ASEE 2022 ANNUAL CONFERENCE
Excellence Through Diversity
MINNEAPOLIS, MINNESOTA, JUNE 26TH-29TH, 2022 ASEE
Paper ID #37769

# Assessment of Ethics and Social Justice Aspects in Data Science and Artificial Intelligence

**Franz Kurfess (Dr.)**

**Katya Nadine Vasilaky (Assistant Professor)**

**Tina Cheuk (Assistant Professor)**

Assistant Professor, Cal Poly, San Luis Obispo

**Ryan Jenkins (Associate Professor)**

**Grace Nolan**

# Assessment of Ethics and Social Justice Aspects in Data Science and Artificial Intelligence


## *Abstract*

This work aims to develop a set of materials and tools, both quantitative and qualitative, for two purposes: First, for the assessment of ethical and social justice (ESJ) considerations in research projects, and second, as a pedagogical toolkit that allows users to improve their understanding of these aspects of data ethics. Below we describe three existing assessment methodologies for evaluating ESJ in data science research projects: a scoring rubric, a questionnaire, and a *canvas sheet* (i.e., a user-friendly template and tool that captures data), and we propose one additional method, a predictive machine learning model. This document describes an evaluation of the feedback from 124 students in two different classes who used the questionnaire and canvas sheet to assess their team projects. This data set is also being used to test a proof of concept for the machine learning model. Our emphasis at this stage is to improve the instruments, with a quantitative analysis of the numerical and scale-based responses, and a qualitative evaluation of the text-based suggestions from participants. The primary insights from this first round of evaluations indicate that students showed no strong preference between the questionnaire and the canvas sheet, with slight advantages on "Perspective" and "Further Research" for the *canvas sheet*, and a similar advantage for "Group Discussion" for the questionnaire.


## Introduction

The impact of research activities on human participants or animals is subject to certain constraints, regulations, and expectations. At many institutions, such work is subject to review and approval by Institutional Review Boards (IRB). In our work, we are expanding the scope of these considerations to the wider area of ethics and social justice that may not be currently present in the scope of IRBs' review (Datenethikkommission, 2019; European Commission, 2021; Ferretti et al., 2022; Hao, 2021; Huh-Yoo et al., 2021; Kitchens, 2019; Rose et al., 2018; Tranberg et al., 2018). In other words, there are significant challenges that are specific to Big Data that include "computational complexity, methodological novelty, and limited auditability" that may result in potential physical, emotional, social, legal, and economic harms (Ferretti et al., 2022). Our goal is to go beyond ethical concerns where the goal is to 'do no harm,' but to move towards more just outcomes. That is, our contribution to the field acknowledges that the world we live in, and those who hold decision making power hold biases, and these biases have resulted in systematic and historic inequities for those on the margins (i.e., Black, indigenous, people of color, disabled, etc.), especially pertaining to the area of data sciences and artificial intelligence (Benjamin et al., 2019; Charitidis, 2019; The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019; United Nations, 2015). To this end, our team aims to design solutions that not only mitigate harm but do good — and improve the lives of these marginalized communities.

Within the context of a strategic research initiative at our university, California Polytechnic State University in San Luis Obispo (Cal Poly), a group of about thirty researchers collaborated on a Data Science Initiative (DSI), with an informal vision of 'Data for All, Data for Good'. A working group, Ethics and Social Justice (ESJ), emerged from this larger team effort and was tasked to examine ways to assess the consistency and outcomes of activities with the collective's stated vision. This notion of 'Data for All, Data for Good' captures the deeper and widely held convictions that data scientists, like other professionals, minister to the well-being of society, broadly construed (AI for Good Foundation, 2019; AI for Good Foundation & Syngenta, 2017; Bloomberg, 2015; Desai, 2021; KDNuggets, 2015; Koschinsky, 2015; McGregor & Banifatemi, 2018; Spanache, 2020; Syngenta & AI for Good Foundation, 2017; Vieweg, 2021). The main goal of the proposed toolkit is to create a broad and accessible framework that can be used by researchers who do work in the fields of applied data sciences. This toolkit is intended to serve as a planning, evaluation, and reflection guide for research teams who leverage data sciences methods and tools in their work. Developed by the ESJ group, the toolkit provides guidance on the development of data science research projects that moves toward more ethically and socially just processes and outcomes, while generating new knowledge, opportunities, and ideas for the field at large (Brown & Mecklenburg, 2021, 2021; Ethics & Compliance Initiative (ECI), 2021; Forsberg, 2004; Marshall, 2009; Open Data Institute, 2021; Thereaux, 2021).

The toolkit includes a set of questions team members should ask and reflect on throughout the research development process: ideation, proposal writing, sampling/data collection, instrumentation, analysis, implementation/interpretation of outcomes & products, etc. As a first step, we noticed a need to create a basis of understanding and vision at the institutional level among students, instructors, and scholars on the significance of the ethical and social justice impacts of data science. One of the objectives of this toolkit is to establish a common language among individuals and teams to communicate ethical and justice concerns on data science projects that have direct and indirect implications to barriers to an inclusive, just, and sustainable world.

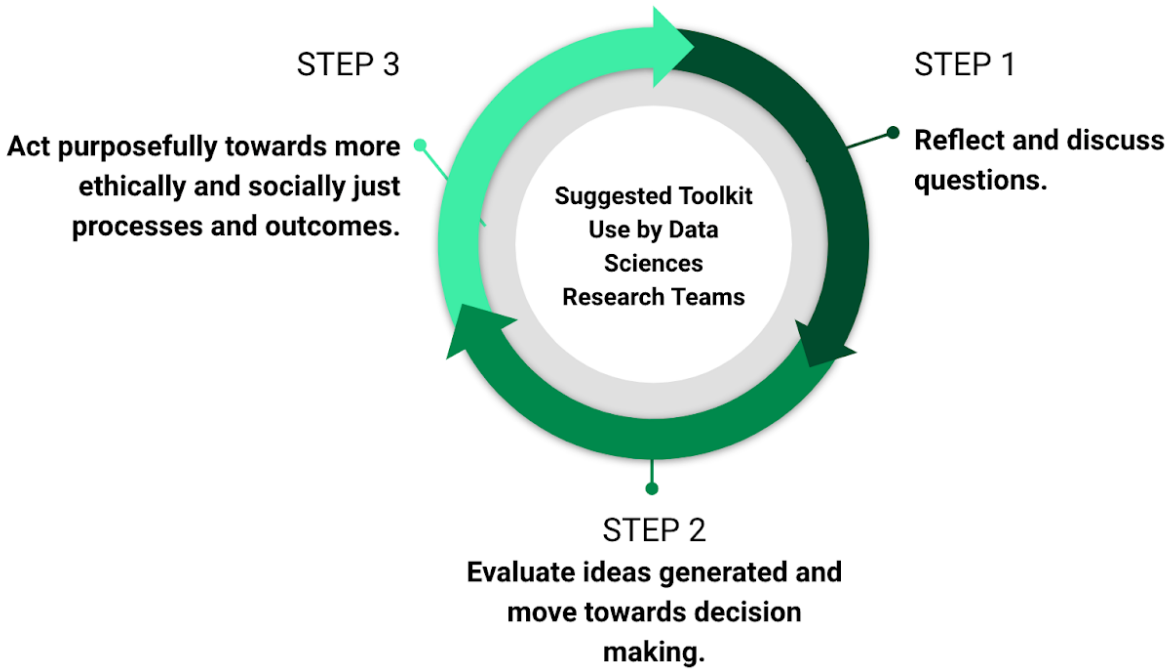


Figure 1: ESJ Reflection, Evaluation, Action Cycle

In Figure 1, we diagram an iterative cycle of questioning, reflection, and discussion that allows team members to build a shared knowledge base on the ethical and social justice implications in the conceptualization, design, and application of the knowledge, resources, and tools that emerge from data science projects. We envision different components of our toolkit to be used in different steps, and are currently evaluating the suitability of the components for different points within the life cycle of a research project.

As a basis for this assessment, we reviewed related work and assembled educational material for information sessions, workshops, and self-study (ACM, 2018; Athey, 2017; Chatila & Havens, 2019; Datenethikkommission, 2019; European Commission, 2021; Gebru, 2020, 2021; IEEE Standards Association, 2021; Open Data Institute, 2021; The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019; Tranberg et al., 2018). As an interdisciplinary group, we met weekly for over a year sharing ideas and coming to a consensus on ways we can center research so that this toolkit could be used to have the largest positive impact on communities that have been systematically and historically marginalized in our societies.

**Analysis and Assessment Methods**

Guided by the charge of 'Data for All, Data for Good', we explored a variety of methods to assess relevant ethical and social justice aspects within these two inter-related goals. Our work includes the following four approaches:

1. The ESJ Scoring Rubric (Appendix A) with distinct descriptors and criteria levels where evaluators select the most appropriate statement to express the degree to which a project

meets the respective criterion. Our initial rubric design was intended to be a tool for external reviewers (e.g., funding committees) used to give guidance to the project team as a form of summative feedback. However, our team recognized that this toolkit could be used as a formative tool for research teams as they conceptualize and develop their project proposals. In its current form as a spreadsheet, this rubric is not practical for large groups of participants. Conversion into a rubric for use in a Learning Management System like Moodle or Canvas[1] is under development.

2. The ESJ Questionnaire (Appendix B), derived from the rubric, with a mixture of quantitative Likert-scale questions and open responses for comments. Here, our initial assumption is that researchers use it to assess their own projects, or for peer evaluations of projects.
3. The ESJ Canvas Sheet (Appendix C), consistent with the rubric to capture essential ESJ aspects for a project on a single page. This tool looks suitable for initial group discussions about project ideas.
4. A predictive model for ESJ assessment based on Natural Language Processing and Machine Learning methods. This tool is intended to support the decision-making process of individuals or groups throughout the ESJ Reflection, Evaluation, and Action Cycle, not as a stand-alone system for the automated assessment of projects or activities.

At the beginning of our project, we explored the idea of quantifying data ethics qualities in research proposals via "data ethics scores". We defined various dimensions of 'Data for All' and 'Data for Good' and proposed a scoring system for each criterion, which we describe below. Within our iterative design processes, we realized that the quantitative measures alone might not be sufficient in empowering project teams to engage with the more nuanced and challenging conversations that are critical in the ethical and justice applications of our work in data sciences. As a result, we built in a protocol (e.g., reflection questions) that aims to catalyze and frame discussions among team members so that the work could better realize our collective vision of 'Data for All' and 'Data for Good'. With the rubric, we put the emphasis on text-based responses in order to reduce the temptation of using a quantification process of ethical and social justice qualities, for two main reasons. First, it was not possible to validate the scoring system. For example, weighting various dimensions would create confusion and controversies about why certain dimensions are receiving a higher score. Additionally, the criteria are not of equal value and cannot be reduced to a numerical value. Second, we didn't want to reduce the purpose of the toolkit to a perfunctory checklist. In our view, it is more important that scholars critically reflect on each criterion and take action in revising the project proposal and plans so that the work better reflects the vision of 'Data for All' and 'Data for Good'. However, we recognize that text-centric assessments are significantly harder to process and consolidate. They can work well in a setting with a panel of highly qualified evaluators who examine a small number of projects

[1] Please note that the term Canvas is used as the product name of the Learning Management System by Instructure, and also as a descriptor for a "canvas sheet" used as an assessment instrument. To emphasize the difference, we will use "canvas sheet" or "ESJ Canvas" for the latter.

individually and then convene for an overall assessment. In our educational context, we wanted to see how this set of ESJ tools could be used by undergraduate students who may be embarking in research and careers that use and apply data sciences across various disciplines. The use of this toolkit is paired with student-generated proposals that students themselves practice evaluating. As a result, students deployed this toolkit with both quantitative and qualitative elements. This included the ability to 'score' or evaluate each criteria against a numerical scale and an open response text box where students can provide the presence or absence of evidence that supports their quantitative evaluation. In addition, we are exploring the use of an ESJ canvas, where important aspects are examined in a limited space equivalent to one sheet of paper (ADAPT Centre for Digital Content Technology, 2017; Open Data Institute, 2019).

These methods can be used for the assessment of existing or proposed projects and activities by an ESJ panel or similar group for self-assessment by individuals and groups engaged in or considering an activity, and as an educational method to raise awareness for ESJ issues that emerge from Data Sciences research. At this point, we are examining their suitability through usability testing with different potential user populations: Students in project-centric classes use them for educational purposes to raise their awareness of these issues, and researchers apply them to their own projects and those of others for self-assessment. At later stages, we are planning to use them for the assessment of proposals submitted to funding sources interested in collaborating with our group. Since in most settings it is impractical to apply all of these methods in parallel or sequentially, we will use the results from initial evaluations to identify scenarios for their use. Once we have stable versions of all instruments, we will examine multiple scenarios, such as

- *Ideation and Brainstorming*: ESJ Canvas sheet by the development team (Step 1 in Figure 1)
- *Project Proposal Self-Assessment:* ESJ Questionnaire by the development team or a mock evaluation team (Step 2)
- *Project Proposal Submission Check*: Predictive model used by project management or institutional review board (Step 3)
- *Proposal Assessment*: ESJ Rubric used by a funding source, or for self-assessment by the development team after project milestone completion (Step 3)

Our experiences gathered from usability evaluations (see the Experimental Design section) will guide us through further explorations of the above and other situations.

***ESJ Rubric***

The working group mentioned above developed the set of questions to be used for evaluating projects in terms of Data for All and Data for Good. An initial version of this rubric is included as Appendix A: ESJ Rubric.

Our primary aim for the rubric is to have evaluators answer these questions with open-ended text to encourage and allow for reflection on each of these topics. The rubric can be used for a summative evaluation, e.g. by external evaluators to make comments based on the evidence presented or absent in the proposal. It can also be used as a formative tool for research teams to reflect on their own work to see the degree by which their proposed project meets the vision of 'data for all, data for good'. Considering that there are circumstances when a numeric score is expected or required, the open-ended text responses may be accompanied by a numerical score. Through a few iterations of usability testing on sample projects with small numbers of participants, we arrived at the version described below.

Regarding **Data for Good**, we broke down our questions into the following categories:

- Social Good: research objectives, research participants, policy implications.
- Data Methods: project selection and scope, data curation, methodology, and transparency and replication.

The aim with Social Good is that researchers prioritize studies that improve the well-being of individuals, communities, or the spaces that they live in. In particular, we want researchers to consider the populations they are highlighting or emphasizing in their work and how that may have external consequences for, or potentially marginalize, other populations who are not participating in the study. Along these same lines, the project should have policies that aim to distribute results and resources equally among the general population and in an accessible way. The aim with Data Methods is for researchers to consider the quantitative and qualitative techniques being used. This can include how and what data are being collected and if these collection methods pose risks to participants in or outside the study. It asks researchers to consider the various ways in which "biases" can influence their work. This includes human bias that may be compounded by machine learning algorithms and data sampling biases that influence the predictive models (Benjamin et al., 2019; D'Ignazio & Klein, 2020; Noble, 2018; O'Neil, 2016). Finally, Data Methods also asks researchers to consider the documentation of their work – whether the work can be replicated and if the processes used were transparent.

Regarding **Data for All**, we broke down our questions into the following categories:

- General Population,
- Local Community,
- Community within the organization (Cal Poly in our case)
- Historically Underrepresented Minorities,
- Indigenous Populations,
- Environment, and
- Urban and/or Lived Spaces.

While this is not an exhaustive list of the categories that a data science project may target, we felt it covered specific enough categories to help an evaluator focus on the relevant literature and

historically relevant facts for that category. In addition to asking the evaluator to apply our questions regarding Data for Good in reference to one or more of the groups listed, we also ask evaluators to consider the researchers' expertise with these populations or environments either through past work or lived experiences.

In sum, the resulting rubric (see Appendix A: ESJ Rubric) serves multiple purposes. It captures our insights from background research, discussions, and usability testing for the assessment of projects regarding ESJ aspects. It also serves as a reference for other tools using different ways of capturing responses from evaluators, such as questionnaires and canvas sheets. Furthermore, it acts as a stepping stone for the development of a predictive model based on open-ended responses that require reflection rather than categorization.

***ESJ Questionnaire***

The structure of this questionnaire is derived from the rubric, with some modifications to accommodate limitations of the questionnaire format and the tool used (Google Forms). An anonymized version of the one used for the first iteration of usability tests in the winter term is included as Appendix B: ESJ Questionnaire.
It contains three assessment components together with administrative elements including an overview, an Informed Consent Form, and project selection information. The Data for Good and Data for All sections consist of a series of Likert scales, one for each subcategory identified in the ESJ Rubric section above. At the end of each section is a text box with room for a longer response. While we experimented with a setup that included a textbox for each category, user feedback from pilot studies indicated that the majority preferred the significantly shorter version with a single textbox for each category. These sections are formulated to collect objective feedback (to the degree possible) from assessors: Did the authors of the documents examined address the issues for the respective category? In order to explicitly collect the subjective opinion of the reviewers, we also included a section, "Your Opinion," with one scale and one textbox each for Data for Good and Data for All. In this section, we ask the evaluators how well the project meets the criteria for these two main aspirations.

***ESJ Canvas Sheet***

Both the rubric and the questionnaire can be time-consuming; some participants in pilot studies expressed strong opinions about this. We also developed two variations of a *canvas sheet*, often used in entrepreneurial settings to explore the viability of product ideas (Stubben et al., 2014). Such canvas sheets have been used for ethics and social justice evaluations as well (ADAPT Centre for Digital Content Technology, 2017; Open Data Institute, 2019). In our first iteration of usability experiments, we suggested the use of a paper-based version for group discussions, followed by individual submissions through a Web-based, interactive version. The content of both versions is identical, but the formatting was different for the Web version to facilitate the use of input options ranging from mobile phones to desktops. Included as Appendix C: ESJ

Canvas Sheet is the paper-based version; an interactive version is undergoing internal testing but is not ready yet for wider distribution.

We believe that this tool, which requires users to enter essential aspects on one sheet of paper (or its digital equivalent), is especially suitable for educational purposes. In pilot studies, we found that users had a significantly higher awareness of ESJ aspects in the initial stages of a project after using an ESJ canvas sheet. Especially in a setting where a group has to make a choice among several project alternatives, the use of such a canvas sheet encourages them to consider the potential impacts and consequences of their choice. In its current incarnation, this tool is intended to foster discussion but does not assist in the resolution of conflicting opinions among group members. For future versions, we are considering the inclusion of elements for conflict resolution similar to those used in the swarm AI platform of Unanimous.ai (L. Rosenberg, 2016; L. Rosenberg et al., 2021; L. B. Rosenberg, 2015).

Again with the goal of balancing qualitative, text-based responses that encourage reflection, with quantitative, scale-based ones that are easier to process and consolidate, we developed two versions: One with text boxes only, the other with four scales (also referred to as "sliders") on each side of each text box: *Relevance* of the category, *competence* of the team members, *effort* spent by the team, and overall *impact* of the project.

With these three tools to assess ESJ aspects of projects, we are in the process of conducting user studies with two primary goals: First, to collect feedback on the usefulness, usability, and effectiveness of the tools in order to improve them. Second, we are using them as data collection tools for explorations of predictive models for ESJ assessment described in the next section. At the time of writing (May 2022), we completed the first phase of user studies with students and a set of workshops with participants from the Data Science Initiative at Cal Poly. Later phases will use revised versions of the instruments and include other user groups such as researchers or evaluators of project proposal submissions.

***Predictive Models for ESJ Assessment***

Below we describe predictive models for scoring data science research projects based on evaluators' open responses to questions regarding ethics and social justice. Using open text responses, our initial machine learning model will parse an evaluator's responses and predict the sentiment of those responses, which will reflect the degree to which the project adheres to ethics and social justice criteria. Alternative models using other Machine Learning methods are in the early stages of exploration.

The features of this system include the materials that are distributed to researchers for data collection and the tool that is used to analyze the collected data. The ESJ Questionnaire and the ESJ Canvas Sheet are used as the method of data collection. The interface for entering information about research projects depends on the system being usable and featuring a holistic curation of questions. The user interface should be intuitive and work both on standard

computers and mobile formats. Each question should be clear and open-ended enough to get a succinct, complete answer. The collection of questions should sanction the user to provide all relevant information about the project in question. This data will then be processed using methods relying on computational power and machine learning. Humans will be asked to analyze the same data and, essentially, process it in the same way to provide a subjective score. If the computer-generated scores are statistically similar to the human scores, it could be concluded that this tool may be beneficial for the assessment of projects. We would like to emphasize that we do not propose to automate this assessment process. In our view, it is essential that human judgment is part of this process, and should be used when making decisions about judging projects. In addition, a project document receiving a high score does not necessarily confirm that it is maintaining ethical and social justice standards in research. The same, of course, can be said of human evaluators. While we are aware of human bias and selection bias in this model, at this point we do not have a clear path toward overcoming them.

In order to create a system that can provide ethical input, one must define some ethical standards and associations with numerical values that reflect these sentiments as accurately as possible. The questionnaire tool is straightforward to convert to numerical values because it features a scoring system that clearly mirrors a quantitative system. However, all versions of the canvas yield data that is in the form of text and corresponds to each of the provided questions. This kind of raw data is more complex to convert into useful quantitative values. To address this, our project will explore multiple strategies to extract empirical information. As the system develops, it will become clear which, if any, of these approaches are accurately reflecting the sentiments of the original canvas responses. To begin, a model will be developed using natural language processing (Bird et al., 2009; Jurafsky & Martin, 2009; Manning & Schütze, 1999) methods and tools (Honnibal et al., 2020; Montani et al., 2021; NLTK Project, 2022; Petrochuk, 2022; Řehůřek & Sojka, 2011/2010; Stanford NLP Group, 2022). The early stages of this subsystem will likely rely on simple word-sentiment associations and be trained further as more data is collected. In the case of the canvas with sliders, the sliders will provide an element of purely quantitative information to be used in conjunction with the written responses.

### *Classification of Project Proposals with NLP*

Over the timeline of a quarter, a team of students from an Introduction to Artificial Intelligence class developed a system that attempts to quantify how relevant research project proposals are to data science and ethics research. This automated system's goal is to provide utility by helping individuals determine a project's relevance to ethics & data science. The team used an LDA (Latent Dirichlet Allocation) model (Njagi et al., 2016) to examine project proposals to determine their relevance to data science and ethics research. In this exploratory project, the team used flexible conditions to determine whether it is possible to classify the relevance of a given project proposal. Ultimately, they found that with subjective qualitative testing, they were able to achieve a system that is capable of producing a relevance score when determining if a given project abstract is relevant to data science and ethics research.

In the system implementation, the model was trained on project abstracts instead of project proposals. During this training phase, the hand-compiled training dataset takes in project abstracts and performs pre-processing via the NLTK library (NLTK Project, 2022). This pre-processing consists of word stemming, stopword removal, and tokenization. This is then fed into the LDA model for training. When given a project abstract or project proposal to test, the system performs the pre-processing steps above and feeds this matrix into the model for comparison. This comparison consists of comparing the topics extracted during the training phase to the topics that were extracted from the testing document. At the end of this comparison stage, a report is generated describing the quantitative similarities for the top N topics extracted from the testing document.

In this exploratory project, the team relaxed development conditions and created project assumptions to see if the goal was achievable. In doing so, they introduced their own biases in the areas of data collection and system evaluation. In the decision to train our model on project abstracts instead of project proposals, the team evaluated example project proposals and compared these proposals to abstracts.

They found that the keywords and topics relevant to data science and ethics research were similar to each other in their subjective evaluation. With this discovery, they moved forward with the assumption that training the system on project abstracts instead of project proposals would not heavily decrease accuracy. The data collection could be expanded to include more sources instead of relying on project abstracts collected from past IEEE International Symposiums on Ethics in Engineering, Science, and Technology. Finally, the system evaluation criteria consisted of subjective tests on arbitrary testing documents that were composed of documents related to data science & ethics research and documents that were not related at all. This subjective analysis introduced human bias due to human intervention and analysis.

In the table below, the system was trained on a dataset from past IEEE Symposiums project abstracts that heavily contained topics regarding engineering and ethics-related research. In the example below, the team tested the system on two sentences: one related to engineering and ethics and one sentence about a completely arbitrary topic such as potatoes. During parameter tuning for this example, they chose the LDA model to extract three subtopics.

The subtopic comparison performs a vector comparison of the extracted subtopics from the given input's subtopics and the LDA model's subtopics. The numeric score represents the relevance of each subtopic extracted from the input compared to the subtopics extracted from the LDA model. In the input related to potatoes, there is a score of 0.333. Since the number of subtopics was set to three, this score means that the three subtopics extracted from the input have an equal chance of being relevant to the three subtopics within the LDA model. This means that the input's subtopics have no relevance to the LDA model's subtopics.

| | | |
|---|---|---|
| Example Input | `Modern machine learning tools are so complex, they are difficult for humans to interpret and understand. That makes it difficult to determine appropriate inputs and ethical implications of results.` | `The potato is a starchy tuber of the plant Solanum tuberosum and is a root vegetable native to the Americas.` |
| Subtopic Comparison | `(0, 0.94575685), (1, 0.026976354), (2, 0.027266765)` | `(0, 0.33333334), (1, 0.33333334), (2, 0.33333334)` |

As a disclaimer and reminder, the LDA model in this system is subject to the biases introduced in a limited training dataset.

**Experimental Design**

Since we are still in the process of developing our instruments, we identified three primary phases for our experiment.

***Phase 1: Testing with Students for Usability and Increase of Awareness***

The focus of this phase is on the Questionnaire and Canvas instruments. They are designed to be used with minimal instructions by participants with limited or no background in ethical and social justice aspects of Data Science. Our goal was to use the two instruments with multiple classes in different disciplines during the winter term (early January through mid-March). The initial plan was to use the two instruments at the beginning of the term, and again at the end. One of the classes, Introduction to Artificial Intelligence, consisted of two sections, where we planned to examine temporal dependencies between the two instruments by having one section do the Questionnaire first, followed by the Canvas sheet, with the reversed sequence for the other section. Another class was a General Education class with a focus on virtual environments and digital twins. In both classes, students worked on team projects that involved datasets of significant size, typically in the GByte range.

Due to changes in the required Human Subjects review process, the approval of our experiments was delayed, and we were not able to follow through with this plan and could use the instruments only at the end of the term.

The data collected from this first iteration will be the basis for a proof-of-concept predictive model. The intention here is to test the NLP and ML methods for this model; we do not expect to have a model that produces meaningful assessments at this point. Since the training set data are collected from participants with limited expertise in ethics and social justice aspects, and little training in the use of the instruments, we rely mostly on their own interpretation.

The second iteration of this phase with improved versions of the instruments will be conducted in the fall term in similar classes.

***Phase 2: Testing with Researchers for Usability, Awareness and Effectiveness***

During the spring and summer terms, we are conducting experiments with researchers, using the ESJ Questionnaire, Canvas, and Rubric instruments. In the first iteration, a small group of about 3-5 researchers participated in an information session and workshop. After an introduction of the methods and tools, we asked the participants to evaluate an example project on the development of tools to assist the World Bank with the assessment of school building safety in poor countries. Common to all participants was an interest in Data Science and related disciplines. Their background in ethics and social justice aspects varied considerably, ranging from practically none to longstanding research activities in those areas. Later iterations will consist of similar workshops with participants that include external collaborators from other academic institutions, research organizations, companies, and other interested parties. Due to the small number of participants, we obtained valuable feedback through discussions during the workshop. Several of them found the canvas sheet very useful, and are considering its use during the formative stages of team-based class projects, with the main goal of raising the awareness of students regarding ethical and social justice aspects of potential project topics.

As we assemble more data from participants with deeper expertise and better training in the use of our instruments, we will use the insights gained from earlier iterations to improve the predictive model.

***Phase 3: Testing with Evaluators for Usability and Effectiveness***

This phase will use a further refined set of the ESJ Questionnaire, ESJ Canvas, and ESJ Rubric instruments for the evaluation of project proposals submitted to funding sources. We are in initial discussions with internal and external partners who have expressed interest in the use of our instruments. We expect this phase to take place in the Academic Year 2022-23.

**Participants**

In the first iteration of Phase 1, we asked students from two different classes to participate in the experiment. Two sections of one class, Introduction to Artificial Intelligence, consisted mostly of students in Computer Science and related fields at the junior, senior, and graduate levels, with an overall enrollment of 71 students. We did not collect demographic data. While not required, most

students at that stage would have taken a Professional Responsibility class, which addresses ethical issues in Computer Science. All of the students in this class did a self-assessment of their own team project in the class as a lab assignment, using both the Questionnaire and the Canvas instrument. They had the option of declining to participate and doing a similar lab assignment instead; no students chose this option. Students also had the option of submitting up to two additional assessments of other team projects as make-up options for other labs they may have missed or not done well on.

The other class was a large section of a General Education class open to students from all majors, with an emphasis on virtual reality and digital twins; 75 students were enrolled. The students in this class used the Questionnaire instrument for a self-assessment of their own team project.

**Results**

As of March 2022, we completed the first iteration of Phase 1. In total, we received 190 responses from two groups of students:

- 71 participants from Introduction to Artificial Intelligence used the Questionnaire for an assessment of their own projects; an additional 53 responses in this class were for projects by other teams, for a total of 124 responses.
- 72 participants completed 130 Canvas Sheets.
- 63 Participants completed a comparison of the Questionnaire and Canvas Sheet.
- 57 participants from the General Education class completed self-assessments of their own projects with the Questionnaire. These students did not use the Canvas Sheet.

Students had the option to decline participation (and perform another activity with a similar scope); five students from the General Education class chose this option for the Questionnaire.

Based on the information gathered from the responses described above, participants expressed a slight preference for the ESJ canvas sheet over the questionnaire. Most prominently this was found in the Perspectives and Further Research categories, with values of 3.29 and 3.30, respectively, on a scale from 1 through 5. The only situation in which they preferred the Questionnaire was Group Discussion. This was surprising because we expected the Canvas sheet to foster group discussion based on initial feedback and observations in pilot studies. Interestingly, a pilot printable version of the Canvas sheet was also introduced to Artificial Intelligence students and they noted that that version was the tool of choice in a group discussion setting. The varying effects of typing on a computer and writing on paper have affected the feedback from our initial testing.
The participants were more straightforward in assessing the 'Data for Good' aspects of projects. Most projects received a better score for 'Data for Good' than "Data for All." About 70 percent of proposals received a score equal to or more than 4 out of 5 for the overall 'Data for Good' score, while 60 percent for the overall 'Data for All' score. The anonymity of participants is protected so that they might feel more inclined to provide honest answers.

As far as usability, the user interface for the ESJ Canvas Sheet was clearly inferior to the one of the ESJ Questionnaire, which is based on the widely used Google Forms tool. We assume that this may have caused some bias against the ESJ Canvas Sheet, but did not attempt to control for it. Some students also pointed this out in their comments. Despite this, the Canvas sheet was still overall slightly more preferred, with many students noting that they appreciated the space to describe their projects in their own words, allowing for them to provide more holistic accounts of their respective projects.

We asked participants to write their thoughts about the 'Data for All' and 'Data for Good' criteria. In an initial review of the text-based answers for qualitative analysis, we found the following aspects:

- We noticed an interesting dichotomy between those who evaluated their proposals and those who assessed other proposals. It is not rare to see that self-evaluators suggest that their "project is rather disconnected from any demographic issues, in the data or the application aspect. As a result, it does not reckon with these issues". On the other hand, it was more common that external reviews would expect more detailed thoughts regarding 'Data for All' in the proposals, with statements such as the "project could have discussed bias and marginalized groups more throughout its documentation." The researchers for a general-purpose project/application don't see the need or appropriateness of opening discussion on the demographic issues. Not all reviewers agree as such.
- The other disputed aspect of 'Data for All' was whether data should be shared with all or not—the paradox of empowering the public with open-source data vs. privacy and security concerns.

Overall, most students found the questionnaire a valuable tool for future projects and ethical and social justice considerations in their projects (75 percent scored four or more out of 5). A comment such as "I think this survey was a great use of time and made me look at my project in a different way!" is a testimony that this rubric responds to its purpose.

The format of questions was acceptable for most students (63 percent scored four or more out of 5). However, we received comments such as "the wording of the questions was very confusing at times." Some requested a better explanation for various categories. We will revise the questions and explanations for the next version of the tools.

Another common comment was having the option of NA in the answers. Selecting 0 scores for many questions didn't seem logical when these topics did not apply to the project. This is an issue that we are aware of; resolving it may require a transition to a different tool: Google Forms, which we used in this iteration, does not facilitate NA as an option, and we are exploring alternatives.

Some respondents liked the combination of multiple-choice questions and open-ended questions. In contrast, others suggest that “these questions require more free-form responses because they are highly complicated issues and ranking just doesn't give much information.”

Based upon a thorough review of the ESJ canvas sheet responses, there is an expected and apparent difference between responses that were completed by participants who worked on the project in question themselves (primary researchers), than by participants recently exposed to it (third-party participants). Not all authors’ roles are apparent, but many use language like “they” or “we” when describing methods and implementations. Responses that are clearly from a primary researcher are often longer and more detailed. They very rarely leave a question blank or write “I don’t know.” Responses from a third-party participant often contain short and choppy responses to questions. “I don’t know” and related phrases are much more common. Clearly, this tool is immediately more useful for primary researchers and other people involved closely with the project because they already have the knowledge to answer each question thoroughly. In answering these questions, the participant must consider the potential effects of their research. Further, the canvas sheet serves a different purpose when being completed by a third-party participant. The canvas would be a valuable tool to use in settings like workshops and courses that deal with data science and ethics. Participants can become familiar with the layout of the canvas sheet while taking time to review projects. Primarily, it will oblige them to think in the mindset that encompasses Data for All and Data for Good.

There were two classes that made up the bulk of participants from the initial testing. There were very few obvious differences between the responses with respect to the different classes. Because canvas responses are evaluated anonymously, besides project topics, there was no clear indicator in the textual responses that indicated a difference in the way participants used the Canvas.

After the initial round of testing and gathering feedback, the following improvements for the Data Ethics Canvas have been identified:

- The next round of testing should include an experiment that organized canvas responses by project team. All responses that pertain to a specific project would be evaluated together to get a composite score for the project. This approach pertains specifically to the online version of the canvas sheet. Instead of each response mapping to one score, a group of responses written about the same project will map to one collective score for the project. This way, participants are protected by anonymity and multiple perspectives can be considered. Responses can be collected anonymously because participants are completing the canvas sheet online. The goal of evaluating multiple perspectives at once is to get a holistic and accurate representation of the project.
- Similarly, the printable version has its own use in in-person, group settings. This version should be tested with groups that are assessing their own work as well as groups that are assessing the work of others. The tactile component of this Canvas version has shown to be beneficial to these kinds of environments.

- An objective, third-person description of the project, like its abstract, may be beneficial to include in the assessment inputs. When collecting human assessment data to compare against the model's predictions, multiple assessors noted that a project abstract would be beneficial to use in their evaluations. It follows that the abstract would provide more context before going straight to the purpose or methods.
- The interface of the canvas sheet can be improved in a number of ways. The second iteration will have two pages, the first consisting of only questions and text boxes. The second page shows the user the sliders for quantitative assessment and an option for more detailed instructions and an option to go back and view previous answers. This design eliminated any need for scrolling. In the future, the sliders could be replaced by a heat-map or other form of input method for comparison to the slider data that is being collected.

**Next Steps**

Due to the delay in Human Subjects approval, we had to modify our original plan to use a revised version of the assessment rubric and related tools at the end of the winter term. The upcoming spring term will be used for a thorough analysis of the results so far and a revision of the instruments used (ESJ Questionnaire, ESJ Canvas). At the beginning of the fall term, we will use the revised instruments during the period when students select topics for their upcoming class projects. In classes from Data Science, Computer Science, Architecture, and other disciplines, they will assess projects from prior terms, do a self-assessment of potential project topics for their team, and evaluate projects proposed by other teams.

In addition to working with students as participants, we are also reaching out to researchers in Data Science and related fields to conduct self-assessments of their own projects and research activities, and assessments of past proposals submitted to an internal round of seed funding (see Phase 2 above). In Phase 3, we are planning to collaborate with evaluators of funding proposals to test our instruments with actual proposals submitted to funding opportunities where ethical and social justice aspects are of importance.

**Conclusions**

What started out as an attempt to better understand the implications of the vision 'Data for All, Data for Good' when applied to research activities in Data Science and related fields has led us to the development of a suite of instruments for the assessment of project proposals and research activities. While these instruments and the accompanying material are still under development, our initial experiments with students indicate that this effort is worth pursuing. An analysis of the results reveals that students and researchers valued the opportunity to reflect on the ethical and social justice aspects of their research projects. They found the instruments used overall to be very helpful and an effective use of their time, and provided us with ample feedback on possible improvements.

**References**


ACM. (2018). *ACM Code of Ethics and Professional Conduct* [Association for Computing Machinery]. https://www.acm.org/code-of-ethics

ADAPT Centre for Digital Content Technology. (2017). *The Ethics Canvas*. https://www.ethicscanvas.org/

AI for Good Foundation. (2019). Projects—AI for Good Foundation. *AI for Good Foundation*. https://ai4good.org/active-projects/

AI for Good Foundation, & Syngenta. (2017). *AI for Agriculture: Help Feed the World with AI*. https://ai4good.org/ai-for-agriculture/

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, *355*(6324), 483–485. https://doi.org/10.1126/science.aal4321

Benjamin, M., Gagnon, P., Rostamzadeh, N., Pal, C., Bengio, Y., & Shee, A. (2019). Towards Standardization of Data Licenses: The Montreal Data License. *ArXiv:1903.12262 [Cs, Stat]*. http://arxiv.org/abs/1903.12262

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit* (1st edition). O'Reilly Media.

Bloomberg. (2015). Data for Good Exchange 2015. *Bloomberg L.P. Pages*. https://www.bloomberg.com/lp/d4gx-2015/

Brown, S., & Mecklenburg, A. (2021, February 1). The ODI & Consequential to research the second generation of ethics tools – The ODI. *Open Data Institute Blog*. https://theodi.org/article/the-odi-consequential-to-research-the-second-generation-of-ethics-tools/

Charitidis, C. A. (2019). Quest for Ethical and Socially Responsible Nanoscience and Nanotechnology. In R. Iphofen (Ed.), *Handbook of Research Ethics and Scientific Integrity* (pp. 1–15). Springer International Publishing. https://doi.org/10.1007/978-3-319-76040-7_42-1

Chatila, R., & Havens, J. C. (2019). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. In M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, & E. E. Kadar (Eds.), *Robotics and Well-Being* (Vol. 95, pp. 11–16). Springer International Publishing. https://doi.org/10.1007/978-3-030-12524-0_2

Datenethikkommission. (2019). *Opinion of the Data Ethics Commission* (p. 238). Data Ethics Commission of the Federal Government Federal Ministry of the Interior, Building and Community; Federal Ministry of Justice and Consumer Protection. https://datenethikkommission.de/wp-content/uploads/DEK_Gutachten_engl_bf_200121.pdf

Desai, S. (2021, April). *Stanford University CS 21SI: AI for Social Good*. https://explorecourses.stanford.edu/search?view=catalog&filter-coursestatus-Active=on&page=0&catalog=&academicYear=20192020&q=CS%2021SI%3A%20AI%20for%20Social%20Good&collapse=

D'Ignazio, C., & Klein, L. F. (2020). *Data Feminism*. MIT Press.

Ethics & Compliance Initiative (ECI). (2021). *The PLUS Ethical Decision Making Model—Ethics & Compliance Toolkit*. Ethics and Compliance Initiative. https://www.ethics.org/resources/free-toolkit/decision-making-model/

European Commission. (2021, March 8). *Ethics guidelines for trustworthy AI | Shaping Europe's digital future*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Ferretti, A., Ienca, M., Velarde, M. R., Hurst, S., & Vayena, E. (2022). The Challenges of Big Data for Research Ethics Committees: A Qualitative Swiss Study. *Journal of Empirical Research on Human Research Ethics*, *17*(1–2), 129–143. https://doi.org/10.1177/15562646211053538

Forsberg, E.-M. (2004). The Ethical Matrix—A Tool for Ethical Assessments of Biotechnology. *Global Bioethics*, *17*(1), 167–172. https://doi.org/10.1080/11287462.2004.10800856

Gebru, T. (2020). *Black in AI*. https://ai.stanford.edu/~tgebru/

Gebru, T. (2021, January 25). *MIDAS, U-M AI Lab, IT Dissonance Series, and Ethics, Society, and Computing Present: Timnit Gebru—YouTube*. https://www.youtube.com/watch?v=23MxOh99N54&ab_channel=MichiganInstituteforDataScience

Hao, K. (2021, April 13). *Big Tech's guide to talking about AI ethics*. MIT Technology Review. https://www.technologyreview.com/2021/04/13/1022568/big-tech-ai-ethics-guide/

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. https://doi.org/10.5281/zenodo.1212303

Huh-Yoo, J., Kadri, R., & Buis, L. R. (2021). Pervasive Healthcare IRBs and Ethics Reviews in Research: Going Beyond the Paperwork. *IEEE Pervasive Computing*, *20*(1), 40–44. https://doi.org/10.1109/MPRV.2020.3044099

IEEE Standards Association. (2021). *IEEE SA - The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS)*. https://standards.ieee.org/industry-connections/ecpais.html

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed). Pearson Prentice Hall.

KDNuggets. (2015). The Definitive Guide to doing Data Science for Social Good. *KDnuggets*. https://www.kdnuggets.com/the-definitive-guide-to-do-data-science-for-good.html/

Kitchens, W. S. (2019, June 26). *"EU Ethics Guidelines for AI Are Just the Beginning," Law360, June 26, 2019. | News & Insights | Alston & Bird*. Lw360. https://www.alston.com/en/insights/publications/2019/06/eu-ethics-guidelines

Koschinsky, J. (2015). *Data Science for Good: What Problems Fit?* 10. https://ocw.mit.edu/courses/comparative-media-studies-writing/cms-631-data-storytelling-studio-climate-change-spring-2017/readings/MITCMS_631s17_koschinsky_2015.pdf

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing* (1st edition). The MIT Press.

Marshall, J. (2009, November 3). Ethical Decision-Making Tools. *Ethics Alarms*. https://ethicsalarms.com/rule-book/ethical-decision-making-tools/

McGregor, S., & Banifatemi, A. (2018). First Year Results from the IBM Watson AI XPRIZE: Lessons for the "AI for Good" Movement. In S. Escalera & M. Weimer (Eds.), *The NIPS '17 Competition: Building Intelligent Systems* (pp. 233–249). Springer International Publishing.

Montani, I., Honnibal, M., Van Landeghem, S., Boyd, A., Peters, H., McCann, P. O., Samsonov, M., Geovedi, J., O'Regan, J., Orosz, G., Altinok, D., Kristiansen, S. L., Roman, Explosion Bot, Fiedler, L., Howard, G., Wannaphong Phatthiyaphaibun, Tamura, Y., Bozek, S., … Dubbin, G. (2021). *explosion/spaCy: V3.2.1: doc_cleaner component, new Matcher attributes, bug fixes and more* (v3.2.1) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.1212303

Njagi, D., Zuping, Z., Hanyurwimfura, D., & Long, J. (2016). Incorporating Lexical Knowledge via WordNet to Latent Dirichlet Allocation in Offensive Message Detection. *Journal of Computational and Theoretical Nanoscience*, *13*, 3464–3471. https://doi.org/10.1166/jctn.2016.5243

NLTK Project. (2022). *NLTK: Natural Language Toolkit* (3.7) [Computer software]. https://www.nltk.org/

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy* (First edition). Crown.

Open Data Institute. (2019). *Data Ethics Canvas*. Open Data Institute. https://theodi.org/wp-content/uploads/2019/07/ODI-Data-Ethics-Canvas-2019-05.pdf

Open Data Institute. (2021, October 19). Data ethics: Practical and benchmarking tools. *Open Data Institute Blog*. https://theodi.org/article/data-ethics-practical-and-benchmarking-tools/

Petrochuk, M. (2022). *Welcome to Pytorch-NLP's documentation! —PyTorch-NLP 0.5.0 documentation*. https://pytorchnlp.readthedocs.io/en/latest/

Řehůřek, R., & Sojka, P. (2010). *Software Framework for Topic Modelling with Large Corpora* (4.1.0, pp. 45–50) [Python]. University of Malta. http://is.muni.cz/publication/884893/en (Original work published 2011)

Rose, E., Edenfield, A., Walton, R., Gonzales, L., McNair, A., Zhvotovska, T., Jones, N., Mueller, G., & Moore, K. (2018). *Social Justice in UX: Centering Marginalized Users*. 1–2. https://doi.org/10.1145/3233756.3233931

Rosenberg, L. (2016). Artificial Swarm Intelligence vs human experts. *2016 International Joint Conference on Neural Networks (IJCNN)*, 2547–2551. https://doi.org/10.1109/IJCNN.2016.7727517

Rosenberg, L. B. (2015). Human swarming, a real-time method for parallel distributed intelligence. *2015 Swarm/Human Blended Intelligence Workshop (SHBI)*, 1–7.

https://doi.org/10.1109/SHBI.2015.7321685
Rosenberg, L., Willcox, G., Palosuo, M., & Mani, G. (2021). Forecasting of Volatile Assets using Artificial Swarm Intelligence. *2021 4th International Conference on Artificial Intelligence for Industries (AI4I)*, 30–33. https://doi.org/10.1109/AI4I51902.2021.00015
Spanache, I. (2020, December 21). *Data Science for Social Good*. Medium. https://towardsdatascience.com/data-science-for-social-good-a88838bc8ed0
Stanford NLP Group. (2022). *Stanford CoreNLP* (4.4.0) [Computer software]. https://stanfordnlp.github.io/CoreNLP/
Stubben, S., van Tilburg, T., Olesen, T. S., Liengard, S., Vyrostko, M., & Breum, N. B. (2014). *Project Canvas—Visual project communication and overview*. Project Canvas. http://www.projectcanvas.dk/
Syngenta, & AI for Good Foundation. (2017). *2017 Syngenta AI Challenge*. https://www.ideaconnection.com/Syngenta-AI-Challenge/
The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). *Ethically Aligned Design—A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* (p. 294) [White Paper]. IEEE. . https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html
Thereaux, O. (2021, May 28). The next generation of data ethics tools – The ODI. *Open Data Institute Blog*. https://theodi.org/article/the-next-generation-of-data-ethics-tools/
Tranberg, P., Hasselbalch, G., Olsen, K., & Byrne, C. S. (2018). *DATAETHICS – Principles and Guidelines for Companies, Authorities & Organisations*. DataEthics.eu.
United Nations, D. of E. and S. A. (2015). *Transforming our World: The 2030 Agenda for Sustainable Development | Department of Economic and Social Affairs*. https://sdgs.un.org/publications/transforming-our-world-2030-agenda-sustainable-development-17981
Vieweg, S. (Ed.). (2021). *AI for the good: Artificial intelligence and ethics*. Springer. https://doi.org/10.1007/978-3-030-66913-3

## Appendices

The following three appendices are attached as separate documents.

***Appendix A: ESJ Rubric***

***Appendix B: ESJ Questionnaire***

***Appendix C: ESJ Canvas Sheet***