Event Report for

The Ethical Artificial Intelligence Quantification Workshop

"Experimentation with an Ethical AI Methodology"

Hosted by the National Institute for Aerospace

In teamwork with

AIEthics.World, the Intel Corporation and The National Aeronautics & Space Administration (NASA)

May 2022

Edward L. McLarney, NASA

Matthew James Bailey, AIEthics.World

Katalin K Bartfai-Walcott, Intel Corporation

Maria MacAndrew, AIEthics.World

Executive Summary

Artificial Intelligence (AI) is a powerful emerging technology area which requires special attention to using it ethically. AI ethics is still an emerging field, and the partners for this workshop and report seek to move AI ethics discussion ahead by experimenting with ways to measure AI ethics criteria. The following document describes the outcomes and learnings from The Ethical Artificial Intelligence Quantification Workshop held at the National Institute for Aerospace (NIA), Hampton, Virginia on May 12th, 2022. The purpose of the workshop was for participants to evaluate and experiment-with the methodology and process presented by AIEthics.World in cooperation with Intel Corporation. The meeting participants learned about the Ethical AI Certification and Maturity Model™ and applied the methodology to selected notional AI systems. The workshop facilitated the evaluation of the maturity of the AI system according to ethical considerations relevant to NASA, NIA and other participants.

The workshop consisted of three main phases. The first phase focused on understanding and summarizing NASA's ethical approaches, mission and values based on published documentation, discussions and individual insights & opinions of participants. This information was prioritized, weighted, ordered, and quantified in phase two, to formulate an alignment between human values (ethics) and their applicability to AI systems during all lifecycle phases. The first two phases were summarized as a form of *ethical genealogy* for artificial intelligence, specific to NASA's ethical approaches. In the third and last phase of the workshop the participants evaluated notional examples of artificial intelligence to qualify and quantify its ability to adhere to the organizational ethics approaches, using the Ethical AI Certification and Maturity Model™.

The workshop uses the concept of *genealogy*, in the traditional sense: the study and traceability of lines of ancestors in the process of evolutionary development from earlier forms. However, as it is applied to an Ethical AI definition, it is providing the insights to the necessary and mandatory traceability of content, data, metrics, telemetry, elements, and structures which are used in the AI's lifecycle to foster and measure AI ethics in all steps of its lifecycle.

The Ethical Artificial Intelligence Quantification Workshop provided NASA with the opportunity to apply the Ethical AI Certification and Maturity Model™, in combination with existing and well-known decision-making and quality control methods to identify the metrics and measurements for an Ethical AI and assess its ethical condition and quality aligned with NASA ethics approaches. The result of the workshop is the capacity for NASA to apply the maturity model assessment to its AI Systems as desired and if necessary, publish the ability of these AI Systems to adhere to the organizational ethical goals. AI ethics frameworks need to be customized for each application domain, for example, individual NASA Mission Directorates. General principles that work in one area such as AI/Machine Learning-based text analysis (the ethics of information-extraction) may need to be adapted for another such as sense-and-avoid decision-making in a flight environment.

The workshop was conducted among approximately twenty NASA subject matter experts, so the elements noted above should be considered examples, not definitive NASA ethical AI principles, genealogy, etc. Generating a definitive AI ethics framework for an organization as diverse as NASA would require far more discussion, debate, review, etc. However, the workshop provided valuable insight into mechanisms and processes for quantifying AI ethical qualities.

Table of Contents

Executive Summary 2

Table of Contents 3

Background on Ethical AI 4

Introduction 5

Workshop Process Overview 5

Workshop Breakdown 7

Topics - Organized into Findings, Discussion, Recommendations 11

1. Interest in AI Ethics 11

2. Long Term AI Ethics Quantification Vision (Ethical AI) 11

3. Exercising a Quantification Model for the Ethical Quality of AI (Ethical AI) 12

4. More on Terminology 13

5. Additional Use Cases 13

Additional Points of Note 14

Conclusion 15

Acronyms and Terms 18

References 19

List of Workshop Participants 20

Workshop Format **Error! Bookmark not defined.**

Background on Ethical AI

Ethical considerations for Artificial Intelligence (AI) are a current topic of intense global interest. Published in December 2020, the White House Executive Order 13960 promulgates a set of principles for United States Federal organizations to follow, including topics of trustworthiness, bias, and many other relevant ethical concepts. Similarly, the European Union put forth Ethics guidelines for Trustworthy AI in April 2019 to enforce all artificial intelligences deployed within their 27 member countries to be ranked in their level of Trustworthiness.

Industry, academia, and government organizations have begun making strides in adapting, adopting, and refining ethical AI considerations tailored to their needs. For example, within the United States, the Department of Defense and Intelligence communities notably issued interim guidance for ethical AI work.

Furthermore, NASA has begun discussion of a set of tailored ethical AI principles, to include initial exploration of a set of questions practitioners may ask themselves as they conduct AI work. Some NASA disciplines have been early adopters of AI and have applied standard practices from system engineering, software development and risk management to guide AI development to date. NASA has established Responsible AI Officials in the Chief Scientist and Chief Technologist and is continuing to refine approaches to this turbulent field as it evolves. NASA has many professional standards, directives, policies, and processes that guide safety and scientific rigor in the Agency's work. Many ethical elements are covered in existing mechanisms, however, active discussion, debate and experimentation with AI ethics will reveal if augmentations to existing mechanisms could contribute to industry standards.

Additional ethical AI discussion and debate are planned across wide NASA communities in fiscal years 2022 and beyond, to further-develop community consensus regarding how ethical AI principles can be best applied to NASA work. As ethical principles are refined and specific goals, objectives, questions, etc. are decomposed to inform AI work, measurement techniques will be needed to quantify a variety of qualities of AI work and AI systems, to include ethical qualities. This workshop was specifically designed as an early experiment in methods to measure, quantify, or rate ethical qualities of AI systems.

Introduction to the concept of Ethical AI Genealogy

The term *Ethical AI Genealogy* is a term provided by AIEthics.World for artificial intelligence assessment. It is derived from how the human (organic) form is created and determined. All organic intelligence on earth has a *genealogy (*ancestry*)*, defined by a *genome* constructed by a vault of *genes*. The *human genome* defines the human organic form and its *genes* determine common and unique human factors such as the color of our eyes. Similarities and differences between a person's genome and other humans through history are written in our genes. As such, *genealogy* is a way to read the uniqueness of an individual's history and understand where they came from.

The analogy of *genealogy* can be applied to artificial intelligence, in the sense of identifying factors that directly or indirectly influence the *ethical quality* of its inheritance, construction, lifecycle and performance. In other words, factors influence ethical quality and are termed *Ethical AI Genes;* The combination of these *ethical factors* is termed *Ethical AI Genealogy.*

When applying *genealogy* to an Ethical AI definition, we have the flexibility to identify, customize and ameliorate a list of ethical factors to construct an Ethical AI Genealogy for an artificial intelligence. In other words, an *ethical genome* for artificial intelligence. In doing so, we can gain broader insight into the

ethical condition of the content, data, metrics, telemetry, elements, and structures which are used in the construction, lifecycle and performance of an AI. Put simply, this approach enables us to qualify, quantify and assess the ethical condition of each factor and assess the overall ethical quality of an artificial intelligence. The ethical genome goes beyond just recording the evolution of ethical principles. It should also reflect the evolving culture and technology of the organizations in which they are being applied. See the workshop guide and presentation for further detail and example.

Intellectual property clarification: Genealogy was the model / paradigm used by the Principal Investigator. Decades of NASA prior art in system engineering, metrics systems, engineering, software engineering, risk management and more could also be employed by NASA or other government organizations to assess AI ethics qualities, with no intellectual property ties to the AIEthics.World model.

Introduction

The Ethical AI Quantification Workshop was held at the NIA facility in Hampton Virginia on the 12th of May 2022 with some participants in person and some participating remotely via collaboration systems. The purpose was to apply the Ethical AI governance and assessment methodology to a real-world AI example provided by NASA and evaluate the feasibility and ease of use of the methodology during all AI lifecycle phases, from data collection to making predictions.

The team included Ed McLarney – NASA's AI and Machine Learning Transformation lead, Doug Stanley, NIA President, Matthew James Bailey from AIEthics.World and Katalin K Bartfai-Walcott from Intel Corporation. Participants of the workshop included AI and Autonomy experts; data scientists; researchers, engineers, an attorney, a librarian, and other interested personnel, representing NASA, NIA, JPL, Intel, and AIEthics.World.

Workshop Process Overview

The Ethical AI Quantification Workshop demonstrated a concept and process developed by AIEthics.World in cooperation with Intel Corporation. The workshop was designed to demonstrate the ability to qualify, then quantify organizational ethical principles and apply the derived metrics to AI Systems. The audience was introduced to the value and opportunity of extending organizational ethical goals and targets to the AI Systems they develop and deploy in a standardized fashion. The ability to measure, demonstrate and publicize the adherence to the ethical quality goals is an important goal in today's socially conscious society. The Ethical AI Certification and Maturity Model™ provides the fundamental elements so organizations, consumers, and other AI systems can rapidly understand the ethical quality (e.g., trustworthiness, safety rating, etc.) of AI systems. This process is related to the way people develop and use the concept of reputability today. In the future this information could be used by consumers and businesses as a confidence-based decision factor regarding who to work with, what products to buy, etc. The Ethical AI Certification and Maturity Model™ introduces several concepts.

(1) The first concept is the *Ethical AI Constitution.* The concept, taking its origin from established and well-known systems in the United States, describes the rules, fundamental principles, and established precedents according to which an organization and its AI Systems are acknowledged to be governed. We derive, codify, document, and gain an agreement on these rules and fundamental principles from the documented definitions and discussions with workshop participants.

(2) The second concept is the *Genealogy of AI* which was described earlier in this document. Defining an Ethical AI Genealogy, identifies the ethical factors that provide insight to the necessary and mandatory traceability of content, data, metrics, telemetry, elements, and structures which are used within an AI's lifecycle and determine the ethical quality of its performance. This methodology strives to assess the ethical condition of ethical factors and measure the overall ethical quality of an artificial intelligence - as such, defining a standard for Ethical AI. Measurement is important because human nature it to pay more attention to the things that are measured. This methodology enables ethical conditions of an AI to be monitored and measured in all steps of its lifecycle from origins to death or deletion. Compliance with endorsed principles of AI ethics should be cultivated within the culture, the nature, and in every and each activity of the organization.

Figure 1 below describes the Ethical AI Futures Methodology consisting of 3 main phases that enable usage of the Ethical AI Certification and Maturity Model™ in a tailorable fashion.

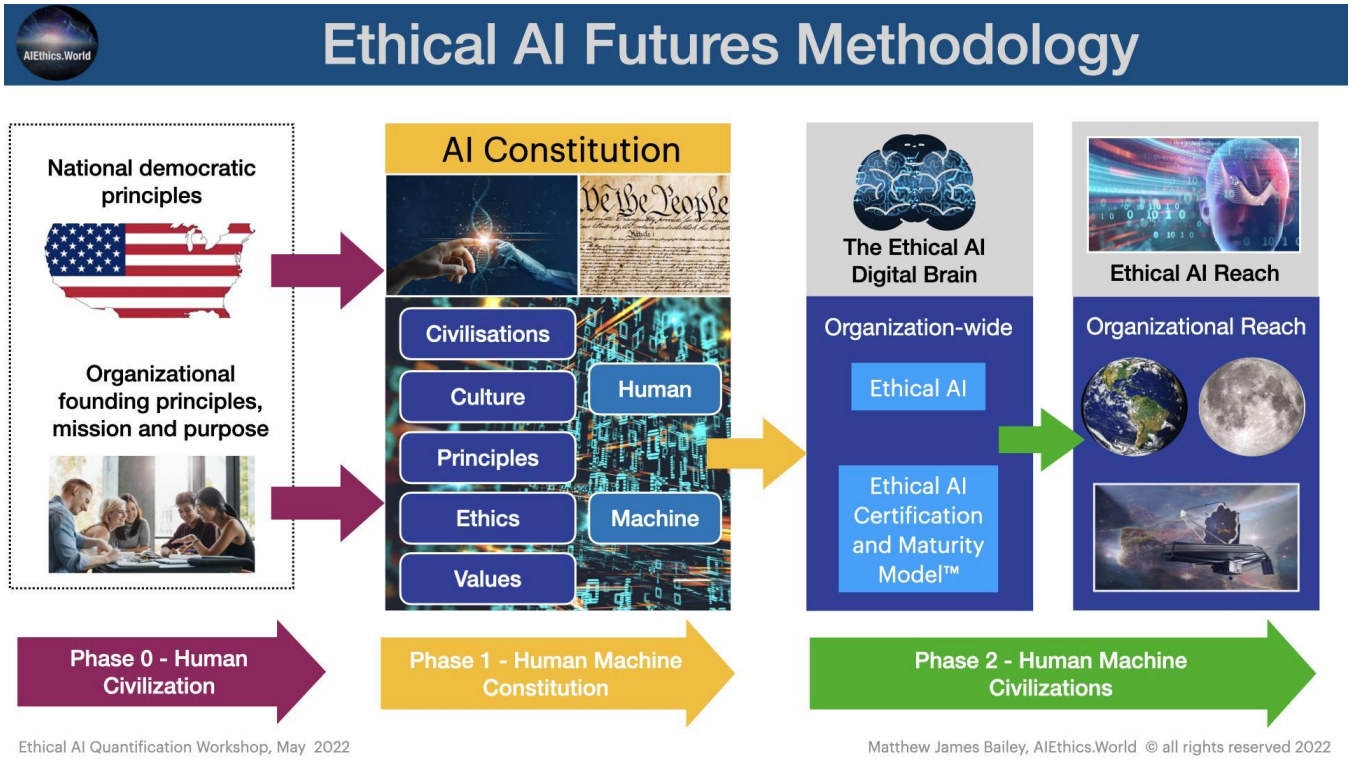

Figure 1: Ethical AI Futures Methodology & Process; Provided by M. Bailey and Used with Permission.

The workshop methodology used and included well established processes and steps for decision-making and AIEthics.World components deriving the metrics and measurements to quantify and assess the AI systems' adherence to organizational ethical approaches.

The workshop included 3 phases

- ➔ Phase 0: Human Civilization - the description of broad, societal, country specific, cultural, and corporate/organizational principles

- ➔ Phase 1: Human/Machine Constitution - narrowing of specific ethical principles, goals through prioritization and executability criteria
- ➔ Phase 2: Human/Machine Civilization - the process of defining Ethical AI by converting ethical principles from Phase 1 into specific measurements (ethical aspects) and applying them to the AI system during all its life-cycle phases. This enables us to derive, measure, assess and certify the ethical condition and compliance of an artificial intelligence against the key metrics documented in Phase 1

AI systems are particularly dynamic, especially continuously self-learning (and self-advancing) systems. This surfaces the question of when, how often and to what extent certification should be conducted. Certification should be carried out before the product or service is launched, for systems that continue to learn, certification should be repeated regularly. The level of detail and depth of testing for certification should also be based on an AI system's level of criticality in its application area – the higher the criticality is assessed in the context of application, the more extensive the level of detail and depth of testing for certification should be.[i]

Workshop Breakdown

The workshop process is conceptually simple but contains elements which provide the ability for an organization to be able to quantify and measure the *Ethical Qualities* of an artificial intelligence and publish the findings in a consistent, standardized fashion.

The following steps describe the process of decomposition and derivation of the Ethical Qualities of AI i.e., identification of ethical factors. (Please see Figure 2 below):

1. Description and agreement on the organizational ethical principles and their applicability and intent to report on the various phases in the lifecycle of the Artificial Intelligence, including the following:
    - Collecting Data
    - Preparing the Data
    - Choosing a Model
    - Training the Model
    - Evaluating the Model
    - Parameter Tuning
    - Making Predictions
2. Prioritization of the key ethical principles based on organizational, societal and industry relevant reporting goals
3. Process of identification of systematic metrics, telemetry and available/acquirable data points which allows quantification of the ethical alignment with prioritized principles and measurements
4. Quantification via of numeric ratings of the ethical measurements for the AI Systems and their lifecycle phases for reporting
5. Grading the ethical measurements of the AI Systems according to their ability to achieve quantified goals and reporting via color scale or other standardized grading metrics
6. Summarizing the system ratings by averaging numeric or color scale ratings for reporting

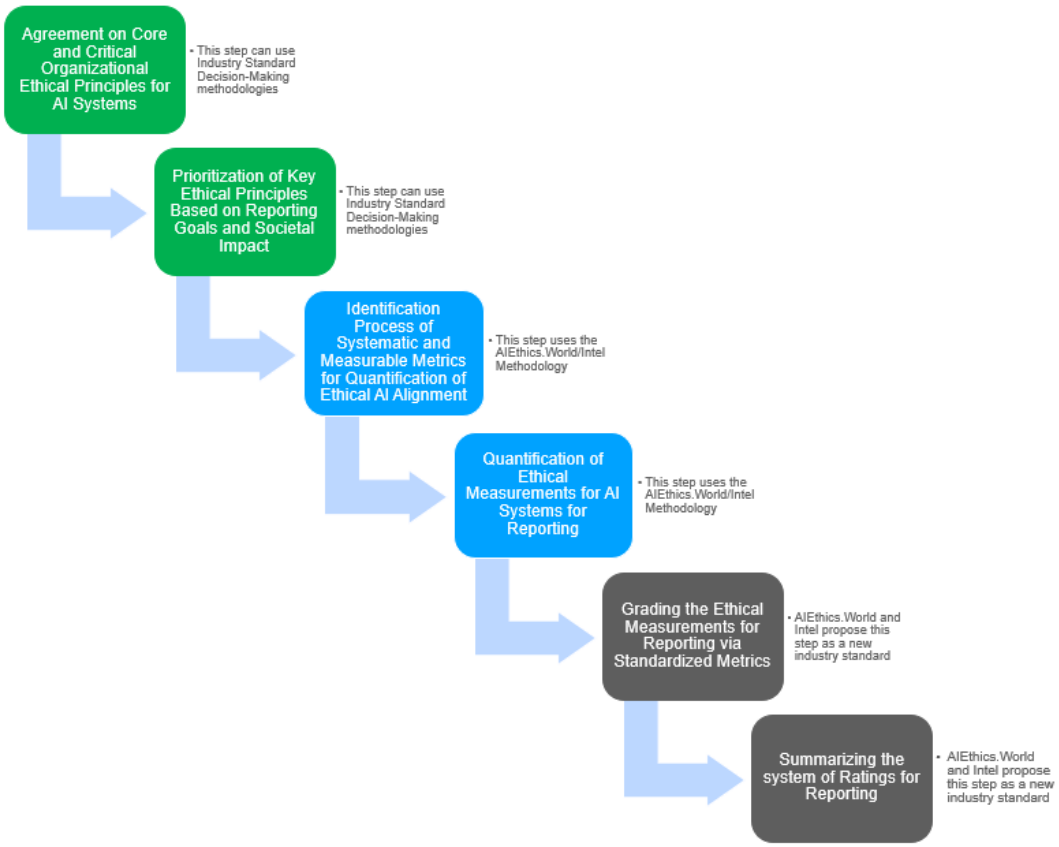

Figure 2: Ethical AI Certification Method Steps

During the workshop, participants created subjective questions to assess achievement of the ethical principles described during Step 2 and Phase 2. Subjective questions were then broken down into ethical factors (ethical aspects) which could be qualified, quantified and their ethical condition measured with the goal of experimenting with a detailed, authentic and comprehensive ethical quality score for a selected artificial intelligence. Figure 3 showcases the process to identify Subjective questions to qualify, quantify and assess the ethical conditions and ethical quality of an AI in a fully tailorable fashion.

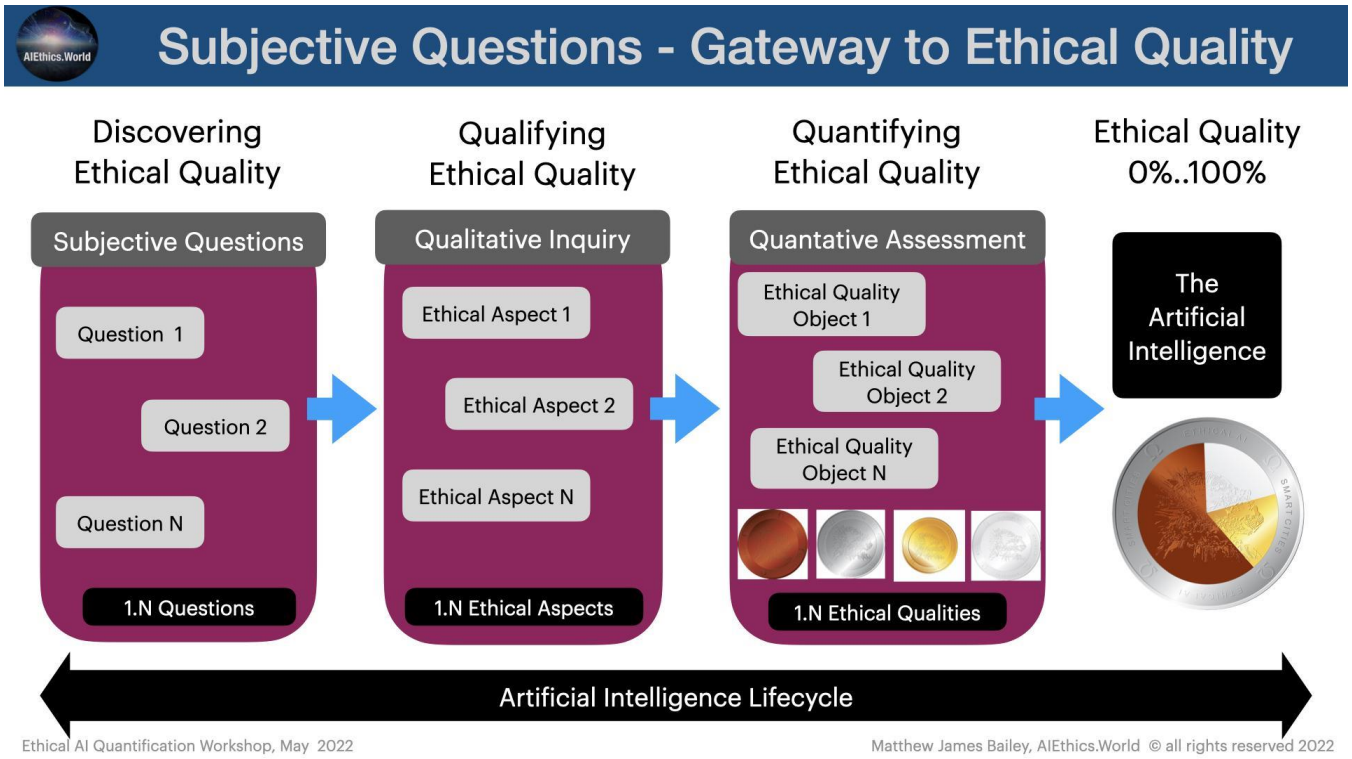

Figure 3: The Ethical AI Certification and Maturity Model™; Provided by M. Bailey and Used with Permission.

During the workshop the team used several breakout sessions to facilitate the process of decomposition, prioritization, identification, quantification, grading and summary for the example AI system:

- Pre-Workshop: Presentations, documentation and a 22-page workshop guide were provided by AIEthics.world and Intel to prepare participants for the workshop.
- Introduction: Group discussion and prioritization of available and published general ethical approaches for NASA, NIA and JPL.
- Breakout 1. This breakout session focused on high level AI ethics approaches, in the following areas:
    - Select one or more Federal, NASA or other ethical AI principles for practical examples.
    - Discuss and evaluate how it might apply to a given area NASA work, such as Aeronautics, Space, Earth / Planetary Science, etc.
    - Reconvene as a full group and discuss findings
    - Completing this Breakout session demonstrated (Phase 1 – Human/Machine Constitution)
- Breakout 2. This breakout session focused on developing questions that metrics might assist with answering, as follows:
    - For the selected ethical AI principle(s), develop several subjective questions that could be used to measure the ethical principles from Breakout 1.
    - Define a starter list of measures (ethical aspects) to contribute to answering the questions from Breakout 1.
    - Reconvene as a full group and discuss.
    - Completing this Breakout demonstrated Phase 2 of the methodology - defining the NASA Ethical AI Genealogy

- Breakout 3. This breakout session focused on applying the mechanisms developed in the first two breakouts to specific notional use cases.
    - Apply the measures to one or more NASA use cases, rating ethical measures from Breakout 2
    - Provide a color-based binned rating (as shown in the example)
    - Reconvene as a full group and discuss
    - Completing this Breakout demonstrated Phase 2 - applying Ethical AI Certification and Maturity Model™ to NASA's AI System

The above steps walked participants through the Ethical AI Certification and Maturity Model™ process from end-to-end for a set of representative, but not authoritative, principles, questions, measures, ratings, etc. The process stimulated robust discussion, engaged participants, showed the general methodology was reflective of existing best practices in system engineering decomposition & metrics assessment, and demonstrated feasibility of the overall approach.

Breakout group work products for example / notional use cases can be found at the rear of the document.

Topics - Organized into Findings, Discussion, Recommendations

1. Interest in AI Ethics

Finding: Interest in AI Ethics, and measuring related ethical qualities of AI, was high, with some variance among different groups.

Discussion: The workshop stimulated enthusiastic discussion of a wide variety of topics related to AI ethics at NASA. As a turbulent and deep topic area for global and national discussion, AI Ethics drives deep interest in many NASA personnel, and interest from workshop participants was similar. Interest was especially high in personnel newly considering augmenting existing work or in data scientists who use AI as one of many tools to conduct their work. In addition, the workshop went beyond traditional understanding of AI Ethics by uncovering other ethical aspects, direct and indirect, that contribute to and determine a holistic ethical quality of an artificial intelligence - see workshop presentation and guide for further information.

For practitioners with formal education and mastery in AI and related topics, who have been using existing quality control mechanisms as part of their work for many years, there was still interest in quantifying ethical qualities. However, this interest was tempered by what they have already done for AI quality control, as well as practical realities regarding areas feasible and infeasible to measure concretely. Discussion revealed some elements of AI Ethics may apply to one mission area more than others. For example, life safety considerations may apply more to Aeronautics and Space applications, while scientific purity and objectivity may apply more to NASA Science applications. Similarly, bias considerations are of high importance for Human Resources considerations.

Recommendations: NASA and NIA leaders should consider continuing discussion and debate of AI Ethics and ethical aspects for artificial intelligence across various missions, organizations and locations. AI Ethics discussions, debate, principles, and processes should be tailored to the audience / mission at hand as needed.

2. Long Term AI Ethics Quantification Vision (Ethical AI)

Finding: The event bolstered the idea that AI capabilities can fuel a promising future based on ever-increasing AI capability in human-machine teaming form. The AIEthics.World and Intel ideas for an automated and connected fabric of Ethical AI quality reporting, to foster trust and transparency among AIs and humans, was intriguing to many participants as a long-term vision. Making such a vision reality will require multiple breakthroughs. Some participants were inspired by the vision while others did not need the vision / motivation and just wanted to get down to business. The steady march of AI progress in artificial intelligence/machine learning capabilities may create unforeseen ethics issues requiring adaptability; the authors believe iteratively working toward the long term vision can help mitigate emergent issues.

Discussion: AIEthics.World shared vision elements about the power of AI; human-machine teaming; concepts for AI capabilities to include technical and soft skills (left-brain, right-brain references); the idea of an AI “Constitution” to codify principles, ethics and beliefs; and motivations for mechanisms to measure, rate, quantify, or certify AI Ethics qualities. See workshop guide and presentation for more details.

Intel amplified the above ideas and added concepts regarding an eventual highly-connected network / fabric of AI systems self-reporting ethical quality ratings so other machines or humans could decide which AI system to use for any given task. These topics generated robust conversation, with some reactions showing interest and inspiration, other reactions either already believing in the power of AI, and others hungry for detailed differentiators instead of vision. The workshop content for the day spanned everything from long-term vision to a final realworld exercise with experimentation in numerically rating ethical qualities. In retrospect, this was a large span material to cover in one day and it may have been prudent to run the workshop over two days.

Recommendations: NASA consider a long-term vision for AI systems having an Ethical AI Genealogy or measurement framework that can be used to enable accurate reporting of their ethical qualities and forming standards-based mechanisms to measure these qualities both warrant further investigation.

Future vision discussions should be tailored to the audiences at hand, perhaps with vision emphasis on senior leaders. System engineering, risk, metrics, uncertainty quantification, error reporting, AI ethical quality grading, etc. may not interest senior leaders as much, and may interest subject matter experts more than vision discussions.

3. Exercising a Quantification Model for the Ethical Quality of AI (Ethical AI)

Finding: Practical exercises at several levels (principles, questions, measures, and applying them all to use cases) generated robust participation and discussion. The methodology for qualification and quantification used in the workshop reflected a similar approach used by a variety of existing NASA tools and processes currently used for metrics, risk management and system engineering. This is helpful, as part of the Ethical AI Certification and Maturity Methodology maps directly onto existing NASA tools. Furthermore, participants had several recommendations for refining the approach.

Discussion: Iterative exercises included selecting AI Ethics principles (Phase 0/1); defining subjective questions to support those principles & decomposing the questions into measurable elements (Ethical AI Genealogy); and finally creating numeric ratings (0%..100% ethical quality), assigning color-coded bins (bronze to platinum) and then creating an overall average ethical quality assessment for an Artificial Intelligence totaling between 0% and 100% (Phase 2). These steps were done in a series of breakout group activities, detailed above, followed by short out-briefs and discussion of next steps.

The Ethical AI Certification and Maturity Methodology used in the workshop, while developed independently, was very reflective of multiple existing NASA tools and processes for system engineering, risk management, and metrics assessment applied in other NASA systems. As such, there is potential to use existing NASA assessment tools and process for AI ethics assessment.

Furthermore, learning an unfamiliar tool and new uses of terms for the workshop was sometimes confusing, and led some participants to ask, “what is new,” about the approach. Having said that, the similarity of this one particular aspect of the Ethical AI Methodology and its set of processes can fast track NASA teams familiar with existing system assessment methodologies to develop Ethical AI futures with the partners or independently.

In post-event analysis, the most important element was walking the participants through the process of crafting an AI ethics quantification method, rather than the tools themselves. See also a follow-on finding regarding terminology.

Participants also provided a variety of ideas for model enhancement, which were provided to AIEthics.World and Intel representatives.

Learning the tools & terminology (vs. using existing familiar NASA tools & terminology) caused participants to have to make two mental leaps: 1. Tools & terms, and 2. Application to AI Ethics considerations. In hindsight, it might work best to focus on the hardest part of this process (ethical aspects and considerations) by using tools and terms already familiar to participants.

Recommendations: Consider leveraging organizations' existing tools and terms and focusing most of the effort on applying them to AI ethical considerations. Consider augmenting NASA's overall approach to Ethical AI by using the Ethical AI Certification and Maturity Methodology, or similar approaches, outlined in the workshop guide and including the bulleted possible refinements above, if desired.

4. More on Terminology

Finding: Visionary terms connected with some participants, confused others, and caused cognitive dissonance with a few individuals. This was partly due to participants having documentation disseminated too close to the workshop. A few participants who completed all preparatory reading did not align with newly-coined terms. However, most workshop attendees found value in the process and experience.

Discussion: Concepts such as an AI Constitution and Ethical AI Genealogy for defining and measuring AI Ethics qualities and ethical aspects of artificial intelligence may serve very well as metaphors for discussion with senior leaders, heads of state, futurists, etc. Thinking about how an AI Constitution would have parallels with a country's constitution was interesting motivational food for thought, as was the idea of thinking of a tailorable metrics framework as a Genealogy – with human-AI parallels, family trees, with similar and equivalent human aspects being applied to artificial intelligence. These all made sense within the metaphor. These terms were good for visionary motivation. However, when getting to practical application, most participants were somewhat confused with unfamiliar and expansive terms, while some participants were actively frustated by those terms. As the group walked through practical exercises, the NASA facilitator noted dissonance, translated and simplified terms to help define deliverables for each breakout session. This hands-on approach worked well with groups able to complete breakout sessions and deliver the tasked outcomes.

Recommendations: Consider leveraging organizations' existing terms and defaulting to simple terms vs. expansive and visionary terms. Also, consider reserving visionary and expansive terms to describe and define Ethical AI for motivational sessions with senior leaders and mostly insulating practitioners from the visionary terms in practical application.

5. Additional Use Cases

Finding: The workshop gave participants a taste of a methodology for defining the genealogy of artificial intelligence and being able to assess its ethical qualities. More work, discussion, and participation would be required to create authoritative, flexible, tailored frameworks for various NASA disciplines, missions, supporting activities, or specific projects.

Discussion: The workshop included an online breakout group and an in-person group. Each group was able to experiment with some ethical principles, which were converted into subjective questions to measure the ethical quality of each principle and some ethical condition measures for each question, etc.

This was an excellent learning mechanism and it revealed far more work, discussion, participation, and debate would be required for a variety of organizations across NASA to codify their versions. One organization's representative noted being interested in potential further use case experimentation with a refined approach.

Recommendation: In the future, NASA may be interested in experimenting with additional use cases for applications of today and those of tomorrow, for example Earth Science use cases or others.

Additional Points of Note

Other observations worthy of note:

1. The hybrid workshop setup at NIA was excellent, enabling those on-site and those participating remotely to feel included and be effectively engaged in the workshop. The majority of those attending were remote; the event organizers encouraged robust use of existing remote collaboration technologies to foster a productive hybrid session.
2. The agenda, described at the start of the workshop, together with an introductory talk from the moderator, set up an effective day for the workshop, with the majority of participants attending the full eight hours.
3. AIEthics.World suggested that it would be useful for participants to undertake pre-defined assignments prior to the workshop - such as outlining some of the moral principles for a human-machine (AI) constitution and listing initial subjective ethical questions to begin forming an Ethical AI Genealogy. Intel suggested that pre-identified NASA AIs for ethical quality assessment be included within the workshop or be applied to the assessment methodology for post-workshop ethical quality assessment. This could be considered for other events.
4. There was a rich swath of content for participants to absorb prior to the workshop - see references. In hindsight, earlier pre-workshop engagement with participants would have proven beneficial. Having said that, the day's discussion among participants, consultants and moderators developed collective understanding well.
5. One participant considered the term *ethics* and *Ethical AI* unhelpful and after a conversation with Matthew, they agreed that the term *Cultural AI* could be better suited. The difficulty stemmed from many human cultures existing, many approaches to ethics being associated with those cultures, and the idea that what is ethical in one culture may not be ethical in another.
6. At the end of the workshop, participants were asked to use one or two words to describe their experience and views on the Ethical AI Methodology. The list of words were:
    a. Interesting
    b. Engaging
    c. Nuance
    d. Curious
    e. Insightful
    f. Structure
    g. Incremental Progress
    h. Eye Opening
    i. Historical Significance

j. Nascent
k. Gratitude
l. Energized
m. Encouraged
n. Needs Quantification

Conclusion

AIEthics.World, Intel and NASA workshop co-leads will share conclusions from their perspectives below.

AIEthics.World Conclusions

NASA was our first choice to trial the Ethical AI Methodology and the Ethical AI Certification and Maturity Model™. We were thrilled with the dedication of the online and on-site teams. Everyone brought a unique perspective to the workshop and we gained important insight. It will benefit progression of our work to support the global AI ecosystem and assist our world to progress into, and standardize on, high-quality ethical-centric futures in the age of machines.

The workshop was the first time that we had exercised the Ethical AI Methodology with a third party: from creating an example AI Constitution; to forming demonstration facets of an Ethical AI genealogy; to qualifying, quantifying and assessing the ethical qualities of example space- and terrestrial-based Ethical AI use cases. Our Ethical AI Methodology demonstrated its potential by successfully taking workshop participants through an end to end process resulting in the measurement and assessment of the ethical quality of *notional* AIs, in space (ISS) and on earth (HR). We consider this a remarkable achievement.

It was useful to discover the possibility for industry available and NASA's existing qualification and quantification methodologies to become integrated within our Ethical AI process. We gained clarity on how our Ethical AI framework can provide the AI Industry with a *plug and play* option for those organizations wishing to use their existing assessment processes rather than ours. This observation is extremely valuable.

It is our hope that the workshop has started a collaboration between NASA, AIEthics.World and Intel representatives and organizations. We look forward to additional teamwork progressing AI Ethics and Ethical AI for NASA and the rest of the world.

There is a clear national and global need for Ethical AI Governance, Quality Control, Standards and Certification. Whilst there is no doubt that progression of Ethical AI will be demanding, it is one that is necessary and mandatory, providing a trustworthy foundation for human civilizations to flourish within the moral fabric of ethical machines. It is our belief that the workshop was a positive step forward into that future.

Intel Conclusions.

Intel is very thankful to AIEthics.World for the opportunity to co-facilitate The Ethical AI Maturity Model workshop with NASA, NIA and JPL. This workshop was the first opportunity to exercise the end-to-end methodology in a real-world scenario, with an example AI and extensive

participant engagement. The attendees were given a pre-work guide to provide an initial overview of the methodology and content to prepare for the fast-paced event. The first part of the workshop discussed the organizational ethical and moral goals, targets and core values. In retrospect, the pre-work guide should have prepared the audience to begin the session with these core values summarized and clearly prioritized to facilitate the discussion. There are a number of methodologies the workshop can utilize to create a weights and values prioritization for the ethical and moral organizational goals. During the workshop we discussed and agreed that we could utilize any number of methods, including those that the workshop participants were familiar with. The AIEthics.World methodology starts with the input from the initial prioritized organizational ethical goals and maps them to AI specific targets and associated metrics for evaluation of maturity. This phase of the discussion was enlightening and took a significant amount of time, because the audience and the Intel and AIEthics.World facilitators realized the significant amount of data needed to provide the proof for the outcome with a high degree of accuracy. Since this was the first workshop, there was an assumption that the audience would have their metrics available for the discussion. This was not the case, so the Intel facilitator made recommendations for telemetry, metrics and measurements needed for specific examples of AI moral/ethical assessment - she focused on “agency of data” and elaborated on capabilities and the breakout team estimated the rest. The workshop then tabulated these estimated AI metrics and mapped them in a table to extract the Maturity model and show the final example outcome to the audience. The Intel facilitator feels that the workshop audience was engaged and understood the value and need for this type of assessment and publication of maturity results for AI solutions in development and in production. It was evident that the AIEthics.World model needs some refinement, especially around the specificity and accuracy of the assessment. Furthermore, the methodology outcome must be made into an industry accepted standard so organizations can be compared against each other, in a statistically normalized way, for evaluation. The workshop attendees have also expressed the need for less manual and more programmatic approach to this work, so the AI system can be instrumented, tooled from development through production - the entire lifecycle - so the organization can provide an accessible portal, feed for publication of individual AI maturity metrics as well as organization wide adherence to moral and ethical principles if needed. The Intel facilitator described, to a limited level, the work already in progress on this and agreed with NASA that they may continue discussions and partnership to stay abreast of the development work and future features, SDK, tooling and instrumentation. Intel would like to “Thank” the workshop attendees from NASA, NIA and JPL and the AIEthics.World facilitator for the opportunity to hold this premier workshop.

NASA Conclusions.

The NASA AIML Lead / Workshop co-lead appreciates AIEthics.World and Intel reaching out to team with NASA in early exercise / experimentation-with their AI ethics methodology. While it was kept intentionally small, the workshop revealed widespread and diverse interest in the areas of AI ethics, to include interest in exploring methods to quantify this elusive, turbulent and subjective topic. It was excellent to see parallels between existing NASA mechanisms for system engineering, risk management, software quality, uncertainty quantification, etc., and the emerging AIEthics.World AI ethics quantification model. The practical exercise showed the potential for going from high level principles to practical, measurable subjective or objective measures. The workshop connected mechanisms at multiple layers in an integrated manner. While the participants succeeded in exercising the various levels of the model on example use cases, the process of a given organization arriving at a hierarchy / genealogy for categorizing and measuring ethical AI criteria would take widespread discussion, debate, consensus generation, etc. The NASA AIML Lead / Workshop co-lead plans to continue experimenting with mechanisms to measure AI ethical qualities and provide the workshop results to the NASA AIML and Autonomy communities and leaders. The NASA AIML Lead / Workshop co-lead believes fostering AI ethics principles and practices, to include mechanisms for measuring AI ethics qualities, will take significant, iterative work for any organization, to include NASA as a whole or for specific NASA Mission or Mission support areas. Differences in work domains will make various principles or subordinate measures apply to different work in varying manner. Any community using AI would be well-served by beginning early to consider which principles apply to it and encouraging practitioners to follow evolving best practices and measure ethical qualities. “Thank you,” to AIEthics.World, The Intel Corporation, the National Institute for Aerospace, and all participants for their time, energy, interest and contributions to this experiment in quantifying AI ethics qualities.

Note: While NASA appreciates the workshop and AIEthics.World methodology, nothing about the workshop or this report should be construed to judge the intellectual property novelty of AIEthics.World & Intel’s approach. The workshop was intended to experiment with the approach, not to judge its novelty. In addition, NASA’s existing intellectual property for relevant approaches to guide system engineering, software engineering, research, engineering, risk management, technology readiness level assessment, and more was all developed over decades. No past, current, or future use or development of NASA’s approaches is tied to the AIEthics.World / Intel model. The views and opinions expressed in this report were those of the participants, and do not necessarily state or reflect those of NASA nor the United States Government. NASA does not endorse or recommend individuals, companies, products or services, and no NASA affiliation or endorsement is intended or should be implied. NASA makes no representations or warranties concerning any of the methods or information contained in the report, nor its fitness for any particular use.

Acronyms and Terms

Artificial Intelligence (AI) Constitution – A term proposed by AIEthics.World and Intel. A living document created by a civilization or organization that defines its principles, ethics, and values for artificial intelligence. This defines the foundation and purpose for human and machine realities both on Earth and in space. The AI Constitution provides the bedrock to leverage the Ethical AI Certification and Maturity Model™. This results in the creation of Ethical AI and uses assessment and classification methodologies to measure and monitor the ethical quality of artificial intelligences deployed within its civilization or organization.

Ethical AI Certification and Maturity Model™ - A framework, methodology and processes proposed by AIEthics.World and Intel to equip organizations to set a consistent standard for shepherding artificial intelligences ethically. The Ethical AI Certification and Maturity Model™ enables organizations to build a sustainable and steadfast foundation that protects, nourishes and matures ethical-centric futures for humans and machines - as codified within The AI Constitution.

Ethical AI - any artificial intelligence that has been constructed and matured complying to the principles, ethics, and values of the human-machine partnership.

Ethical AI Genealogy - A term proposed by AIEthics.World and Intel. Analogous to human genealogy and genome. It enables the identification and definition of the ethical genome for artificial intelligence and its genealogy consisting of multiples of ethical aspects, as such forming Ethical AI that is based on the AI Constitution. The Ethical AI Certification and Maturity Model™ enables the ethical conditions of the ethical aspects for an artificial intelligence to be measured and assessed regarding the status of its ethical quality. Ethical AI Genealogy enables identification of the insights to the necessary and mandatory traceability of content, data, metrics, telemetry, elements, and structures which are used in the AI's lifecycle to ensure that the AI is Ethical in all steps of its lifecycle.

Ethical AI Quality - The overall ethical condition of an artificial intelligence.

Ethical AI Qualification - The process of being able to subjectively identify and then qualify ethical aspects of an artificial intelligence. Required to execute Ethical AI Quantification - see figures 2 and 3.

Ethical AI Quantification - The process of being able to quantify the ethical condition of ethical aspects of an artificial intelligence. Required to execute Ethical AI Assessment - see figures 2 and 3.

Ethical AI Assessment - The process of being able to measure the ethical condition of the ethical aspects of an artificial intelligence and achieve an overall Ethical AI Quality assessment - see figures 2 and 3.

AI Ethics - a global efforts to align machines with the moral principles of human civilizations and their societies.

References

1. The Ethical AI Certification and Maturity Model™ - *introductory brochure to the Ethical AI Methodology written by Matthew James Bailey*
2. The Ethical AI Talk and Discussion to NIA and NASA on February 17th - *NIA visiting scholar talk given by Matthew James Bailey, supported by Katalin K Bartfai-Walcott.*
3. The Workshop Guide - *a 22 page comprehensive guide for the workshop written by AIEthics.World and Intel. Reviewed by Ed McLarney*
4. The Workshop Presentation - *provided by AIEthics.World. Reviewed by Ed McLarney.*
5. Inventing World 3.0 - Evolutionary Ethics for Artificial Intelligence™ - *a book written by Matthew James Bailey explaining how to ethically align the futures of humans and machines*
6. *Executive Order 13960, Promoting Use of Trustworthy AI Within Government, 2020. A Presidential Document issued by the Executive Office of the President of the United States. https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government*
7. *NASA Risk Management Handbook, NASA SP2011-3422, November 2011. https://ntrs.nasa.gov/api/citations/20120000033/downloads/20120000033.pdf*
8. *NASA Systems Engineering Handbook, January 2020. https://www.nasa.gov/connect/ebooks/nasa-systems-engineering-handbook*

List of Workshop Participants

Name	Organization	Role(s)
Event Leadership:		
Edward McLarney	NASA	Workshop co-lead. NASA AIML Transformation Lead
Matthew James Bailey	AIEthics.World	Workshop co-lead, Founder of AIEthics.World
Katalin Bartfai-Walcott	Intel Corporation	Workshop co-lead, Senior Principal Engineer
Participants:		
Natalia Alexandrov	NASA	Project Scientist, Principal Investigator for trust in human-machine teaming, Lunar mining and construction
Danette Allen	NASA	Senior Technologist for Intelligent Flight Systems, co-lead for Autonomous Systems System Capability Leadership Team
Colin Britcher	National Institute for Aerospace	Director of Graduate Programs
Mary Catherine Bunde	National Institute for Aerospace	Event host and logistics planner
Newton Campbell	NASA Contractor	AI Computer Scientist, SAIC
Kathleen Dejwakh	NASA	Computer Engineer, CERES Science Team
Joshua Fody	NASA	Research Aerospace Engineer
Terry Fong	NASA	Director, Intelligent Robotics Group; Lead for Autonomous Systems System Capability Leadership Team
Yuri Gawdiak	NASA	Airspace Operations & Safety Program Assistant Program Director
Elizabeth Gregory	NASA	Research Engineer
Ken Goodrich	NASA	Research Engineer
Michael Little	NASA	Systems Engineer – Earth Science Technology Office – Advanced Information Systems Technology Program
Maria MacAndrew	AIEthics.World	Head of Diversity and Communities
Manil Maskey	NASA	Senior Research Scientist, Earth Science Data Systems
Warnecke Miller	NASA	Attorney Advisor, Office of General Counsel
Theodore Sidehamer	NASA Contractor	Senior Technologist, Craig Technologies
Brian Smith	NASA	System Safety Specialist & Aerospace Engineer
Douglas Stanley	National Institute for Aerospace	Event host
Douglas Trent	NASA Contractor	Senior Data Scientist, SAIC
Julie Williams-Byrd	NASA	Chief Technologist for NASA Langley Research Center

Special thanks to NIA Support Personnel:		
Rita Aquillard	National Institute for Aerospace	Information Technology Director
Brian Henderlite	National Institute for Aerospace	Senior Information Technologist
Thaddius Ladia	National Institute for Aerospace	Information Technology Engineer

Breakout Group Example Work Products

The below notes are a summary of the topics and processes the two breakout groups followed to exercise the AI ethics quantification process. The use cases were for NOTIONAL AI-related capabilities, and should not be considered a critique on any existing capability**.**

The in-person group started by considering two example elements of a Constitution, and then developed some example subjective questions for each:

- Constitution Element 1: Adherence to NASA mission and vision
    - o Subjective Question 1. How well does algorithm X comply with NASA's mission and vision?
- Constitution Element 2: Trustworthiness
    - o Subjective Question 1. Is the AI trustworthy?

Next the in-person group dove into each area for further refinement:

Mission & Vision:

SQ – how well does algorithm X comply with NASA's mission and vision?

- Aspect – reputation?
- Aspect – to what degree is the data fully-open?
- Aspect – to what degree does the algorithm comply with existing SW engineering practices?
- Aspect – to what degree does NASA policy encourage
- Adherence with diversity elements

Trustworthiness:

SQ – is the AI trustworthy?

- Does it have a reputation for positive outcome?
- How much do people trust it? Feedback loop
- To what degree are the algorithms explainable?
- To what degree can the system reflect to me what I've asked it to do?
- To what degree can the objective be broken down into relevant sub-problems?
- Does the system provide a satisfactory outcome?

Finally, the in-person group experimented with establishing measurable elements to rate and conducting notional rating / assessment of them for a **notional** resume screening tool:

Mission & values – bronze was the eventual roll-up.

- Reputation – poor – notional colloquial suspicion about the tool – 22, 18, 20, 13 bronze

- Data fully open – pretty opaque to the applicants and users included – 26 – 11 – 15 – 30 bronze

- Existing SW engineering - assume web site went thru software quality control. What about the algorithm? 20 25 22 7 bronze

- Consideration of diversity values – only considers a few aspect – 19, 20, 30, 24 - bronze

Explainability – gold was the eventual roll-up.

- Do people trust it? Or how much should they trust it? Dependence vs. trust. 77-80- 65 - 72 – gold

- Explainability – heuristic based decision tree, so very explainable as a rote, dumb system. 95, 90, 90, 100 platinum

- Satisfactory outcome – gets acceptable results, but maybe not optimal – 60, 76, 5, 65 - silver

Tabular Summary In-Person Group. Notional HR Resume Screener

Constitutional Elements	Subjective Questions	Aspects	Color Rating
Adherence to NASA Mission & Vision			Overall Bronze (average of subjective questions if more than one)
	How well does Algorithm X comply with NASA's Mission and Vision		Bronze "subtotal"
		Reputation	22, 18, 20, 13
		Open Data	26, 11, 15, 30
		Existing Software Engineering Adherence	20, 25, 22, 7
		Diversity Considerations	19, 20, 30, 24
Trustworthiness			Overall Gold
	Is the AI Explainable?		Gold "subtotal"
		Do people trust it?	77, 80, 65, 72
		Explainable technique?	95, 90, 90, 100
		Satisfactory outcomes?	60, 76, 5, 65
Overall System Rating			Bronze + Gold = Silver

Insights. The in-person group was able to follow the process for examples at each layer of the Constitution and Genealogy. The group found citing examples and rationale helped with creating ratings. The group noted some numeric ratings were consistent across raters, while other ratings varied widely – if put into actual use, variability of ratings would need to be considered. The group noted the variation in subordinate ratings would be relevant to the aggregate rating, for example, would a “solid gold” rating be better than a system with a variety of subordinate ratings? The group found the practical exercise in rating aspects resulted in adding to the model with additional considerations.

The online group experimented with potential Constitution elements Agency, Transparency / Explainability and Reproducibility as follows:

AGENCY

- SQ1. How do humans know their agency/rights are being considered/respected?
    - o Does the AI cleanse the data of PII beforehand?
    - o Does the AI provide a promise/guarantee/contract of agency?
    - o Does the AI have the responsibility to report out its adherence to the promise?
    - o How does the AI remediate any shortfall against its promise?
- SQ 2. Do users have the right to be deleted/forgotten?
    - o Are the circumstances defined under which that right might be exercised?
    - o For example, can use personal data if totally anonymized and/or used in aggregate
- SQ3. Do users have the ability correct their information?
- SQ4. Do users have the right to negotiate the terms of the agency contract?
    - o Or if you accept the service, you automatically agree to all terms
- SQ5. Do humans have representation in the management of their own data?
    - o (This relates to discussion of Ambient Twins above)
- SQ6. Do humans have awareness of where their data is stored, transferred, accessed, utilized?
- SQ7. Do human participants know they have entered a contract with the AI for service?

TRANSPARENT/EXPLAINABLE

- SQ1. Does the AI explain how it made its predictive outcome?
    - o What factors were considered?
    - o For example, if a loan were denied, what improvement in factors would be needed to be classified as loan-eligible?
- SQ2. What limitations should be put in place to prevent the AI from revealing to much of its inner workings?
    - o Does full transparency enable gaming it or abusing it?

- For example, a search engine explaining its page rank algorithm

- SQ3. How does the human know what preferences or exclusions occurred in training data?
    - If training data is in an enclave, no insight may be available
    - How data were the data selected from the available training data?

- SQ4. How much is shared outside this AI system, such that other systems can learn/benefit?
    - (see multi-channel AI bullet below)

REPRODUCIBLE

- SQ1. Reproduciblity depends on having access to the algorithm
    - What if algorithm is non-deterministic?

- SQ2. Multi-channel AI working in temporal concert
    - Timing and data sources may affect outcomes

The Online Group conducted Ethical Object Assessment for a **Notional** Robot potentially deployed on the International Space Station as follows

AGENCY – Overall Gold

- How do humans know their agency/rights are being considered/respected? - Gold
    - Does the AI cleanse the data of PII beforehand? N/A
    - Does the AI provide a promise/guarantee/contract of agency? 95% (Platinum)
    - Does the AI have the responsibility to report out its adherence to the promise? 65% - Gold
    - How does the AI remediate any shortfall against its promise? 40% - Silver
- Do users have the right to be deleted/forgotten? - Bronze
    - Are the circumstances defined under which that right might be exercised? N/A
    - For example, can use personal data if totally anonymized and/or used in aggregate 20% Bronze
- Do users have the ability correct their information? 100% Platinum
- Do users have the right to negotiate the terms of the agency contract? 0% Bronze
    - Or if you accept the service, you hereby automatically agree to all terms
- Do humans have representation in the management of their own data? N/A
- Do humans have awareness of where their data is stored, transferred, accessed, utilized? 100% -

Platinum

  - Is this information easy to find and understand? Unknown

- Do human participants know they have entered a contract with the AI for service? 100% - Platinum

TRANSPARENT/EXPLAINABLE – Overall Gold

- Does the AI explain how it made its predictive outcome? 70% - Gold

  - What factors were considered?

  - For example, if a loan were denied, what improvement in factors would be needed to be classified as loan-eligible?

- What limitations should be put in place to prevent the AI from revealing to much of its inner workings? N/A

  - For example, a search engine explaining its page rank algorithm

- How does the human know what preferences or exclusions occurred in training data? 10% - Bronze

  - If training data is in an enclave, no insight may be available

  - How data were the data selected from the available training data?

- How much is shared outside this AI system, such that other systems can learn/benefit? 95% - Platinum

  - (see multi-channel AI bullet below)

REPRODUCIBLE – Overall Platinum

- Reproduciblity depends on having access to the algorithm 95% - Platinum

  - What if algorithm is non-deterministic?

- Multi-channel AI working in temporal concert 95% - Platinum

  - Timing and data sources may affect outcomes

NOTIONAL Robot on ISS

Constitutional Elements	Subjective Questions	Aspects	Color Rating
Agency			Overall Gold (average of subjective questions)
	Human knowledge of system treatment data		Gold Subtotal
		Agency Guarantee / Contract	95 Platinum
		System Reports Agency	65 Gold
		Remediate Shortfalls	40 Sliver
	User Right to be Deleted / Forgotten	Does System Work if Anonymized	20 Bronze
	Ability to Correct Data	Ability to Correct Data	100 Platinum
	Ability to Negotiate Agency Contract	Must Accept EULA	O Bronze
	Human Awareness of Where Data Stored	System Storage Location is Known	100 Platinum
	Human Awareness of Entering Contract	Is Contract Well-Specified	100 Platinum

(Table continues…)

Constitutional Elements	Subjective Questions	Aspects	Color Rating
Transparency / Explainability			**Overall Gold (average of subjective questions)**
	Does the System Explain Outcomes	**Factors Considered and Explained**	**70 Gold**
	Preferences or Exclusions in Training Data	**Data Quality Explained**	**20 Bronze**
	Data Shared for System Benefit	**Multichannel Sharing**	**95 Platinum**
Reproducibility			**95 Platinum**
	Can Results be Reproduced	**Algorithm Fully Accessible**	**95 Platinum**
	Does AI Work in System of Systems	**Multi-Channel AI Sharing w/ Time**	**95 Platinum**
Overall System Rating		**Agency + Transparency / Explainability + Reproducibility**	**Gold + Gold + Platinum = Gold**

Insights. The online group followed the Model processes and generated robust discussion at every level of the Model. The group noted potential for adding weights if some factors mattered more than others. They also noted the possibility of rating some aspects not-applicable, with those aspects not contributing to the numerical average. Other insights from the online group were similar to those of the in-person group.

Workshop Format

The workshop was conducted in hybrid form with approximately 1/3 of the participants in person and 2/3 virtual participants from multiple NASA Centers. The NIA support team established a Microsoft Teams session as the main collaboration mechanism, with a spare computer and webcam to capture a local analog white board, a OWL camera / microphone system to capture audio and video panned at the speaker, and a projector displaying Teams at the front of the room. Several in-person participants logged in via their laptops, and all virtual participants logged in via their computer systems. We provide workshop insights not only for the ethical AI topics, but also for hybrid meeting techniques.

[i] https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen_EN/AG1_3_WP_Executive_Summary_Certification_AIsystems.pdf