading practices are needed.
- ❖ **Attribute and acknowledge** all contributions to your research, including data and model sources.
- ❖ **Clarify the protections taken** in your research around privacy, vulnerable populations, and proprietary rights with AI/ML training data, modeling, and reporting of results

Draft

Documentating

- ❖ **Document AI/ML decisions and associated digital products (software, etc.)** throughout the entire lifecycle of your research.
- ❖ **Document the life-cycle stages** (e.g., use case and data understanding, feature selection, model selection and development (with documentation of model assumptions and implication for use case), quality control safeguards, deployment, adoption and democratization).
- ❖ **Ensure documentation of provenance** with sources of data and adjustments to the data, as well generations, versions, and sources of models, and other digital objects.
- ❖ **Provide clear access** to relevant information about the AI/ML algorithms and methods.

Reporting

- ❖ **Communicate** the limitations and uncertainties in your research.
- ❖ **Disseminate** the findings to achieve appropriate impacts.

Additional supporting information on Module 1 principles:

Transparency is an ethical goal; a mark of the trustworthiness of model predictions. It can be achieved in different ways but ideally should follow leading practices and implies convenient access to relevant information about a research project for those having a legitimate interest in that project.

- Tradeoffs between transparency and other values must sometimes be made, including but not limited to: proprietary rights and privacy. These should be documented.
- Where there is a high risk of harm to individuals and communities requiring measures of security and privacy it may not sometimes be appropriate to be fully transparent
- Transparency implies documenting and communicating the limitations and uncertainties inherent in a given research project. Where there are reasons to be opaque, it should be acknowledged.
- Code attribution and acknowledging other contributions made by those outside the circle of the project are required to facilitate transparency.

Aims of transparency*:*

- The principal aim of transparency is the establishment of trust in the ends and means of a project.

Draft

- To establish trust, transparency should contribute to the facilitation of explainability, interpretability and replicability. Explainability, interpretability and replicability are integral aspects of transparency.

Module 1 Responsibilities and Leading Practices

- **Researchers are responsible for providing transparency** with AI/ML research design decisions, limitations of training data and models, and other key choices throughout the research life cycle, including as indicated in the other modules.
- **Verification and validation methods** should be reported; **evaluation metrics** should be documented and explained and **errors, and uncertainty** should be quantified and explained to the extent possible.
- **Input parameters should be reported,** including associated levels of confidence.
- **Report potential biases in training data** and implications for individuals and groups who might be at risk due to these biases.
- **Data and code should be available** following leading practice for FAIR data and software and cited in any publications or outputs.
- **Publishers should provide guidelines and instructions** to ensure transparency following leading practices including additional practices for AI/ML work as outlined here.
- **Funders of AI/ML work should require transparency plans** and that proposed methodology and data management and sharing plans comply with these leading practices.
- **The methodology should be explained as plainly and completely as possible,** including model training, and other steps to inform AI/ML results.
- **Experts and stakeholders should be acknowledged and credited**, and their input described.

Module 1 Use Cases and Illustrative Examples

- When AI/ML is utilized in modeling complex weather patterns, indicating the uncertainty and assumptions for the model helps experts and non-expert users make informed decisions.

Module 1 FAQs

- How do we convey quality information about the model?
    - It is standard practice to report the evaluation of the model following a defined evaluation metric or framework.

Draft

- How do we quantify/ensure/verify trustworthiness of ML model predictions, especially when the model will be used to inform decisions of particular consequence?
- How much information needs to be provided in order to qualify as being transparent?

Draft

Module 2: Intentionality, Interpretability, Explainability, Reproducibility, and Replicability

Module 2 Focus

Ensuring Intentionality, Interpretability, Explainability, Reproducibility, and Replicability with AI/ML in research

Module 2 Key Points

First, it is important to specify and justify the method chosen, and when possible, include alternatives considered. Model specification and documentation are needed, along with evidence that the model is operating as intended, and it is applied to the data and to solve problems it was developed to.

For a model to be used, it should be both reproducible and replicable. In general, this implies that results can be obtained again by the group who first developed the model, or by independent researchers that adopted it. Setting aside a verification dataset along with the expected output, can be used to ensure the replicability of results. Documentation of steps in model development and testing is important both for replicability and explainability.

In some cases, pre-registration of hypotheses is helpful as an indication of explainability. However, many AI/ML applications involve exploratory, discovery science in which pre-registration of hypotheses is not possible. Even in these cases, some specification and documentation of research intent are important so that unexpected or negative findings are recognized as such and further analysis can be conducted to determine the degree to which the findings are indeed robust and trustworthy.

Module 2 Learning Objectives

- Understand the key concepts related to replicability and explainability
- Build skills in the leading practices on how to ensure an AI/ML system is robust, explainable, and replicable.

Module 2 Vision

AI/ML is undergoing rapid development, and new algorithms are often rapidly available. In many cases, their statistical qualities and uncertainties are not fully known. As a result, we need a foundational approach that encourages understanding and testing of algorithms. Ideally, a scientific question should ground the justification of the method choice and application. We prioritize an open science approach to enable replicability. We define this as an approach that provides clear model specification incorporating domain knowledge and keeping hypothesis driven motivation at the forefront. We encourage the application and development of methodologies for model explainability of AI/ML models that includes post and ad hoc exploration of data and results. Remember that replicability is a map to lead other people to where you are now while explainability helps lead other people to understand why the model performs in a certain way, and helps them develop better routes.

Module 2 Definitions

- Following the definition of National Academies of Sciences, **replicability** refers to when a new study is conducted and new data are collected to achieve the same or a similar scientific question as a previous one.[Add reference here]
- As suggested in National Institute of Standards and Technology, **explainability** refers to the ability of a system to supply accompanying evidence or reason(s) for outputs produced from an AI/ML system.

Module 2 Principles

Intentionality

❖ **Indicate the intent of AI/ML applications and steps to purposefully address ethical concerns.,** even if research hypotheses are not specified in exploratory applications.

Interpretability

❖ **Always provide the interpretation of the model and findings,** including areas of uncertainty or limitations.

Explainability

❖ **Ensure that the results can be understood by expert and non-expert users** of the research.

Reproducibility

❖ **Take necessary measures to ensure that results can be reproduced** if the same data and approach is taken.

Draft

Replicability

- **Provide considerations for researchers seeking to replicate the results** with comparable data.

Additional supporting information on Module 2 principles:

Aim towards incorporating the following elements in our thinking when developing and deploying AI/ML models.

- **Intentionality:** what is the intended research question that we want to address? Taking purposeful steps to address the ethical concerns of AI/ML development and applications.
  - Is this research undertaken with a testable hypothesis in mind?
  - Are the results intended to inform decision making? If so, how well can you use the results to inform decision making?
  - How well have the results addressed the research question or the original hypothesis?
  - Have we taken the time to address aspects of explainability and interpretability at all stages of the ethical data science lifecycle?.

- **Interpretability:** *How the data connects to and influences the output/results/conclusions. Generated from the implementation of the model itself, not from post hoc exploration.*
  - What are the limitations of our data? How does the type of our data (spatial, network based, temporal, observational, experimental …) influence our model choices?
  - How well does the model provide intuition into behavior, physics laws, etc.?
  - Is our model well specified? Why was this model specification chosen?
  - Do we understand how the model is regressing or classifying the data?
  - Does our training set represent a ground truth or is it biasing our results?
  - Can we quantify the uncertainty in the model?

- **Explainability:** *High-level, simplified understanding of the data, model, and results, able to be conveyed through verbal/written descriptions*
  - Have we explored the latent space of what our model has actually learned?
  - Have we clarified our methods in such a way that other scientists understand their application?
  - How have we made our results understandable to experts and/or non-experts?

Draft

- **Reproducibility and Replicability:** *The ability for an independent investigator to repeat methods and results*
    - If someone uses the same or similar data, will they reach the same or similar conclusion? Does this hold for different models?
    - Have we adhered to open science practices? Are data, metadata, and code made appropriately public?

Module 2 Responsibilities and Leading Practices

- **Researchers employing AI and ML techniques in their research strive to ensure that their research is explainable and reproducible.** This involves both understanding, documenting, and communicating the nature of the data, models, and any assumptions or biases inherent in selecting the data and methodology.
- **Researchers intentionally and from the start, design an explainable model.** This includes defining the research question and/or testable hypotheses and developing a model that will provide insight into the nature of the relationship between the model input and output (i.e. not simply throw data at a problem and accept the model output as truth).
- **Researchers provide documentation of both low-level explanations for a scientific audience and high-level explanations for non-technical audiences.** Low-level explanations define the model and its assumptions and parameters, specify how the model uses the data to reach its result/conclusion, and describe how changing the data (may) affect the model output. High-level explanations describe the data, the model, the results, and known assumptions and biases.
- **Researchers test their models for robustness against randomness in both parameter initialization and training methodology** and verify that their results hold regardless of initial parameter values and methodology.
- **Researchers provide uncertainty quantification for their models.** This includes exploring both the efficacy of the model and the robustness of the results according to the state of the art. Understanding the meaning of the model confidence.
- **Researchers should adhere to open science practices**, ensuring that their training data and code are publicly available to the highest possible extent. Journals could provide a set of requirements to receive an "open science" label.
- **Researchers and Educators lean on expertise in other fields.** Research teams are cross-disciplinary, including expertise in computer science and

statistics. Graduate level training in statistics and/or computer science is routinely incorporated into the Geology/Geophysics degree path.

- **Journals encourage or require adhering to accepted AI/ML community standards.** This may look like recommending that the methods section address ethical concerns. A steering committee of AI Ethics researchers could provide a living document that guides these community standards, and stays updated on the current pitfalls in state-of-the-art AI.
- **Journals assign AI/ML fluent editors and reviewers.** Publishers maintain a database of qualified reviewers for AI/ML submissions across domain expertise. Out-of-domain AI/ML experts are paired with subject matter experts when appropriate domain specific AI/ML reviewers and/or editors are not available.
- **Journals routinely publish negative results.** Well-defined, hypothesis driven work is valuable regardless of the outcome. These results can add clarity and understanding of AI/ML methods and reduce repeated, unfruitful efforts.
- **Funding agencies appropriately support the effort involved in ethical AI/ML.** Opportunities expressly request adherence to ethical standards and provide funds for the time and expert personnel required to do so.
- **Funding agencies offer regular opportunities for verification and validation.** Reproducibility and replicability studies are commissioned.
- **Funding agencies prioritize funding for groups providing their science in an open manner where possible.**

Module 2 Use Cases and Illustrative Examples

- Reproducibility crisis (also seen in Module 3) https://reproducible.cs.princeton.edu/#rep-failures
- National Academies' Report on Replicability and Reproducibility
- Reproducibility Challenge by NeurIPS
- NIST Four Principles for Explainable AI

Module 2 FAQs

- How do we ensure that we understand how the model is reaching its conclusions?
- How do we ensure that other scientists are able to recreate our work? (low-level knowledge required for reproduction)
- How do we ensure that other people can understand what we have done? (high-level understanding)

Draft

Module 3: Risk, Bias, Impacts

Module 3 Focus

Identifying risks, bias, intended and unintended consequences with AI/ML ethics in research

Module 3 Key Points

Mitigating AI/ML bias, risk, and harm will enable AI geoscientists to promote impactful, transformative, beneficial research. This involves a responsibility for researchers to anticipate potential disparities in the application of models and algorithms, as well as the assessment of early and continuing results for negative impacts. The mitigation work is both proactive and reactive.

The responsibility for mitigating bias, risk and harm lies with researchers, users of the models, and funders of the research. Typically, the harm is unintentional but deeply embedded in the data, such as disparities in communities with robust weather data and others with less warning of weather events because of gaps in sensors and tracking systems that correlate with low income communities. Training data that doesn't reflect the diversity of society possess particular risks in AI/ML applications. Mechanisms to provide voice to vulnerable populations who might be impacted by the application of AI/ML in research are especially important. This can happen through advisory committees, community forums, and ongoing multi-stakeholder consortia associated with research initiatives. Funders are encouraged to build voice and mitigation mechanisms into the budgets for funded AI/ML research.

Investments in tools and methods to identify bias in geoscience data are encouraged. Leadership from AGU can be embodied in the appointment of a chief AI/ML risk officer serving on a broader ethics committee or in the form of other resources that can provide the needed consultation and advice to AGU members and other as appropriate. A consortium of relevant professional societies may provide the needed set of shared resources in this domain.

Module 3 Learning Objectives

1. Appreciate the key sources of risk and bias in AI/ML applications.
2. Build capability in mitigating or at least reducing risk and bias in AI/ML applications.

Draft

Module 3 Vision

AI/ML can benefit the Earth, geospace, space, biological, environmental, and related sciences in both knowledge generation and decision-making. However, to achieve these benefits, we must develop a set of specific, actionable, and inclusive ethical principles and responsibilities that will guide developers and users of AI. This module elucidates the biases and risks of AI/ML use by the Earth and space science research communities and develops principles to identify and address those biases and risks. These principles will also include the ability to communicate the capacity of AI/ML predictions to promote transformative justice, fairness, and the flourishing of life and the sciences.

Module 3 Definitions

- AI/ML systems include datasets, models, and deployments

Module 3 Principles

Risk

❖ **Identify risks of AI/ML applications for relevant stakeholders,** with particular attention to vulnerable communities and fragile ecosystems.

Bias

❖ **Identify and document potential sources of bias** in training data, algorithms, and other aspects of AI/ML applications

Impacts

❖ **Identify and advance the public good** as appropriate with AI/ML applications.

Additional supporting information on Module 3 principles:

To minimize the risk of AI/ML systems causing harm, intentionally or unintentionally, AI/ML developers should:

- Acknowledge that Earth, humanity, and society are linked. As such, AI/ML researchers should give comprehensive and thorough evaluations of the AI/ML systems and their impacts.
- Ensure that the public good is the central concern throughout the development of AI/ML systems.
- Work to address historic injustices and ensure such injustices do not continue to propagate further because of the AI models

Draft

- Aim toward using AI/ML systems to benefit people, ecosystems, and groups that have historically been excluded from or harmed by technological advances
- Recognize and take special care of AI systems that become integrated into the infrastructure of society.
- Ensure that the AI model is developed to protect natural systems, including Earth and its environment.
- Follow overarching guidelines that govern research activities as discussed within AGU's general AGU Scientific Ethics Policies and Integrity Policy.

Module 3 Responsibilities and Leading Practices

- Earth, environmental, and space science researchers will ensure that AI/ML systems developed for Earth, Geospace, Space and related sciences avoid harm throughout the AI/ML lifecycle by:
    - Taking responsibility for AI/ML systems and datasets and ensure that there is always a valid point of contact for all deployed and shared models and datasets
    - Ensuring that models and data are transparent to relevant parties who will use, or otherwise be affected by, the AI/ML system
        - Documenting known biases in the data and model and expected uses of the model (e.g., datasheets, model cards, or other avenues of sharing information which are publicly accessible)
    - Ensuring that AI/ML models are regularly assessed for:
        - Biases stemming from computational, human, or systemic causes
        - Fair and transparent outputs
        - Non-discriminatory practices
        - Privacy protection of individuals
    - Ensuring that if an AI/ML model or dataset is found to be actively causing harm after deployment, adjusting or removing (retracting) the result and publicly notifying users that the system is deprecated.

- Earth, environmental, and space scientists will ensure that AI/ML systems developed for Earth, Geospace, Space and and related sciences avoid harm throughout the AI/ML lifecycle by ensuring that:

    - The development team is diverse, including but not limited to members of the communities where the model will be deployed or otherwise impact
    - Training, testing, and all other data critical to the development or assessment of the model is thoroughly documented and vetted for potential biases including computational, human, and systemic biases

Draft

- Potential risks and benefits of AI/ML are identified, and a plan is developed to address the risks.
- Relevant parties are clearly identified, and the risks and mitigation plan are shared publicly.

Module 3 Use Cases and Illustrative Examples

The below examples try and describe situations where bias can start to leak into the lifecycle of AI Systems:

- (dataset bias) In situ or remote observations used for training data that do not cover the full spectrum of social-economical conditions
    - E.g. comparing city districts/regions/countries to each other might not come with the full same spectrum data used to train a model thus leaking bias into the final outcomes.
- (dataset bias) “For example, I have traced algorithmic-driven water development projects in the U.S. southwest dating back a century and have uncovered the explicit ways in which algorithmic frameworks contribute to the settler colonial function and environmental racism of water policy in the region. This is to say that the disavowal of Native American water rights is literally encoded in the technical function of U.S. state-run automated decision systems, many of which grew out of resource capture and allocation projects.” Source
- (model bias) Setting thresholds and model tuning based on historical/agreed rules of thumb where that history is dominated by one segment of the community.
    - Similar to the issue with seatbelts and crush dummies, based on a certain height and a male physique, there is a similar problem in substorms definition. Substorms - hard to characterize, people have used a long time “you know it if you see it”; many different ways to define it. Those that are chosen are from the 1970s - 1990s led by senior white men, instead of younger generations who have used to define it more systematically.
- (model bias) Model evaluation and selection
    - Reproducibility crisis: In a recent talk by Arvind Narayanan and others, there is an ongoing debate on the difficulties of model evaluation. https://twitter.com/random_walker/status/1542879661331345408
- (deployment bias)
    - Cost of redeployment to address biases in light of e.g. new datasets is prohibitive and thus doesn't get done. (from the researcher on a time sensitive grant to a commercial company with operational funding constraints)
- (general coverage of bias and other ways AI can go wrong)

Draft

- Data collected for geosciences often suffers from a variety of biases, including data rarity, skew in measurements and instruments, humans causing adversarial issues in the data and more. The bias impacts the model throughout the lifecycle from development to deployment. Reference: McGovern, A., Ebert-Uphoff, I., Gagne, D., & Bostrom, A. (2022). Why we need to focus on developing ethical, responsible, and trustworthy artificial intelligence approaches for environmental science. Environmental Data Science, 1, E6. doi:10.1017/eds.2022.5

Module 3 FAQs

- What does the chief AI ethics officer do?
    - Provide strategic guidance across professional organizations
    - Interface with funding agencies
    - Facilitate and develop leading practices for responsible conduct of AI/ML research
- What do we do if we identify that our model is causing harm or a dataset we have released has bias?
    - Amend any published papers
    - Add disclaimer to data, products, and software
    - Notify the chief ethics officer if the work is published in AGU, notify the funding agency as appropriate, plus your home institution as appropriate
- What happens if we ran out funding but an issue has been identified?
    - See answer to having identified harm
    - In addition: Notify the funding agency and users about the issue.
- What can funding agencies do to help mitigate harm from AI?
    - We recommend funding agencies facilitate addressing any issues of AI risk and harm throughout the AI system lifecycle.
    - We recommend funding agencies set aside a pool of money set to redress any issues, thus issues can be addressed even if funding has finished

Module 4: Trust in AI/ML

Module 4 Focus

Issues related to the complexities of "trust" and AI/ML systems

Module 4 Key Points

Trust in AI/ML is not something we can prescribe or guarantee, but there are ways we can work to increase the likelihood AI/ML models, systems, and developers are perceived as trustworthy. Trust in AI/ML is context-dependent and can be influenced by factors across the entire AI/ML lifecycle: We need to consider trust from the questions we ask, the data we are using, and the models we develop to how the output is communicated, interpreted, and used.

Building trust in AI/ML systems requires open and transparent research (to the extent feasible). We need to communicate and quantify uncertainty, be able to explain what models do and do not do, and communicate successes and failures. Taking into account multiple perspectives, especially those of potential users, in AI/ML research, development, and deployment will increase the likelihood that the AI/ML systems are trusted. There are broader dimensions of trust in technology and trust in science that underlie trust-development with AI/ML systems.

Draft

Module 4 Learning Objectives

- Understanding that trust and trustworthiness are subjective and perceptual, yet part of established value systems in society.
- Appreciating that trust in AI/ML systems is highly dependent on the context surrounding the system and the potential trustor.
- Developing relationships with potential users and affected communities with the aim of developing trust.

Module 4 Vision

To incentivize and provide infrastructure for co-developing trust throughout the entire life cycle of scientific endeavors that rely on AI/ML.

Module 4 Definitions

Trust: The willingness to assume risk by relying on or believing in the actions of another party (AI2ES, 2022).

Module 4 Principles

Trust

- ❖ **Foster equity and engaging relationships** across stakeholders in all phases of the AI/ML research life cycle.
- ❖ **Provide open and direct communications** with all stakeholders associated with the AI/ML research, including knowns and unknowns, strengths, and limitations.
- ❖ **Acknowledge and appreciate the context** for the research, including how the context impacts the AI/ML research and how the research impacts the context.
- ❖ **Engage in interactive co-development** to learn and adapt the AI/ML research design and methods.
- ❖ **Emphasize knowledge transfer** among the research team, users, and affected communities through education, training, and co-learning.

Additional supporting information on Module 4 principles:

- **Equitable and engaging relationships**: Building trust requires building and maintaining equitable relationships among all involved with and with those potentially impacted by the research at hand. This relationship building will require a strong emphasis on engagement among these groups.
- **Open and direct communication**: Trust will also require open and direct communication with all stakeholders. This involved communicating the history of the field and the state of current efforts. What are the knowns and unknowns? What are the strengths and weaknesses? This transparency is key for setting expectations and facilitating strong user-AI teams.
- **Acknowledgement and appreciation of context**: Context comes up in many different ways throughout the research and operational processes. Knowing and appreciating the challenges and opportunities this context will generate and being ready to work with it will help make more useful and trusted end products.
- **Iterative and flexible codevelopment over time**: Together, the above principles demand an iterative and flexible codevelopment process that gives space for changes over time for AI to be trusted by end users.
- **Emphasize knowledge transfer among the research team, users, affected communities**. Education, training, and learning from one another are key foundations for establishing trust.

Draft

Module 4 Responsibilities/Leading Practices

- **Follow leading practices for AI/ML development and reporting while also being transparent about this process and making the technical components explainable and FAIR (Findable, Accessible, Interoperable, Reusable).** This will involve adhering to the ethics code principles and making sure that you are communicating and explaining them effectively to all stakeholders.
- **The research team engages stakeholders throughout the entire research process**: This will involve engaging with communities and end users when defining problems, collecting and using data, model design and development, communicating the results and uncertainties. This also involves taking an interactive approach to co-development and relationship building examining both the data inputs and outputs.
- **Have a multi-way conversation about the context of the problem, the model, and its intended applications.** This will involve following the CARE principles (Collective Benefit, Authority to Control, Responsibility, Ethics) and making sure there is knowledge transfer throughout the entire research and stakeholder team.
- **Communicate often and openly within the research team, with end users and stakeholders, and with communities who are potentially affected by your research.** This will require finding shared understandings and values for these conversations. Use relatable and approachable examples that can build on past context, history, successes and failures of AI. This will involve communicating uncertainties, failure modes, and risks associated with the research.

Draft

Module 4 Use Cases/Illustrative Examples

- As researchers we tend to want a “litmus paper” for our models and work - is this *good* or *bad* AI/ML? If it’s *bad*, what do we need to do to make it *good*? In the case of AI/ML trust, there are no guarantees for “making it good” or making people trust your work. But, there are leading practices for establishing the relationships and understandings that *may* facilitate trust.
- For example, say you have a model that predicts the need to evacuate before a hurricane in a given neighborhood. If you live in this neighborhood and get an alert on your phone saying you need to evacuate your home because an AI model says so, would you? Most of us would not trust that information alone. But say you get a notification from the National Weather Service that suggests the same thing? What about your local TV meteorologist or your neighbor? Each of

these sources are different but could all rely on an AI model. This shows how contextual and relational trust in AI is, as well as how important the principles and values above are.

Module 4 FAQs

- Why are we using the word trust?
- How is AI/ML similar to and different from other science issues?
- What applications of AI/ML do *we* as a research community trust AI/ML to do alone? How do we see humans and AI/ML models working together?
- What and who are we asking people to trust? AI/ML models? Developers? The interpreters of AI/ML output?
- How do we address changes in systems over time?

See also: Guidelines on reporting on AI

Draft

Module 5: Outreach, Training, and Leading Practices

Module 5 Focus

Ensure researchers, practitioners, funders, and the broader AI/ML community have awareness, understanding, and access to training for ethical use of AI/ML.

Module 5 Key Points

Ethical AI/ML practices are essential for high-quality science and positive public impact. Increasing awareness of ethical AI/ML and advocating for its inclusion in all AI/ML work, must be a central tenet of any work by the data science community.

Adoption of ethical AI/ML practices requires a deliberate action on behalf of the researchers and others relevant to the research. Training and access to resources enables the development of these essential skills. Professional societies must commit to providing access to resources and training, and advocating for researchers' time to learn these practices and develop curricula to train the next generation.
Resources are not "one size fits all;" a broad, inclusive community with a wide variety of activities requires a commensurate breadth of training and educational materials. A modular approach to training materials is recommended so that materials can be combined in multiple ways. The training needs vary across early-career, mid-career and more senior researchers, and the time to participate in training and development is a key factor. A "leader as teacher" model is recommended where Principal Investigators (PIs) and mentors can bring modular material to research teams on a timely basis. "Pre-mortems" and post-mortems are recommended to anticipate what might go wrong in the planning of research involving AI/ML and subsequently to learn from outcomes.

Module 5 Learning Objectives

- Ensuring that early career, mid-career and senior researchers employing AI/ML methods have the knowledge, skills and expertise to mitigate bias, risk, and harm.
- Building awareness and capability to include in the research process representatives from vulnerable populations and others at risk from the use of AI/ML methods.

Draft

Module 5 Vision

The implementation of ethical use of AI/ML requires an awareness of the concepts, an understanding of the practices, and access to training resources. AI/ML work requires the full participation of the broader community of practice, including ethicists and humanists as well as the public, to ensure contributions are diverse, inclusive and comprehensive. To realize this vision, practitioners require the skills and knowledge to implement Ethical AI/ML and evaluate their efforts from an Ethical AI/ML standpoint.

Module 5 Definitions

- Open science (partial list)
    - UNESCO Open Science Recommendation
    - NASA Transform to Open Science (TOPS)
    - NSF Open Science Alliance
    - NSF FAIR and Open Science (FAIROS) Research Coordination Network Investment
- Principles
    - The FAIR Guiding Principles for scientific data management and stewardship
    - The CARE Principles for Indigenous Data Governance
    - The TRUST Principles for digital repositories

Module 5 Principles

Training

❖ **Provide training, resources, and support** for AI/ML Ethics to all researchers and institutions.

❖ **Include the principles, importance, and benefits** to both science and humanity in all training and resources for AI/ML Ethics.

Outreach

❖ **Make available the resources and expertise to support training and resources for AI/ML ethics** to all researchers and stakeholders through scientific societies, institutions, and other organizations.

Leading Practices

❖ **Manage and update training and resources for AI/ML Ethics** to ensure the current state of practice.

Additional supporting information on Module 5 principles:

- Ethical AI/ML is a non-optional and fundamental part of AI/ML research
- Practitioners of AI/ML should be aware of: 1) the principles of Ethical AI/ML, 2) why they are important, 3) how Ethical AI/ML benefits both science and humanity
- Training and access to resources to understand and apply ethical AI/ML are necessary to achieve this. [though we may not be providing these directly]
- There are a broad range of constituencies, and resources and training materials should be responsive to the needs of the different constituencies
- Ethical AI/ML is not a goal or an end result; it provides a set of principles to guide research. As such, training and outreach resources must reflect the evolving state of Ethical AI/ML.

Module 5 Responsibilities/Leading Practices

- Ethical AI/ML should mitigate both the potential for negative impacts on people and on the quality of the science
- Communication of the principles and practices of Ethical AI/ML to all constituents (outreach)
- Access to training resources so practitioners can perform ethical AI/ML research and report results consistent with these principles
- Ensure inclusivity/comprehensiveness of community resources
- Work to identify resources and tools that facilitate the adoption and inclusion of Ethical AI for all constituencies using AI/ML.
- Promote the inclusion of Ethical AI/ML in all aspects of AI/ML training, outreach, discussions and publications.
- Develop and provide considerations on how to use the framework for self-evaluation with consistent application to the intent of the principle.
- Ensure that Ethical AI/ML is included in all training, outreach, and general discussions of AI/ML. Promote Ethical AI/ML as integral to AI/ML practice.
- Work to replace the Data Science lifecycle with an Ethical Data Science Lifecycle.

Module 5 Use Cases/Illustrative Examples

- A researcher using a publicly available dataset uses a model they obtained from an open source repository. The model produces a result that is somewhat controversial. The authors want to ensure that the result is valid before publication. By learning the Ethical AI/ML practices of interpretability and explainability, the authors can perform additional analysis of the model's performance and results to ensure robustness and validity.

Draft

- A reviewer receives a paper from an editor and is asked to provide an anonymous review. The reviewer is concerned about the provenance and the appropriateness of the data used, and is furthermore concerned that the result may have a negative impact if interpreted incorrectly. What practices can the reviewer recommend to the author to mitigate potential impacts?
- Scientific results that are open/reproducible/ethical can be used as a training example of how to evaluate/audit results as a third party. Can also train authors on how to produce papers that facilitate this.
- “AI/ML Fails” (i.e. inappropriate, faulty, or reckless use of AI/ML) cause negative impacts and erode trust in AI/ML practices overall. This can be turned into a beneficial learning experience by examining high-profile “AI Fails” and demonstrating how practices of Ethical AI could have prevented them.
- Potential use case: NASA Transform to Open Science (TOPS) trainings - could add one on use of ethical AI/ML (https://github.com/learnopenscience)
- Hugging Face community, training Hugging Face – The AI/ML community building the future.
- FastAI/Kaggle fast.ai · Making neural nets uncool again (practical ethics)
- ADSA’s forthcoming Data Science Ethos Lifecycle tool will gather use cases and present them to a researcher or learner to understand the societal and ethical implications of the work. (see the paper)

Module 5 FAQs

- How do we ensure that all Earth, environmental, and space science meeting sessions, topical meetings, town halls etc. on AI follow the principles of Ethical AI/ML?
- How do we ensure that all relevant constituencies using AI/ML are aware of Ethical AI/ML practices?
- How do we offer access to Ethical AI/ML? Who does the training? At what level? (What is ethical AI/ML versus How to apply and practice ethical AI/ML - h/t Barbara)
- What are the indicators (antennas) for signs of success (evaluation of the community’s progress)?

Additional Creative Ideas:
- Gather use cases discreetly (leverage ADSA, AGU community)
- Ignoble prize for AI/ML models could generate compelling use cases
- Incentivize team reviews of manuscripts

Draft

Module 6: Participatory Methods and Domain Expertise

Module 6 Focus

Inclusive research design and conduct with AI/ML – ensuring voice for diverse communities, domain expertise, and context

Module 6 Key Points

A key guiding principle comes from the disability movement: "Nothing about us without us." No research should be conducted that impacts individuals and groups in society without their consent. This requires the formation of advisory groups, the utilization of stakeholder and rightholder mapping surveys, the democratic selection of community representatives, and other mechanisms for input.

A key practice involves the co-production of knowledge. This is valuable with stakeholders and essential with what are termed "rights holders" such as first nations. This input is important in the planning and conduct of research, as well as on a continuing basis after the research is complete to address continuing implications of the research.

Open science principles are key, even if not all data can or should be open (e.g., asking researchers to publish data, NASA Information Policy NASA SPD-41).

Extra resources are needed for participatory practices. Institutional Review Boards (IRBs) need to be informed about participatory methods, which may involve a balancing of benefits and risks associated with the use of AI/ML (not just the elimination of risk). Note that participatory methods vary with scale, from AI/ML applications that are local, regional, national, and international.

Module 6 Learning Objectives

- Appreciate the value and impact of participatory methods in AI/ML research.
- Identify ways to ensure domain expertise and integration across relevant fields and disciplines.

Module 6 Vision

Draft

Ensuring participatory design as the leading practice of AI/ML research and applications to ensure the development is inclusive of users and affected groups from the beginning. ("Nothing about us without us").

Module 6 Definitions

- Participatory - engaging people who will be affected from the very beginning of the work, and through all phases of the work
- Inclusive
- Stakeholders:
- Strong need for a distinction between equality and equity

Module 6 Principles

Participatory Methods

- ❖ **Ensure voluntary and continuing consent** from individuals or communities who may be impacted by AI/ML research.
- ❖ Respect the **autonomy of associated stakeholders and ensure representation in decision-making.**
- ❖ **Research teams should be designed with inclusion and diversity** in mind at all stages, from conceptual design, data collection, method development, analysis, publication, and deployment.
- ❖ **Research teams should intentionally search for gaps in representation** to ensure all end-users and impacted groups are represented.

Domain Expertise

- ❖ **Diversity is part of domain expertise,** reflected in the team design, community participation, project design, and data collection and analysis

Additional supporting information on Module 6 principles:

- "No" research impacting a group without their continuous consent maintaining their autonomy and representation at decision-making level
    - Under what condition, may one deviate from this principle?
- Research teams should be designed with inclusion and diversity in mind at all stages, from conceptual design, data collection, method development, analysis, publication, and deployment.
    - Diversity is part of the team design, community participation, project design, and data collection and analysis
    - Who gets a seat at the table and who is included in the conversations about compute, education, research/development/deployment

Draft

participation points to the importance of public engagement in research design?

- Research teams should intentionally search for gaps in community representation to ensure all end-users and impacted groups are represented.

Module 6 Responsibilities and Leading Practices

Leading Practices:

- Knowledge co-production: engage stakeholders including affected groups in all research stages from designing questions to validation and deployment. Relevant stakeholder community groups who can lead and engage stakeholders should be identified which can continue to engage the stakeholder groups after the research team may have broken up.
- Enact an actionable framework that enable users and affected groups to provide feedback regarding potential risks and harms of the research input at all stages
- During the research design phase, implementing a similar process like Institutional Review Board (IRB) process to ensure the design is inclusive
- Regarding data collection and usage, research team should follow the leading practice in data sovereignty and governance (i.e., CARE principles)
- Maintain a transparent development and reporting framework to allow stakeholders including potentially affected groups to monitor the process and provide real time feedback.
- Data ownership and usage rights: during data reuse research teams should also engage the data owner and affected communities.
- During the development process, choose the most appropriate AI methods for the applications. If the general AI model does not fit the purpose, the research team should actively work with domain experts and end users to develop new AI models (e.g., Physics-aware AI, Geo-statistics aware AI).

Responsibilities:

- Throughout the lifecycle, various actors/participants have inclusivity responsibilities
    - Developer/researcher:
        - To be alert and protect against bias and exclusion.
        - Actively question which groups are not included and should be.
    - Data owners and stewards: to ensure regular permission and consent from impacted groups and maintain a record of interactions.
    - Professional societies: providing and implementing guidelines that promote participatory design in the research and society journals

Draft

    - Auditor/credentialing organization (objective third party): review and audit research framework to minimize and mitigate potential risk of the research
    - Users: engage in the research development process to provide real time feedback to the research team
    - Policy makers:
    - Procurer/funder: require inclusive development and regular reporting during the research process

Module 6 Use Cases and Illustrative Examples

- OECD Large language Models inclusion of more than English language in development of language technologies
- Predicting What We Breathe (http://airquality.lacity.org), a NASA grant with the City of Los Angeles, was designed with residents of neighborhoods impacted by environmental injustice, has ongoing community engagement, team members from those neighborhoods, and distributes sensors to residents to become community scientists
- Voice Assistant on use of non-traditional English vernacular/accents
- Lacuna Fund for inclusive datasets for agriculture in Africa - https://lacunafund.org/datasets/agriculture/

Module 6 FAQs

- How can we ensure the research team is diverse and inclusive? What research infrastructure is needed?
- What are the implications of ethics (such as data ownership, sovereignty, or privacy) for open science (e.g., asking researchers to publish data, NASA Information Policy NASA SPD-41a)?
- How is individual data protected?
    - Researchers are responsible for anonymizing the data so that individuals or sensitive data cannot be identified. This includes personally identifiable data, as well as data that identifies structures or locations that the community wants to be anonymous (such as burial sites). Researchers should ask the community during engagement what they consider sensitive and document those responses.
- How may one (ethically) reuse data from another researcher? What restrictions are implied by ethics?
    - Yes, but you must adhere to the norms and sensitivities identified by the researcher in their community engagement. If the intended use is different from the original use, then the community should be re-engaged.

Draft

Module 7: Considerations for Organizations, Institutions, Publishers, Societies, and Funders

Module 7 Focus

Organizations have a responsibility to define their approach to establishing and administering AI/ML ethics policies, including codes of conduct, principles, reporting methods, resolution processes, and other categories; values articulation and governance design at levels above the individual and including fostering a culture around ethical AI/ML.

Module 7 Key Points

Professional societies, universities, federal labs, industry labs, publishers, funders, and other organizations and institutional actors have a leadership role when it comes to AI/ML ethics. AI and ML technologies are developing at rapid rates, calling for flexible and adaptive approaches by these organizations and institutions.

Community-driven principles require sponsorship and hosting of forums, town halls, and other engagement mechanisms by leading organizations and societies. This is key to surfacing and considering current practices and making necessary updates as practices evolve. There will be tensions that surface, such as the tensions between transparency and privacy, with institutional leaders playing key roles in naming these tensions and fostering constructive dialogue about the tensions.

Professional societies and other publishers have a particular responsibility to promulgate policies and practices relevant to the publication of research involving AI/ML models and algorithms. Federal agencies in the United States, European Union, and other settings operate under directives to ensure the ethical use of AI/ML, which can be a model for others. While industry typically treats aspects of AI/ML as proprietary, there are community liability issues that point to the carving out of “pre-competitive” spaces in which AI/ML practices, applications, and risks are shared and evaluated.

Module 7 Learning Objectives

- Identify opportunities and responsibilities within organizations, societies, and communities to advance AI/ML ethics.
- Explore how best to influence the relevant fields and disciplines utilizing AI/ML in research

Draft

Module 7 Vision

To facilitate the creation of timely and iterative mechanisms and approaches, with respect to AI/ML ethics, to guide the organization or society AGU community to foster positive outcomes, and mitigate risks, and provide means to resolution or reconciliation.

Module 7 Definitions

- Mindfulness – a choice and an unfolding; includes personal agency on the part of researchers and others to shape the organizations, societies, and other communities of which they are members.
- Encourage responsible innovation where research is designed and delivered for the benefit of all -
    - The processes for how we deliberate together as guidance for how we act together
    - A process of anticipating, reflecting, engaging, and acting that promotes socially desirable creativity and opportunity (https://www.ukri.org/about-us/epsrc/our-policies-and-standards/framework-for-responsible-innovation/) Here is a supporting quote from the Australian context:
        - ”Responsible innovation is where researchers consciously and critically assess the potential risks, benefits and uncertainties of the future science and technology they are developing. In doing so, this aims to deliver as a way of addressing those challenges with a view to ensuring socially and ethically responsible science and technology that is designed and delivered for the benefit of all Australians. This program of research assesses the potential risks, benefits and uncertainties of future science and technology” (From Data61/CSIRO - Responsible Innovation Platform)

Module 7 Principles

Organizations and Institutions

- ❖ **Align new and existing programs objectives and approaches** across the AI/ML Ethics Modules.
- ❖ **Partner with multiple organizations** to help broaden awareness, education, adoption, and other engagement.
- ❖ **Include ethical AI/ML into courses and other ethical training.**
- ❖ **Include ethical AI/ML into grant processes**

Draft

Societies and Communities

- ❖ **Provide workshops and education for society members** on the AI/ML Ethical Framework.
- ❖ **Collectively provide governance of this AI/ML ethics framework;** Support development and updates **to** leading practices related to the AI/ML Ethics Framework.
- ❖ **Measure the effectiveness of the efforts** specific to implementing the AI/ML Ethical Framework.
- ❖ **Adopt the AI/ML framework** into the organization's ethical guidance.
- ❖ **Promote the importance and adoption of the AI/ML Ethical Framework** in relevant communities.
- ❖ **Ensure all affected communities are part of the development and updates** to the AI/ML Ethics Framework.

Funders

- ❖ **Include the AI/ML Ethical Framework** in expectations and guidance for grants, including in data management and sharing plans. Encourage broader outreach plans to address ethical AI/ML as appropriate.
- ❖ **Include experts in AI/ML ethics as reviewers and panelists** for AI/ML grants. Provide training from program officers around ethical AI/ML.
- ❖ **Support continued governance** of this framework.

Publishers

- ❖ **Develop reviewer and editor guidance** for handling AI/ML papers, including on inclusion of appropriate reviewers; inform editors and staff of expectations.
- ❖ **Develop author guidelines** consistent with the Ethical Framework, including around FAIR data and software, recognizing contributions, reporting uncertainties, and methods sections.
- ❖ **Follow leading practices** regarding data and software citations, including guidance for authors.

Additional supporting information on Module 7 principles:

- Establish a process that encourages and facilitates conversations
    - Consider communication vs. control
- Iterate - start with "timely good enough" vs. "late & perfect" or "rapid & wrong"
    - 'Iterate': a process of responding to feedback (e.g., from stakeholders, from critical internal reflection within the organization)
    - Criteria along which you assess during iteration - the ethical checklist/risk assessment - dynamic, evolving criteria, instead - actively seeking out new

    voices (identify marginalized) and sensing and accommodating changing situations
  - Appreciate and make explicit value systems within situational contexts: for example, choices/actions taken in “emergency” vs “Business as Usual”; prototype (beta) vs deploy (scale)
- Beyond the standard AI ethics considerations… Openness, honesty, inclusion, flexibility, evolving, adaptability, kind/humane/thoughtful, acknowledgement of the human experience / human context, resilience, choice for mindfulness, accountable, explainable, innovative
- Balance philosophical exploration with practicalities
  - Engage different communities with different levels of abstraction or concreteness
- Work values and principles in parallel with concrete questions, rules of thumb, etc. for practitioners to consider, etc.
- Governance
  - Feedback that funnels into update process
  - Ongoing management
- Support to organization members -- before / during / after

Module 7 Responsibilities and Leading Practices

- Connect with policy makers to embed AI/ML ethics as part of their processes and conversations.
- Encourage publishers to promote a review of scholarly submissions for alignment with these principles.
- Explicitly encourage wide diversity in scholarly society ethics leadership, alignment, and guidance.
- Encourage AI ethics conversations across the broad stakeholder community to elicit principles, etc.
- Introduce new concepts such as mindfulness, agency and ‘otherness’ (this concept includes people and environment).
- Acknowledge and value that some principles may involve judgment, intangibles, and a variety of choices while others may be clear and concrete.

Module 7 Use Cases/Illustrative Examples

- Scientific societies and other organizations that have science integrity guidance and/or scientific code of conduct policies would benefit from considering a future update using the AI/ML Ethics Principles and Responsibilities to help support their researchers.

Draft

- Funders considering AI/ML related grants could value proposals that include using an AI/ML ethical framework for designing and managing their project.
- Publishers with journals receiving AI/ML related research could provide review guidance to value the use of a relevant AI/ML ethical framework in the research approach.

Module 7 FAQs

- How do we form timely, iterative mechanisms & approaches to guide organizations and societies regarding AI ethics to foster positive outcomes and mitigate systemic risks? (see Responsibilities/Leading Practices)
- How do we help communities understand how to have AI ethics conversations using listen first? Community centric, ethnographic approaches

Draft

References (in development)

- WEF Inclusive Research White Paper
- Ada Lovelace Institute. (2021). "Participatory data stewardship: A framework for involving people in the use of data." Ada Lovelace Institute.
- Delgado, F., Yang, S., Madaio, M., & Yang, Q. (2021). "Stakeholder Participation in AI: Beyond 'Add Diverse Stakeholders and Stir.'"
- Sloane, M., Moss, E., Awomolo, O. & Forlano, L. (2020). "Participation is not a design fix for machine learning."
- Leslie, D., Katell, M., Aitken, M., Singh, J., Briggs, M., Powell, R., ... & Burr, C. (2022). "Data Justice in Practice: A Guide for Developers."
- Saulnier, L., Karamcheti, S., Laurençon, H., Tronchon, L., Wang, T., Sanh, V., Singh, A., Pistilli, G., Luccioni, S., Jernite, Y., Mitchell, M. & Kiela, D. (2022). "Putting Ethical Principles at the Core of the Research Lifecycle."
- "Citizen science" biases in populations (SciStarter): https://academic.oup.com/bioscience/article/72/7/651/6605713?login=false
- Citizen science biases in populations (Zooniverse): http://eprints.lse.ac.uk/81320/1/Woodcock_Doing%20Good%20Online%20.pdf
- University of Washington Tech Policy Lab "Diverse Voices" project
- Federal Citizen Science Toolkit
- National Academies, Learning Through Citizen Science, Enhancing Opportunies by Design https://nap.nationalacademies.org/read/25183/chapter/1
- Documenting Data Production Processes: A Participatory Approach for Data Work, [2207.04958v2] Documenting Data Production Processes: A Participatory Approach for Data Work (arxiv.org)
- National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303.
- NASA Science Mission Directorate (SMD), (2022, September 26) SMD Policy Document SPD-41A, https://science.nasa.gov/science-red/s3fs-public/atoms/files/SMD-information-policy-SPD-41a.pdf
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots:Can Language Models Be Too Big?. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency. ACM. https://doi.org/10.1145/3442188.3445922
- UNESCO, (2021) UNESCO Recommendation on Open Science, https://unesdoc.unesco.org/ark:/48223/pf0000379949
- NASA Transform to Open Science (TOPS) Mission, https://science.nasa.gov/open-science/transform-to-open-science
- NSF Open Science Alliance, https://openscialliance.github.io/
- NSF FAIR and Open Science RCN Investment (2022), https://www.nsf.gov/pubs/2022/nsf22553/nsf22553.htm
- NSF FAIR and Open Science cohort ARL announcement, https://www.arl.org/news/arl-applauds-nsf-open-science-investment/

Draft

- Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
- Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., … Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. Data Science Journal, 19(1), 43. DOI: http://doi.org/10.5334/dsj-2020-043
- Lin, D., Crabtree, J., Dillo, I. et al. The TRUST Principles for digital repositories. Sci Data 7, 144 (2020). https://doi.org/10.1038/s41597-020-0486-7

Draft

Appendix A: AI/ML Ethics “Pulse” Stakeholder Survey

In preparing the AI/ML Ethics Modules, a diverse set of researchers, policy makers, students, industry representatives, and others were surveyed to more fully understand the broader context. The results from this surrey are summarized here.

Introduction

Across scientific domains, Artificial Intelligence (AI) and Machine Learning (ML) are playing increasingly important roles in research. Existing standards for reproducibility and ethics in research can be challenged by AI and ML. There are concerns in society about bias and other adverse impacts of AI and ML. In this context, considerations for AI/ML ethics in research is needed.

This report is based on a “stakeholder pulse survey” of researchers, administrators, and others in order to provide situational awareness that can inform the development of AI/ML ethics. This report is designed to indicate where stakeholders are aligned, where views are particularly intense, and where there is variance in their views. Both qualitative and quantitative data are provided, each of which informs dialogue in different ways.

This is part of a 2022 project convened by the American Geophysical Union (AGU), funded by the National Aeronautic and Space Administration (NASA), and this portion has been conducted by WayMark Analytics.

Overview

There is wide support for 1) having clear ethical standards and guidelines for the use of AI/ML in research, as well as for ensuring 2) explainability/interpretability and 3) replicability when AI/ML is used in research. These are three of the sixteen indicator issues that were selected by leading experts, covering many aspects of AI/ML ethics. At the same time, most of the indicator issues are major "pain points" – rated as very important and also as very difficult to do by more than half of the respondents. Importantly, there is very little support for researchers using AI/ML in any way they choose – without attention to ethical standards or guidelines. There are minority views on many of the indicator issues, indicating a need for engagement and dialogue.

A set of qualitative "must haves" involve well-conducted research, conscious of bias, yet there are considerable barriers in the quality of the training data, the lack of knowledge and skills in addressing bias, the lack of governing bodies, and other factors. Qualitative success visions and "anything else?" comments are extensive, poignant, and compelling.

Although the report is comprehensive, these should still be treated as preliminary findings designed to generate dialogue, point to needed additional confirmation, and then action.

Meet the Respondents (n=118)

What is your primary role when it comes to the use of Artificial Intelligence (AI) and Machine Learning (ML) in research? Please answer all questions from this perspective.

- Researcher who uses AI/ML in research -- 39.8% (n=47)
- Researcher who does not use AI/ML in research, but is knowledgeable about the technologies -- 26.3% (n=31)
- Researcher who does not use AI/ML in research & is not knowledgeable about the technologies -- 9.3% (n=11)
- Research Computing and Data Professional -- 22.9% (n=27)
- Student (graduate or undergraduate) -- 10.2% (n=12)
- Administrator/leader in university -- 6.8% (n=8)
- Administrator/leader in government -- 7.6% (n=9)
- Administrator/leader in government contractor -- 5.1% (n=6)
- Administrator/leader in commercial organization -- 2.5% (n=3)
- Administrator/leader in not-for-profit organization -- 1.7% (n=2)
- Other - Write In -- 14.4% (n=17)

What is your general level of knowledge of and experience with Artificial Intelligence (AI) and Machine Learning (ML)

- Limited or no knowledge -- 1.7% (n=2)
- Awareness of how AI and ML works, but no direct experience -- 28.0% (n=33)

Draft

Some direct experience using AI and ML in research or other applications -- 39.8% (n=47)

Extensive direct experience using AI and ML in research or other applications -- 19.5% (n=23)

Expert able to lead theory development and innovation with AI and ML in research and other applications -- 9.3% (n=11)

What is your general level of knowledge of and experience with ethics in research

Limited or no knowledge 3.4% (n=4)

Awareness of the role of ethics in research, but no direct experience 36.2% (n=42)

Some direct experience applying ethical standards to decisions and actions in research projects 39.7% (n=46)

Extensive direct experience applying ethical standards to decisions and actions in research projects 15.5% (n=18)

Expert able to lead theory development and innovation applying ethical standards to decisions and actions in research projects 5.2% (n=6)

Which of the professional societies participating in this research are you a member of? select all that apply

Association for Computing Machinery (ACM) -- 11.9% (n=14)

American Geophysical Union (AGU) -- 55.1% (n=65)

American Meteorological Society (AMS) -- 26.3% (n=31)

American Astronomical Society (AAS) -- 11.0% (n=13)

Geological Society of America (GSA) -- 3.4% (n=4)

American Association for the Advancement of Science (AAAS) -- 11.0% (n=13)

Institute of Electrical and Electronics Engineers (IEEE) -- 14.4% (n=17)

None of the above -- 17.8% (n=21)

Please indicate your years of experience

1 year or less	1.7% (n=2)
2-4 years	4.2% (n=5)
5-10 years	16.1% (n=19)
11-20 years	21.2% (n=25)
21-30 years	25.4% (n=30)
Over 30 years	29.7% (n=35)
It’s complicated	1.7% (n=2)

What is your gender identity?

Woman	25.4% (n=30)
Man	66.1% (n=78)
Non-binary, two-spirit, gender queer, or agender	4.2% (n=5)
Prefer not to answer	4.2% (n=5)

Draft

Pulse Results for “Indicator” Issues

Scale: 0 1 2 3 4 5 6 7 8 9 10 (Not Important – Very Important) | Scale: 0 1 2 3 4 5 6 7 8 9 10 (Very Difficult – Very Easy)

Important or Very Important *(7-10)*	Part 1: Establishing/implementing ethical standards	Difficult or Very Difficult *(0-3)*
95%	Establishing **clear ethical standards and guidelines** for the use of AI/ML in research.	68%
90%	Researchers having **sufficient knowledge of what AI/ML algorithms are designed to do** in research.	61%
88%	Implementing/ensuring **compliance with ethical standards and guidelines** for the use of AI/ML in research.	75%
18%	Researchers being able to use AI/ML in any way they find appropriate, **without being limited by any ethical standards or guidelines.**	25%

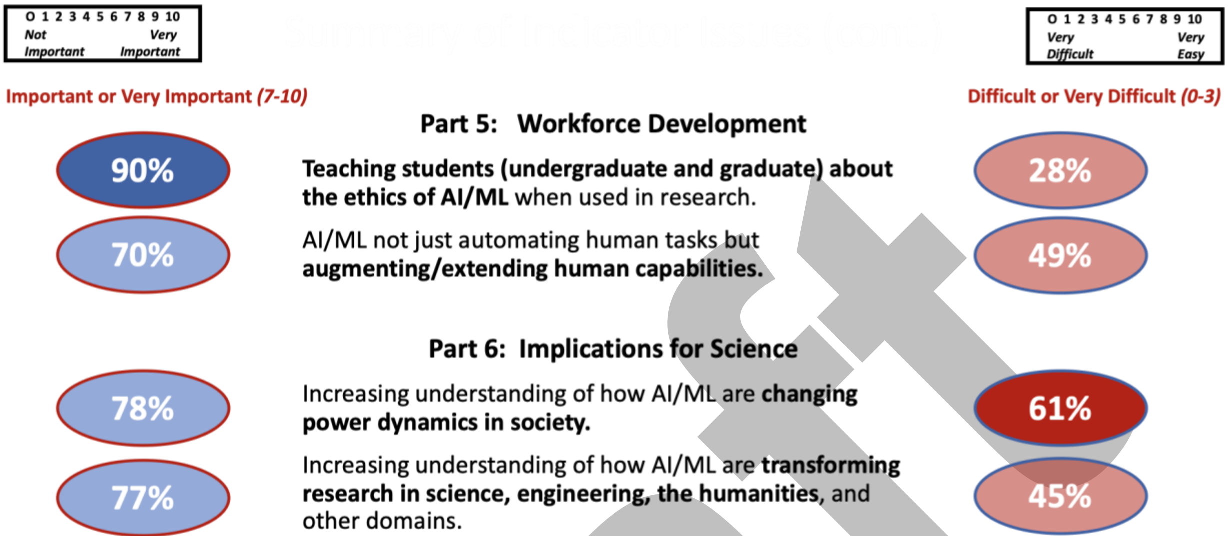
Important or Very Important *(7-10)*	Part 2: Interested parties/stakeholders	Difficult or Very Difficult *(0-3)*
87%	Knowing **who are the interested parties** likely to be impacted by the use of AI/ML in research.	59%
74%	**Interested parties associated with research involving AI/ML having sufficient knowledge and input** into what the algorithms are designed to do.	78%

Scale: 0 1 2 3 4 5 6 7 8 9 10 (Not Important – Very Important) | Scale: 0 1 2 3 4 5 6 7 8 9 10 (Very Difficult – Very Easy)

Important or Very Important *(7-10)*	Part 3: Potential Bias, Risk, and Harm	Difficult or Very Difficult *(0-3)*
88%	Having/developing tools and methods to **audit AI/ML results for potential biases.**	75%
90%	Having/developing tools and methods **to assess the risks** when it comes to the use of AI/ML in research.	57%
78%	Clarifying **who is responsible for any harm** that results from recommendations or findings based on the use of AI/ML.	69%
84%	Guidance on the use of AI/ML directly or indirectly with **sovereign data in tribal communities and/or with respect to vulnerable populations.**	51%

Important or Very Important *(7-10)*	Part 4: Explainability/replicability	Difficult or Very Difficult *(0-3)*
93%	Ensuring **explainability/interpretability** when AI/ML is used in research.	69%
90%	Ensuring **replicability** in science when AI/ML is used in research.	64%

Draft



Selected quotes from respondents

"AI/ML is not about replacing humans, but about empowering them."

"We must build upon both our successes but also our failures in AI/ML. In some cases, such as chatbots that become racist, the failures are easy to see. However, in many cases when bias is introduced, the failures of AI/ML will be more subtle and harder to see. It is more important than ever for practitioners of AI/ML to be inclusive and reflective on their work."

"Experts in any field simply want to advance their field and ignore ethics. This human tendency is problematic…"

"When machine learns, who possess the knowledge? Who combines that knowledge for further research?"

"Most users who provide code used to analyze data do a bad job of explaining and documenting it."

"AI/ML must not be allowed to result in devaluing human beings by other human beings."

"Industry has overtaken government and most higher learning in sheer capacity; similar circumstances are hard to find in history; the USA despite its rhetoric, is building an environment more similar to modern China than the EU. Dangerous times."

"If there's a big knowledge gap between the scientific understanding and the common understanding of a technology, but the technology is transformational and ubiquitous in daily life, it is important to build trust, ensure transparency, and develop a general basic standard of understanding of how the technology can impact and affect people."

"I am deeply concerned about this doing lasting damage to already vulnerable populations."

"…nothing about us without us (from the accessibility community…"

"Solve ethics issues before it is too late."

Draft

Appendix B: Existing AI and Data Principles and Frameworks

OECD AI Principles

1. **Inclusive growth, sustainable development and well-being:** Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being.
2. **Human-centered values and fairness:**
    a. AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights.
    b. To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.
3. **Transparency and explainability:** AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art:
    a. to foster a general understanding of AI systems,
    b. to make stakeholders aware of their interactions with AI systems, including in the workplace,
    c. to enable those affected by an AI system to understand the outcome, and,
    d. to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.
4. **Robustness, security and safety:**
    a. AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk.
    b. To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art.
    c. AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias.
5. **Accountability:** AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.

Draft

Principles of Trustworthy AI in Government (Executive Order 13960)

1. **Lawful and respectful of our Nation's values**. Agencies shall design, develop, acquire, and use AI in a manner that exhibits due respect for our Nation's values and is consistent with the Constitution and all other applicable laws and policies, including those addressing privacy, civil rights, and civil liberties.
2. **Purposeful and performance-driven.** Agencies shall seek opportunities for designing, developing, acquiring, and using AI, where the benefits of doing so significantly outweigh the risks, and the risks can be assessed and managed.
3. **Accurate, reliable, and effective.** Agencies shall ensure that their application of AI is consistent with the use cases for which that AI was trained, and such use is accurate, reliable, and effective.
4. **Safe, secure, and resilient.** Agencies shall ensure the safety, security, and resiliency of their AI applications, including resilience when confronted with systematic vulnerabilities, adversarial manipulation, and other malicious exploitation.
5. **Understandable**. Agencies shall ensure that the operations and outcomes of their AI applications are sufficiently understandable by subject matter experts, users, and others, as appropriate.
6. **Responsible and traceable**. Agencies shall ensure that human roles and responsibilities are clearly defined, understood, and appropriately assigned for the design, development, acquisition, and use of AI. Agencies shall ensure that AI is used in a manner consistent with these Principles and the purposes for which each use of AI is intended. The design, development, acquisition, and use of AI, as well as relevant inputs and outputs of particular AI applications, should be well documented and traceable, as appropriate and to the extent practicable.
7. **Regularly monitored.** Agencies shall ensure that their AI applications are regularly tested against these Principles. Mechanisms should be maintained to supersede, disengage, or deactivate existing applications of AI that demonstrate performance or outcomes that are inconsistent with their intended use or this order.
8. **Transparent**. Agencies shall be transparent in disclosing relevant information regarding their use of AI to appropriate stakeholders, including the Congress and the public, to the extent practicable and in accordance with applicable laws and policies, including with respect to the protection of privacy and of sensitive law enforcement, national security, and other protected information.
9. **Accountable.** Agencies shall be accountable for implementing and enforcing appropriate safeguards for the proper use and functioning of their applications of AI, and shall monitor, audit, and document compliance with those safeguards. Agencies shall provide appropriate training to all agency personnel responsible for the design, development, acquisition, and use of AI.

Department of Defense Ethical Principles for AI

1. **Responsible**. DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities.
2. **Equitable**. The Department will take deliberate steps to minimize unintended bias in AI capabilities.

Draft

3. **Traceable**. The Department's AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with transparent and auditable methodologies, data sources, and design procedure and documentation.
4. **Reliable**. The Department's AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles.
5. **Governable**. The Department will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.

The Five Safes Framework

1. **Safe data**: data is treated to protect any confidentiality concerns.
2. **Safe projects**: research projects are approved by data owners for the public good.
3. **Safe people**: researchers are trained and authorized to use data safely.
4. **Safe settings**: a SecureLab environment prevents unauthorized use.
5. **Safe outputs**: screened and approved outputs that are non-disclosive

FAIR Principles

1. **Findable**: Metadata and data should be easy to find for both humans and computers.
2. **Accessible**: Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorisation.
3. **Interoperable**: The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.
4. **Reusable**: The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

CARE Principles

1. **Collective benefit:** Data ecosystems shall be designed and function in ways that enable Indigenous Peoples to derive benefit from the data.
2. **Authority to Control:** Indigenous Peoples' rights and interests in Indigenous data must be recognised and their authority to control such data be empowered. Indigenous data governance enables Indigenous Peoples and governing bodies to determine how Indigenous Peoples, as well as Indigenous lands, territories, resources, knowledges and geographical indicators, are represented and identified within data.
3. **Responsibility:** Those working with Indigenous data have a responsibility to share how those data are used to support Indigenous Peoples' self determination and collective benefit. Accountability requires meaningful and openly available evidence of these efforts and the benefits accruing to Indigenous Peoples.
4. **Ethics:** Indigenous Peoples' rights and wellbeing should be the primary concern at all stages of the data life cycle and across the data ecosystem.

NSF AI Institute on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES) has a code of ethics that covers AI as part of the code:

1. When creating AI systems, members will:
    - Ensure that the public good is the central concern during all professional computing work
    - Give comprehensive and thorough evaluations of AI2ES AI algorithms and their impacts, including analysis of possible risks.
    - Recognize and take special care of AI systems that become integrated into the infrastructure of society.
2. Members will create AI systems that will:
    - Avoid harm
    - Protect the Earth and its environment including human and animal welfare.
    - Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing.
    - Be fair and take action not to discriminate.
    - Respect privacy.
    - Honor confidentiality.
    - Avoid creating or reinforcing bias.
    - Uphold high standards of scientific excellence.

Existing Data Protection Regulations

Listed below are GDPR and CCPA principles. Though these were created primarily to address data about individuals, and the rights that individuals have with their data, several of the principles could also be interpreted and applied in the context of open data. Needless to say, if the data does have PII and other information about individuals, then it must conform to GDPR and/or CCPA, wherever those may apply.

The 7 Principles Of EU General Data Protection Regulation (GDPR)

(https://www.privado.ai/post/what-are-the-7-principles-of-gdpr)

1. **Lawfulness, Fairness & Transparency**
    a. **Lawfulness**
        i. **Consent**- if the client provides consent, you can collect their data
        ii. **Contract**- if you are drawing up an agreement with the client and the contract requires you to have their data, (e.g. you need staff data for payroll purposes)
        iii. **Legal obligation**- to process a legal obligation
        iv. **Protection of vital interest**- if the data processing is essential for the survival of the subjects or another individual, for instance, if you need staff data for an emergency medical condition
        v. **Public task**-if the data processing is necessary for a task relating to the public interest
        vi. **Legitimate interest**- if the processing is necessary to carry out a legitimate interest
    b. **Fairness**: Adhering to the promise you made with the subject while collecting the data.
    c. **Transparency**: Notifying the subject about what you will do with the data and who can potentially access the data.

Draft

2. **Purpose Limitation:** data should be used only for the purpose for which it was collected. Else, requires additional consent from the data provider.
3. **Data Minimization:** collect only the minimal amount of data needed for a purpose.
4. **Accuracy:** data stored should be accurate and up to date.
5. **Storage Limitation:** every data item has an expiration date, after which you lose the right to store the data.
6. **Integrity & Confidentiality:** data user is responsible for ensuring integrity and confidentiality of the data.
7. **Accountability:** data user is accountable for its use. Should document and justify each step.

California Consumer Privacy Act (CCPA)

1. **Right to Access:** consumers have a right to access their data
2. **Right to Notice**: data cannot be collected without notification.
3. **Consent**: consumer must consent.
4. **Right to Opt-out**: consumers can say, “no”.
5. **Equality**: service providers must promise not to discriminate against customers, i.e. provide lower quality service if they decided to not provide their data for non-essential purposes, such as marketing needs or similar. In other words, service provides shouldn’t make it difficult for consumers to exercise their right to protect their data.
6. **Right to Deletion:** have the right to be “forgotten”.

Ethics Principles for Access to and Use of Veteran Data (https://www.oit.va.gov/about/ethical-data-use/index.cfm?)

1. The primary goal for use of Veteran data is for the good of Veterans.
2. Veteran data should be used in a manner that ensures equity to Veterans.
3. The sharing of Veteran data should be based on the Veteran’s meaningful choice.
4. Access to and exchange of Veteran data should be transparent and consistent
5. De-identified Veteran data should not be reidentified without authorization.
6. There is an obligation of reciprocity for gains made using Veteran data.
7. All parties are obligated to ensure data security, quality and integrity of Veteran data.
8. Veterans should be able to access their own information.
9. Veterans have the right to request amendments to their own information.

MAKING AUTOMATED SYSTEMS WORK FOR THE AMERICAN PEOPLE

Among the great challenges posed to democracy today is the use of technology, data, and automated systems in ways that threaten the rights of the American public. Too often, these tools are used to limit our opportunities and prevent our access to critical resources or services.

Draft

These problems are well documented. In America and around the world, systems supposed to help with patient care have proven unsafe, ineffective, or biased. Algorithms used in hiring and credit decisions have been found to reflect and reproduce existing unwanted inequities or embed new harmful bias and discrimination. Unchecked social media data collection has been used to threaten people's opportunities, undermine their privacy, or pervasively track their activity—often without their knowledge or consent.

These outcomes are deeply harmful—but they are not inevitable. Automated systems have brought about extraordinary benefits, from technology that helps farmers grow food more efficiently and computers that predict storm paths, to algorithms that can identify diseases in patients. These tools now drive important decisions across sectors, while data is helping to revolutionize global industries. Fueled by the power of American innovation, these tools hold the potential to redefine every part of our society and make life better for everyone.

This important progress must not come at the price of civil rights or democratic values, foundational American principles that President Biden has affirmed as a cornerstone of his Administration. On his first day in office, the President ordered the full Federal government to work to root out inequity, embed fairness in decision-making processes, and affirmatively advance civil rights, equal opportunity, and racial justice in America.[i] The President has spoken forcefully about the urgent challenges posed to democracy today and has regularly called on people of conscience to act to preserve civil rights—including the right to privacy, which he has called "the basis for so many more rights that we have come to take for granted that are ingrained in the fabric of this country."[ii]

To advance President Biden's vision, the White House Office of Science and Technology Policy has identified five principles that should guide the design, use, and deployment of automated systems to protect the American public in the age of artificial intelligence. The Blueprint for an AI Bill of Rights is a guide for a society that protects all people from these threats—and uses technologies in ways that reinforce our highest values. Responding to the experiences of the American public, and informed by insights from researchers, technologists, advocates, journalists, and policymakers, this framework is accompanied by From Principles to Practice—a handbook for anyone seeking to incorporate these protections into policy and practice, including detailed steps toward actualizing these principles in the technological design process. These principles help provide guidance whenever automated systems can meaningfully impact the public's rights, opportunities, or access to critical needs.

From Principles to Practice

Safe and Effective Systems

You should be protected from unsafe or ineffective systems. Automated systems should be developed with consultation from diverse communities, stakeholders, and domain experts to identify concerns, risks, and potential impacts of the system. Systems should undergo pre-deployment testing, risk identification and mitigation, and ongoing monitoring that demonstrate

Draft

they are safe and effective based on their intended use, mitigation of unsafe outcomes including those beyond the intended use, and adherence to domain-specific standards. Outcomes of these protective measures should include the possibility of not deploying the system or removing a system from use. Automated systems should not be designed with an intent or reasonably foreseeable possibility of endangering your safety or the safety of your community. They should be designed to proactively protect you from harms stemming from unintended, yet foreseeable, uses or impacts of automated systems. You should be protected from inappropriate or irrelevant data use in the design, development, and deployment of automated systems, and from the compounded harm of its reuse. Independent evaluation and reporting that confirms that the system is safe and effective, including reporting of steps taken to mitigate potential harms, should be performed and the results made public whenever possible.

Algorithmic Discrimination Protections

You should not face discrimination by algorithms and systems should be used and designed in an equitable way. Algorithmic discrimination occurs when automated systems contribute to unjustified different treatment or impacts disfavoring people based on their race, color, ethnicity, sex (including pregnancy, childbirth, and related medical conditions, gender identity, intersex status, and sexual orientation), religion, age, national origin, disability, veteran status, genetic information, or any other classification protected by law. Depending on the specific circumstances, such algorithmic discrimination may violate legal protections. Designers, developers, and deployers of automated systems should take proactive and continuous measures to protect individuals and communities from algorithmic discrimination and to use and design systems in an equitable way. This protection should include proactive equity assessments as part of the system design, use of representative data and protection against proxies for demographic features, ensuring accessibility for people with disabilities in design and development, pre-deployment and ongoing disparity testing and mitigation, and clear organizational oversight. Independent evaluation and plain language reporting in the form of an algorithmic impact assessment, including disparity testing results and mitigation information, should be performed and made public whenever possible to confirm these protections.

Data Privacy

You should be protected from abusive data practices via built-in protections and you should have agency over how data about you is used. You should be protected from violations of privacy through design choices that ensure such protections are included by default, including ensuring that data collection conforms to reasonable expectations and that only data strictly necessary for the specific context is collected. Designers, developers, and deployers of automated systems should seek your permission and respect your decisions regarding collection, use, access, transfer, and deletion of your data in appropriate ways and to the greatest extent possible; where not possible, alternative privacy by design safeguards should be used. Systems should not employ user experience and design decisions that obfuscate user choice or burden users with defaults that are privacy invasive. Consent should only be used to

Draft

justify collection of data in cases where it can be appropriately and meaningfully given. Any consent requests should be brief, be understandable in plain language, and give you agency over data collection and the specific context of use; current hard-to-understand notice-and-choice practices for broad uses of data should be changed. Enhanced protections and restrictions for data and inferences related to sensitive domains, including health, work, education, criminal justice, and finance, and for data pertaining to youth should put you first. In sensitive domains, your data and related inferences should only be used for necessary functions, and you should be protected by ethical review and use prohibitions. You and your communities should be free from unchecked surveillance; surveillance technologies should be subject to heightened oversight that includes at least pre-deployment assessment of their potential harms and scope limits to protect privacy and civil liberties. Continuous surveillance and monitoring should not be used in education, work, housing, or in other contexts where the use of such surveillance technologies is likely to limit rights, opportunities, or access. Whenever possible, you should have access to reporting that confirms your data decisions have been respected and provides an assessment of the potential impact of surveillance technologies on your rights, opportunities, or access.

Notice and Explanation

You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you. Designers, developers, and deployers of automated systems should provide generally accessible plain language documentation including clear descriptions of the overall system functioning and the role automation plays, notice that such systems are in use, the individual or organization responsible for the system, and explanations of outcomes that are clear, timely, and accessible. Such notice should be kept up-to-date and people impacted by the system should be notified of significant use case or key functionality changes. You should know how and why an outcome impacting you was determined by an automated system, including when the automated system is not the sole input determining the outcome. Automated systems should provide explanations that are technically valid, meaningful and useful to you and to any operators or others who need to understand the system, and calibrated to the level of risk based on the context. Reporting that includes summary information about these automated systems in plain language and assessments of the clarity and quality of the notice and explanations should be made public whenever possible.

Human Alternatives, Consideration, and Fallback

You should be able to opt out, where appropriate, and have access to a person who can quickly consider and remedy problems you encounter. You should be able to opt out from automated systems in favor of a human alternative, where appropriate. Appropriateness should be determined based on reasonable expectations in a given context and with a focus on ensuring broad accessibility and protecting the public from especially harmful impacts. In some cases, a human or other alternative may be required by law. You should have access to timely human consideration and remedy by a fallback and escalation process if an automated system fails, it produces an error, or you would like to appeal or contest its impacts on you. Human

Draft

consideration and fallback should be accessible, equitable, effective, maintained, accompanied by appropriate operator training, and should not impose an unreasonable burden on the public. Automated systems with an intended use within sensitive domains, including, but not limited to, criminal justice, employment, education, and health, should additionally be tailored to the purpose, provide meaningful access for oversight, include training for any people interacting with the system, and incorporate human consideration for adverse or high-risk decisions. Reporting that includes a description of these human governance processes and assessment of their timeliness, accessibility, outcomes, and effectiveness should be made public whenever possible.

Applying the Blueprint for an AI Bill of Rights

While many of the concerns addressed in this framework derive from the use of AI, the technical capabilities and specific definitions of such systems change with the speed of innovation, and the potential harms of their use occur even with less technologically sophisticated tools.

Thus, this framework uses a two-part test to determine what systems are in scope. This framework applies to (1) automated systems that (2) have the potential to meaningfully impact the American public's rights, opportunities, or access to critical resources or services. These Rights, opportunities, and access to critical resources of services should be enjoyed equally and be fully protected, regardless of the changing role that automated systems may play in our lives.

This framework describes protections that should be applied with respect to all automated systems that have the potential to meaningfully impact individuals' or communities' exercise of:

Rights, Opportunities, or Access

Civil rights, civil liberties, and privacy, including freedom of speech, voting, and protections from discrimination, excessive punishment, unlawful surveillance, and violations of privacy and other freedoms in both public and private sector contexts;

Equal opportunities, including equitable access to education, housing, credit, employment, and other programs; or,

Access to critical resources or services, such as healthcare, financial services, safety, social services, non-deceptive information about goods and services, and government benefits.

A list of examples of automated systems for which these principles should be considered is provided in the Appendix. The Technical Companion, which follows, offers supportive guidance for any person or entity that creates, deploys, or oversees automated systems.

Considered together, the five principles and associated practices of the Blueprint for an AI Bill of Rights form an overlapping set of backstops against potential harms. This purposefully overlapping framework, when taken as a whole, forms a blueprint to help protect the public from

Draft

harm. The measures taken to realize the vision set forward in this framework should be proportionate with the extent and nature of the harm, or risk of harm, to people’s rights, opportunities, and access.

[i] The Executive Order On Advancing Racial Equity and Support for Underserved Communities Through the Federal Government. https://www.whitehouse.gov/briefing-room/presidential-actions/2021/01/20/executive-order-advancing-racial-equity-and-support-for-underserved-communities-through-the-federal-government/

[ii] The White House. Remarks by President Biden on the Supreme Court Decision to Overturn Roe v. Wade. Jun. 24, 2022. https://www.whitehouse.gov/briefing-room/speeches-remarks/2022/06/24/remarks-by-president-biden-on-the-supreme-court-decision-to-overturn-roe-v-wade/

Draft

Appendix C: AI/ML Ethics Steering Committee

- Ayris A Narock, NASA / Adnet, 0000-0001-6746-7455
- Micaela Parker, Academic Data Science Alliance, 0000-0003-1007-4612
- Yuhan “Douglas” Rao, NOAA / North Carolina Institute for Climate Studies, 0000-0001-6850-3403
- Thomas Donaldson, The Wharton School
- Guido Cervone, Pennsylvania State University, 0000-0002-6509-0735
- Lance Waller, Emory University, Life Sciences/ NASEM, 0000-0001-5002-8886

Draft