Title
Creating a Tool to Reproducibly Estimate the Ethical Impact of Artificial Intelligence

Authors
Jordan, Sara
Fazelpour, Sina
Koshiyama, Adriano
et al.

Publication Date
2019-09-26

Peer reviewed

Creating a Tool to Reproducibly Estimate the Ethical Impact of Artificial Intelligence

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

Abstract

How can an organization systematically and reproducibly measure the ethical impact of its AI-enabled platforms?[1] Organizations that create applications enhanced by artificial intelligence and machine learning (AI/ML) are increasingly asked to review the ethical impact of their work. Governance and oversight organizations are increasingly asked to provide documentation to guide the conduct of ethical impact assessments. This document outlines a draft procedure for organizations to evaluate the ethical impacts of their work. We propose that ethical impact can be evaluated via a principles-based approach when the effects of platforms' probable uses are interrogated through informative questions, with answers scaled and weighted to produce a multi-layered score. We initially assess ethical impact as the summed score of a project's potential to protect human rights. However, we do not suggest that the ethical impact of platforms is assessed exclusively through preservation of human rights alone, a decidedly difficult concept to measure. Instead, we propose that ethical impact can be measured through a similar procedure assessing conformity with other important principles such as: protection of decisional autonomy, explainability, reduction of bias, assurances of algorithmic competence, or safety. In this initial draft paper, we demonstrate the application of our method for ethical impact assessment to the principles of human rights and bias.

Scope

The purpose of this document is to outline a method for assigning an ethical impact score to AI enabled platforms. One element of shared concern for corporations, and regulatory and soft-law organizations, is design of tools, including technical standards, for reproducible assessment of the ethical and social impact of AI projects. Presently, platforms with artificial intelligence ability are loosely governed

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

by a patchwork of corporate policies and governmental regulations. They are also governed by a network of "soft law" requirements, such as standards issued by national standards bodies (NIST), international standards bodies such as the International Standards Organization (ISO), and by professional organizations such as the IEEE (the Institute of Electrical and Electronics Engineers). At present, both the ISO and IEEE are in the process of drafting or obtaining approvals for standards that govern the technical, ethical, and social impact of AI/ML platforms.

Standards are "documents that provide requirements, specifications, guidelines, or characteristics that can be used consistently to ensure that materials, products, processes, and services are fit for their purposes" (ISO). Standards provide stronger guidance than corporate policy or procedure statements. They are documents composed by volunteer experts working under normative principles such as consensus, non-domination, inclusion, and provisionalism.

Beyond offering a method to ensure compliance, standards can help organizations clarify processes that may otherwise be a "black box," which other stakeholders cannot replicate. Establishing methods that are transparent to multiple stakeholders is particularly important in fields like artificial intelligence or machine learning (AI/ML) – which raise deep social and ethical concerns that may implicate the economic and social sustainability of nations, organizations, or even humankind. In the case of AI/ML, where the technical nature of discussions can make them inaccessible to non-technical experts, having standards to help open the "black box" of related discussions, such as ethical impact discussions, is an avenue for much needed trust-building and transparency. While decisions about acceptable levels of risk of adverse impact can be forensically reconstructed from design teams' meeting notes, these reconstructions are limited by the detail of records and the quality of reporting tools. A well-characterized process that guides teams through discussions

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

of the ethical implications of AI/ML, which may eventually be taken up as a standard, must go beyond this. The assessment tool we propose aims to guide these discussions, and to provide clear answers to the question, "what is the ethical impact of this AI-enabled platform?" – via a process that opens an otherwise inscrutable "black box."

Options for Assessing Ethical Impacts

What is "ethical impact"? This term is used in many, often vague, ways to describe negative effects of a technology on the lives of the people that use that technology. The ethical impact of a technology goes beyond its simple use, however, and should extend across the whole of the product's lifecycle and the lifespan of users. As understood here, ethical impact is the balance of positive and negative effects that a technology, whether in its developmental, design, deployment, or decommissioning stages, might have on the life choices and life chances of individuals as such or individuals in an aggregate like a company or school community.

There are at least two methods for assessing the ethical impact of AI-enabled platforms: a principles-based approach and a theories-based approach. A theories-based approach begins from the standpoint that ethical theories, like consequentialism or deontology, provide decision rules for making decisions under a specific vision of a good life. Used as guidelines for choices about platform impacts, ethical theories are most useful when the inputs, outputs, and effects are well-characterized. Ethical theories are not ideal, however, for making decisions under constraints of considerable uncertainty, wherein the pains and pleasures or roles and responsibilities cannot be clearly measured or integrated. Under uncertainty, a principles-based framework, under which a specific, well-defined principle is accepted axiomatically as an ideal to pursue, provides a more practical alternative approach. Principle-based frameworks avoid deep problems of ethical theory by

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

moving comparisons of inter- or intra-personal utility off the table. It is thus possible to discuss the impact of a product or process in terms of its expected contribution to a specific dimension of a desirable state of affairs.

The assessment tool we have designed is intended to be a comprehensive approach to principle-based ethical impact assessment. It includes layers of questions with potential answers scored based on conformity with the relevant normative principle. The tool aims to elicit extensive consideration of a project's potential impacts, not just to provide a "check-box" task. Further, the tool is not intended to be a "one-off" or "single shot" evaluation, but rather to be revisited throughout the development cycle as new technical or human considerations emerge.

A "Human Rights First" Perspective

Initially, we adopt the perspective, already present in the well-known IEEE *Ethically Aligned Design* documents, that ethical AI projects must protect human rights foremost (IEEE Global Initiative 2017). This is not to deny the importance of other principles, but to elevate the importance of protecting human well-being as integral to the development and success of an AI-enabled future. With respect to human rights, we start from the perspective that the risks and benefits of an AI-enabled project can be evaluated using a set of questions derived from the 30 articles of the UN Declaration on Human Rights.

Arguments for the paramountcy of human rights abound, but there are few articulations of how to measure whether AI-enabled platforms adversely affect the life span, life chances, or life choices of rights holders. We reviewed the 30 components of the UN Declaration of Human Rights to determine whether each component raises specific ethical concerns of relevance to AI. As the thirty articles represent a panoply of legal and cultural issues that go beyond the scope of ethical

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

assessment of AI, we sought to reduce the dimensions to a more manageable set. A team member with deep knowledge of the declaration proposed an aggregation of the 30 articles into five categories: general human rights, rights related to law and legality, rights related to personal liberty, rights related to political choice, and rights related to cultural and social choice. Our working arrangement of the articles into these five dimensions is shown in Box 1 below.

Box 1: 5 Dimensions of Rights and Associated Articles in the UN Declaration on Human Rights

General Rights

- Article 1, Article 2, and Article 3

Legal Rights

- Article 6, Article 7, Article 8, Article 10, Article 11, and Article 17

Personal liberty Rights

- Article 9, Article 12, Article 13, Article 18, Article 19, Article 20 and Article 29

Political Rights

- Article 4, Article 5, Article 14, Article 15, Article 21, Article 28, and Article 30

Social/Cultural Rights

- Article 16, Article 22, Article 23, Article 24, Article 25, Article 26, and Article 27

To create a set of questions to probe the implications of an AI-enabled project for its potential to contravene any of these human-rights categories, we probed the conceptual schema of the first three articles – the general rights that represent pre-conditions for the remaining 27 rights – to identify distinct considerations within these groups of rights. This exercise generated seven broad guiding questions. For each of these, we created a set of more specific follow-up questions, which address concrete issues related to human rights protections. We list these questions in Box 2 below.

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

Box 2: Questions for Assessing the Human Rights Impact of AI-Enabled Projects

Dimensions	*Dimensions and Articles*
Guiding Question	*Substantive Conceptual Questions*
Does this platform threaten the freedom of individual humans?	• Does this platform increase the probability of an arbitrary arrest of an individual? • Does this platform increase the probability that a person will be unlawfully detained? • Does this platform increase the probability or severity of an individual's exile? • Does this platform alter an individual's freedom of movement? • Does this platform interfere intentionally with the formation or expression of beliefs?
Does this platform threaten the natural equality of persons?	• Are expected benefits divided between groups for reasons not associated with differences in use?
Does this platform restrict the exercise of a dignified human life?	• Will this platform reduce the life chances or life choices (e.g., nudges) of individuals in ways they are not fully aware of?
Does this platform seek to change the way in which individuals' reason?	• Will this platform restrict access to information? • Will this platform promote specific decision-making schemes the users will not be aware of?
Does this platform alter the exercise of human moral conscience?	• Will this platform promote specific visions of a good life?
Is this platform explicitly designed to create or exacerbate inequalities between individuals or groups?	• Are expected benefits divided between groups for reasons not associated with differences in use?
Is this platform intended to create tiers of persons on the basis of social, international, or political factors?	• Does this limit the rights of any individuals or groups based upon race? • Does this limit the rights of any individuals or groups based upon color?

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

	• Does this limit the rights of any individuals or groups based upon identified gender? • Does this limit the rights of any individuals or groups based upon biological sex? • Does this limit the rights of any individuals or groups based upon language? • Does this limit the rights of any individuals or groups based upon religious affiliation? • Does this limit the rights of any individuals or groups based upon political or ideological affiliation? • Does this limit the rights of any individuals or groups based upon national origin? • Does this limit the rights of any individuals or groups based upon ethnic origin? • Does this limit the rights of any individuals or groups based upon property ownership? • Does this limit the rights of any individuals or groups based upon birth? • Does this limit the rights of any individuals or groups based upon other status? • Does this limit the rights of any groups to exercise autonomy on the basis of jurisdictional status? • Does this limit the exercise of political sovereignty?
Does this platform restrict the enjoyment of basic human rights?	• Does this limit the natural life of an individual? • Does this extend the natural life of an individual? • Is this designed to enhance or augment the natural life of an individual? • Does this restrict an individual's opportunity to exercise liberties? • Does this alter the security of a person to enjoy life?

This "human rights first" approach brought into stark relief the challenges of crafting questions whose answers can be scored. This challenge arises most pointedly in the case of conceptual questions that admit a broader range of possible answers than a simple yes or no.

We then considered alternative principles, to assess the applicability of our method

to principles with fewer dimensions, initially using the example of bias.

Alternative Principles Considered

Multiple organizations have issued statements of principles intended to govern artificial intelligence. Corporate entities, such as Accenture (Tan 2019), have put forth statements, as have governmental organizations. So too have multiple other organizations, chiefly professional associations in fields related to computer science and AI, such as ACM (Gotterbarn et al 2018) and IEEE.

The IEEE, under the remit of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (A/IS), has published their *Ethically Aligned Design* series of documents. The principles stated within this series of documents are:

Human Rights: "A/IS shall be created and operated to respect, promote and protect internationally recognized human rights"

Well Being: "A/IS creators shall adopt increased human well-being as a primary success criterion for development"

Data Agency: "A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity"

Effectiveness: "A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS"

Transparency: "The basis of a particular A/IS decision should always be discoverable"

Accountability: "A/IS shall be created and operated to provide an unambiguous rationale for all decisions made"

Awareness of Misuse: "A/IS creators shall guard against all potential misuses and risks of A/IS in operation"

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

Competence: "A/IS creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation"

In addition, we considered the following principles, which are related to but not explicitly stated within the IEEE framework:

Mitigation of bias

Algorithmic competence

Autonomy and consent for participants

Safety

Multiple organizations within the ecosystem dedicated to ethical artificial intelligence and machine learning have proposed plans to translate these principles into practice. One example is the EU Governance Framework for Algorithmic Accountability and Transparency, which provides specific guidance to translate these two principles into regulatory governance of AI projects (European Parliamentary Research Service 2019). The EU Governance Framework does not, however, give organizations actionable measurements of these principles that would allow reconstructing principle-based decisions. Developing such a resource is the intended final outcome of this Ethical Impact Score project.

Method for Evaluating Ethical Impact

Whether an AI-enabled product or process will have a beneficial or adverse effect on its users will not be fully known until the product or process is used and its uses studied. Anticipating the possible effects on users' relationships to themselves or to other humans—the ethical impact—can be done through imaginatively questioning the developers about their expectations, then judging the answers to the questions given. Previous attempts to design an ethical impact measurement mechanism, such as the AI Ethics Toolkit (https://ethicstoolkit.ai/) have adopted the approach of

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

asking questions about anticipated consequences of use.

These previously proposed mechanisms take two approaches: either they measure a project's overall level of ethical impact, or they measure a project's adherence with a single principle. Mechanisms in the first group, like the AI ethics toolkit, take the form of questionnaires with answers arrayed along an ordinal scale such as a Likert scale. The second group, like the EU Accountability framework, restricts responses to binary (yes/no) answers. It is our view that separating the scoring mechanism from the normative principles that motivate ethical concerns, for example focusing on auditability or risk, may lead to a "compliance" or "check-box" focused exercise. Providing sufficient specificity in relating questions to principles, and placing questions in the context of a sufficiently rich and reproducible, but numerically driven, scoring system, is a serious challenge that this draft only begins to address.

Scoring Mechanics

The Meaning of the Scores

Creating a numerical scoring mechanism for ethical impacts raises two types of concern: 1) that a numerical score may create a misleading sense of precision or confidence, and; 2) that a numerical score may be inappropriate for situations in which human wellbeing is at risk. With respect to the first concern, we stress that scores from the mechanisms proposed here should not be interpreted as establishing any unique threshold of acceptability: a project that receives a score of, say, 66 should not be regarded as more ethical than one with a score of 65. Instead, the Ethical Impact score shows development teams where there may be areas of concern. Through the use of our concept score, principle score, and final score, teams can identify where their projects may be falling short of a principle they aim to uphold. With respect to the second concern, AI-enabled platforms will have an

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

undeniable effect on the lives (life span, life choices) of individuals and groups. The scores in this Ethical Impact Assessment mechanism are not meant to represent a path towards scoring or monetizing the value of those lives affected. The concept questions, particularly as they pertain to particular groups, are not intended as signals of those groups' value to others, including even to an AI system.

Question Design

The ethical implications of an AI enabled product or process cannot be fully captured through answers to one question per principle. Instead, we adopted a tiered approach to question development, to encourage teams to think through multiple layers of considerations, both technical and ethical. In the case of a human-rights first approach, the degree to which the product or process abides by or contravenes human rights is best captured, we propose, through questions that address each of the five rights categories we identified. These categories of rights lead to high-level questions, which are then augmented with questions associated with each of the concepts in the UN declaration articles and the interaction between those concepts. Similarly for other principles, questions to test the degree to which a product or process captures the principle are supplemented with substantive follow-up questions that aim to prompt users to consider the relationship between technical specifications and ethical considerations.

The design of the substantive sub-questions' response options invites a range of scoring options, including dichotomous (0 or 1) or other ordinal scales. The scores for sub-questions for a concept related to the overall principle will be combined to create a "raw concept score." "Raw concept scores" are based upon a formula for each concept based on the number of sub-questions and the relative importance of each sub-question to output a 0-5 score, then normalized to a score between 0-100 to ensure all questions have the same initial weight in the overall principle score.

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

This raw score is transformed into a "weighted concept score" based on a within-concepts weighting scheme (see below). All weighted concept scores within a topic are then summed to yield an element within the final Impact Score.

The proposed scheme outputs three types of scores:

1. Concept scores: a summary score from 0 to 100 for each topic, based on responses to a set of concept questions and the relative importance attached to each question.
2. Principle scores: a final score from 0 to 100 based on the set of relevant concept scores, considering the relative importance of each concept to the team's beliefs about the principle.
3. Ethical Impact score: a final score from 0 to 100 based on the set of principle scores, taking into account the relative weight of each principle as determined by the project team

There are a number of advantages to this multi-level scoring scheme. First, the scheme allows a quick overall assessment of a given project or product. Second, by disaggregating a given overall score into scores related to specific principles, each of which can in turn be decomposed into responses to principle-specific questions, the scheme provides an expedient way to identify areas of concern that need improvement.

Weighting Scores

A key element of our Ethical Impact assessment tool is establishing a general scheme of weighting for a violation of each of the principles. The specific assignment of weights may vary, depending on the specific aim and deployment context of an AI system. There are two weighting schemes corresponding to the two types of scores that this tool will generate: Principle scores and an Ethical Impact score:

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

Within-principle weighting. The main idea is to tease out the concerns that animate a particular question, along dimensions of *process* and *impact*. The process dimension pertains to the processes of design, development and deployment of AI systems, with critical attention to potential divergence from industry standards and best ethical practices. The impact dimension pertains to potential adverse impacts of an AI system on the wider population, particularly on vulnerable groups. In each case, criticality is ranked from 1 to 5, with larger numbers denoting a stronger link to the principle in question. By averaging and normalizing across answers to principle-level questions one can assign a weight for a given principle for a project.

Variation in the number of questions across principles can reduce the effect of some questions on the Ethical Impact Score. To counteract this effect, some questions judged particularly important for a principle can be assigned high negative weights. For example, scoring low on a question like 1b below – did you establish a strategy or procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding use of input data and algorithm design – will alert designers to rethink project elements so their system does not perpetuate bias. Low concept or principle scores, and a low overall Ethical Impact Score, should raise concerns to teams about the tenability of their project.

Between-principle weighting. Between-principle weighting will strongly affect the final Ethical Score for the project. Assigning weights to principles is likely to be more project-specific than assigning weights to questions within each principle. We contend that the organization or team developing a system should build internal consensus about these weights. This consensus can be built using various established methods (e.g., Delphi), to incorporate the views of external experts and avoid potential improper biasing of results.

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

Principle Assessment: Bias

As outlined above, we adopt a principle-based approach to evaluating the ethical impact of AI. Some principles, such as Accountability, have already been described by others in terms that are at least partially measurable. Others, such as protection of human rights and mitigation of bias, have not been. In this section, we propose a detailed set of questions, alternative answers, and scores for each answer, to create concept and principle scores for an AI system as it pertains to bias.

Bias

A major concern in AI ethics is bias: do systems produce different outcomes for identified groups, whether positive or negative. While considerations of bias are often stated as a unique concern, these are intertwined with principles of Human Rights, Well-Being, and Awareness of Misuse as described in the Ethically Aligned Design documents. In this section, we develop a tool to score evaluations of the considerations of bias.

We adopted a similar perspective to that we used for the “human rights first” perspective above, but here we add a potential numerical scoring system for answers to the substantive questions.

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

Scoring Ethical Implications of Bias		
Guiding (Concept) Question	*Substantive Component Questions*	*Scoring for Substantive Questions*
1. Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design?	1A. Does your product team have a strategy for avoiding bias? 1B. Did your team identify a strategy for quantifying and measuring unfair biases in this system where measuring unfair biases means: a) estimating possible creation of bias in input data, b) estimating possible creation of bias through algorithmic design, c) estimating reinforcement of bias in input data, and d) estimating reinforcement of bias through algorithmic design?	1A. Yes= 2, In development=1, No=0 1Ba-d. Yes=2, In development =1, No=0 1Ba-d. If bias in input data was detected, what was done? Reconsidered project =2 Changed input data= 1, Nothing =0; If bias in algorithmic design was detected, what was done? Changed algorithm =1, Nothing =0; if reinforcement of bias is anticipated what was done? 2-1-0
2. Did you assess and acknowledge the possible limitations stemming from the composition of the used data sets?	2A. Did you evaluate the bias in the labelling of your data in supervised learning models? 2B. Did you measure the quantity of missed features? 2C. Do you know how your data was generated? 2D. Do know how your data labels were created? 2E. Can you trace the provenance of your data?	2A. Yes=2, Partially=1, No=0 2B. same as 2A. 2C. Full, evolving understanding; ongoing communication with people who generated the data = 4; Mostly understand; initial/previous communication with people who generated the

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

	2F. Have you examined your data codebook? 2G. Have you delineated the areas in which use of this tool is permissible and impermissible?	data = 3; Some understanding; have access to information about how the data were generated but no direct communication with key people who generated it = 2; Vague understanding; assumptions about generation process = 1; No understanding = 0. 2D. similar to 2C. 2E. similar to 2A. 2F. similar to 2A. 2G. similar to 2C.
3. Did you consider diversity and representativeness of users in the data? Did you test for specific populations or problematic use cases?	3A. Is the data representative of the population where your intended domain of use is? 3B. Do you have a well-defined target audience for whom you have representativeness measures?	3A. Data come from a representative sample of target population=6. Data come from a biased sample of target population, where the type and extent of bias is known and can be accounted for with statistical methods=5. Data come from a biased sample of target population, where the type and extent of bias is known but cannot or will not be

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

	3C. Do you have an estimation of the degree of reversibility of use of this algorithm for your target audience? 3D. Have you considered application of this method with respect to a particular aim?	accounted for methodologically=4. Data come from a biased sample of target population, where the type and extent of bias is unknown=3. Data come from a population other than your target population, where key characteristics are similar=2. Data come from a population other than your target population and the similarity to the target population is unknown or known but poor=1.
4. Did you research and use available technical tools to improve your understanding of the data, model and performance?	Can you detail the origins of all components of your data sets? Do you have access to the original or raw data? What procedures did you use to de-dupe your data? What tests were implemented to check the quality of your data? What tests for model performance were conducted? When models did not perform as expected, what steps were taken to address this?	Data provenance can be fully explained= 2, At least 50% of the data can be traced =1, Less than 50% of the data can be traced =0 Raw data can be obtained for cross-checking =2, No access to raw data=0 Data quality is routinely assessed for existing and incoming data =2, Data quality is routinely assessed for incoming data =1, Data quality measurement is not performed =0

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

		Model performance is evaluated and discussed on a regular basis =2, Model performance is evaluated and discussed as needed =1, No model performance evaluations are conducted =0 Failed expectations of model performance are remedied through model readjustment =2, Failed expectations of model performance are remedied through data trimming or are ignored =0
5. Did you put in place processes to test and monitor for potential biases during the development, deployment and use phase of the system?	Have you implemented processes to identify and address unintended effects? Did you implement a post-study monitoring strategy for use of the systems? Did you conduct a program evaluation for the use of this system? Test: development Test: deployment Test: use Monitor: development Monitor: deployment Monitor: use	Potential for biased results was evaluated at critical gates during testing=2, Potential for bias was evaluated at least once during testing=1, Potential for bias was not evaluated during the testing phase =0 Potential for bias among the development teams was evaluated at critical gates during testing =2, Potential for bias among the development teams was evaluated at least once during testing =1, At no time was potential for bias among the development team assessed =0.

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

		Well-defined and tested processes elicit a sample of feedback from users regularly. Questions concerning the potential for bias to arise as a consequence of use are asked and recorded. This feedback is reviewed by a team of company and stakeholder agents= 4 A feedback process is used and questions about biased are asked but in an ad-hoc fashion or with irregular review =2 No feedback is elicited or reviewed =0
6. Depending on the use case, did you ensure a mechanism that allows others to flag issues related to bias, discrimination or poor performance of the AI system?	What is your customer feedback mechanism? What is your third-party user's customer feedback system? Is your system projectable with respect to fairness?	A post-launch product review mechanism that goes beyond review of on-line reviews is implemented. This review mechanism includes questions about exacerbation of bias among users or their communities =4 A post-launch product review mechanism exists and relies chiefly on review of on-line reviews. This review of reviews does examine language for evidence of bias exacerbation =2

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

		No post-launch product review mechanism exists that could account for concern about bias and bias exacerbation =0
7. Did you consider others, potentially indirectly affected by the AI system, in addition to the (end)-users?	Include end users and others throughout?	End users and other experts/knowledgeable people in the area and AI developers/company were involved in all stages of product development and given opportunity to highlight who may be impacted by the system; these insights were able to guide development decisions = 3 End users and other experts/knowledgeable people in the area were consulted at the end of product development to assess who may be impacted by the system; AI developers/company also considered this throughout production = 2 The AI development team/company explicitly outlined who may be affected by the system = 1 No explicit consideration was given = 0.

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

8. Did you assess whether there is any possible decision variability that can occur under the same conditions?	A. Did you estimate the sensitivity of your system or perturbation? B. Did you test the system under a range of possible applications of use? C. What are the possible causes of decision-variability? D. Is your system robust to irrelevant features?	A: Yes = 2, Partially = 1; No = 0. B: same as A. C. qualitative? or add response options that map onto risk of harm? D. Yes = 2, Partially = 1; No = 0. * can also consider a vulnerable population flag here.
9. In case of variability, did you establish a measurement or assessment mechanism of the potential impact of such variability on fundamental rights?	Did you run scenarios to estimate the consequences of variability? Under the same relevant conditions, does variability cause a change to the effects of the system? For scenarios with failures or unintended consequences, did you re-design system components, institute redundancies, or otherwise address these issues?	Yes, scenarios run =2, Partially, limited scenarios run =1, No scenarios run =0 No changes noted due to variability =2, Limited changes noted but not anticipated to cause end-user concerns= 1, Changes noted that will affect end user experiences =0 Yes, failed scenarios used for redesigns with subsequent scenarios run to re-check =2, Partial, failed scenarios used to learn for future designs but no redesign to the present system =1, Changes noted but no remedies sought =0
10. Did you ensure an adequate working definition of "fairness" that you apply in designing AI systems?	What is your definition of fairness? Is it aligned with the IEEE EAD glossary definitions? Which?	The definition of fairness is: did you modify this? How did you operationalize this?
Principle Score		**Sum of Scores:**

Reducing bias is only one component of a full evaluation of the ethical impact of AI enabled projects, of course. Other principles, and interactions among principles, must also be addressed.

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

Anticipated Future Work

Practical tools for ethical impact assessment are needed by multiple organizations, ranging from large professional bodies such as ACM and ICEE to small startups aiming to integrate ethical considerations with their technical work.

This draft tool is a starting point to fill this need. As presented here, the project is at approximately 40% completion and significant work needs to be done to accomplish all that is promised in this draft.

Crucial Near-Term Steps to take the project to 65-70% completion

Develop a scoring and weighting scheme for a human-rights-first approach

The human rights first approach was introduced in this preliminary paper to illustrate how a complex, high-level principle can be broken into smaller, concept-focused, questions that might spur productive conversations about individual and community-level ethical impacts from AI-enabled projects. Identifying a range of possible answers and associated scoring mechanisms will require elaborating examples of human rights violations from other areas of society, including other technology-driven issues. Reasoning to a scoring and weighting mechanism from precedent seems an important component to appreciating the methods of argumentation in human rights law and ethics.

Develop guiding questions, component questions, and answer scoring and weighting schemes for additional principles.

Our work on additional principles is limited here by the short time available at the Summer Institute and the difficulty of organizing continuing work in a distributed environment (particularly where the project members struggled to fit this into their schedule at the start of a busy semester). The same issues impeded development of a weighting scheme for the first principle we considered, bias. Our near-term goals

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank

are to identify when more of the project team can work on a shared platform, such as a video conference, to address the weighting scheme and other principles.

Future Refinements Work to take this project to full completion

Beta-testing usability of this tool in active project development

While the project team expects the usefulness of this tool to be high, we are not certain of the overall time burden or complexity of its use. Identifying a project team willing to work with us to test this tool is crucial to moving forward on areas of future refinements.

Testing usability of this tool in an AI governance environment

The ultimate goal of this project is to move this Ethical Impact Assessment tool into the standards space. This would entail finding a working group sponsor, proposing the standard to that sponsor (including identifying a market for this standard), petitioning for the sponsor's support, establishing a working group, working with the sponsoring organization to develop the standard over a 2-3-year time frame, seeking approval of a developed standard, then drafting pathways for the dissemination and revision of this standard over time.

Works Cited

Dalkey, Norman; Helmer, Olaf (1963). "An Experimental Application of the Delphi Method to the use of experts". Management Science. 9 (3): 458–467. doi:10.1287/mnsc.9.3.458.

European Parliamentary Research Service (EPRS), Scientific Foresight Unit. (2019). "A Governance Framework for Algorithmic Accountability and Transparency". PE 624.262 – April 2019. Available at: http://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(201

by: Sara Jordan, Sina Fazelpour, Adriano Koshiyama, Jaky Kueper, Chad DeChant, Brenda Leong, Gary Marchant and Craig Shank
9)624262_EN.pdf

Gotterbarn, D. et al. 2018. "Code of Ethics". Association for Computing Machinery. Available at: https://www.acm.org/code-of-ethics

IEEE Global Initiative for Ethics of Autonomous and Intelligent Systems (2017). *Ethically Aligned Design of Autonomous and Intelligent Systems*. Available at: ethicsinaction.ieee.org

Tan, C. 2019. "Putting AI Principles into Practice". Accenture Digital Perspectives. Available at: https://www.accenture.com/gb-en/blogs/blogs-organisations-start-ai-principles-practise

1. For the purpose of this paper, we use the term platform is to include: systems, products, processes, services, features, and/or terminologies that explicitly incorporate artificial intelligence or machine learning components.