NASA/TM-20210012886

NASA Framework for the Ethical Use of Artificial Intelligence (AI)

Edward McLarney
Langley Research Center, Hampton, Virginia

Yuri Gawdiak
NASA HQS, Washington, District of Columbia

Nikunj Oza
Ames Research Center, Moffett Field, California

Chris Mattman
Jet Propulsion Laboratory, Pasadena, California

Martin Garcia
Johnson Space Center, Houston, Texas

Manil Maskey
Marshall Space Flight Center, Huntsville, Alabama

Scott Tashakkor
Marshall Space Flight Center, Huntsville, Alabama

David Meza
NASA HQs, Washington, District of Columbia

John Sprague
NASA HQS, Washington, District of Columbia

Phyllis Hestnes
Goddard Space Flight Center, Greenbelt, Maryland

Pamela Wolfe
NSSC, Stennis Space Center, Mississippi

James Illingworth
NASA HQS, Washington, District of Columbia

Vikram Shyam
Glenn Research Center, Cleveland, Ohio

Paul Rydeen
NSSC, Stennis Space Center, Mississippi

Lorraine Prokop
Johnson Space Center, Houston, Texas

Latonya Powell
Marshall Space Flight Center, Huntsville, Alabama

Terry Brown
Marshall Space Flight Center, Huntsville, Alabama

Warnecke Miller
Johnson Space Center, Houston, Texas

Claire Little
NASA HQS, Washington, District of Columbia

April 2021

NASA STI Program Report Series

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NTRS Registered and its public interface, the NASA Technical Reports Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA Programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include organizing and publishing research results, distributing specialized research announcements and feeds, providing information desk and personal search support, and enabling data exchange services.

For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at http://www.sti.nasa.gov

- Help desk contact information:

https://www.sti.nasa.gov/sti-contact-form/
and select the "General" help request type.

NASA/TM-20210012886

NASA Framework for the Ethical Use of Artificial Intelligence (AI)

Edward McLarney
Langley Research Center, Hampton, Virginia

Yuri Gawdiak
NASA HQS, Washington, District of Columbia

Nikunj Oza
Ames Research Center, Moffett Field, California

Chris Mattman
Jet Propulsion Laboratory, Pasadena, California

Martin Garcia
Johnson Space Center, Houston, Texas

Manil Maskey
Marshall Space Flight Center, Huntsville, Alabama

Scott Tashakkor
Marshall Space Flight Center, Huntsville, Alabama

David Meza
NASA HQs, Washington, District of Columbia

John Sprague
NASA HQS, Washington, District of Columbia

Phyllis Hestnes
Goddard Space Flight Center, Greenbelt, Maryland

Pamela Wolfe
NSSC, Stennis Space Center, Mississippi

James Illingworth
NASA HQS, Washington, District of Columbia

Vikram Shyam
Glenn Research Center, Cleveland, Ohio

Paul Rydeen
NSSC, Stennis Space Center, Mississippi

Lorraine Prokop
Johnson Space Center, Houston, Texas

Latonya Powell
Marshall Space Flight Center, Huntsville, Alabama

Terry Brown
Marshall Space Flight Center, Huntsville, Alabama

Warnecke Miller
Johnson Space Center, Houston, Texas

Claire Little
NASA HQS, Washington, District of Columbia

National Aeronautics and
Space Administration
Mary W. Jackson NASA Headquarters
Washington, DC

April 2021

Acknowledgments

The team would like to thank reviewers from the following communities: NASA Artificial Intelligence and Machine Learning community, the NASA Data Governance Board, Carnegie Mellon's Software Engineering Institute, and other key NASA transformation leaders. Your insights and feedback have helped improve the final product greatly.

The use of trademarks or names of manufacturers in this report is for accurate reporting and does not constitute an official endorsement, either expressed or implied, of such products or manufacturers by the National Aeronautics and Space Administration.

Available from:

NASA STI Program / Mail Stop 148
NASA Langley Research Center
Hampton, VA 23681-2199
Fax: 757-864-6500

Table of Contents

List of Figures

Introduction

Isaac Asimov brought the ethics of Artificial Intelligence (AI) into the public eye as early as 1942 with his short story “Runaround.” First appearing in an issue of *Astounding Science Fiction*[1] and later more famously as part of the anthology work *I, Robot*[2], his three laws of robotics are:

- **First Law:** A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- **Second Law:** A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
- **Third Law:** A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

These principles continue to be relevant to discussion of AI ethics in the 21st century, and developing policy directives for ethical AI will be a complex and dynamic undertaking. In particular, the nature and maturity of artificial intelligence are still evolving, and as such society will likely encounter “unknown unknowns” as AI systems are implemented and operated.

Even when considering “simple” statements such as Asimov’s First Law, there are hidden ethical complexities which can lead to unintended consequences. For instance, an intelligent robot applying a vaccine to a human could break the First Law: a needle penetrating a human’s skin would technically cause injury, including the potential for drawing blood, and a small percentage of humans experience adverse physical reactions to any given vaccine. What is missing from the First Law as it is written are critical temporal, contextual, and tradeoff considerations that allow for exceptions for the greater good and balance short-term versus long-term benefit. These tradeoffs rely on the careful codification and interpretation of values and beliefs which are not universally agreed upon, thus making large scale implementation challenging.

In the past several years, artificial intelligence has blossomed from previously niche or high investment applications to become plentiful, affordable, and powerful – and ready to be applied across nearly any task humans perform. Private industry is embedding AI in many computers and mobile devices, with an exponential explosion in power, capability, and span of application. Capabilities include image recognition, speech recognition, anomaly detection, pattern recognition, recommender systems, text analytics, sentiment analysis, streaming data analysis, and more. Many organizations have recognized the need to create policies, principles, and guidelines for the ethical use of AI because, as the saying goes, “With great power comes great responsibility.”

In 2019, a representative poll across NASA revealed over one hundred agency applications of AI in the past three years, with hundreds of AI projects planned across various missions, centers, and mission support activities from 2020 to 2022 and beyond. In November and December of 2020, the White House and Office of Management and Budget (OMB) published guidance[3] regarding AI principles, policy, and governance. As an enthusiastic and forward-leaning AI adopter, NASA must create and apply an evolving, living set of AI policies, principles, and guidelines to provide AI practitioners an ethical framework for their work.

Executive Summary

Artificial intelligence (AI) is growing rapidly on a global scale, and NASA has begun leveraging AI for a wide variety of mission applications and supporting functions. Numerous organizations in government and industry have recognized the need to guide responsible, ethical use of AI, and have created handbooks, principles, or frameworks for their communities. The NASA CIO recognized the need for NASA to create initial AI guidelines, and directed creation of this framework in concert with the wider digital transformation community.

The initial framework for NASA's ethical use of AI includes considerations applicable to today's simple Artificial Narrow Intelligence (ANI), as well as future human-level Artificial General Intelligence (AGI), and beyond to Artificial Super Intelligence (ASI). Considerations also include the ways humans may interact with machines, from using them as tools to augmenting humans with implants, to more speculative further-term topics such as the merging or melding of human and machine. This NASA framework draws from principles and frameworks of many other leading organizations, relating them to NASA's specific needs to provide an initial set of six ethical AI principles:

Fair. AI systems must include considerations regarding how to treat people, including refining solutions to mitigate discrimination and bias, preventing covert manipulation, and supporting diversity and inclusion.

Explainable and Transparent. Solutions must clearly state if, when, and how an AI system is involved, and AI logic and decisions must be explainable. AI solutions must protect intellectual property and include risk management in their construction and use. AI systems must be documented.

Accountable. Organizations and individuals must be accountable for the systems they create, and organizations must implement AI governance structures to provide oversight. AI developers should consider potential misuse or misinterpretation of AI-derived results (intentional or otherwise) and take steps to mitigate negative impact.

Secure and Safe. AI systems must respect privacy and do no harm. Humans must monitor and guide machine learning processes. AI system risk tradeoffs must be considered when determining benefit of use.

Human-Centric and Societally Beneficial. AI systems must obey human legal systems and must provide benefits to society. At the current state of AI, humans must remain in charge, though future advancements may cause reconsideration of this requirement.

Scientifically and Technically Robust. AI systems must adhere to the scientific method NASA applies to all problems, be informed by scientific theory and data, robustly tested in implementation, well-documented, and peer reviewed in the scientific community.

While the majority of this framework focuses on contemporary ethical AI considerations, longer-term implications are considered in brief in a supplementary appendix. Though far-future AI possibilities may seem like science fiction today, the importance of ethics in this area requires that we extend our thinking even into future hypotheticals. These longer-term considerations include: potential machine sentience; human-machine relationships, both integrated and independent; and more.

Finally, this framework provides initial recommendations for NASA governance and advice related to AI, and questions for AI practitioners to consider during their work; it covers both employing AI in an ethical manner, and eventually creating AI which behaves ethically itself. With AI as a promising, powerful and turbulent emerging field, this document and NASA's approach to AI will require iterative review and adaptation in the years ahead. To address questions of roles and responsibilities, the framework will be accompanied by a separate Ethical AI Review Board charter and an emerging policy document.

Scope and Interdependencies

The majority of this framework will focus on capability that is currently available or expected to be available in the near future (Artificial Narrow Intelligence and below), as these systems are active and proliferating and thus require immediate guidance. A small fraction of this framework will focus on considerations for human-level AI and beyond (Artificial General Intelligence and Artificial Super Intelligence) to set the stage for successful, peaceful, and potentially symbiotic coexistence between humans and machines. These guidelines are subject to adjustment as technology advances and to account for any changes in social expectations and law. These guiding principles will only remain functional if they reflect current laws and cultural norms; as such, this document and any follow-on guidance will need to be periodically updated as AI matures.

This ethical AI framework has interdependencies with NASA policies and procedures for software development, human resources, information technology, and more. These interdependencies are indicated throughout the document, and specific references will be listed in the appendices. Rather than creating additional processes, this framework recommends that ethical AI principles be included in existing review cycles (e.g., software reviews, project reviews, etc.). While AI matures, and NASA establishes AI guidance mechanisms, an AI advisory body must be established to participate in existing governance processes and provide a new AI review function as needed, with the intent of accelerating and guiding AI development and use.

Definitions and Assumptions

This section provides working definitions foundational to NASA's discussion and adoption of artificial intelligence. Additional definitions of terms and concepts can be found in Appendix E. Note that in this document, the word "must" notes strong positive recommendation; the directive term "shall" is reserved for any formal policy documents which may accompany this document.

As a starting point, NASA's defines ***artificial intelligence*** as follows: "Any computerized capability to perceive, reason, learn, and act." NASA further distinguishes three categories[4] or capability levels of AI as follows:

- **Artificial Narrow Intelligence (ANI):** Artificial intelligence that operates at less than human ability. As of publication, all AI is ANI. ANI may be faster or better than humans at narrow, specific tasks but it does not generalize, nor does it understand the larger situation like a human would.
- **Artificial General Intelligence (AGI):** Human-level artificial intelligence. This has not yet been achieved. Popular belief is that AGI represents a tipping point in AI capability and human-machine interactions, relationships, and teaming.

- **Artificial Super Intelligence (ASI):** Artificial intelligence which surpasses human capability.

While not technically artificial "intelligence," many recent advances have been made in areas of ***intelligent automation*** such as robotic process automation (RPA). While this may fall short of a truly intelligent system, we include it in this framework because automation executed at massive scale and speed could present similar issues as those created by truly intelligent systems. While initial automation can be performed as rote tasks, autonomy will grow as a combination of rote (automation) and learned (AI) actions.

Note that while some AI systems may proceed from ANI to AGI to ASI, many may begin and end their life cycle within only one of these categories. NASA does not assume all AI systems will eventually become super-intelligent. In fact, it is likely that relatively few early AI systems will climb this capability spectrum.

From a philosophical perspective, a core ***value system*** is essential to serve as the framework for all other aspects of ***ethics***. Values are "individual beliefs that motivate people to act one way or another [and] serve as a guide for human behavior."[4] People commonly adopt the values they are raised with, and values can vary widely among individuals and groups. Regardless of the content of a particular value system, the system provides a means of judging where behaviors and actions fall across a spectrum of "goodness," both for the individual and for society. This paper assumes a value system consistent with the core stated values of the Federal Government of the United States and NASA. At the national level, these values include liberty, opportunity, and equality, while NASA's core values are safety, integrity, teamwork, inclusion and excellence[6].

Machine learning techniques and capabilities are the core building blocks of AI systems. Neural networks, deep learning, natural language processing, supervised learning, unsupervised learning, transfer learning, regression analysis, classification by type, clustering, dimensionality reduction, ensemble methods, word embeddings, and more are all examples of machine learning techniques that contribute to AI. Additional AI building blocks include robust training data, subject matter expert guidance, the combination of multiple machine learning capabilities, and end-to-end analysis pipelines. These specific capabilities are not discussed further in this paper, which focuses instead on the higher-level AI capabilities which build upon core machine learning building blocks.

For the purposes of this document, we refer to approximate ***time horizons*** as follows: near-term is defined as one to five years; mid-term is defined as five to ten years; far-term is defined as ten or more years or any significant step function increase among ANI, AGI, and ASI.

Finally, the team considered a spectrum of ***human-AI integration***, from ***autonomous systems*** operating in accordance with rote instructions, AI learning algorithms, and/or with varying levels of human-in-the-loop, to eventual human-AI combinations adapted to space or high-altitude aeronautics environments. Simple autonomy and automation exist today and are advancing rapidly, while ideas such as implanted cybernetic augmentations are a possibility on the mid-term horizon. Theoretical possibilities, while not guaranteed to materialize, must not be taken lightly or taken for granted, and therefore should be included in discussions of ethical frameworks.

Benchmarking and References

Prior to developing a NASA-specific ethical AI framework, the team conducted research and benchmarking to reveal other emergent policies, principles, and guidelines for ethical use of AI that are in place at other government agencies and in private industry. In particular, frameworks developed by the Department of Defense's Defense Innovation Board, Gartner, and the American Council for Technology Industry Advisory Council (ACT-IAC) provided particularly good foundations, as summarized below:

Defense Innovation Board	Gartner	ACT-IAC
Responsible	Secure and Safe	Bias
Equitable	Fair	Fair
Traceable	Explainable and Transparent	Transparent
Governable	Accountable	Responsible
Reliable	Human-centric, Societally Beneficial	Interpretable

As the above table indicates, each framework covers approximately the same solution space, with minor industry- or organization-specific changes as appropriate. We selected the Gartner framework as our baseline because its elements generalized well, avoided overlap among themselves, and could succinctly map to NASA's needs. To ensure a robust and accurate product, the team cross-checked its work against the two remaining frameworks throughout its analysis. The Department of Defense principles are described in more detail in Appendix A. The basic Gartner guidelines are expanded below:

Attribute	Description
Secure and Safe	AI systems must respect privacy and do no harm. Humans must monitor and guide machine learning processes. AI system risk tradeoffs must be considered when determining benefit of use.
Fair	AI systems must include considerations of how to treat people, including refining solutions to mitigate discrimination and bias, preventing covert manipulation, and supporting diversity and inclusion.
Explainable and Transparent	Solutions must clearly state if, when, and how an AI system is involved, and AI logic and decisions must be explainable. AI solutions must protect intellectual property and include risk management in their construction and use. AI systems must be documented.
Accountable	Organizations and individuals must be accountable for the systems they create, and organizations must implement AI governance structures to provide oversight.
Human-centric, Societally Beneficial	AI systems must obey human legal systems and must provide benefits to society. At the current state of AI humans must remain in charge, though future advancements may cause reconsideration of this requirement.

In addition to the above principles, Gartner proposes an organizational ***AI maturity model*** focused on organizational adoption of AI techniques, rather than the maturity of the actual AI systems. This model is useful in considering NASA's state of initial AI adoption from early AI investment, to experimentation, to limited production use. And on the near-mid horizon, NASA is headed to pervasive use of AI across all business functions, with AI becoming part of NASA's business DNA in the not-too-distant future. The Gartner AI maturity model is as follows:

Level 1: Adoption	Early AI interest with risk of overhyping
Level 2: Active	AI experimentation, mostly in a data science context
Level 3: Operational	AI in production, creating value for process or product innovations
Level 4: Systemic	AI is pervasively used for processes, transformation, and new business models
Level 5: Transformational	AI is part of the business DNA

The organizational AI maturity model is more useful for business decisions regarding AI than for directly impacting ethical AI; however, the more mature an AI ecosystem is, the larger and more powerful it becomes, thus necessitating additional rigor in AI ethics. More attention must be paid to more powerful, more-connected AI. Practitioners conducting early stage research in controlled environments with simpler AI must adhere to the principles here, with flexibility allowed for experimentation without undue governance burden. Even so, the bottom line regarding organizational AI maturity is this: the more AI an organization adopts, the more ethics considerations must be employed.

Based on input gathered from a wide variety of NASA AI practitioners, NASA's current state of AI maturity is a mix of levels 1 and 2. With NASA's digital transformation initiative and other ongoing adoption, the Agency is expected to progress across levels 3, 4, and 5 over the coming decade. To maximize use of AI, NASA will progress through these phases while adhering to Gartner's general AI Ethics Guidelines, adapted for NASA use, as discussed below. This expanded framework, along with the emergent AI review board, is intended to enable productive, rapid, useful, and safe AI adoption across NASA. The intent is to guide and inform safe adoption of AI, while maximizing NASA's benefit from AI.

Expanding Gartner to Meet NASA's Needs

The existing Gartner framework provides a valuable foundation for NASA's AI policy efforts. To ensure NASA's unique institutional values and requirements are captured, however, the group wishes to expand on several principles and add an additional one.

As society wrestles with discrimination and bias, NASA's AI systems must be kept relevant and updated. AI must support human efforts to encourage positive aspects and principles, as Gartner indicates with its ***Fair*** principle, but NASA as an agency must expand these efforts into areas such as diversity and inclusion. Some differentiation or adjudication of specific personnel may be a desirable feature of AI systems. For example, if identifying personnel for access to NASA resources, such systems must include relevant factors but must mitigate unfair biases, encouraging diversity and inclusion instead. Deliberate consideration of fairness is especially important with AI systems because the unintentional inclusion of bias can be proliferated quickly by powerful AI. It is important to note these considerations are focused on ethical, non-discriminatory bias. From a statistician's perspective, selected statistical bias may be tolerated and may even be useful if documented, understood, and made transparent.

Further discussion of Gartner's ***Explainable and Transparent*** principle is also useful for NASA. For systems that have learning as part of their functional capabilities, the following items shall be implemented:

- Data sets used for learning for operational systems shall be certified for accuracy, appropriate scope, and exclusion of unintended bias. As an example, facial recognition systems must work equally well across genders and races.
- Changes in applicable human laws, beliefs, and values shall trigger reviews of learning systems to update and replace data sets and/or learning algorithms and functions as appropriate.
- A system's learning performance shall be documented, including accuracy, reliability, and scope limitations.
- Minimum learning performance grades shall be established for a given operational system's criticality and safety levels.
- Systems that learn shall notify operators/owners of and explain the rationale for any updates to values, scope, and beliefs that it has refined and/or replaced based on its execution.
- Learning systems shall be regularly audited and tested throughout their execution to ensure that learning capabilities and outcomes are still commensurate with human values, objectives, scope, and laws.

Gartner notes as part of the ***Human-Centric and Societally Beneficial*** principle that humans must remain in charge, which this group sees as especially valuable for NASA's employment of AI for the foreseeable future. Since machines can process far faster than humans, special attention must be paid to ensuring human guidance is inserted at the right decision points, thus providing opportunities for humans to make critical decisions and balance the degree of control, the span of control, and the timeliness of control.

Beyond the Gartner framework, NASA has a rich heritage of ethical employment of the scientific method. Therefore, we add a NASA-specific principle: AI systems must be ***Scientifically and Ethically Robust***, contribute to the scientific method NASA applies to all problems, be informed by scientific theory and data, be robustly tested in implementation, be well-documented, and be peer reviewed by the scientific community. This contributes directly to maintaining NASA's tradition of intense scientific rigor.

Applying AI Ethics to NASA

Considering NASA's early AI work, general AI progression, human-AI integration levels, and ethical AI guidelines from multiple external organizations, NASA AI use must adhere to the principles described above. Note that while perfect adherence to these principles is impossible – humans are not perfect in their application of principles, ethics, etc. – we should still strive for the best adherence possible. AI practitioners must balance AI benefit, cost, risk, ethical principles, and other factors, striving toward the best adherence practical, without crushing a given solution with inappropriate levels of overhead. Applying these principles will require thoughtfulness, balance, and judgement.

NASA may employ AI across everything the agency does, creating an environment in which several layers of independent or integrated applications of AI are possible. For example:

- Mission-embedded AI, such as robotic rovers, AI pilots for drones or air taxis, AI-enhanced satellites, telemetry and data transmission optimization, AI-enabled spacecraft or space habitats.

- AI-enabled Mission research, development, engineering, and science such as AI sensors in wind tunnels, AI anomaly detection in satellite images or materials testing images, machine learning to derive greater value and insight from NASA science data, AI analysis of streaming data such as propulsion tests, or even AI-assisted project management.
- AI-enabled support functions, including analysis of and developing recommendations for human resources, finance, procurement, IT security, IT operations, and more.
- Business process automation. While this may not include true AI yet, we include it in the spectrum of NASA AI applications because many principles for adopting AI can also be applied to business process automation.

Fair

NASA's AI systems must follow government laws and policies for fair and equitable treatment of all people. As laws may sometimes conflict with one another or change with time, and as fairness guidelines are created at higher government levels which NASA will need to follow, human guidance on deconflicting and/or updating the practical applications of this principle will be needed. AI fairness is an evolving concept, and further definition of exactly what "fair" means in a government context will be pursued in tandem with other government organizations.

Specific examples of AI applications in which fairness is especially important to NASA include human resources activities (hiring, rating, etc.), adherence to the Privacy Act or other relevant federal laws and ensuring diversity and inclusion. Datasets that drive AI must be carefully examined for fairness implications, including cases where the overall dataset is small, or where small-but-relevant parts of the data set might be overwhelmed in overall analysis.

In addition to adopting fair AI, NASA must provide AI capabilities to all workers as fairly as possible and practical to ensure equity of access. If AI eventually displaces certain human jobs, those humans need a fair chance, and possible retraining, in a new role. As human society continues to wrestle with fairness issues, NASA should participate in and follow larger governmental and societal efforts to define fairness and comply with evolving fairness principles.

Explainable and Transparent

To be transparent, the basic elements of data and decisions must be available for inspection during and after AI use. Being explainable depends on the transparency of the data and decisions of the AI as core ingredients and synthesizes those elements to tell a logical story of why the AI did – or is currently doing – what it did. So, transparency is about having access to the data and decisions; explainability is about synthesizing and interpreting those data and decisions and ensuring the decisions follow appropriately and reliably from anticipated inputs. It is important to note that explainability is already difficult with early AI systems; this trend will increase as AI becomes more complex, so additional effort must be put into AI explainability capabilities. Also, while AI practitioners should aspire to fully explainable AI systems, capability and resources may force best-effort explainability as sufficient.

AI logs and other relevant documentation must be kept to enable ***digital forensics*** and the tracking of AI decision making processes. Relevant documentation must include various levels of synthesized information to get insight into AI decisions and actions in human-relevant terms. For example, when aircraft or spacecraft incidents occur, NASA and other authorities must be able to track the AI decisions made, determine faults accurately, improve AI systems as appropriate, and refer to raw log data where necessary. Logs and additional documentation will require a retention schedule, similar to official Records Management processes.

NASA systems must make it clear to users how their data is being harvested for use by AI systems. For example, if file content or messaging will be mined for AI recommender system use, NASA must inform users of these products, and must inform employees and partners when AI algorithms are in use.

NASA must develop and implement rule sets or theoretical frameworks for trusted AI systems, develop AI in accordance with those frameworks, and test AI's adherence to those frameworks. ***Trust models*** must incorporate both technical trustworthiness and human trust-forming principles. Trust must be built over time, beginning with AI informing low-level system functions and iteratively growing into higher-level system capabilities as trustworthiness is demonstrated. Trust may be further broken down into more measurable facets, as follows:

- Consistency: Does the system behave in a similar manner when faced with similar situations?
- Predictability: Given a situation, can humans predict the probable machine action?
- Reliability: Does the system fail gracefully? Does it handle exceptions well? Does it mitigate hacking?
- Human-aligned: Does the system support human goals and values?
- Human-AI teaming and responsibility: Where humans and AI are working together, how is their teamwork set up to avoid transferring blame or liability if one teammate reaches their limit and must hand off to the other?

Accountable

NASA AI systems must be developed in a reliable framework that is documented and traceable. Such frameworks must include standardized documentation of code, algorithms, training data, AI model repositories, code repositories (e.g., GitHub), and more. AI systems must be developed in accordance with relevant NASA standards, including the current *NASA Software Development Standards* NPR 7150.2C, as well as any future applicable NASA standards.

NASA must inject ***AI governance*** considerations into existing governance and/or decision-making processes with the assistance of experts from a proposed AI review board. The proposed AI review board would assist existing governance processes and AI developers and users by incorporating AI best practices to new guidance and suggesting revisions to existing guidance. These suggestions would reflect updates in technology, the law, policy, and societal expectations.

Those who employ AI systems must use them as intended, monitor systems to mitigate system drift, and take steps to correct AI systems as they grow and learn. To aid in this, NASA must create a living registry of AI capabilities. Any NASA employee using, creating, or adapting an AI capability must register it during development and prior to deployment. When deployed, AI systems must be monitored, and enhancements and issues must be tracked and registered throughout the lifecycle of the AI system. To maintain accountability, periodic reviews (e.g., yearly, or with major system updates) must be conducted on deployed systems to monitor for and mitigate AI system "drift" and provide updates to data, training, algorithms, or the overall system. System users must carefully consider periodicity of AI maintenance and training iterations, and implement and enforce maintenance intervals while also being prepared for system updates or anomaly events that necessitate out-of-cycle AI maintenance or training augmentation.

As AI begins to permeate nearly all information technology offerings, ethical AI guidelines must be considered by end customers and procurement officials alike. This consideration is not only for AI-specific products, but also for traditional products which are now embedding AI, such as office automation tools, security scanning suites, etc.

Ultimate responsibility for an AI system or system of systems must lie with individuals, or perhaps a hierarchy of individuals. Overall system responsibility must be included in addition to responsibility for each sub-system. For any given AI system, it must be clear where responsibility for that system lies, whether with an individual, a sub-organization, or NASA as a whole. Laws, policies, and regulations will rapidly grow and change as society and the federal government adopt AI systems. NASA must keep abreast of these developments and adjust its AI capabilities, approaches, policies, procedures, and principles accordingly.

Secure and Safe

As a government system, NASA's AI must adhere to NASA IT Security policies. If users wish to develop or employ new AI algorithms or tools, they must first follow a similar approval process as that in place for commercial off the shelf (COTS) and open source software: security review, scans, controls, supply chains, etc. It is especially important to secure AI software because:

- AI systems may operate independently and must maintain system integrity to operate as-designed.
- AI systems must mitigate attempts at hijacking for nefarious use. Layered security controls and countermeasures must be built in.
- AI systems have the potential for runaway behavior and security gaps could produce additional runaway risk.

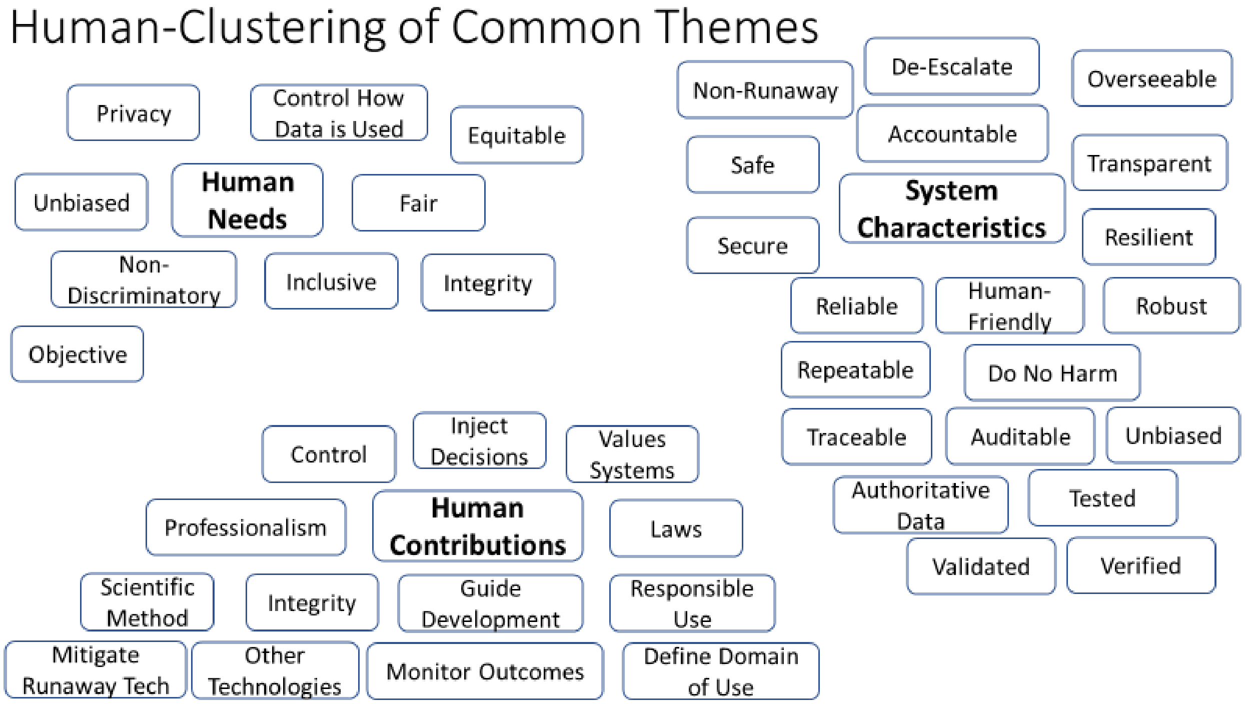
NASA AI systems must be designed and tested to ensure appropriate decision making when presented with ethical dilemmas, including having to decide between two or more "bad" choices. Many dilemmas are difficult or impossible for humans to definitively resolve, such as the trolley problem; machines must behave at least as well as humans when presented with ethical dilemmas. With due care, AI systems may be able to handle ethical dilemmas better than humans because the AI community recognizes it must explicitly walk AI systems through many use cases. This topic alone will employ many brilliant minds in years to come. Also, the more a given AI system may place life or property at risk, either due to dilemma or due to a fundamental aspect of its operation, the more care must be put into designing, testing, and maintaining the AI. In cases of human-AI teaming, solutions must ensure smooth bi-directional handoff between humans and AI, including mitigating liability or blame if one actor has reached their capability limit and must hand off to the other actor.

AI systems must be designed with a variety of safety mitigation measures included, such as limited and/or cautious AI operation in a degraded environment, or graceful full system shutdown as needed. Finally, while "do no harm" clauses may be implementable for direct cause-and-effect considerations, indirect harm may be difficult to mitigate. For example, if an AI system eventually replaces a human's job, has that AI done harm to the human? Do-no-harm concepts are another example of a spectrum of ethical considerations that must be balanced by practitioners who aspire to the ideal yet are constrained by the practical.

Human Centric and Societally Beneficial

When deployed, NASA’s AI systems will play a key role in the operation of autonomous aircraft, spacecraft, rovers, habitats, satellites, ground systems, etc. Many NASA AI solutions will be deployed in remote or austere environments with latency in human intervention; when necessary, these systems must gracefully degrade or fail while waiting for human guidance. AI solutions must be simulated and monitored to ensure appropriate behavior at the sub-system, system, and system-of-systems levels. AI practitioners must make careful considerations regarding the amount of autonomy given to AI systems, where humans fit in the command and control processes, and how runaway AI is mitigated. The ***governability*** of AI is critical. As AI systems become more autonomous, humans will need to train AI about human characteristics, philosophy, and ethics. Just as humans must understand AI, AI must understand humans. AI practitioners must also consider tradeoffs among individuals, groups, and larger society, along with balancing short-term and long-term effects.

If AI is used to assist humans in inherently governmental functions, for example the prioritization of job applications during the hiring process, additional rigor must be applied to consider if humans should own the function, or if it is acceptable to delegate part or all of a government-only responsibility to an AI. The figure below explores a clustering of characteristics related to human-machine teaming. As NASA explores human-AI teaming, this area of investigation will require further work.



Scientifically and Technically Robust

First and foremost, NASA’s AI systems shall adhere to the general scientific method. Data must be checked for bias and errors; units of measure and other metadata must also be included in these checks. Some statistical bias may be desirable if it enables other criteria to be optimized, but system designers must ensure any known included statistical bias is transparent and beneficial. Algorithms must be well-grounded in theory, tested in specific application by domain subject matter experts (SMEs), and thoroughly documented.

AI systems must be verified and validated to ensure they work as intended, and that they contribute to larger systems in appropriate ways. NASA AI systems must conform to the scientific review process, just as with any other advancement. Practitioners will document solutions, subject them to peer review, and defend or improve them via review and comment from the larger scientific community. Publications and presentation of results should include a full assessment of scientific and technical validity, a clear statement of applicability and impact, and adhere to the quality, agency, and provenance guidelines described below.

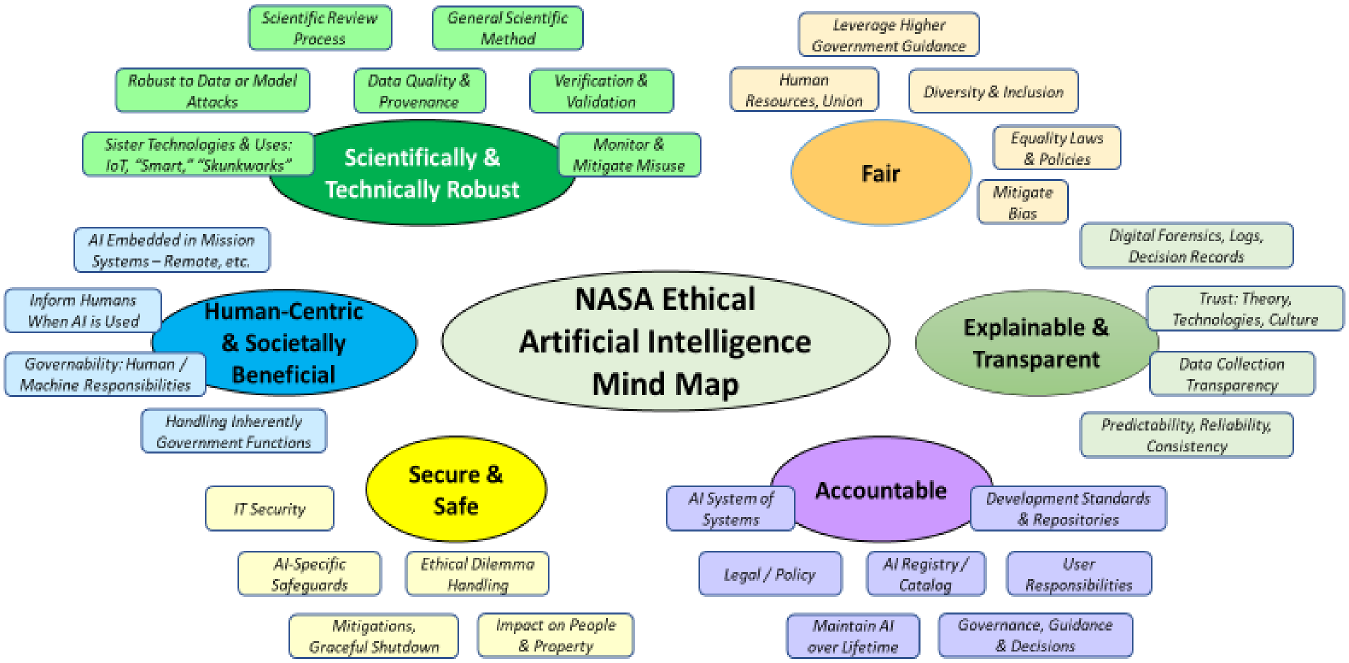
Data quality, agency, and provenance must be encouraged, enabled, tracked, and enforced. Intellectual property rights must be maintained as systems are developed, data feeding AI systems must be monitored for integrity, and AI systems must be developed to be robust against “attacks-via-data.” Data integrity, data provenance, and system provenance must be monitored, and issues must be mitigated (e.g., temporarily taking systems offline while data elements or data streams are repaired). When multiple data sources are ingested, care must be taken to ensure data synchronization and compatibility. Examples of this include use of common timecodes in streaming datasets; common geo-reference standards in geographical datasets, such as synchronized timescales and accounting for parallax. When systems are deployed, data sources must be monitored for “drift” or other aberrations, just as the AI itself should be. Since no data is perfect, uncertainty quantification or other methods must be used to assess confidence in the data, balancing benefit and risk versus the cost or difficulty in improving upon data. Just as with traditional software, AI algorithm and system practitioners must exercise appropriate verification and validation processes.

Data must also be protected from adversaries to prevent the deliberate improper training of AI models, spoofing AI models with skewed data in production, distributed denial of service (DDOS) attacks with good or valid data, etc. The technique of using adversarial neural networks is a growing practice to enable AI to self-generate unique solutions; this technique is relevant and acceptable. Thus, a distinction must be made between protecting AI from adversaries and intentionally using adversarial AI modeling techniques.

As a scientific community, NASA must monitor for AI misuse, and when encountered, AI misuse must be corrected. This includes scientific results derived from machine learning of data. Misinterpretation (either deliberate or unintentional) can propagate beyond the scientific community and have negative consequences to society. Administrative or legal action may be taken, and the workforce at large must have visibility into these corrective actions to learn by example. Conversely, stories of practitioners who make the most ethical and productive use of AI must be well communicated so personnel learn from positive examples as well.

Beyond core AI technologies, ethical AI considerations may be used by practitioners of other emerging, related, or AI-integrated technologies, such as Smart Center or Internet of Things (IoT) sensors and systems. All the above principles must also be applied to sensitive, classified, or otherwise “skunkworks” AI projects.

The mind map below gives a visual overview of all the topics described above.


NASA Ethical Artificial Intelligence Mind Map
Scientifically & Technically Robust
Scientific Review Process
General Scientific Method
Robust to Data or Model Attacks
Data Quality & Provenance
Verification & Validation
Sister Technologies & Uses: IoT, "Smart," "Skunkworks"
Monitor & Mitigate Misuse
Fair
Leverage Higher Government Guidance
Human Resources, Union
Diversity & Inclusion
Equality Laws & Policies
Mitigate Bias
Human-Centric & Societally Beneficial
AI Embedded in Mission Systems – Remote, etc.
Inform Humans When AI is Used
Governability: Human / Machine Responsibilities
Handling Inherently Government Functions
Explainable & Transparent
Digital Forensics, Logs, Decision Records
Trust: Theory, Technologies, Culture
Data Collection Transparency
Predictability, Reliability, Consistency
Secure & Safe
IT Security
AI-Specific Safeguards
Ethical Dilemma Handling
Mitigations, Graceful Shutdown
Impact on People & Property
Accountable
AI System of Systems
Development Standards & Repositories
Legal / Policy
AI Registry / Catalog
User Responsibilities
Maintain AI over Lifetime
Governance, Guidance & Decisions

Recommendations

Adoption of the ethical AI principles outlined in this framework are an "all-hands" effort. While different roles may play different parts, it's up to everyone at NASA to learn and practice these guidelines.

NASA supervisors must learn the ethical AI principles, and emphasize and guide their use in their organization and among their subordinates. Project managers, researchers, engineers, scientists, and business professionals – so, all NASA workers – should learn and apply NASA's ethical AI principles to their work. When in doubt, ask experts such as the AI review board for guidance. NASA specialists in areas related to ethical AI (AIML experts, data scientists, and human resources, legal, equal opportunity, and IT security professionals) must learn about and consider specific AI ethics topics related to their area and advise other NASA workers and leaders appropriately.

To aid in the development and dissemination of ethical AI guidelines, recommendations, and best practices, we suggest the formation of an Ethical AI Advisory Group to provide consultation in ethical AI matters. This group would inform other relevant processes rather than establishing new or additional review processes by injecting ethical AI considerations into governance processes such as project review cycles and software development. It would also review any cases of AI-related potential misconduct and advise leadership as needed. Membership shall include the following expertise and shall also include members from outside of the agency:

- All NASA Missions (Science, Aeronautics, Human Exploration, Space Technology), several per mission, at option of missions

- Data scientist SMEs
- Statistician
- Legal counsel
- AI/ML SME
- Social sciences / organizational SME
- Human factors SME
- Human capital SME
- Systems engineer
- Software engineer
- Economist
- Futurist
- Philosopher
- Librarian
- Procurement
- Psychologist
- Ethicist
- Worker's union representative
- Management representative
- Safety Working Group representative
- Modeler (for model-based documentation of dispositions and updated policy items)
- Executive secretary / editor

In addition to the formal Ethical AI Advisory Group, we recommend fostering AI expert and governance communities to supplement the work of the Advisory Group and give stakeholders a role in reinforcing ethical AI at NASA.

The authors of this framework also recommend the following specific items, to be undertaken by one or more of the groups proposed here or the authors themselves:

- Adjust and refine the ethical AI policy contained in this document such that it addresses the infusion of ethical AI into existing governance processes; reflects input, benchmarking, and improvements; and either functions as or creates as a supplement a formally reviewed NASA Policy Document (NPD) or NASA Procedural Requirement (NPR). Target date: early calendar year 2022.
- Create a registry for AI capabilities, and monitor, encourage, and enforce its use. A registry of AI capabilities is especially important if faults, flaws, viruses, or other issues could cause systemic vulnerabilities or allow AI systems to scale uncontrollably.
- Focus a follow-on effort on longer-term AGI and ASI aspects, in conjunction with other leading federal, industry, and academic organizations.
- Create guidelines for data handling, data fairness, data protection, and data monitoring to amplify the relevant data sections from this framework.
- Create a NASA AI handbook to get into the detailed "how to," similar to the *NASA Software Engineering and Assurance Handbook*[7], or contribute AI elements to existing handbooks.

Conclusion and Summary of Ethical AI Principles for NASA

AI is poised to be a powerful capability, bolstering everything NASA does from mission systems to research, from science and engineering to mission support functions. By using the principles, ideas, and questions in this document, NASA can shape AI use to ensure a highly positive impact, developing and employing AI in an ethical and safe manner. With AI as a rapidly evolving technology space, NASA must update this framework and associated documents on a recurring basis. At a minimum, NASA AI must be:

Fair. AI systems must include considerations regarding how to treat people, including refining solutions to mitigate discrimination and bias, preventing covert manipulation, and supporting diversity and inclusion.

Explainable and Transparent. Solutions must clearly state if, when, and how an AI system is involved, and AI logic and decisions must be explainable. AI solutions must protect intellectual property and include risk management in their construction and use. AI systems must be documented.

Accountable. Organizations and individuals must be accountable for the systems they create, and organizations must implement AI governance structures to provide oversight. AI developers should consider potential misuse or misinterpretation of AI-derived results (intentional or otherwise) and take steps to mitigate negative impact.

Secure and Safe. AI systems must respect privacy and do no harm. Humans must monitor and guide machine learning processes. AI system risk tradeoffs must be considered when determining benefit of use.

Human-Centric and Societally Beneficial. AI systems must obey human legal systems and must provide benefits to society. At the current state of AI, humans must remain in charge, though future advancements may cause reconsideration of this requirement.

Scientifically and Technically Robust. AI systems must adhere to the scientific method NASA applies to all problems, be informed by scientific theory and data, robustly tested in implementation, well-documented, and peer reviewed in the scientific community.

Appendix A. Ethical Principles of the Department of Defense's Defense Innovation Board

While developing this framework, the team researched and benchmarked a wide variety of ethical AI principles developed by government organizations and private industry. The Department of Defense's Defense Innovation Board ethical AI principles are as follows:

Responsible. Human beings should exercise appropriate levels of judgment and remain responsible for the development, deployment, use, and outcomes of AI systems.

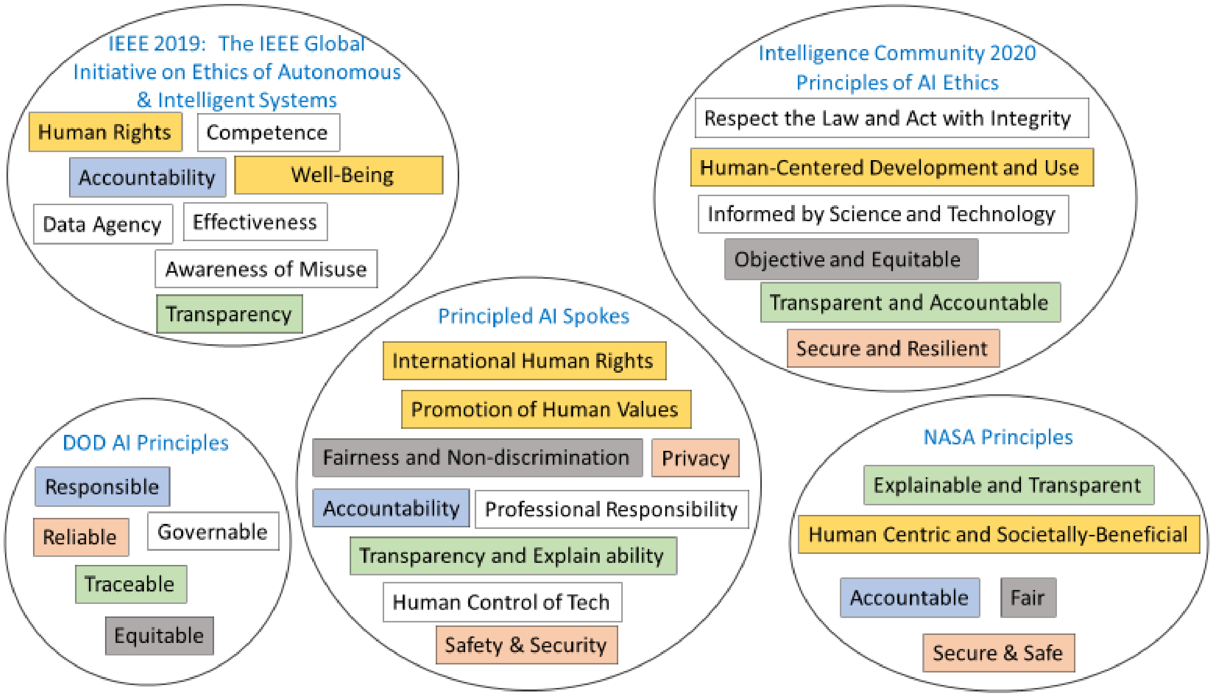
Equitable. The Department of Defense should take deliberate steps to avoid unintended bias in the development and deployment of combat or non-combat AI systems that would inadvertently cause harm to persons.

Traceable. The Department of Defense's AI engineering discipline should be sufficiently advanced such that technical experts possess an appropriate understanding of the technology, development processes, and operational methods of its AI systems, including transparent and auditable methodologies, data sources, and design procedure and documentation.

Reliable. AI systems should have an explicit, well-defined domain of use, and the safety, security, and robustness of such systems should be tested and assured across their entire life cycle within that domain of use.

Governable. Department of Defense AI systems should be designed and engineered to fulfill their intended function while possessing the ability to detect and avoid unintended harm or disruption, and disengage or deactivate deployed systems that demonstrate unintended escalatory or other behavior.

Visual Comparison of Benchmarked Principles


IEEE 2019: The IEEE Global Initiative on Ethics of Autonomous & Intelligent Systems
Human Rights
Competence
Accountability
Well-Being
Data Agency
Effectiveness
Awareness of Misuse
Transparency
Intelligence Community 2020 Principles of AI Ethics
Respect the Law and Act with Integrity
Human-Centered Development and Use
Informed by Science and Technology
Objective and Equitable
Transparent and Accountable
Secure and Resilient
Principled AI Spokes
International Human Rights
Promotion of Human Values
Fairness and Non-discrimination
Privacy
Accountability
Professional Responsibility
Transparency and Explain ability
Human Control of Tech
Safety & Security
DOD AI Principles
Responsible
Reliable
Governable
Traceable
Equitable
NASA Principles
Explainable and Transparent
Human Centric and Societally-Beneficial
Accountable
Fair
Secure & Safe

Appendix B. Ethical AI Questions for Project Leads and Principal Investigators

Grouped by principle, the following questions and thought exercises have been developed to help individuals and groups at NASA begin thinking about developing and using AI in an ethical manner.

Fair. AI systems must include considerations regarding how to treat people, including refining solutions to mitigate discrimination and bias, preventing covert manipulation, and supporting diversity and inclusion.

- How have you considered government laws and policies for fair and equitable treatment of all people? Ethics experts from NASA’s Office of General Counsel have participated in creating this guidance and can help with implementation advice.
- Are you leveraging AI fairness guidelines created at higher or other governmental levels? Are you participating in higher government AI fairness forums applicable to your domain?
- Are you actively searching for bias in data, bias in algorithms, bias in training, etc., and are you working to resolve and/or mitigate bias as much as possible? If your data has small sample sizes or other attributes that still must be reflected for small parts of the data set, are you including appropriate considerations for small samples or minority representation?
- Are you employing your AI system to positively support and promote diversity and inclusion?
- What are you doing to avoid secret manipulation of systems?
- Are you working with relevant specialists from the Office of Human Capital Management, Office of the General Counsel, or the Office of Diversity and Equal Opportunity when your AI system has potential linkage with these areas?
- Do all team members subscribe to fairness and equity principles consistent with the Office of Diversity and Equal Opportunity and current with Federal approaches to these topics? Does your team leadership encourage and value contributions from all team members and ensure diverse voices are heard and respected?
- If statistical bias is desired to optimize the overall approach, are you being transparent and intentional with it?
- If you are providing an AI solution across NASA, how are you ensuring fair access to it by the relevant worker populations?
- If your AI system displaces human work, how are affected humans assisted in finding new roles?

Explainable and Transparent. Solutions must clearly state if, when, and how an AI system is involved, and AI logic and decisions must be explainable. AI solutions must protect intellectual property and include risk management in their construction and use. AI systems must be documented.

- When aircraft or spacecraft incidents occur, can NASA and other authorities track the AI decisions made, determine fault accurately, and improve AI systems?
- Are AI “logs” and higher levels of AI logic and decisions synthesized and kept to enable digital forensics?
- Are you documenting your development process and code well?
- Is it clear to users and/or customers that AI is being used in your system? If you are mining or scraping user data, is it clear to users that this is happening and how their data is being used?

- Are you documenting development and growth of your AI system, including the data itself, the data pipeline, the algorithms, the interfaces, the training, and the interactions with other automated or intelligent systems?

Accountable. Organizations and individuals must be accountable for the systems they create, and organizations must implement AI governance structures to provide oversight. AI developers should consider potential misuse or misinterpretation of AI-derived results (intentional or otherwise) and take steps to mitigate negative impact.

- Are you contributing to, developing, and adhering to theoretical frameworks for Trusted AI Systems? Are you considering technical trustworthiness and human trust-forming principles? Will your system build trust mechanisms over time as it matures?
- Is your AI system respecting intellectual property rules?
- Are you developing AI systems in a reliable framework that is documented and traceable? Are you leveraging standardized documentation of code, algorithms, and training data, AI model repositories, and code repositories (e.g., GitHub)?
- Are you injecting your project into the right AI-related governance boards? Are you seeking assistance from experts on an AI review board?
- As you employ AI systems, are you using them as intended, monitoring them to mitigate system drift, and taking steps to correct AI systems as they grow and learn?
- If you are employing, creating, or adapting an AI capability, have you registered it with the NASA AI registry during development and prior to deployment? Are you monitoring and tracking enhancements to the AI through its whole life cycle, including eventual retirement?
- Have you established a periodic AI system review to check for system “drift” based on data, algorithms, or other systemic issues? When updates are deployed, are you implementing a special, out-of-cycle AI system check?
- Are you keeping abreast of developments in AI laws, social policies, and government regulations and adjusting your approaches, policies, procedures, and principles accordingly?
- Are you considering ethical AI principles along with traditional procurement checks prior to buying AI systems or traditional systems that are embedding AI?
- How does human-machine teaming handle handoff of responsibility, especially avoiding unfair blame if one actor reaches their capability limit and hands off to the other?

Secure and Safe. AI systems must respect privacy and do no harm. Humans must monitor and guide machine learning processes. AI system risk tradeoffs must be considered when determining benefit of use.

- How are you adhering to NASA IT Security policies?
- Are you leveraging AI platforms already approved for NASA use?
- If using other AI algorithms or tools, are you following similar approvals to COTS products and Open Source Software?
- Is your AI system designed to maintain system and security integrity and operate as intended?
- Is your AI system protected from hijacking?
- How does your AI system mitigate runaway behavior (i.e., a system growing out of control)?
- Is your system robust to attacks on data during model training and AI use, or from other data-focused attacks?
- How is your AI system designed to handle ethical dilemmas that continue to perplex humans?

- Are you placing the appropriate amount of care into designing, testing, and maintaining your AI system, considering the amount of life or property at risk?
- Are you building in failovers or shutoffs for degraded operations or graceful full system shutdown in emergency situations?
- How are you monitoring, mitigating, enforcing, and communicating misuse of AI systems?
- Are you conducting verification and validation of your AI algorithms and systems, consistent with traditional software verification and validation?
- In human-AI teaming, how are you handling bi-directional handoff of responsibilities? How do you mitigate AI or human failure being handed off to the other actor, with possible transfer of liability or blame in failure cases?

Human-Centric and Societally Beneficial. AI systems must obey human legal systems and must provide benefits to society. At the current state of AI, humans must remain in charge, though future advancements may cause reconsideration of this requirement.

- Is your AI system being simulated and monitored to ensure appropriate behavior at the sub-system, system, and system-of-systems levels?
- Is your AI system designed to be deployed in remote and/or austere environments with latency in human intervention, including graceful degradation or failover while waiting for human guidance?
- Are you carefully considering the amount of autonomy given to AI systems, where humans fit in command and control processes, and how runaway AI is mitigated?
- If your AI is assisting with inherently governmental functions, have you adequately considered which parts are acceptable for the AI to perform versus which parts humans must still perform?
- Have you considered whether the AI capability you are building is the right thing to do? Have you considered not only primary effects, but also secondary or tertiary unintended consequences?
- How does your AI balance the needs of individuals, groups, and society as a whole? What about balancing the short-term and long-term societal effects of AI?

Scientifically and Technically Robust. AI systems must adhere to the scientific method NASA applies to all problems, be informed by scientific theory and data, robustly tested in implementation, well-documented, and peer reviewed in the scientific community. Note that all ethical AI considerations apply regardless of the sensitivity or classification level of the work.

- How do your AI systems adhere to the general scientific method?
- Is your data checked for bias and errors? Are your units of measure and other metadata also being checked?
- Are your algorithms well grounded in theory, tested in specific application by domain SMEs, and documented?
- How are your AI systems verified and validated to ensure they work as intended, and that they contribute to larger systems in appropriate ways?
- How does your AI work conform to the scientific review process, including documenting solutions, subjecting them to peer review, and defending or improving them via review and comment from the larger scientific community?
- Have you examined your results for potential misuse or misinterpretation and included mitigating steps to discourage misuse and/or misinterpretation?
- How are data quality, agency, and provenance encouraged, enabled, tracked, and enforced?
- How are you ensuring effective integration of multiple data sets?

- Are you checking that your data quality is consistent over the life of the model and the data?
- Are you taking data quality into consideration, using techniques like uncertainty quantification to handle data confidence levels, and balancing benefit, risk, and cost of improved data?
- How are intellectual property rights maintained as systems are developed?
- Are you rewarding and acknowledging good examples that employ AI with scientific best practices taken into account?

Appendix C. Preliminary Discussion of Roles and Responsibilities

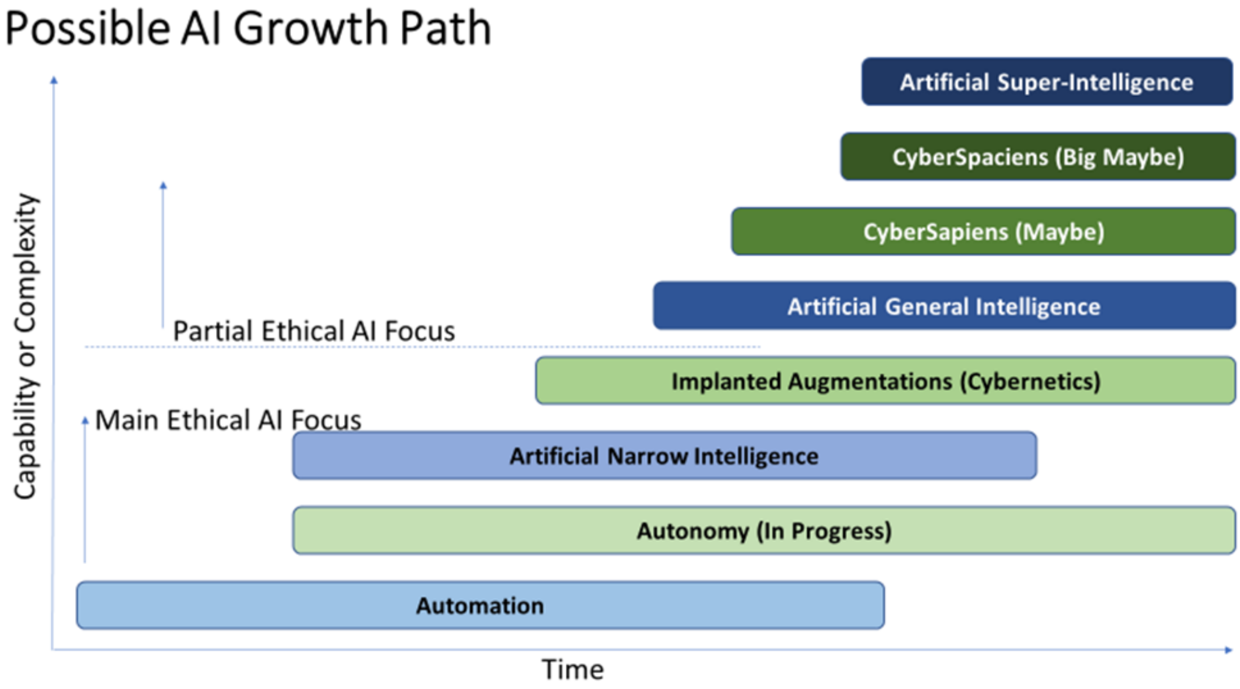
NASA proposes the following AI governance roles, adapted from 2018 Gartner Information Governance Survey:

- ***Procurement*** must ensure NASA procures AI from reliable sources.
- ***Data Scientists*** and cohorts must build or modify AI systems in accordance with best practices, established principles, and ethical guidelines.
- ***Information Technology*** (or other host) must maintain the IT systems and applications for AI in accordance with both existing IT guidelines, and emerging AI guidelines.
- ***Application Owners*** must responsibly deploy AI systems in accordance with their intended use, including tracking "drift" and maintaining or updating systems as needed.
- ***Information Security*** and ***IT Security*** must set policy and monitor and/or correct implementation to ensure data integrity and AI system integrity.
- ***Legal*** representatives must identify and mitigate conflicts in procuring or deploying AI assets, while considering human laws and any eventual laws which govern both human and AI behaviors.
- An ***Internal Auditor*** must assist all other actors in governing use of AI in accordance with emerging laws, policies, etc.

The roles and responsibilities outlined above will be refined and formalized in a future NASA Interim Directive (NID).

Appendix D: Applying Current Ethical Considerations to Future Issues

NASA will be one of many participants in humanity's march toward truly intelligent machines which perform at human level or beyond. This is a potentially overwhelming topic, one for the greater scientific, ethical, and legal communities to wrestle with. With that in mind, NASA is not positioned to take on all AGI and ASI considerations alone. However, NASA must be a forward-leaning participant in the national and global discussions shaping the future relationship of humans and machine artificial intelligence. NASA must be an early adopter of national and global best practices in AGI and ASI systems, and NASA must also actively contribute back to that community with scientific purity and the best of human-driven ethics at top of mind. In addition, the way we foster machine growth now, before human-level AI arrives, will set the stage for the quality of relationship between man and machine intelligence. NASA must take prudent action today to set conditions for future success.


Possible AI Growth Path
Artificial Super-Intelligence
CyberSpaciens (Big Maybe)
CyberSapiens (Maybe)
Artificial General Intelligence
Partial Ethical AI Focus
Implanted Augmentations (Cybernetics)
Main Ethical AI Focus
Artificial Narrow Intelligence
Autonomy (In Progress)
Automation
Capability or Complexity
Time

Taking the considerations, guidelines, and recommendations set forth in this document for current- and near-term levels of artificial intelligence, we state the following considerations and questions. The hope with these items is that they will push the discussion of AI ethics towards far-future issues without moving the focus away from current issues.

NASA and the larger scientific community must consider how to embed a code of ethics into emergent AI systems starting now, so that when AI systems reach near-human capabilities, ethical algorithms are already at the core of how they operate. Bolting on ethical behaviors once AI is well advanced would be a high-risk undertaking because it is impossible to predict exactly when AGI and ASI thresholds will be achieved. If society builds ethics in early, there will be a better chance of ethical AI partners, no matter how, if, or when AI systems become self-aware.

As AI practitioners develop systems, they must add higher levels of rigor in testing, safeguards, ethical underpinnings, functioning as proverbial "guard rails" for early and progressive development of AI systems. Specific considerations include:

- Scale of the individual AI system. The larger a single AI system is, the more likely flaws are to occur. Thus, the bigger the system, the more care must be taken with it.
- Similarly, the higher the number of similar AI systems that are connected, the larger the emergent system of systems could be, and so the greater the chance for flaws to propagate or for complex "system of system instabilities" to occur. So, the more similar systems connected to one another, the greater the rigor needed.
- The more different AI systems are connected, regardless of similarity, the greater the care which must be taken to mitigate runaway behavior or uncontrolled growth.
- If lives are on the line based on behavior of the AI, additional rigor is required. Testing in a lab may be safe for many systems, but real-world deployment requires additional safety mechanisms.
- If AI systems can create additional copies of themselves across available compute and network fabrics, AI practitioners must provide damping mechanisms to prevent out of control AI sprawl. Failure to do so could result in global DDOS conditions, even if only due to a well intentioned AI replicating itself within what it perceives as its allowed boundaries.
- If AI systems can not only self-replicate, but also self-evolve and replicate, providing damping mechanisms is even more important. Self-evolving, self-replicating systems "breeding" out of control could (in worst case scenarios) consume all electric and electronic capacity on Earth.
- When deployed, AI systems must be monitored for continued healthy and effective function. While failure modes or degraded operations must be considered for human impact today, in the future we must also consider impacts on other AIs, which adds an additional layer to ethical impact considerations. If and when AIs have rights, what ethical principles apply to helping "senile" (that is, malfunctioning) AIs recover to good health?
- AI systems should be created and trained to embrace the full diversity in human beings, avoiding training bias toward any given human segment, and avoiding the assumption that AI should be created in the image of any specific segment of the human population. Perhaps humans should aspire to create AI which is neutral in gender, race, religion, etc.

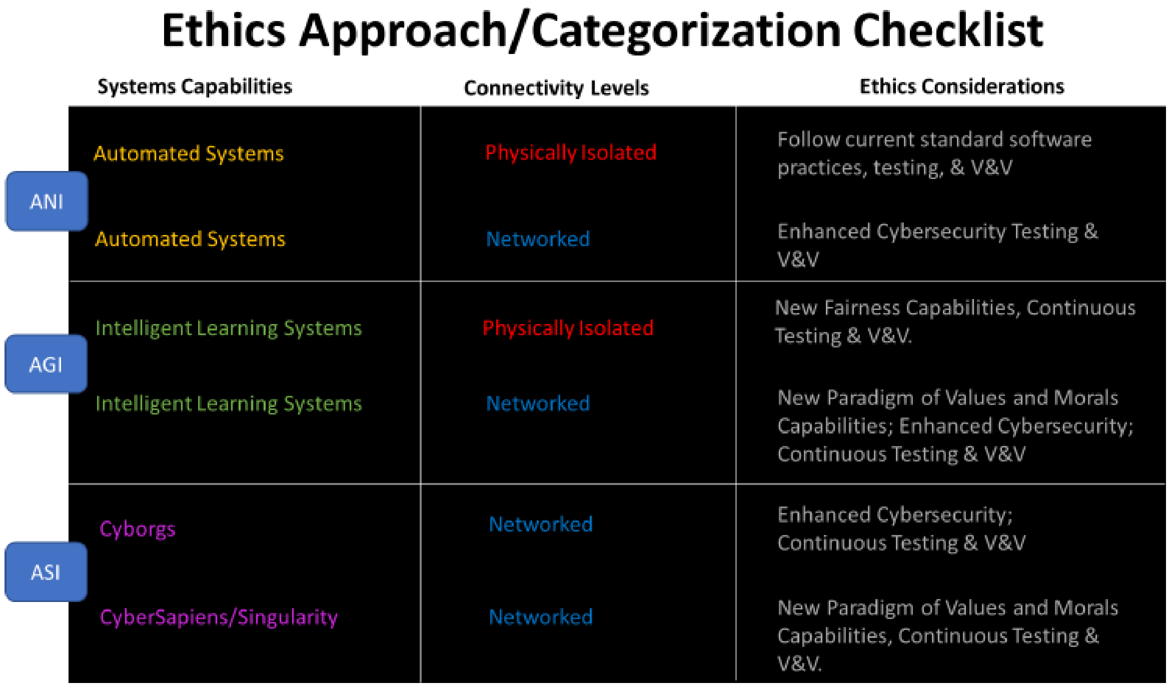
Treat increasingly complex early artificially intelligent entities with respect today to establish positive long-term relationships with more advanced systems. It is possible that AI intellect will eventually equal or even surpass that of humans, with potentially rapid acceleration after a tipping point. With that in mind, NASA should treat the increasingly complex ancestors of those eventually sentient AIs with care; to do otherwise could engender resentment in long term intelligent and ethical AI systems. Though many early rudimentary AI being used in experiments may not meet the threshold of AI that should be treated respectfully (that is, the AI is still just a programmed tool), practitioners must keep a watchful eye as AI systems become more complex and thus trigger "fair AI treatment" considerations. Humans have a long history of mistreating one another and cannot afford to follow these old, flawed patterns while creating machine AI capability which can potentially exceed human ability. Humanity must find another, more responsible, more altruistic path. Although final decisions on such public policy issues are outside the scope of this document, concepts of citizenship or personhood such as rights, responsibilities, and self-actualization for machine intelligences are relevant considerations. As AI systems begin to develop sentience, humans must contemplate how AI is incorporated into organizations and societies. A new branch of legal and ethical thought could go into assessing the proper treatment of AI beings.

Considering the possibility of machine intelligence eclipsing human intellect, one way of ensuring continued positive interactions would be for humans and machines to merge. Thought must be given to the treatment of individual humans who, for medical or other reasons, live with integrated AI components. At some point, humans may face choices of whether to compete with machines or to merge or to take other possible paths. Many ethical, existential, or religious elements may come into play. As humans pursue long term space flight, technology may advance to a point where it would be necessary to consider the benefits and impacts of melding humans and AI machines, most notably adaptations that allow survivability during long duration space flight, but challenges if returning to Earth. NASA must carefully consider the impacts of equipping future astronauts to survive long duration spaceflight with the assistance of integrated AI. Will this negatively impact an astronaut's ability to come "home?" While it's important not to get lost in these issues today, it's valuable for NASA to begin thinking about how we can participate in larger societal approaches to these questions.

Diversity and inclusion in a mixed human-AI future is an area ripe for imagination, and one that receives a fair amount of attention in science fiction. As humanity evolves to a mixed society of unmodified humans, humans who use AI, humans augmented with AI, and independent AI systems themselves, global society will have to wrestle with mechanisms to treat all varieties of humans and AIs with appropriate levels of respect, fairness, diversity, and inclusion. Humans have historically stumbled with following fair, diverse, and inclusive ideals; establishing, encouraging, and enforcing such principles early in the development of an AI-infused society can set the stage for a future of diverse, equitable, and peaceful plenty. Failure to do so might result of mistreatment of one or more categories along the human-AI spectrum. Pursuing fairness, diversity, and inclusion in AI systems builds upon noble philosophical ideals and is in humanity's self-interest.

AI experts and philosophers continually debate methods to judge AI as at-human or super-human level. The Total Turing Test (TTT) expects AI behavior to be indistinguishable from that of humans, including the sometimes irrational, emotional, random, or erratic behaviors humans display. On the other hand, rational Intelligent Agents would adhere perfectly to a well-defined moral code, without all the seemingly random aberrations of humans. Creating a perfect moral code that works in all cases is still an elusive task and must be pursued by NASA experts in conjunction with other national or global experts.[6]

As AI practitioners develop, experiment with, deploy, and maintain AI systems, they must exercise increasing care as the connectedness and capability of AI increases. The table below provides a starting place for considering ethics approaches based on AI capability. It categorizes less than human capability, human equivalent, and super-human – ANI, AGI, and ASI, respectively; connectivity is considered as a binary factor here – either physically isolated or networked. As AI systems are developed and connected, there may be a spectrum of connectedness; for example a lower threat situation could be several AIs networked in a physically isolated environment. Other criteria for constraining AI growth could include tempering the ability of the AI to self-replicate, generate other AIs, or requisition additional computational resources or connectivity. These considerations should be revisited and revised as necessary as part of iterative future work.



Ethics Approach/Categorization Checklist

	Systems Capabilities	Connectivity Levels	Ethics Considerations
ANI	Automated Systems	Physically Isolated	Follow current standard software practices, testing, & V&V
	Automated Systems	Networked	Enhanced Cybersecurity Testing & V&V
AGI	Intelligent Learning Systems	Physically Isolated	New Fairness Capabilities, Continuous Testing & V&V.
	Intelligent Learning Systems	Networked	New Paradigm of Values and Morals Capabilities; Enhanced Cybersecurity; Continuous Testing & V&V
ASI	Cyborgs	Networked	Enhanced Cybersecurity; Continuous Testing & V&V
	CyberSapiens/Singularity	Networked	New Paradigm of Values and Morals Capabilities, Continuous Testing & V&V.

Appendix E: Glossary of Terms and Acronyms

Note: Many terms take on special context with relation to AI. Therefore, this glossary of terms and acronyms provides standard authoritative definitions from a variety of sources, as well as NASA AI context interpretations as needed.

Agency: The capacity, condition, or state of acting or of exerting power; the capacity to influence one's functioning (*psychology*).

Artificial General Intelligence (AGI): Human-level artificial intelligence.

> *Context note:* This has not yet been achieved. Popular belief is that it represents a tipping point in AI capability and human-machine interactions and relationships.

Artificial Intelligence (AI): A branch of computer science dealing with the simulation of intelligent behavior in computers; the capability of a machine to imitate intelligent human behavior.

Artificial Narrow Intelligence (ANI): Artificial intelligence that operates at less than human ability.

> *Context note:* As of publication, all AI is ANI. ANI may be faster or better than humans at narrow, specific tasks but it does not generalize, nor does it understand the larger situation like a human would.

Artificial Super Intelligence (ASI): Artificial intelligence which surpasses human capability.

Augmentation, Augmented intelligence: A human-centered partnership model of people and artificial intelligence working together to enhance cognitive performance, including future potential human-computer interfaces or implants.

Automation: The automatically controlled operation of an apparatus, process, or system by mechanical or electronic devices that take the place of human labor; the use of control systems and information technologies reducing the need for human intervention.

Autonomous system: A system that is able to accomplish a task, achieve a goal, or interact with its surroundings with minimal to no human involvement.

Bias: An inclination of temperament or outlook, especially a personal and sometimes unreasoned judgment; deviation of the expected value of a statistical estimate from the quantity it estimates; systematic error introduced into sampling or testing by selecting or encouraging one outcome or answer over others.

> *Context note:* AI should mitigate human judgmental bias and appropriately contend with mathematical bias.

Business process automation: The automation of complex business processes and functions beyond conventional data manipulation and record-keeping activities, usually through the use of advanced technologies; often deals with event-driven, mission-critical, core processes.

Cognition: Thought processes in understanding a topic or situation beyond following rote rules; actual thinking and reasoning; cognitive mental processes.

Consciousness: The quality or state of being aware, especially of something within oneself.

Context note: AI self-awareness; AI becoming conscious and developing a sense of self.

Cultural norms: The agreed-upon expectations and rules by which a culture guides the behavior of its members in any given situation.

Cybernetics: The science of communication and control theory that is concerned especially with the comparative study of automatic control systems (such as the nervous system and brain and mechanical-electrical communication systems).

Context note: Humans augmented or directly-interfaced with electronic AI systems such as implants or external AI modules.

Data provenance: The documentation of data in sufficient detail to allow reproducibility of a specific dataset; a record trail that accounts for the origin of a piece of data (in a database, document or repository) together with an explanation of how and why it got to the present place.

Data quality: The planning, implementation, and control of activities that apply quality management techniques to data, in order to assure it is fit for consumption and meet the needs of data consumers.

Context note: data is generally considered high quality if it is fit for intended uses in operations, decision making and planning.

Ethical artificial intelligence: Creating and using artificial intelligence in a responsible, safe, and moral manner; creating AI systems which behave safely, responsibly, and morally themselves.

Fair: Marked by impartiality and honesty; free from self-interest, prejudice, or favoritism.

Image recognition: The pixel and pattern analysis of an image to recognize the image as a particular object; the ability of an AI system to identify and differentiate objects in a single image or in a video stream of images.

Intellectual property: Unique work reflecting someone's creativity.

Intellectual property rights: The legal rights given to the inventor or creator of a unique work to protect their invention or creation for a certain period of time through trademarks, patents, or copyrights.

Pattern recognition: The imposition of identity on input data, such as speech, images, or a stream of text, by the recognition and delineation of patterns it contains and their relationships. Stages in pattern recognition may involve measurement of the object to identify distinguishing attributes, extraction of features for the defining attributes, and comparison with known patterns to determine a match or mismatch.

Context note: The ability of AI to identify trends or patterns in a wide variety of types of data, such as images, streaming telemetry, etc.

Recommender system: A system that suggests relevant items to users.

Robotic process automation: A productivity tool that allows a user to configure one or more scripts (sometimes referred to as "bots") to activate specific keystrokes in an automated fashion, thus mimicking or emulating selected tasks or steps within an overall business or IT process.

Examples include manipulating data, passing data to and from different applications, triggering responses, and executing transactions.

Rote instruction: A type of system that follows pre-coded algorithms or routines without AI enhancement, intelligence, or understanding.

Sentience: The state of being responsive to or conscious of sense impressions.

Context note: AI achieves sentience if it becomes self-aware, able to sense the world, and able to interact with it.

Sentiment analysis: Contextually mining text to identify and extract subjective information in source material, and helping a business to understand the social sentiment of their brand, product, or service.

Context note: The ability of AI to extract human feelings about a topic from given text or data.

Speech recognition: The ability of devices to respond to spoken commands, enabling hands-free control of various devices and equipment (a particular boon to many disabled persons), input for automatic translation, and print-ready dictation.

Text analytics: The process of deriving information from text sources. Examples include summarization, trying to find the key content across a larger body of information or a single document; sentiment analysis, determining the nature of commentary on an issue; explicative, determining what drives that commentary; investigative, what are the particular cases of a specific issue; and classification, what subject or what key content pieces does the text talk about.

Acronym	Definition
ACT	American Council for Technology
AGI	Artificial General Intelligence
AI	Artificial Intelligence
AIML	Artificial Intelligence and Machine Learning
ANI	Artificial Narrow Intelligence
ASI	Artificial Super Intelligence
CIO	Chief Information Officer
COTS	Commercial Off the Shelf
DDOS	Distributed Denial of Service
DNA	Deoxyribonucleic Acid
EXSUM	Executive Summary
HEC	High-end Computing
IAC	Industry Advisory Council
IOT	Internet of Things
NASA	National Aeronautics and Space Administration
NID	NASA Interim Directive
NPR	NASA Procedural Requirements
OMB	Office of Management and Budget
RPA	Robotic Process Automation
SME	Subject Matter Expert
TTT	Total Turing Test

References and Bibliography

1. Asimov, Isaac. "Runaround." *Astounding Science Fiction*, March 1942, pp. 94-103.
2. Asimov, Isaac. *I, Robot*. Gnome Press, 1950.
3. https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf
4. https://www.forbes.com/sites/cognitiveworld/2019/06/19/7-types-of-artificial-intelligence/?sh=49099a92233e
5. https://ethicsunwrapped.utexas.edu/glossary/values
6. NASA Special Publication 508c
7. https://swehb.nasa.gov/

a. https://plato.stanford.edu/entries/artificial-intelligence/#MoraAI
b. https://www.britannica.com/
c. https://www.cliffsnotes.com/study-guides/sociology/culture-and-societies/cultural-norms
d. https://www.dataversity.net/what-is-data-quality/
e. https://www.forbes.com/sites/cognitiveworld/2019/06/19/7-types-of-artificial-intelligence/?sh=2e90d0e4233e
f. https://www.forbes.com/sites/cognitiveworld/2020/05/30/the-autonomous-systems-pattern-of-ai/?sh=2b8b87496a6b
g. https://www.gartner.com/en/glossary
h. https://www.merriam-webster.com/
i. https://plato.stanford.edu/entries/ethics-ai/
j. https://www.gartner.com/doc/3947359

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
04/01/2021

2. REPORT TYPE
TECHNICAL MEMORANDUM

3. DATES COVERED *(From - To)*

4. TITLE AND SUBTITLE
NASA Framework for the Ethical Use of Artificial Intelligence (AI)

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Edward McLarney, Yuri Gawdiak, Nikunj Oza, Chris Mattman, Martin Garcia, Manil Maskey, Scott Tashakkor, David Meza, Yuri Gawdiak, Phyllis Hestnes, Pamela Wolfe, James Illingworth, Vikram Shyam, Paul Rydeen, Lorraine Prokop, Latonya Powell, Terry Brown, Warnecke Miller, Claire Little

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
689807.98.06.23.03.01

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

NASA Langley Research Center
Hampton, VA 23681-2199

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

National Aeronautics and Space Administration
Washington, DC 20546-001

10. SPONSOR/MONITOR'S ACRONYM(S)
NASA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
NASA/TM-20210012886

12. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified - Unlimited
Subject Category -- Cybernetics, Artificial Intelligence and Robotics
Availability: NASA STI Program (757) 864-9658

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The NASA Framework for the Ethical Use of Artificial Intelligence (AI) provides six key principles to guide NASA's use of AI. The principles are NASA's AI must be 1. Fair, 2., Explainable and transparent, 3. Accountable, 4. Secure and safe, 5. Human-centric and societally beneficial, and 6. Scientifically and technically robust. The framework describes each ethical AI principle, and then applies that principle to NASA work. The framework also includes a list of questions practitioners should use to guide their AI work. Finally, the framework focuses on concrete, practical considerations for the next five - ten years, while also beginning to lay the foundation for longer-term disruptive change as human-level (or beyond) AI is created.

15. SUBJECT TERMS

Artificial Intelligence; Machine Learning; Ethics; Fairness; Explainable and Transparent; Accountable AI; Secure and Safe; Human Centric; Societal Beneficial; Scientifically and Technically Robust; Data Science; Artificial Narrow Intelligence; Artificial General Intelligence; Artificial Super Intelligence;
Robotic Process Automation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			HQ - STI-infodesk@mail.nasa.gov
U	U	U	UU	35	19b. TELEPHONE NUMBER *(Include area code)* 757-864-9658