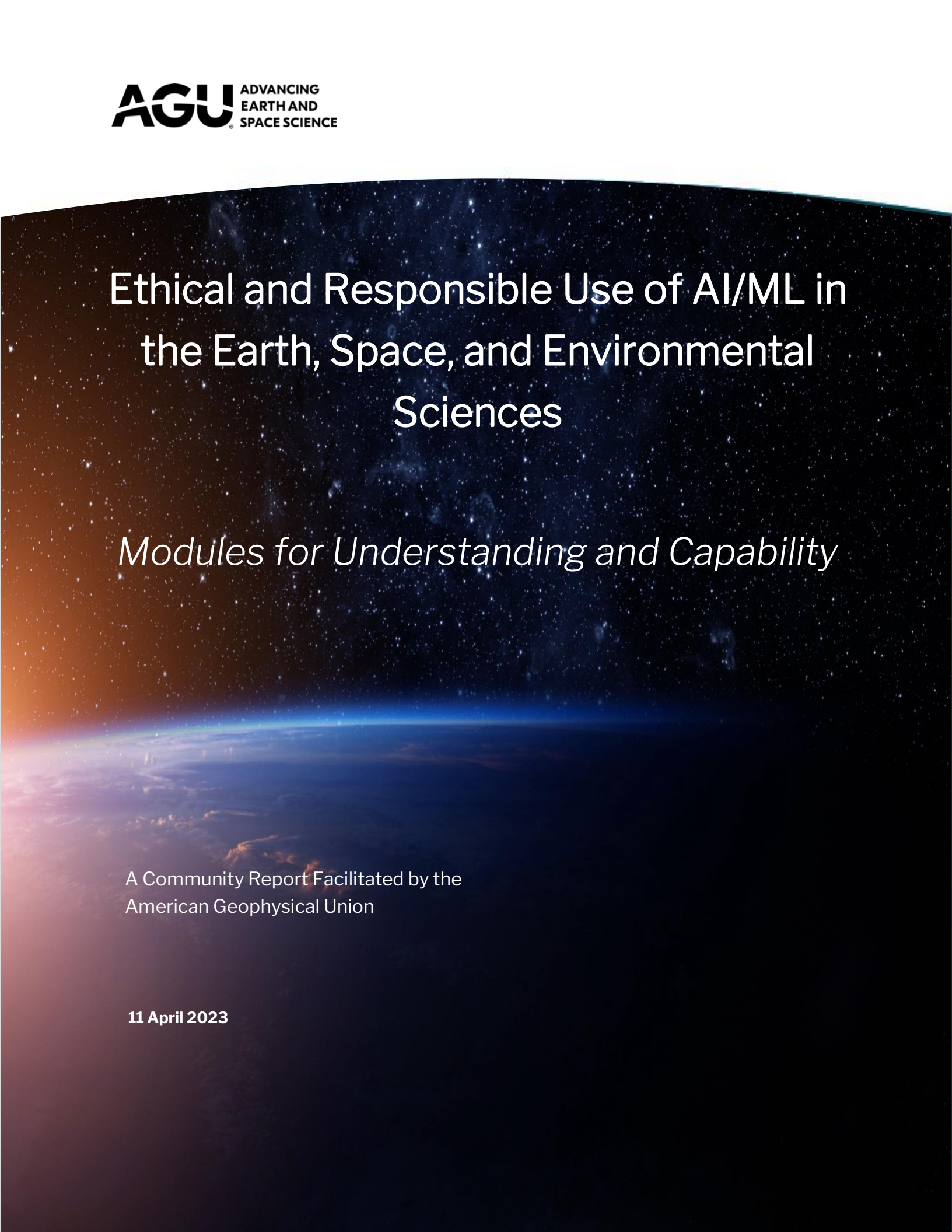
Ethical and Responsible Use of AI/ML in the Earth, Space, and Environmental Sciences

Shelley Stall[1], Guido Cervone[2], Caroline Coward[3,4], Joel Cutcher-Gershenfeld[5], Thomas J Donaldson[6], Chris Erdmann[7], R. Brooks Hanson[1], Jeanne Holm [8], John Leslie King[9], Laura Lyon[1], Delia Pembrey MacNamara[10], Amy McGovern[11], Ryan McGranaghan[3,4], Ayris A. Narock[12], Micaela S. Parker[13], Ge Peng[14], Yuhan "Douglas" Rao[15], Erin Ryan[16], Brian Sedora[1], Shashi Shekhar [17], Kristina Vrouwenvelder[1], Lance A. Waller[18], Christopher D. Wirz[19], and AGU AI/ML Ethics Workshop Participants[20]

[1]American Geophysical Union
[2]Pennsylvania State University
[3]Caltech
[4]NASA Jet Propulsion Laboratory
[5]Brandeis University
[6]The Wharton School, University of Pennsylvania
[7]Michael J. Fox Foundation
[8]City of Los Angeles, UCLA
[9]University of Michigan
[10]International Society for the Systems Sciences
[11]University of Oklahoma
[12]NASA Goddard Space Flight Center / ADNET Systems Inc.
[13]Academic Data Science Alliance
[14]UA Huntsville/MSFC IMPACT
[15]North Carolina State University / North Carolina Institute for Climate Studies
[16]Booz Allen Hamilton
[17]University of Minnesota
[18]Emory University, Life Sciences/ NASEM
[19]National Center for Atmospheric Research (NCAR)
[20]Affiliation not available

April 13, 2023



Ethical and Responsible Use of AI/ML in the Earth, Space, and Environmental Sciences

Modules for Understanding and Capability

A Community Report Facilitated by the
American Geophysical Union

11 April 2023

About AGU

AGU supports 130,000 enthusiasts to experts worldwide in Earth and space sciences.

Through broad and inclusive partnerships, AGU aims to advance discovery and solution science that accelerate knowledge and create solutions that are ethical, unbiased and respectful of communities and their values. Our programs include serving as a scholarly publisher, convening virtual and in-person events and providing career support. We live our values in everything we do, including through our net zero energy renovated building in Washington, D.C. and our Ethics and Equity Center, which fosters a diverse and inclusive geoscience community to ensure responsible conduct.

AGU was established in 1919 by the National Research Council and operated as an unincorporated affiliate of the National Academy of Sciences for more than 50 years. We were independently incorporated in 1972.

Steering Committee

Guido Cervone, The Pennsylvania State University, https://orcid.org/0000-0002-6509-0735

Thomas Donaldson, The Wharton School, University of Pennsylvania

Ayris Narock, NASA Goddard Space Flight Center, ADNET Systems, Inc., https://orcid.org/0000-0001-6746-7455

Micaela S. Parker, Academic Data Science Alliance, https://orcid.org/0000-0003-1007-4612

Yuhan “Douglas” Rao, North Carolina State University / North Carolina Institute for Climate Studies, https://orcid.org/0000-0001-6850-3403

Lance A. Waller, Emory University, Life Sciences/ NASEM, https://orcid.org/0000-0001-5002-8886

Facilitator

Joel Cutcher-Gershenfeld, Brandeis University, https://orcid.org/0000-0001-7659-7024

Editorial Team

Guido Cervone, The Pennsylvania 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, Michael J. Fox Foundation, https://orcid.org/0000-0003-2554-180X

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, NASA Jet Propulsion Laboratory/Caltech | NASA Goddard Space Flight Center, https://orcid.org/0000-0002-9605-0007

Micaela S. 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, https://orcid.org/0000-0002-1986-9115

Yuhan “Douglas” Rao, North Carolina State University / North Carolina Institute for Climate Studies, 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, University of Minnesota, https://orcid.org/0000-0002-9294-4855

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

Amy McGovern, University of Oklahoma, https://orcid.org/0000-0001-6675-7119

Acknowledgements

We appreciate the support of the National Aeronautics and Space Administration (NASA) for funding this initiative (Grant 80NSSC22K0734) as well as the many participants in the workshops and the editorial team generating this report.

Table of Contents

Vision

In the geosciences and across society, Artificial Intelligence and Machine Learning (AI/ML) are becoming ubiquitous. AI/ML tools and methods are enabling advances in understanding the Earth and its systems at all scales, informing critical decisions by researchers, organizations, and government agencies. Automated workflows utilizing AI/ML are, according to the National Academies in the US, accelerating discovery across all of the sciences. This report is designed to support these advances while mitigating potential risks.

The use of AI/ML brings risks, which require the understanding and application of ethical principles, guidelines, and practices to mitigate harm. Moreover, AI/ML algorithms have independent agency, which raises profound questions about who is responsible for findings generated by these models and how biases in models and/or the underlying data used as inputs to models can be anticipated and mitigated. The use of AI/ML in science builds on and requires an extension of the principles associated with open science and the responsible use of computing technologies.

AI/ML tools can deliver results and provide information that cannot be achieved by other methods. Ethical AI/ML tools are essential for high-quality geoscience and planetary science and for addressing and responding to broad societal challenges. AI/ML applications are now being utilized to analyze seismic data, predicting the likelihood of earthquakes; to improve climate and weather models, predicting severe weather; analyzing hydrologic data from sensors and satellites, predicting flooding and water shortages; and assessing countless other types of geologic data.

With the accelerated rates of change enabled by these technologies, the challenges for the geosciences center not just on the ethical responsibilities of researchers, but also on the ethical standards for AI/ML tools that are generating knowledge without full human direction. Ethical standards, principles, and practices associated with AI/ML in geoscience research will be essential to researchers and the broader society in ensuring that the observation, modeling, and forecasting of geo-phenomena (broadly defined) happens in appropriately open and inclusive ways. These applications must consider and mitigate potential adverse impacts on historically marginalized communities and society at large. The challenge is both social and technical.

AGU members, representing a large part of the global Earth, space, and environmental science community, are governed by the AGU Scientific Ethics and Integrity Policy regardless of their research methods, including when using or developing AI/ML tools. These ethical principles align with and are based on guidance for responsible research provided by National Academies and other organizations that include the very role of a scientist in society. The overarching goal of these specific AI/ML Ethics Principles and Responsibilities is to supplement this existing ethical framework by focusing on considerations that researchers and organizations must address in addition to existing requirements. This includes more robust and inclusive research methods,

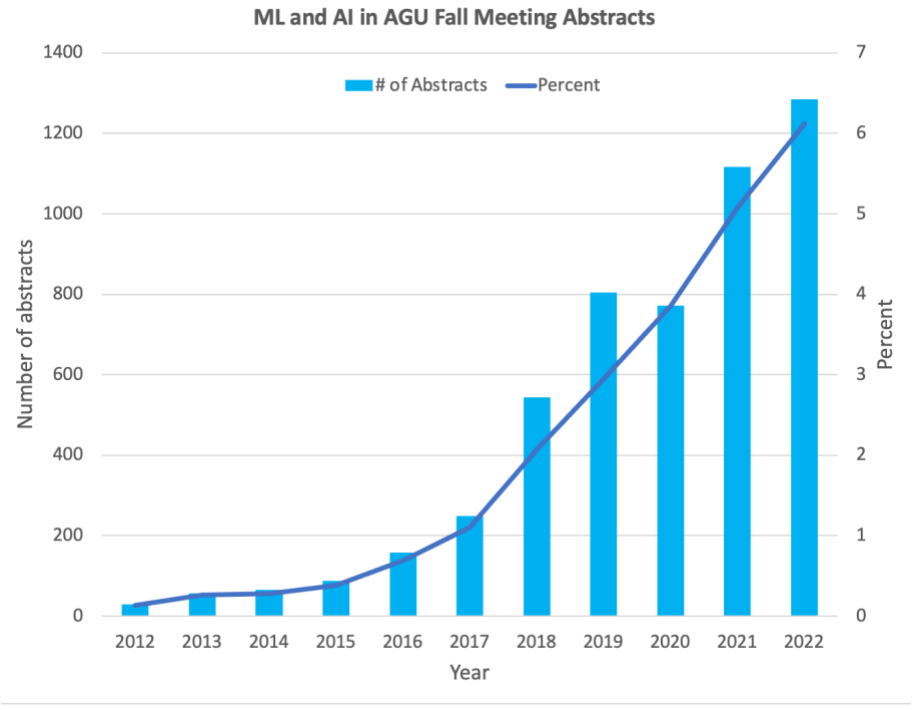
new forms of documentation, new methods for replicability, continuing responsibility for the impacts of research, and proactive expectations of professional societies, funders, and other institutional actors. Our vision is for accelerated discovery that anticipates and mitigates risks.

> *"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")

Introduction and Overview

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

AI/ML research is increasingly prevalent in the geosciences, as illustrated in the chart below tracking AI and ML in AGU fall meeting abstracts. This is still just 6 percent of abstracts, but the rate of change is unlike any other current set of tools and methods. The progression in the past decade is from tens, to hundreds, to thousands of abstracts -- an exponential rate of change. This points to an accelerating use of AI/ML technologies and a need for the social systems around ethics to co-evolve at a matching rate.



The use of any technology or technique such as AI/ML should be understandable and accompanied by documentation on data and tools that allow for the validation and replication of any scientific results. This is a core principle of science and it is complicated in the case of AI/ML methods where the inner workings of models are opaque. Traditionally, the entire method should be explained and accessible and that is possible with AI/ML but requires new ways of thinking about methods. The steps in the process can be documented, but not the actual computation that results. Additionally, the methods 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 is true for any scientific research, but more salient with AI/ML.

This document provides a set of principles and responsibilities for ethical AI/ML and leading practices for AI/ML. These principles and responsibilities were developed through community input and facilitated discussion in the latter part of 2022, and led by a steering committee. 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, implementing this guidance, providing regular updates, and informing and educating researchers about them.

The AI/ML Ethics principles include a Code of Conduct and six modules, each of which is structured to provide responsibilities, description and considerations, and support training and development of needed skills for researchers and scholarly organizations. Although the focus is on AI/ML in the Earth, environmental, and space sciences, many of the principles apply broadly. The six modules include four focusing on researchers and applications (the "Principles for Researchers") and two focusing on practices by scholarly organizations including professional societies, institutions, funders, and publishers. ("Principles for Scholarly Organizations"). These six modules are:

Principles for Researchers:

Module 1: Transparency, Documentating, and Reporting

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

Module 3: Risk, Bias, and Effects

Module 4: Participatory Methods

Principles for Scholarly Organizations:

Module 5: Outreach, Training, and Leading Practices

Module 6: Considerations for Organizations and Institutions, Publishers, Societies, and Funders

For training or education, the modules can each be used separately, or together as a partial (e.g., first four) or full set (with the order flexible). 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.

Each module is organized with the following elements:

Module Focus

Module Objectives

Module Vision

Module Key Points

Module Principles

Module Responsibilities and Leading Practices

Module Use Cases and Illustrative Examples

Instead of a pause in the use of AI/ML, which some have advocated, we support continued advances that co-evolve with the principles for ethical and responsible use of the methods. As a result, this is meant to be a living document, and the principles, responsibilities, and other elements will be regularly reviewed and updated as the technologies, applications, and institutions evolve.

"[W]e all know that each generation has its own test, the contemporaneous and current stand by which alone it can adequately judge of its own moral achievements and that it may not legitimately use a previous and less vigorous test. The advanced test must include that which has already been attained; but if it includes no more, we shall fail to go forward, thinking complacently that we have "arrived" when in reality we have not yet started."

-- Jane Addams, founder of the field of Social work, (quote from Democracy and Social Ethics, page 5)

Code of Conduct for Researchers

The four modules providing Principles for Researchers articulate an overarching set of principles and responsibilities that form a "code of conduct" for researchers using AI/ML in the Earth, environmental and space sciences. Consistent with these principles, researchers should:

1. Ensure compliance with overarching AGU ethical standards and codes of conduct.
2. Integrate AI/ML ethics throughout the research life cycle, from planning, to modeling, to reporting, to anticipating the potential use of AI/ML tools and methods at scale or in unexpected ways.
3. Review data collection methods for potential sources of bias, and implement mitigation methods in the models or in other ways as appropriate.
4. Specify intended use, boundaries, and delimitations on use of AI/ML models (documenting design decisions in the model development) and ensure that these considerations are known when models are adopted for other uses.
5. Assess AI/ML model design and other prediction methods for potential sources of bias, areas of uncertainty and limitations, failure modes, and implement mitigation methods.
6. Ensure that the use of AI/ML tools and methods reflects inputs from communities who might be potentially impacted by these methods, with particular attention to vulnerable and historically underserved populations.
7. Validate and verify results at every stage of the research process, using leading practices appropriate to the applications.
8. Provide sufficient guidance on the algorithms and training data used so that the findings can be replicated or confirmed.
9. Comply with AI/ML reporting and conduct requirements by Institutional Review Boards (IRBs) and other governing bodies.
10. Comply with AI/ML reporting requirements by conference program committees, scholarly journals, and book publishers.

Code of Conduct for Scholarly Organizations

The two modules providing Principles for Scholarly Organizations are designed for organizations that support research, which includes professional societies, institutions, funders, and publishers. Scholarly organizations should:

1. Provide guidance and guidelines for researchers, policy makers, and citizens on the ethical use of research that utilizes AI/ML methods.
2. Support awareness and training in the ethical principles and responsibilities.
3. Provide a means for governance, and review and update processes to ensure compliance and relevance.
4. Support AI/ML communities of practice, multi-stakeholder consortia, public-private partnerships and other collective efforts that enable groups to accomplish together what they can't do individually.
5. Intervene in appropriate ways where AI/ML applications are causing harm in society.
6. Educate societal decision makers on the value and limitations of AI/ML in research in order to enable responsible decisions based on AI/ML findings.

Trust in AI/ML?

Trust is at the core of ethics with Artificial Intelligence and Machine Learning (AI/ML). Ironically, the challenges involve both too little trust and too much trust. The challenge of *too little trust* centers on what is sometimes termed the "black box" or "grey box" in which models and the underlying algorithms are fully or partly opaque, making it hard or impossible to determine if, or to what degree, the results can be trusted. The challenge of *too much trust* centers on the growing prevalence of AI/ML tools in everyday life where it is too easily assumed that the designers of the technology fully anticipated any potential problems with the technology. Given both challenges, we have added a question mark to the title of this section – signaling that trust in AI/ML is still an open issue.

There is an extensive literature on the challenges associated with trust in AI/ML, arguing that building trust must be an incremental process (Ferrario, Loi & Viganò, 2020) and that calls for transparency are hard to put into practice (Babic, Gerke, Evgeniou & Cohen, 2021). A contributor to this study defines the trustworthiness of AI as the willingness to assume risk by relying on and/or believing in the actions of another party (McGovern, 2022). In many ways, the motivation for the Code of Conduct and supporting materials is all about trust and the reliance on the work of others who are employing AI/ML models. The use of AI/ML models requires a measure of belief in the developers and others associated with the models since the models themselves are, in many ways, opaque. Unlike other methods, such as test statistics with multivariate models, AI/ML still lacks some of the core tools and established mechanisms for assessing confidence in the findings. As a result, trust in AI/ML models and associated methods involves human judgment. The Japanese term *jidoka* was developed in the context of high performance work systems to signal the importance of building in human judgment on quality when automating tasks. *Jidoka* has been interpreted as "humans giving wisdom to machines," which is essential if there is to be trust in AI/ML.

In the social and behavioral sciences, human judgements about trust are understood as having multiple dimensions. Trust can be based on reliability, expertise, authenticity, and good will. When it comes to AI/ML in research, these elements apply in the following ways:

- **Reliability** of AI/ML models is reflected in their ability to deliver consistent results when run multiple times.
- **Expertise** is primarily vested in the developers of AI/ML models, which means that authorship and contributors should be documented.
- **Authenticity** is primarily achieved with AI/ML models through transparency, which can happen with the sharing of code, the sharing of design decisions, ensuring access to the underlying data, training data, and test data, testing with use cases that have known outcomes, and even interactive features of a model so that users can do "what if" testing.

- **Goodwill** is hard to establish in the context of AI/ML, but a step in that direction can be achieved if potential sources of bias and other model limitations are identified, which will signal a measure of goodwill in anticipating potential user concerts.

A deeper challenge from the social and behavioral sciences is that the four dimensions of trust listed above are analytically useful, but do not represent the initial way that human beings do or do not trust something. Gigerenzer and other social scientists (Gigerenzer & Gaissmaier, 2011) have documented the use of heuristics in making rapid judgments about trust, with analytic categories only coming later (and having to contend with trust judgments that have already been made). That means that AI/ML models will likely encounter immediate judgements that are either trusting or distrusting and any progress in tempering misplaced trust or reversing distrust will have to overcome these judgements.

A particular complication is when AI/ML models deliver results that comport with our expectations. We need to ask ourselves whether this is validation for our views or if the model and our views have biases based on limitations in the data and the scope of our experience. This is why trust in AI/ML requires input from multiple perspectives and critical thinking about what the models are (and are not) taking into account.

In science fiction, ubiquitous forms of AI have long been depicted as having agency and, as a result, as inviting immediate human judgements on trust. At times these depictions can indicate evil intent (H.A.L. in 2001), supportive intent (J.A.R.V.I.S. in Iron Man), industrious intent (WALL-E), and even friendship among AI-enabled robots (C-3PO and R2-D2 in Star Wars). Today, judgments about these types of agency are beginning to be formed around publicly available AI tools for writing (ChatGPT), drawing (DALL-E 2), and various AI-enabled voice assistance (SIRI, ALEXA, etc.). Even though these AI tools are just an amalgam of human-created content, responses come together in ways that appear original, which then immediately invites human judgments on trust. When these types of tools are used in research and incorporated into what are presented as original findings, the need to understand just what the AI/ML models are doing becomes urgent.

AI ethics are increasingly a focus for research and scholarly publications. The National Science Foundation in the U.S. has launched a series of projects on foundational AI across scientific domains. In the geosciences this includes the establishment of the AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (Mcgovern, 2022). As the founders of this AI Institute note, “there is often a lack of trust by environmental science decision-makers when it comes to relying on “black box” algorithms, especially in life-or-death situations… Developing AI that is trustworthy and useful for environmental risk management requires fundamental natural, mathematical, and social sciences research on the AI needs and perceptions of key users” (Ibid). Among publishers and societies, the Artificial Intelligence Robotics Ethics Society (AIRES) launched *The AI Ethics Journal* in 2019 and Springer launched a new journal, *AI and Ethics*, in February, 2021. When *AI and Ethics* was launched, the

lead editor's note observed, "Disruptive technologies are rarely without some negative consequences and risks. The development of electricity transformed almost every aspect of human life, but also had its casualties, both economically and culturally. The advent of the printing press transformed education across the globe and powered the scientific revolution of the 16th and 17th centuries, but was also responsible for the elimination of the livelihoods of many craftsmen whose artisan skills were replaced by mass production. So it has been throughout history—technology comes with consequences. Artificial intelligence is no exception, but may be exceptional in terms of the unprecedented short time frame of change and the extent of the potential negative consequences if we get it wrong." (MacIntyre, Medsker, & Moriarty, 2021).

Ultimately, trust in AI/ML is not something that can be prescribed or guaranteed. Still, there are ways to increase the likelihood that AI/ML models, systems, and developers will be 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. Journalists have developed Guidelines on reporting on AI that could represent useful guidance for developers and users of AI/ML in research and commercial applications.

Researchers often want a "litmus paper" for our models and work, e.g., an answer to the question: 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, consider the case of 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 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.

As is indicated in these materials, building trust in AI/ML systems requires open and transparent research (to the extent feasible). We need to communicate and quantify uncertainty (again to the extent feasible), 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 underlying issues of trust in technology and trust in science that impact trust-development with AI/ML systems.

Each of the six modules includes guiding principles, supporting information, and a listing of responsibilities for researchers and other stakeholders. We highlight these elements here with respect to trust and AI/ML both as an introduction to these types of materials and as specific guidance with trust and AI/ML. Here are guiding principles, for example, on trust and AI/ML:

- **Foster equity and engaging relationships** across stakeholders in all phases of the AI/ML research life cycle.
- **Engage in 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.

Note that some of these principles just involve unitary decisions and actions by the researcher, such as engaging in open and direct communication or appreciating the context. Other principles involve interactions, such as fostering equitable relationships or engaging in co-development. Here relationship building is needed, not just unitary action. This additional requirement of collaborative interactions typically requires additional time, skills, and resources. In some cases building and sustaining trust will involve forming communities of practice, multi-stakeholder consortia, public private partnerships and other collective efforts that can co-evolve with the AI/ML technologies in agile and adaptive ways (Stakeholder Alignment Collaborative, 2022). Enacting these principles requires guidance, including the following supportive practices for each of the 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.
- **Knowledge transfer among the research team, users, affected communities**. Education, training, and learning from one another are key foundations for establishing trust.

There are specific responsibilities for researchers and others in the research lifecycle when it comes to AI/ML. All six of the modules list researcher and other stakeholder responsibilities and relevant ones are listed here when it comes to trust in AI/ML:

- **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.

Again, some of these responsibilities, such as transparency are within the direct control of a researcher, while others, such as multi-way conversations, involve engagement with others. Both are essential for building trust, but the engagement with others generally involves more time, effort, and resources. The results of these interactions are often insights that can't be achieved any other way.

In summary, trust in AI/ML is a dynamic and interactive process, rather than a one-time event or judgment. As researchers and other stakeholders engage in trust-building, the goal is not just to seek acceptance of the findings, but to advance the science itself by taking into account all dimensions of trust.

Principles for Researchers

Module 1: Transparency, Documentating, and Reporting

Module 1 Focus

Transparency, documenting, and reporting on uncertainties with AI/ML ethics in research are essential. This module describes a key ethical principle 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 Objectives

- Explore how to proactively pursue transparency when using AI/ML in research.
- Clarify considerations in the documentation needed with AI/ML models and data.

Module 1 Vision

Research in AI/ML should be transparent, accessible, and open as possible at every step of the project. This includes the documentation of research design and uncertainties, including data and model biases. Leading practices around open science should be followed for reporting on data collection, data preprocessing, model construction and training (parameter values, etc.), model validation, and results. How and to what extent the results and pretrained models can be used in downstream applications should be explained.

In many cases, researchers in AI/ML are using data that have been sourced and archived by others, in addition to combining diverse data sets and types. Ensuring the quality of these data and following leading community principles such as FAIR, CARE, OCAP, and TRUST, is essential (AGU data position statement). AI/ML researchers should ensure that data sources are cited and available to others, and they should be transparent regarding which datasets were used and how, and alert readers to known biases in data. Any necessary restrictions on access should be documented.

Including subject matter experts (e.g., disciplinary experts in the project science including data sources) can improve and build trust in all steps of an AI/ML project. These can help with identifying quality data and data biases, ensuring explainable and science-informed AI/ML models, providing post-hoc explanation of blackbox/graybox models, and providing sensitivity

analysis for key design decisions. These experts should be recognized and provided credit in project outcomes.

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, which makes transparency hard to achieve. However, there are aspects of AI/ML models that can be described in documentation in ways that indicate intent and limitations. These can be done experimentally or qualitatively with capabilities that enable users to assess how these models 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 due to issues of privacy, proprietary and sovereign data, and related matters. Model results may also disclose proprietary information or cause harm if fully available.

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 5 on outreach) are required to facilitate transparency and trust. Including or consulting additional experts or other stakeholders on the data or code can improve understanding; their roles and contributions should be disclosed. As a practical matter, researchers will need to maintain a record of decisions throughout the research process. They will also need to document interactions with key stakeholders. This is part of the broader principle in research ethics of giving credit to those giving input. Transparency should be considered throughout the whole lifecycle of AI/ML applications from conceptual development for applications.

Module 1 Principles

Transparency

- **Indicate how leading AI/ML practices (listed below) are followed** in your research or where departures from leading practices are needed and why.
- **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

Documentating

- **Document AI/ML decisions and use of any associated model, code or other digital products** 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** including sources of and adjustments to data, as well as generations, versions, and sources of models and other digital objects.
- **Ensure and indicate clear access** to the AI/ML model code to the extent feasible, as well as the data used and created.

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 the leading practices below, and implies convenient access to relevant information about a research 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 sometimes may not 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, they 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.

- 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.
- When datasets follow the FAIR Guiding Principles researchers can better assess potential bias of the data for use in their research.
- Data managed in a repository that supports the CARE Principles supports researchers ensuring proper usage of the data according to the indigenous rights owners of the data.

- In a literature review of provenance, explainable AI (XAI), and trustworthy AI (TAI), the co-authors are critical of post-hoc documentation and call for rigorous tracking of provenance to help accomplish explainability in AI-based systems (Kale, Nguyen, Harris, Li, Zhang, Ma, 2023).
- In a recent Australian study, the authors offer an example of how they explain their transparency: "machine learning is an important approach to synthesise the increasing complexity and sheer volume of Earth science data, and is now widely used in prediction across many scientific disciplines. In this context, we have built a machine learning pipeline, called Uncover-ML, for both supervised and unsupervised learning, prediction and classification. The Uncover-ML pipeline was developed from a partnership between CSIRO and Geoscience Australia, and is largely built around the Python scikit-learn machine learning libraries. In this paper, we briefly describe the architecture and components of Uncover-ML for feature extraction, data scaling, sample selection, predictive mapping, estimating model performance, model optimisation and estimating model uncertainties. Links to download the source code and information on how to implement the algorithms are also provided." (Wilford et al., 2020).
- To support reporting and documentation of AI/ML models in a standard format, practitioners and researchers developed model cards that make information about AI/ML models accessible whenever possible. The suggested model card template includes prompts to report model details, intended use, factors that affect model performance, metrics for evaluation and decision, training and evaluation data used, quantitative analyses, ethical considerations, caveats, and recommendations on model use (Mitchell et al., 2018). Model cards make the technical and other relevant information accessible for various stakeholders to better understand the model and make informed decisions. The practice has been adopted by AI/ML practitioners when sharing models online (e.g., Hugging Face).

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

Module 2 Focus

Ensuring Intentionality, Interpretability, Explainability, Reproducibility, and Replicability (sometimes termed "illities") when employing AI/ML in research.

Module 2 Objectives

- Understand the key concepts related to replicability and explainability.
- Build skills based on 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, with new, readily-available algorithms proliferating. In many cases, statistical qualities and uncertainties will not be fully known. As a result, we maintain a principled approach, enabling understanding and testing of algorithms, as a foundation for the evolution of AI/ML in the geosciences. Scientific questions will ground the justification of the method choice and application. We define this as an approach that provides clear model specification incorporating domain knowledge and keeping hypothesis-driven motivation at the forefront. We prioritize an open science approach to enable interpretability and replicability, and reproducibility where possible. We encourage the application and development of methodologies for model explainability of AI/ML models.

Module 2 Key Points

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

For a model to be used, it should be both reproducible and replicable. In general, this implies that results can be obtained again both by the group who first developed the model and by independent researchers adopting it. Providing a verification dataset for the model alongside the expected output can be used to ensure the replicability of results. Documentation of the steps in model development and testing is also 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 and reported as such. Then, further analysis can be conducted to determine the degree to which the findings are indeed robust and trustworthy.

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 are taken.

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?
- **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 accessible?

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. Readers should be able to understand 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. Researchers should comply with open science requirements of journals and funders.
- **Researchers and Educators lean on expertise in other fields.** Research teams are cross-disciplinary, including expertise in computer science and statistics. Ex: Graduate level training in statistics and/or computer science is routinely incorporated into the Geology/Geophysics degree path.

Module 2 Use Cases/Illustrative Examples:

- Research on National Weather Service Forecasters' perceptions of using AI/ML for forecasting severe weather explored what the concepts of 'explainability' and 'interpretability' meant to them as potential end-users (Cains et al., 2022; McGovern et al., 2022). This work revealed that Forecasters connected AI explainability with how understandable the model was for them and those they served, while AI interpretability

was associated with the need for good visualizations. This case study highlights the need to be clear with what is and is not addressed by our scientific concepts, as well as to engage potential users to better understand how these concepts may or may not translate into practice.

- AI/ML model reproducibility is challenging particularly when the model is developed with large datasets. In 2018, a survey found that only ~10% of accepted papers by top AI/ML conferences share both code and data used to develop the model (Hutson, 2018). This caused concern about the irreproducible outcome. Since then, peer-reviewed journals and conferences have been developing checklists to guide researchers to improve the reproducibility of AI/ML models. More often, research communities started hackathon-style activities to use a crowd-sourcing format to ensure AI/ML models can be reproduced (e.g., reproducibility challenges). These activities not only can be used to investigate published models/research but can also help improve leading practices to ensure model reproducibility (Pineau, J., P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché-Buc & H. Larochelle, 2021).
- When developing the report and documentation of an AI/ML model, it's as important to explain why the model performs poorly in certain conditions as to explain the overall model performance. Explanation of the failed case studies can help users to calibrate their expectations of the model. Rao et al. (2019) used cross-validation and statistical analysis to demonstrate why their ML model, which was designed to fuse station observations and satellite products for climate monitoring, has larger uncertainty in certain geographical regions. The analysis demonstrates the shift between training data and evaluation data due to different local weather conditions.
- Machine learning can help scientists work with large-scale data. The Cassini mission, for example, collected over 600 gigabytes of scientific data from 2004 to 2017. This represents a surge of data on the Saturn system. In comparison, the previous mission to Saturn, Voyager over 20 years earlier, had onboard a ~70 kB 8-track storage ability. Unlike many applications of machine learning, a primary use in planetary space physics applications is to infer behavior about the system itself. This raises three concerns: first, the performance of the machine learning model, second, the need for interpretable applications to answer scientific questions, and third, how characteristics of spacecraft data change these applications. In comparison to these concerns, uses of "black box" or un-interpretable machine learning methods tend toward evaluations of performance only either ignoring the underlying physical process or, less often, providing misleading explanations for it. The present work uses Cassini data as a case study as these data are similar to space physics and planetary missions at Earth and other solar system objects. We build off a previous effort applying a semi-supervised physics-based classification of plasma instabilities in Saturn's magnetic environment, or magnetosphere. We then use this previous effort in comparison to other machine learning classifiers with varying data

size access, and physical information access. We show that incorporating knowledge of these orbiting spacecraft data characteristics improves the performance and interpretability of machine learning methods, which is essential for deriving scientific meaning. Building on these findings, we present a framework on incorporating physics knowledge into machine learning problems targeting semi-supervised classification for space physics data in planetary environments. These findings present a path forward for incorporating physical knowledge into space physics and planetary mission data analyses for scientific discovery. (Azari, Lockhart, Liemohn, Jia, 2020)

Module 3: Risk, Bias, and Effects

Module 3 Focus

Identifying model and design risks, bias, and intended or unintended consequences with AI/ML in research.

Module 3 Objectives

- Appreciate the key sources of risk and bias in AI/ML applications.
- Build capability in managing risk and bias in AI/ML applications while maximizing beneficial effects.

Module 3 Vision

AI/ML can augment and effect the ways we generate knowledge and make decisions across many scientific fields. This module elucidates the model and design biases and risks as they relate to AI/ML use by the Earth, environmental, and space science research communities. We offer principles to identify and address those biases and risks. These principles 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. Through a better understanding of risks and biases, how they come about, and how to identify them, researchers, communities and organizations will be better able to manage and respond to adverse outcomes while maximizing public benefit and effect.

Module 3 Key Points

All AI/ML models and research design involve bias. Mitigating AI/ML bias, risk, and harm enables researchers and organizations to promote impactful, transformative, and beneficial research. Multiple research teams and organizations may be responsible for anticipating potential disparities in the application of models and algorithms, as well as for assessing early and continuing results for negative impacts. The mitigation work is both proactive and responsive.

The responsibility for managing bias, risk and harm lies with individuals and groups involved throughout the AI/ML process, such as researchers, users of the models, and funders of the research. This means responsibility is shared but also demands accountability from all those involved in the process. Harm in AI/ML applications can be deeply embedded in the data itself (for example, arising from training data that doesn't reflect the diversity of society). Mechanisms to provide voice to vulnerable populations who might be impacted by the application of AI/ML in

research are especially important. These mechanisms could include advisory committees, community forums, and ongoing multi-stakeholder consortia or public-private partnerships associated with research initiatives. Funders are encouraged to build voice and management mechanisms into the budgets for funded AI/ML research.

Investments in tools and methods to identify bias in geoscience data are encouraged. Leadership from scientific societies could be embodied in the appointment of a chief AI/ML risk officer serving on a broader ethics committee. This could also take the form of resources providing the needed consultation and advice to society members and others as appropriate. A consortium of relevant professional societies could also provide the needed set of shared resources in this domain.

Module 3 Principles

Risk

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

Bias

- **Identify and document potential sources of bias** in problem identification, training data, algorithms, outputs, and other aspects of AI/ML applications.

Effects

- **Identify potential harms and work to 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, and increase positive effects 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/ML models.
- Ensure the AI/ML system lifecycle intentionally includes the involvement of people and communities that could benefit or be harmed.

- Pay particular attention to AI systems that become integrated into the infrastructure of society.
- Ensure that the AI models are developed to promote sustainable development, 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

- Researchers will ensure that AI/ML systems developed for Earth, space and related sciences avoid harm throughout the AI/ML lifecycle by:
    - Initial conception of the research acknowledges risk and bias as part of the design.
    - Model development involves an explicit step to consider bias and risk associated with the model.
    - Training data is assessed for risk and bias.
    - Taking responsibility for AI/ML systems and datasets and ensure that there is 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, including 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 responsible for developing AI/ML systems will ensure that these systems avoid harm 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 data, testing data, 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.
- 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/Illustrative Examples:

- In a recent study AI/Ml was used for lithology quantification from rock chips analysis with illustrative descriptions of the detection algorithm (Wang, et. al., 2022).
- Given the large volumes of geoscience data that might be utilized with AI/ML algorithms, "containers" represent a way to make code portable across environments (Caraballo-Vega, et al., 2022).
- The bias in AI/ML models can be caused by unintentional factors like low-quality or unrepresentative training data. McGovern, et al. (2022) demonstrated some Earth and environmental science case studies. One example is the lack of geo-diversity of the existing sensor networks that are often used as training data for AI/ML model development and evaluation. Due to the requirement of physical accessibility, existing sensor networks such as the U.S. national radar network has coverage gaps in frontline communities. The demographic coverage gap should be addressed by strategically placing additional sensors to increase the geo-diversity of data used for training and evaluation datasets.
- Potential risk can also be caused by adversarial attacks by ill-intended actors by adding noise to data used by operational AI/ML models or sending faulty information. This can be addressed by assessing and enhancing model robustness. A recent case study of the robustness of a deep learning model that is designed to predict the category of tropical cyclones using satellite images revealed that noise added to satellite images can cause notable underestimation of the cyclone strength. This issue can be addressed by modifying the model training and development strategy to improve the model's robustness.

Module 4: Participatory Methods

Module 4 Focus

Inclusive research design and conduct with AI/ML – ensuring a voice for diverse communities, domain expertise, and context in cases where AI/ML research impacts or is relevant to specific communities (or might reasonably be expected to be).

Module 4 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 4 Vision

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"). Even where there are not directly impacted communities, such as a study of seismic signals on the moon, there may be interested parties whose voices should be included, such as others conducting research on the moon.

Module 4 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. Research that does so, or might be expected to do so, requires the formation of advisory groups, the utilization of stakeholder and rightsholder mapping surveys or focus groups, the democratic selection of community representatives, and other mechanisms for input.

A key underlying assumption is that diverse stakeholders have both common and competing interests. As a result, participatory approaches must simultaneously identify and advance common interests, as well as identifying and addressing competing interests. This will involve both aspirational dialogue and hard conversations. For the hard conversations it may be helpful to remind all to be "hard on the issues, not each other."

One challenge is that awareness of possible impacts may not always be clear at the beginning of a project. In this case, these principles should be applied as soon as possible after such awareness and may affect release of results or context around them, or even further research. In turn, AI/ML research projects should be continually evaluated for possible impacts.

Another challenge is determining how much involvement is sufficient. Just doing stakeholder engagement in the form of communications or meetings with an agenda determined by the researchers is not sufficient. Forms of stakeholder alignment are needed where there is sufficient dialogue so that parties feel heard and, where appropriate, adjustments have been made (or the reasons for not making adjustments are at least understood).

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 "rightsholders" such as First Nations (indigenous peoples), and others with land claims such as the Inuit and Métis peoples. 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 ongoing implications of the research. The recognition of rightsholders as well as stakeholders helps to counteract the negative connotations of the term "stakeholder," which has literal roots in placing a stake in the ground to claim land that was already occupied.

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. Note, however, that not all aspects of CARE or FAIR principles can be fully applied in all AI/ML 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 4 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.

- **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 4 principles:

- "No" research impacting a group without their continuous consent maintaining their autonomy and representation at decision-making level.
- 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 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 4 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 enables 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 and potential harms are mitigated against.
- Regarding data collection and usage, the 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:
        - Be alert to 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.
    - 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.
    - Procurer/funder: require inclusive development and regular reporting during the research process.

Module 4 Use Cases/Illustrative Examples

- The AI/ML Stakeholder "Pulse" Survey data presented in Appendix B is an example of broader outreach in developing this report. The process involved first identifying the relevant stakeholder categories associated with AI/ML in the geosciences (Researcher who uses AI/ML in research, Researcher who does not use AI/ML in research, but is knowledgeable about the technologies, Researcher who does not use AI/ML in research & is not knowledgeable about the technologies, Research Computing and Data Professional, Student (graduate or undergraduate), Administrator/leader in university, Administrator/leader in government, Administrator/leader in government contractor, Administrator/leader in commercial organization, Administrator/leader in not-for-profit organization, and clothes. Then, the process involved identifying the interests that were "at stake" such as establishing ethical standards for AI/ML, educating researchers on

these standards, ensuring compliance with the standards, and about a dozen others. A representative sample of thought leaders on AI/ML in the geosciences were then surveyed on the importance and difficulty of advancing the various interests, along with demographic questions and open-ended qualitative questions. The results from the survey were shared with participants in the workshops that contributed to the development of this report, which both brought more voices into the process and informed the results.

- Crowd the Tap is a citizen science project, the first U.S. Environmental Protection Agency (EPA)-funded project that promotes access to safe drinking water by empowering individuals and groups to investigate the piping infrastructure that delivers drinking water to their homes. (https://ethos.academicdatascience.org/case-studies/)
- It's critical to engage with users of AI/ML model-based applications across the whole lifecycle of model development. A recent initiative that uses AI/ML model to map the urban heat island effect in select cities in U.S. involved citizen scientists to collect training and evaluation data using mobile sensors. The data collection campaigns are led by local organizations and directly involve a large group of local volunteers to collect data for geographical areas that are decided through close consultation with local organizations. The citizen science approach also involves volunteer training and public outreach on how the data will be used and what the final output of the model will be.

Principles for Scholarly Organizations

Module 5: Outreach, Training, and Leading Practices

Module 5 Focus

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

Module 5 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.

Module 5 Vision

The implementation of ethical use of AI/ML in the Earth, environmental, and space sciences 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 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 the inclusion of ethical practices 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 and others 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-planning and post-assessment "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 Principles

Training

- **Provide training, resources, and support** for AI/ML ethics to all researchers, institutional leaders and other key stakeholders.
- **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.
- **Educate societal decision makers** on the value and limitations of AI/ML in research in order to enable responsible decisions based on AI/ML findings.

Leading Practices

- **Manage and update training and resources for AI/ML ethics** to ensure the current state of practice.
- Support AI/ML communities of practice, multi-stakeholder consortia, public-private partnerships and other collective efforts that **enable groups to accomplish together what they can't do individually with respect to AI/ML.**
- Be prepared to **intervene in appropriate ways** where AI/ML applications are causing harm in society.

Additional supporting information on Module 5 principles:

- Ethical AI/ML is a fundamental part of AI/ML research and not optional.

- 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.
- 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

- Mitigate the potential for AI/ML used in research to have negative impacts on people and on the quality of the science.
- Communicate the principles and practices of ethical AI/ML to all constituents (outreach)
- Train practitioners to perform ethical AI/ML research and report results consistent with these principles and make training resources widely available.
- Include communities and community perspectives in training resources.
- Identify resources and tools that facilitate the adoption and inclusion of ethical AI/ML for all constituencies.
- 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 a framework for self-evaluation of AI/ML against the intent of the principles and responsibilities.

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.
- 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?

- The founders of FastAI are focused on making AI understandable by anyone. They provide blogs, training recommendations, and links to Ted talks for all levels of learner. fast.ai · Making neural nets uncool again In 2018 they published a blog AI Ethics Resources with the intention of providing a practical guide and reference.
- ADSA's Data Science Ethos is a tool to help data scientists structure and operationalize their work in a way that accounts for the social and ethical responsibilities of the data science research process, including AI/ML.
- As AI/ML gains popularity as a toolset, organizations can provide proper training for the workforce and engage with stakeholders directly to ensure the responsible use of these evolving tools. U.S. government agencies have developed an active community of practices convened by the General Service Administration to provide a vibrant discussion on promising use cases and potential impacts and risks of various AI/ML technologies. Some organizations are offering tailored training for different groups (e.g., practitioners, downstream users, managers) or developing training materials aiming to improve proficiency in responsible AI/ML. In a recent training for managers of an agency office, the conversation on ethics and risk management received very positive feedback from participants and initiated conversations about further investment in the development of training materials that can directly engage with the general public.

Module 6: Considerations for Organizations, Institutions, Publishers, Societies, and Funders

Module 6 Focus

All scholarly organizations have a responsibility to lead in establishing and administering AI/ML ethics policies, including codes of conduct, principles, reporting methods, resolution processes, training, and other categories. Organizations should articulate values and design governance at levels above the individual, including fostering a culture around ethical AI/ML. In addition, enforcement of these responsibilities is needed to ensure ethical practices, and this lies across organizations and institutions.

Module 6 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.
- Ensure that there is sufficient oversight and enforcement of these principles along the lines of all unacceptable scientific practices or behavior.

Module 6 Vision

Organizations, institutions, publishers, societies, and funders work collaboratively to foster a culture around ethical AI/ML principles and responsibilities in research that builds trust and understanding, fosters community engagement, leads to positive outcomes, mitigates risks, and provides means to resolution or reconciliation when needed.

Module 6 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 and in helping implement the guidance above. 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 6 Principles

Organizations and Institutions

- **Align new and existing programs objectives and approaches** across the AI/ML Ethics Principles and Responsibilities.
- **Partner with other organizations** to help broaden awareness, education, adoption, and other engagement.
- **Include ethical AI/ML principles and responsibilities in courses and other ethical training.**
- **Include ethical AI/ML principles and responsibilities into grant processes.**
- Establish mechanisms to **intervene in appropriate ways** where AI/ML applications are causing harm in society.
- **Educate societal decision makers** on the value and limitations of AI/ML in research in order to enable responsible decisions based on AI/ML findings.

Societies and Communities

- **Provide workshops and education for society members** on the AI/ML Ethics Principles and Responsibilities.
- **Collectively provide governance of these AI/ML Ethics principles and responsibilities;** Support development and updates to leading practices related to the AI/ML Ethics Principles and Responsibilities.
- **Measure the effectiveness of the efforts** specific to implementing the AI/ML Ethics Principles and Responsibilities.
- **Adopt the AI/ML Ethical Principles and Responsibilities** into the organization's ethical guidance.

- **Promote the importance and adoption of the AI/ML Ethics Principles and Responsibilities** in relevant communities.
- **Ensure all affected communities are part of the development and updates** to the AI/ML Ethics Principles and Responsibilities.
- Establish mechanisms to **intervene in appropriate ways** where AI/ML applications are causing harm in society.
- Support AI/ML communities of practice, multi-stakeholder consortia, public-private partnerships and other collective efforts that **enable groups to accomplish together what they can't do individually with respect to AI/ML.**

Funders

- **Include the AI/ML Ethics Principles and Responsibilities** in reviewer guidelines and 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 for program and technical officers around ethical AI/ML principles.
- **Support governance** of AI/ML Ethics Principles and Responsibilities.
- **Provide expectations and supplementary funding for required time and travel** for training, assessment, stakeholder alignment, and professional development.
- Fund mechanisms to **intervene in appropriate ways** where AI/ML applications are causing harm in society.
- **Fund verification and validation studies** which are designed to replicate or reinforce AI/ML fundings, both to increase confidence in the original findings and to advance understanding on how to validate AI/ML findings.
- Support AI/ML communities of practice, multi-stakeholder consortia, public-private partnerships and other collective efforts that **enable groups to accomplish together what they can't do individually with respect to AI/ML.**

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 AI/ML Ethics Principles and Responsibilities, including around FAIR and CARE principles for data and software, recognizing contributions, reporting uncertainties, and in methods sections.
- **Follow leading practices** regarding data and software citations, including guidance for authors.

- **Publish negative or unexpected findings**. Well-defined, hypothesis driven work is valuable regardless of the outcome. AI/ML research that doesn't match expectations is as important as expected findings. These results can add clarity and understanding of AI/ML methods and reduce repeated, unfruitful efforts.

Additional supporting information on Module 6 principles:

- Establish a process that encourages and facilitates discussion, better understanding and trust building.
- Plan for an iterative process that starts with "timely good enough" vs. "late & perfect" or "rapid & wrong".
- Provide multiple ways for participants to give feedback and be prompt in responding to the feedback.
- Have methods in place to track individual contributions in order to provide accurate attribution and credit.
- Assess progress, make adjustments aligned with governance processes, and review assessment criteria periodically.
- Actively seek out new and diverse voices. Monitor and accommodate in 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).
- Balance philosophical exploration with practicalities.
- Establish governance before it is needed. Ensure governance structure, processes, decision making, feedback, and improvements support inclusiveness, understanding, and trust building.

Module 6 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.
- 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.
- Be prepared to intervene in appropriate ways where AI/ML applications are causing harm in society.

Module 6 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 (this document) to help support their researchers.
- 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.
- The Blueprint for an AI Bill of Rights issued by the Office of Science, Technology and Policy in the U.S. White House is an example of policy leadership in this issue.
- The SiteRite project is a machine learning model that provides a predictive model of alternative energy development potential (probability) of wind and solar in India. The idea is to provide a resource that developers, planners, natural resource and land managers, conservationists, etc. can use, when going through the scoping process of alternative energy projects. This tool provides a means of understanding potential impacts and conflicts around ecological resources. An NGO, Center for Science, Technology, and Policy (CSTEP), assisted in leveraging numerous socioeconomic indicators to develop metrics indicating impacts to communities. The web site provides quite a bit of

background information and was the product of working with many international stakeholders and foundations (with major portions of the funding coming from the MacArthur Foundation). (https://www.tncindia.in/what-we-do/siteright/)

- Organizations can consider creating roles, offices, or groups to support and coordinate the responsible use of AI/ML. Recently, NIST launched the Trustworthy and Responsible AI Resource Center, which will facilitate the implementation of, and international alignment with the recently published AI Risk Management Framework. The center provides a set of materials and training events that can be used to facilitate the responsible development of AI/ML models across the organization and adopted by other institutions. (see also: https://www.nist.gov/trustworthy-and-responsible-ai)

Implementation Considerations

This is an initial list of implementation questions to be used by researchers and scholarly organizations to determine how they will implement the principles described for each module. These are not intended to be exhaustive, but instead to begin conversations that lead to change in processes and policies.

Note that implementation will likely involve a combination of change initiatives led by researchers, funders, professional societies, policy makers, community groups, industry leaders, data repositories, and others. Conversations and actions prompted by these questions could involve any of these AI/ML stakeholders.

Module 1: Transparency, Documentating, and Reporting

- How do we convey quality information about the model?
- What is our standard practice to report the evaluation of the model following a defined evaluation metric or framework?
- 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?
- How will negative or un-expected results be reported?

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

- 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)

Module 3: Risk, Bias, and Effects

- What does the chief AI ethics officer do?
    - Note: It is important to provide strategic guidance across professional organizations; interface with funding agencies; and 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?
    - Note: It is important to 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; and take steps to address the root cause of the problem.
- What happens if we ran out funding but an issue has been identified?
    - Note: It is still important to minimize possible harm and notify the funding agency and users about the issue.
- What can funding agencies do to help mitigate harm from AI?
    - Note: We recommend funding agencies facilitate addressing any issues of AI risk and harm throughout the AI system lifecycle. We also 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: Participatory Methods

- 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?
    - Note: 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?
- How do we 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 how should the community be re-engaged?

Module 5: Outreach, Training, and Leading Practices

- 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?
- What are the indicators (antennas) for signs of success (evaluation of the community's progress)?

Module 6: Considerations for Organizations and Institutions, Publishers, Societies, and Funders

- How do we form timely, iterative mechanisms and 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.
- How can we best become aware of harm to society in order to assess and potentially take action?
- When harm to society is detected, what are responsible steps to take to mitigate or end the harm?

Glossary

AI Modeling: Using complex algorithms or layers of algorithms that interpret data and make decisions based on the data. A successful AI model can act as a surrogate for human expertise in any given use case, (intel.com)

AI Model Training: Model training involves processing large amounts of data through the AI model in iterative test loops, checking the results to ensure accuracy, and that the model is behaving as expected and desired. Engineers are on hand during this process to modify and improve the AI model as it learns. (intel.com)

AI Inferences: Deployment of the AI model into its real-world use case, where the AI model routinely infers logical conclusions based on available data. (intel.com)

CARE Principles: A set of governing principles guiding the use of Indigenous data (Carroll, 2020).

Consortia: A formal or informal assembly of individuals, groups, organizations, and other stakeholders operating laterally to accomplish together what they can't do separately. (The Consortia Century, 2023, forthcoming)

Documentating: All scientific research involves documentation of methods sufficient to support explainability and reproducibility. In the context of AI/ML the documentation centers on recording steps in the development and implementation of algorithms, the assembly and utilization of training data, the forms of validation utilized with the findings, and other relevant records of the research.

Equity and Equality: Equality involves treating everyone the same; Equity involves taking into account relevant differences. In some cultures and cases, equal treatment is seen as fair, while in other cultures and cases, equitable treatment is seen as fair. Both are relevant in the context of AI/ML since the mitigation of bias in the data typically calls for some form of equitable or equal treatment -- both are important and they are not the same.

Explainability: As suggested in the 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 (Phillips, 2021).

FAIR Guiding Principles: Four foundational principles—Findability, Accessibility, Interoperability, and Reusability—that serve to guide data producers and publishers to align with overall Open Science objectives (Wilkinson, 2016).

Inclusive: Valuing the contributions of diverse individuals, groups, organizations, and other parties to a project, organization, or community. An inclusive approach respects the individual identities of participants engaged in the work (AGU Diversity and Inclusion Strategic Plan, 2018).

Intentionality: Intentionality involves turning beliefs, knowledge, hopes, dreams, desires, and other "intentions" into action. Developing an AI/ML model involves intentional choices that must be documented in order to know what outcomes are intended and what are unintended.

Interpretability: Interpretability of an AI/ML model encompasses various components, but can be expressed as the extent to which a human can understand the cause of a decision the model makes (Miller, 2017).

Research Lifecycle: The research lifecycle refers to the entire research process from conception to completion, which can involve initial planning and design, execution of the design, identification and publication of results, data curation and sharing, as well as ongoing responsibility for the findings. In the context of AI/ML there are key steps in the process where the work is done by a machine without direction by the researcher, which has implications for the ongoing responsibilities associated with the research.

Machine Learning: Machine Learning is a subfield of artificial intelligence. ML is broadly defined as the capability of a machine to learn from data without being explicitly programmed. (MIT/Sloan Management Review)

Mindfulness: A choice and an unfolding approach that includes personal agency on the part of researchers and others to shape the organizations, societies, and other communities of which they are members.

Open Science: Open science, as defined by UNESCO, is an inclusive construct that combines various movements and practices aiming to make multilingual scientific knowledge openly available, accessible, and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society, and to open the processes of scientific knowledge creation, evaluation, and communication to societal actors beyond the traditional scientific community (UNESCO, 2021).

Participatory: Engaging people who will be affected by the use of AI/ML in research, ideally from the very beginning of the work and through all phases. Forms of participation can range from informal consultation, to focus groups or surveys, to formal meetings, to ongoing forums.

Replicability: Following the definition of National Academies of Sciences, when a new study is conducted and new data are collected replicability involves achieving the same or a similar results to earlier studies on the same scientific question (NASEM, 2019). Replicability is a map to lead other people to where you are now. (Contrast with explainability, which helps lead other people to understand why the model performs in a certain way).

Reproducibility: Following the definition of National Academies of Sciences, reproducibility refers to computational reproducibility— obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis (NASEM, 2019).

Responsible innovation: When research is designed and delivered for the benefit of all. A process of anticipating, reflecting, engaging, and acting that promotes socially desirable creativity and opportunity (UKRI Framework for Responsible Innovation). From Data 61/CSIRO - Responsible Innovation Platform: "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".

Stakeholders: Individuals, groups, organizations, and other actors with an interest or stake in the conduct and impact of AI/ML research. Note that the term "stakeholder" is used in the commercial context as a contrast with shareholders. The term has been criticized for connotations to individuals putting a stake in the ground to claim land that was previously held by first nations and others. In some contexts, the status of original holders of the land are recognised as "rights holders" having a greater claim then stakeholders, who, in turn, have more of a claim than "interested parties."

Stakeholder Alignment: The dynamic process by which individuals, groups, organizations, and other actors discuss common and competing interests to achieve sufficient understanding for collective action. This contrasts with stakeholder management and stakeholder engagement, both of which are from the point of view of one party either managing or engaging other stakeholders. In the contact of AI/ML there is a responsibility for researchers to engage in an ongoing process of alignment with relevant parties, not just trying to engage or manage them. (The Consortia Century, 2023, forthcoming)

Transparency: State of making information available for others to see what has been done (National Academies Press, 2019). Transparency includes documenting and reporting as a part of research methods. It is important to provide donvenient access to relevant information about a research project for those having a legitimate interest in that project.

TRUST Principles: Transparency, Responsibility, User focus, Sustainability, and Technology are the five TRUST Principles. Together they provide a common framework to facilitate discussion around implementing best practice in this critical area of digital preservation. (Lin, 2020).

References

A Blueprint for Equity and Inclusion in Artificial Intelligence. (2022). World Economic Forum. https://www3.weforum.org/docs/WEF_A_Blueprint_for_Equity_and_Inclusion_in_Artificial_Intelligence_2022.pdf

Aiwuyor, J. (n.d.). ARL Applauds NSF Open Science Investment. *Association of Research Libraries*. Retrieved March 10, 2023, from https://www.arl.org/news/arl-applauds-nsf-open-science-investment/

Allf, B. C., Cooper, C. B., Larson, L. R., Dunn, R. R., Futch, S. E., Sharova, M., & CAVALIER, D. (2022). Citizen Science as an Ecosystem of Engagement: Implications for Learning and Broadening Participation. *BioScience*, *72*(7), 651–663. https://doi.org/10.1093/biosci/biac035

Azari, A.R., J.W. Lockhart, M.W. Liemohn, X. Jia. (2020) Incorporating Physical Knowledge Into Machine Learning for Planetary Space Physics, Frontiers in Astronomy and Space Sciences. 7. DOI=10.3389/fspas.2020.00036. https://www.frontiersin.org/articles/10.3389/fspas.2020.00036

Babic, B., S. Gerke, T. Evgeniou & I.G. Cohen. 2021. Beware explanations from AI in health care: The benefits of explainable artificial intelligence are not what they appear, Science, 16 July, 373: 6552, 284-286. DOI: 10.1126/science.abg1834

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Caraballo-Vega, J.A. et al. (2022) Remote Sensing Powered Containers for Big Data and AI/ML Analysis: Accelerating Science, Standardizing Operations, IGARSS 2022 - 2022 **IEEE International Geoscience and Remote Sensing Symposium**, Kuala Lumpur, Malaysia, 2022, pp. 4034-4037, doi: 10.1109/IGARSS46834.2022.9883436.

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal*, *19*(1), Article 1. https://doi.org/10.5334/dsj-2020-043

Cox, J., Oh, E. Y., Simmons, B., Graham, G., Greenhill, A., Lintott, C., Masters, K., & Woodcock, J. (2018). Doing good online: The changing relationships between motivations, activity and retention among online volunteers. *Nonprofit and Voluntary Sector Quarterly*. http://journals.sagepub.com/home/nvs

Delgado, F., Yang, S., Madaio, M., & Yang, Q. (2021). *Stakeholder Participation in AI: Beyond "Add Diverse Stakeholders and Stir"* (arXiv:2111.01122). arXiv. https://doi.org/10.48550/arXiv.2111.01122

Fast.ai—Fast.ai—Making neural nets uncool again. (n.d.). Retrieved March 10, 2023, from https://www.fast.ai/

Federal Citizen Science Toolkit

Ferrario, A., Loi, M. & Viganò, E. (2020) In AI We Trust Incrementally: a Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions. Philosophy and Technology. 33, 523–539. https://doi.org/10.1007/s13347-019-00378-3

Findable Accessible Interoperable Reusable Open Science Research Coordination Networks (FAIROS RCN) (nsf22553) | NSF - National Science Foundation. (n.d.). Retrieved March 10, 2023, from https://www.nsf.gov/pubs/2022/nsf22553/nsf22553.htm

Framework for responsible research and innovation. (n.d.). Retrieved March 10, 2023, from https://www.ukri.org/about-us/epsrc/our-policies-and-standards/framework-for-responsible-innovation/

Gigerenzer, G. & W. Gaissmaier (2011) Heuristic Decision Making. *Annual Review of Psychology* 62:1, 451-482

Haupt, S. E., et. al., (2022) The History and Practice of AI in the Environmental Sciences. Bull. Amer. Meteor. Soc., 103, E1351–E1370, https://doi.org/10.1175/BAMS-D-20-0234.1

Hutson, M. (2018) Artificial intelligence faces reproducibility crisis: Unpublished code and sensitivity to training conditions make many claims hard to verify. Scinece. 16 Feb. 359(6377): 725-726. DOI: 10.1126/science.359.6377.725

Kale, A., T. Nguyen, F.C. Harris, C. Li, J. Zhang, X. Ma, (2023) Provenance documentation to enable explainable and trustworthy AI: A literature review. Data Intelligence; 5 (1): 139–162. doi: https://doi.org/10.1162/dint_a_00119

Leslie, D., Katell, M., Aitken, M., Singh, J., Briggs, M., Powell, R., Rincón, C., Perini, A., Jayadeva, S., & Burr, C. (2022). *Data Justice in Practice: A Guide for Developers*. https://doi.org/10.5281/zenodo.6428185

Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., De Giusti, M., L'Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M. E., Mokrane, M., Navale, V., Petters, J., Sierman, B.,

Sokolova, D. V., Stockhause, M., & Westbrook, J. (2020). The TRUST Principles for digital repositories. *Scientific Data*, *7*(1), Article 1. https://doi.org/10.1038/s41597-020-0486-7

MacIntyre, J., Medsker, L. & Moriarty, R. Past the tipping point?. *AI Ethics 1, 1–3 (2021). https://doi.org/10.1007/s43681-020-00016-1*

Making AI Inclusive: 4 Guiding Principles for Ethical Engagement. (2022, July 20). Partnership on AI. https://partnershiponai.org/paper/making-ai-inclusive-4-guiding-principles-for-ethical-engagement/

McGovern, A., I. Ebert-Uphoff, D. Gagne, & A. Bostrom. (2022). Why we need to focus on developing ethical, responsible, and trustworthy artificial intelligence approaches for environmental science. Environmental Data Science, 1, E6. https://doi.org/10.1017/eds.2022.5

McGovern, A., R. Lagerquist, D. John Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith. (2019) Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning. Bull. Amer. Meteor. Soc., 100, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1

McGovern, A. et. al., (2022) NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES). Bull. Amer. Meteor. Soc., 103, E1658–E1668, https://doi.org/10.1175/BAMS-D-21-0020.1

Miceli, M., Yang, T., Garcia, A. A., Posada, J., Wang, S. M., Pohl, M., & Hanna, A. (2022). *Documenting Data Production Processes: A Participatory Approach for Data Work* (arXiv:2207.04958; Version 2). arXiv. https://doi.org/10.48550/arXiv.2207.04958

Mitchell, M., S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I.D. Raji, T. Gebru. (2019) Model Cards for Model Reporting

FAT* '19: Conference on Fairness, Accountability, and Transparency, January 29--31, 2019, Atlanta, GA. https://doi.org/10.1145/3287560.3287596

Miller, T. (2017). *Explanation in Artificial Intelligence: Insights from the Social Sciences*. https://doi.org/10.48550/arXiv.1706.07269

National Academies of Sciences, Engineering, and Medicine. (2018). *Learning Through Citizen Science: Enhancing Opportunities by Design*. The National Academies Press. https://doi.org/10.17226/25183

Open Science Alliance. (n.d.). Retrieved March 10, 2023, from https://opensciallliance.github.io/

Participatory data stewardship: A framework for involving people in the use of data. (2021). Ada Lovelace Institute. https://www.adalovelaceinstitute.org/report/participatory-data-stewardship/

Phillips, P. J., Hahn, C., Fontana, P., Yates, A., Greene, K. K., Broniatowski, D., & Przybocki, M. A. (2021). *Four Principles of Explainable Artificial Intelligence. NIST Interagency/Internal Report (NISTIR),* National Institute of Standards and Technology, Gaithersburg, MD. https://doi.org/10.6028/NIST.IR.8312

Pineau, J., P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché-Buc & H. Larochelle, (2021). Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). The Journal of Machine Learning Research, 22(1), 7459-7478.

Reproducibility and Replicability in Science. (2019). National Academies Press. https://doi.org/10.17226/25303

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. *Hugging Face*. https://huggingface.co/blog/ethical-charter-multimodal

Science Mission Directorate. (2022). *SMD Policy Document SPD-41a*. NASA. https://science.nasa.gov/science-red/s3fs-public/atoms/files/SMD-information-policy-SPD-41a.pdf

Scientific Ethics Policies and Resources | AGU. (n.d.). Retrieved March 10, 2023, from https://www.agu.org/Learn-About-AGU/About-AGU/Ethics

Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2020). *Participation is not a Design Fix for Machine Learning* (arXiv:2007.02423). arXiv. https://doi.org/10.48550/arXiv.2007.02423

Stakeholder Alignment Collaborative: Joel Cutcher-Gershenfeld, Karen S. Baker, Nicholas Berente, Paul Arthur Berkman, Helen Burman, Pat Canavan, F. Alex Feltus, Alysia Garmulewicz, Alyson Gounden Rock, Ron Hutchins, John Leslie King, Christine Kirkpatrick, W. Christopher Lenhardt, Peter Levin, Spencer Lewis, Michael Maffie, Matthew Mayernik, Barbara Mittleman, Sarah Nusser, Beth Plale, Rajesh Sampath, Namchul Shin, Shelley Stall, John Towns, Susan Winter, Pips Veazey, Kim Zarecor. (2022), When Launching a Collaboration, Keep It Agile, *Stanford Social Innovation Review.* 20(2): 40-47.

Stakeholder Alignment Collaborative (Joel Cutcher-Gershenfeld, Ken Anderson, Karen S. Baker, Nicholas Berente, Helen M. Berman, Alan Blatecky, Chrsistine Borgman, Patrick Canavan, Alysia Garmulewicz, Bobby Clark, Alyson Gounden Rock, Brandon Grant, Michael Haberman, Ron Hutchins, John Leslie King, Spencer Lewis, Christine R. Kirkpatrick, John C Klensin, Kim Leary, W.

Christopher Lenhardt, Michael Maffie, Lauren Michael, Barbara B. Mittleman, Rajesh Sampath, Sarah Soroui, Namchul Shin, Miya Ward, Susan Winter, Kimberly E. Zarecor. 2023 (forthcoming). *The Consortia Century.* New York: Oxford University Press.

The Data Science Ethos—Operationalizing Ethics in Data Science. (n.d.). The Data Science Ethos. Retrieved March 10, 2023, from https://ethos.academicdatascience.org/

Transform to Open Science (TOPS). (n.d.). NASA Transform to Open Science (TOPS). Retrieved March 10, 2023, from https://science.nasa.gov/open-science/transform-to-open-science

UNESCO Recommendation on Open Science—UNESCO Digital Library. (2021). https://unesdoc.unesco.org/ark:/48223/pf0000379949

University of Washington Tech Policy Lab "Diverse Voices" project

Wan, H., M. Pervukhina, V. Shulakova, R. Beloborodov, R. Kempton, C. Delle Piane, M.B. Clennell, T. Nanjo, K. Miyoshi, (2022) Ai/Ml Approach to Lithology Quantification from Rock Chips Analysis. EarthDoc. (June). DOI: https://doi.org/10.3997/2214-4609.202210316

Wilford, J., S. Basak, R. Hassan. B. Moushall, L. McCalman, D. Steinberg, & F. Zhang, (2020) Uncover-ML: a machine learning pipeline for geoscience data analysis, Commonwealth of Australia (Geoscience Australia) 2020 eCat: 134466, doi: 10.11636/134466

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), Article 1. https://doi.org/10.1038/sdata.2016.18

Appendix A: Background on Materials Development

A set of two workshops in 2022, 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.
- The need for a leadership individual or group that can provide consultation and advice for researchers utilizing AI/ML, with the AGU Ethics Committee as a further resource.

A third workshop in February 2023 was held to invite stakeholder feedback on the draft prepared out of the first two workshops and writing groups. During this workshop, stakeholders edited and finalized the draft principles and responsibilities.

In addition to workshops, a stakeholder "pulse" survey of a cross section of geoscientists (n=118) 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%) opposed researchers using AI/ML in any way they chose – without attention to ethical standards or guidelines.

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.

Workshop 1 and 2 Participants

- Abby Azari, Space Sciences Lab, UC Berkeley, 0000-0002-8665-5459
- Abhinav Sharma, Cofounder Insight Browser,
- 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, Partnership on AI, 0000-0003-4985-4376
- Christine Kirkpatrick, San Diego Supercomputer Center, UC San Diego, 0000-0002-4451-8042
- Christopher Luwanga, NTU Singapore, 0000-0002-6723-5563
- Christopher Wirz, NCAR, 0000-0002-8990-5505
- Daisuke Nagai, Yale University, 0000-0002-6766-5942
- Daniel Crichton, Jet Propulsion Laboratory, Caltech, 0000-0002-5487-7719
- 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, Booz Allen Hamilton, 0000-0001-5981-9537
- 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, Pennsylvania State University, 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 Cutcher-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 Charleston, 0000-0002-4677-3995
- Kathleen Creel, Northeastern University, 0000-0001-7371-2680
- Kevin Coakley, CellLink Corporation
- Lance A. 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/Space Biosciences Division, NASA Ames Research Center, 0000-0001-9393-0861
- Lekha Patel, Sandia National Laboratories, 0000-0003-3508-0672
- Louis Barbier, NASA, 0000-0003-0378-6830
- Luis Vega, Meta
- 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 J. Molina, University of Maryland, 0000-0001-8539-8916
- Matthew Argall, University of New Hampshire
- Melanie Sharif, University of Colorado Boulder
- Micaela S. 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
- Richard Tran Mills, Argonne National Laboratory, 0000-0003-0683-6899
- Robert Morris, Koko
- Ryan McGranaghan, NASA Jet Propulsion Laboratory/Caltech, 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, Optum
- 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, 0000-0002-5701-760X
- Steven Crawford, NASA
- Susan J Winter, University of Maryland, 0000-0002-4524-0927
- Sylvain V. Costes, Space Biosciences Division, NASA Ames Research Center, 0000-0002-8542-2389
- Tae Wan Kim, Carnegie Mellon University
- Thomas Donaldson, The Wharton School, University of Pennsylvania
- Victoria Da Poian, NASA, 0000-0003-1175-3078
- Yuhan (Douglas) Rao, North Carolina State University / North Carolina Institute for Climate Studies, 0000-0001-6850-3403

Workshop 3 Participants

- Abby Azari, Space Sciences Lab, UC Berkeley, 0000-0002-8665-5459
- Adolfo Inza, Research in Volcano Seismology – Instituto Geofisico del Peru, 0000-0001-5381-9042
- Ayris Narock, NASA Goddard Space Flight Center, ADNET Systems, Inc., 0000-0001-6746-7455
- Barbara Thompson, NASA Goddard Space Flight Center, 0000-0001-6952-7343
- Bill Howe
- Brooks Hanson, American Geophysical Union, 0000-0001-6230-7145
- Caroline Coward, NASA Jet Propulsion Laboratory, 0000-0001-9848-5912
- Christopher Wirz, NCAR, 0000-0002-8990-5505
- Daniel Berrios, NASA Ames Research Center, 0000-0003-4312-9552
- Daniel Duffy, NASA Goddard Space Flight Center, 0000-0003-0155-5019
- De Canberra

- Yuhan (Douglas) Rao, North Carolina State University / North Carolina Institute for Climate Studies, 0000-0001-6850-3403
- Ekaterina Verner, NASA, 0009-0006-3196-5552
- Ge Peng, University of Alabama Huntsville, 0000-0002-1986-9115
- Jeffrey S. Evans, The Nature Conservancy and University of Wyoming, 0000-0002-5533-7044
- Joel Cutcher-Gershenfeld, Brandeis University, 0000-0001-7659-7024
- Kristina Vrouwenvelder, American Geophysical Union
- Lance A. Waller, Emory University, 0000-0001-5002-8886
- Laura Lyon, American Geophysical Union
- Lauren M. Sanders, Blue Marble Space Institute for Science/Space Biosciences Division, NASA Ames Research Center, 0000-0001-9393-0861
- Manil Maskey, NASA, 0000-0002-5087-6903
- Matthew Argall, University of New Hampshire
- Micaela S. Parker, Academic Data Science Alliance (ADSA), 0000-0003-1007-4612
- Mike Little, WordPress
- Noah Conley
- Ryan McGranaghan, NASA Jet Propulsion Laboratory/Caltech, 0000-0002-9605-0007
- Ryan T. Scott, KBR/Space Biosciences Division, NASA Ames Research Center, 0000-0003-0654-5661
- Shelley Stall, American Geophysical Union, 0000-0003-2926-8353
- Sylvain V. Costes, Space Biosciences Division, NASA Ames Research Center, 0000-0002-8542-2389
- Thomas Donaldson, The Wharton School, University of Pennsylvania
- Tony Boese

Appendix B: 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 survey 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 are 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)
- 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)

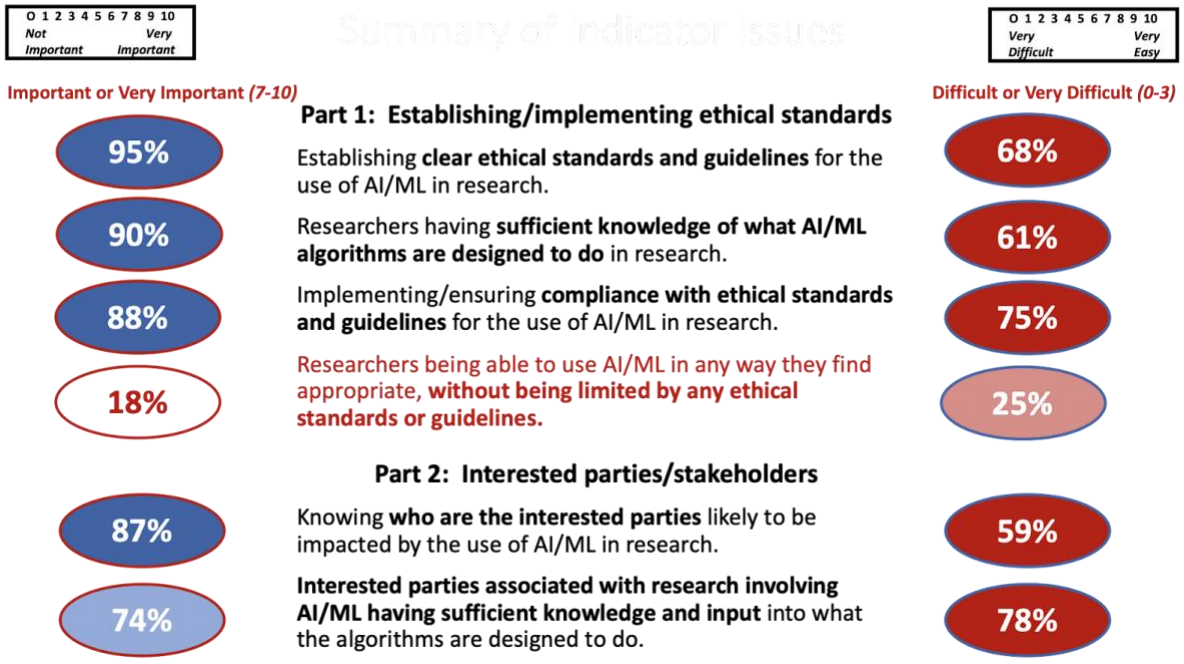
Please indicate your years of experience

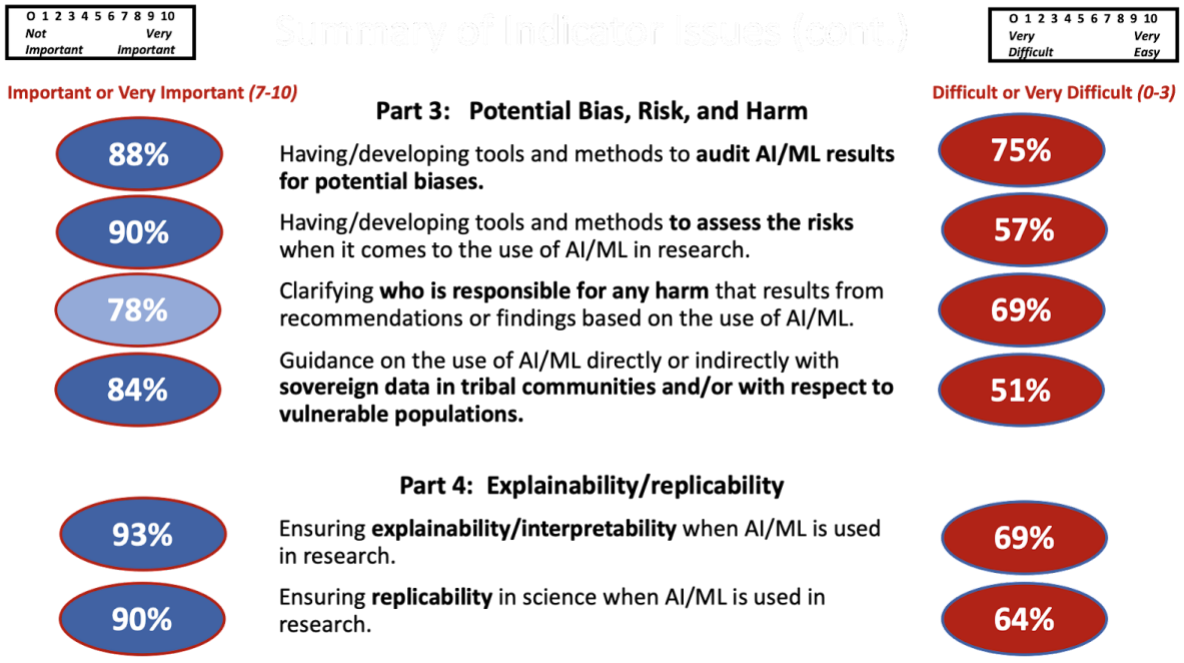
- 1 year or less 1.7% (n=2)
- 2-4 years 4.2% (n=5)
- 5-10 years 16.1% (n=19)
- 11-20 years 21.2% (n=25)
- 21-30 years 25.4% (n=30)
- Over 30 years 29.7% (n=35)
- It's complicated 1.7% (n=2)

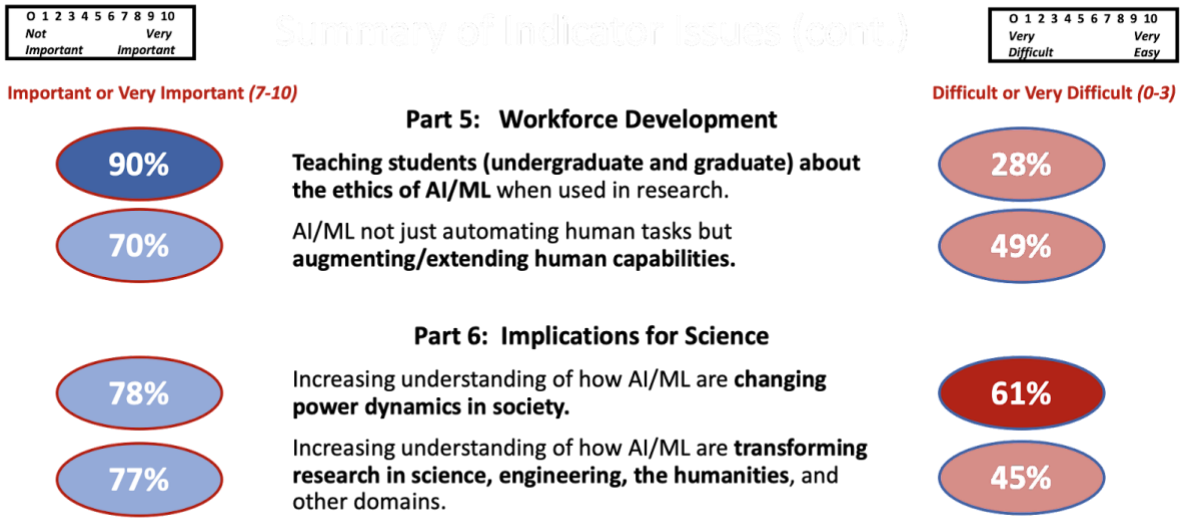
What is your gender identity?

- 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)

Pulse Results for "Indicator" Issues







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."

Appendix C: 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.

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.
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.

NIST AI Risk Management Framework 1.0

1. **Govern** is a cross-cutting function that is infused throughout AI risk management and enables the other functions of the process.
    a. Policies, processes, procedures, and practices across the organization related to the mapping, measuring, and managing of AI risks are in place, transparent, and implemented effectively.
    b. Accountability structures are in place so that the appropriate teams and individuals are empowered, responsible, and trained for mapping, measuring, and managing AI risks.
    c. Workforce diversity, equity, inclusion, and accessibility processes are prioritized in the mapping, measuring, and managing of AI risks throughout the lifecycle.
    d. Organizational teams are committed to a culture that considers and communicates AI risk.
    e. Processes are in place for robust engagement with relevant AI actors.

f. Policies and procedures are in place to address AI risks and benefits arising from third-party software and data and other supply chain issues.

2. The **Map** function establishes the context to frame risks related to an AI system.
   a. Context is established and understood.
   b. Categorization of the AI system is performed.
   c. AI capabilities, targeted usage, goals, and expected benefits and costs compared with appropriate benchmarks are understood.
   d. Risks and benefits are mapped for all components of the AI system including third-party software and data.
   e. Impacts to individuals, groups, communities, organizations, and society are characterized.
3. The **Measure** function employs quantitative, qualitative, or mixed-method tools, techniques, and methodologies to analyze, assess, benchmark, and monitor AI risk and related impacts.
   a. Appropriate methods and metrics are identified and applied.
   b. AI systems are evaluated for trustworthy characteristics.
   c. Mechanisms for tracking identified AI risks over time are in place.
   d. Feedback about efficacy of measurement is gathered and assessed.
4. The **Manage** function entails allocating risk resources to mapped and measured risks on a regular basis and as defined by the GOVERN function.
   a. AI risks based on assessments and other analytical output from the MAP and MEASURE functions are prioritized, responded to, and managed.
   b. Strategies to maximize AI benefits and minimize negative impacts are planned, prepared, implemented, documented, and informed by input from relevant AI actors.
   c. AI risks and benefits from third-party entities are managed.
   d. Risk treatments, including response and recovery, and communication plans for the identified and measured AI risks are documented and monitored regularly.

GAO AI Accountability Framework

An Accountability Framework for Federal Agencies and Other Entities

Fast Facts

As a nation, we have yet to grasp the full benefits or unwanted effects of artificial intelligence. AI is widely used, but how do we know it's working appropriately?

This report identifies key accountability practices—centered around the principles of governance, data, performance, and monitoring—to help federal agencies and others use AI responsibly. For example, the governance principle calls for users to set clear goals and engage with diverse stakeholders.

To develop these practices, we held a forum on AI oversight with experts from government, industry, and nonprofits. We also interviewed federal inspector general officials and AI experts.

What GAO Found

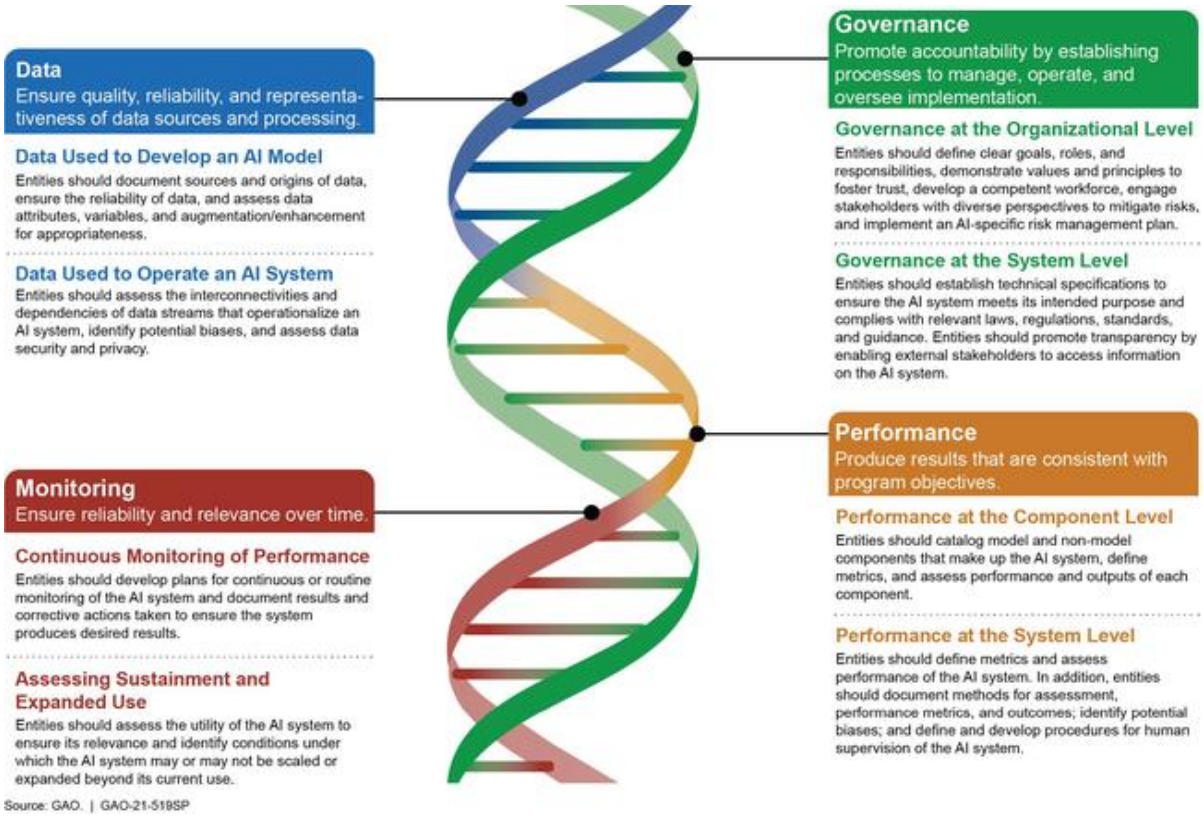
To help managers ensure accountability and responsible use of artificial intelligence (AI) in government programs and processes, GAO developed an AI accountability framework. This framework is organized around four complementary principles, which address governance, data, performance, and monitoring. For each principle, the framework describes key practices for federal agencies and other entities that are considering, selecting, and implementing AI systems. Each practice includes a set of questions for entities, auditors, and third-party assessors to consider, as well as procedures for auditors and third- party assessors.

Why GAO Developed This Framework

AI is a transformative technology with applications in medicine, agriculture, manufacturing, transportation, defense, and many other areas. It also holds substantial promise for improving government operations. Federal guidance has focused on ensuring AI is responsible, equitable, traceable, reliable, and governable. Third-party assessments and audits are important to achieving these goals. However, AI systems pose unique challenges to such oversight because their inputs and operations are not always visible.

GAO's objective was to identify key practices to help ensure accountability and responsible AI use by federal agencies and other entities involved in the design, development, deployment, and continuous monitoring of AI systems. To develop this framework, GAO convened a Comptroller General Forum with AI experts from across the federal government, industry, and nonprofit sectors. It also conducted an extensive literature review and obtained independent validation of key practices from program officials and subject matter experts. In addition, GAO interviewed AI subject matter experts representing industry, state audit associations, nonprofit entities, and other organizations, as well as officials from federal agencies and Offices of Inspector General.

Artificial Intelligence (AI) Accountability Framework


Data
Ensure quality, reliability, and representativeness of data sources and processing.
Data Used to Develop an AI Model
Entities should document sources and origins of data, ensure the reliability of data, and assess data attributes, variables, and augmentation/enhancement for appropriateness.
Data Used to Operate an AI System
Entities should assess the interconnectivities and dependencies of data streams that operationalize an AI system, identify potential biases, and assess data security and privacy.
Monitoring
Ensure reliability and relevance over time.
Continuous Monitoring of Performance
Entities should develop plans for continuous or routine monitoring of the AI system and document results and corrective actions taken to ensure the system produces desired results.
Assessing Sustainment and Expanded Use
Entities should assess the utility of the AI system to ensure its relevance and identify conditions under which the AI system may or may not be scaled or expanded beyond its current use.
Source: GAO. | GAO-21-519SP
Governance
Promote accountability by establishing processes to manage, operate, and oversee implementation.
Governance at the Organizational Level
Entities should define clear goals, roles, and responsibilities, demonstrate values and principles to foster trust, develop a competent workforce, engage stakeholders with diverse perspectives to mitigate risks, and implement an AI-specific risk management plan.
Governance at the System Level
Entities should establish technical specifications to ensure the AI system meets its intended purpose and complies with relevant laws, regulations, standards, and guidance. Entities should promote transparency by enabling external stakeholders to access information on the AI system.
Performance
Produce results that are consistent with program objectives.
Performance at the Component Level
Entities should catalog model and non-model components that make up the AI system, define metrics, and assess performance and outputs of each component.
Performance at the System Level
Entities should define metrics and assess performance of the AI system. In addition, entities should document methods for assessment, performance metrics, and outcomes; identify potential biases; and define and develop procedures for human supervision of the AI system.

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.
    1. **Transparency**: Notifying the subject about what you will do with the data and who can potentially access the data.
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.

Blueprint for an AI Bill of Rights

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. 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 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 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 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 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/

Appendix D: AI/ML Guidance from ChatGPT

There are a number of publicly available AI/ML tools that are drawing considerable attention. There is still a great deal of work needed to guide the public use of the tools, addressing issues of privacy, intellectual property, originality, citations for use, and other matters. Thus, it was with some caution that we asked one of the tools, ChatGPT (Chat Generative Pre-trained Transformer) the following question: "Please provide guidance on the ethical use of AI/ML in the geosciences (with citations)." When asked the same question a second time (a few days later from the same computer) the response was similar, but not identical, and the references were different.

We note three things about the response. First, the five themes highlighted correspond nicely to the themes in this report, which we take as a form of validation both for the report and for ChatGPT. Second, the addition of citations reveals an interesting mix of sources -- conference presentations, leading journals, less well-known journals, etc. -- with most coming from outside of the geosciences and some that could not be found on Google Scholar. Third, the focus is primarily from the perspective of researchers, with less of a focus on the implications for professional societies, government agencies, and other organizations or consortia. This is, of course, reflective of the data on which ChatGPT draws. In this way, the use of AI/ML in this case reveals gaps in the literature as well as common themes. Here is what was generated:

> The field of geosciences has increasingly turned towards artificial intelligence and machine learning to aid in understanding complex earth systems, but with this shift comes the need for ethical considerations in the use of AI/ML. Here are some guidelines and considerations for the ethical use of AI/ML in geosciences, supported by relevant literature:
>
> 1. **Fairness and Bias:** AI/ML models must be designed and trained to ensure fairness and minimize bias, particularly when used in decision-making processes. Fairness refers to the absence of systematic disparities in the outcomes of a model, while bias refers to the presence of unfair treatment towards certain groups or individuals. Careful consideration must be given to the data used to train models, and it is important to actively monitor and address potential biases that may arise. (Gebru et al., 2018; Buolamwini & Gebru, 2018; Crawford et al., 2020)
> 2. **Transparency and Interpretability:** AI/ML models should be transparent and interpretable, meaning that the reasoning behind the model's decisions should be clear and understandable. This is particularly important in the geosciences, where decisions based on AI/ML models can have significant consequences. (Lipton, 2018; Mittelstadt et al., 2019)
> 3. **Privacy and Security:** AI/ML models must be designed and implemented with appropriate safeguards to protect individual privacy and data security. This includes

careful consideration of data storage and sharing, as well as measures to prevent unauthorized access or misuse of data. (Bertino et al., 2020)

4. **Accountability and Governance:** AI/ML models must be accountable to ensure that they are used ethically and effectively. This includes establishing appropriate governance structures and oversight mechanisms to ensure that the models are used responsibly and in compliance with relevant laws and regulations. (Floridi et al., 2018; Jobin et al., 2019)
5. **Community Engagement:** AI/ML researchers and practitioners should engage with impacted communities to ensure that their needs and concerns are addressed. This includes involving local stakeholders in the design and implementation of AI/ML projects, and being transparent about the potential impacts of these projects. (Scheuermann et al., 2019)

References:

- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 77-91.
- Crawford, K., et al. (2020). AI Now Report 2019. AI Now Institute.
- Floridi, L., et al. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, 28(4), 689-707.
- Gebru, T., et al. (2018). Datasheets for Datasets. arXiv preprint arXiv: 1803.09010.
    - *Note: This is a preprint; the actual article is 2021.*
- Jobin, A., et al. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.
- Lipton, Z. C. (2018). The mythos of model interpretability. arXiv preprint ar Xiv: 1606.03490.
- Mittelstadt, B. D., et al. (2019). Explaining explanations in AI. Proceedings of the Conference on Fairness, Accountability, and Transparency, 279-288.
- Scheuermann, A., et al. (2019). Citizen engagement in the governance of AI and data: Opportunities and challenges. Paris: Organisation for Economic Co-operation and Development.
    - *Note: Citation not found in Google Scholar*