Assessing AI Impact Assessments: A Classroom Study

Nari Johnson
Carnegie Mellon University
narij@andrew.cmu.edu

Hoda Heidari
Carnegie Mellon University
hheidari@andrew.cmu.edu

Abstract

Artificial Intelligence Impact Assessments ("AIIAs"), a family of tools that provide structured processes to imagine the possible impacts of a proposed AI system, have become an increasingly popular proposal to govern AI systems. Recent efforts from government or private-sector organizations have proposed many diverse instantiations of AIIAs, which take a variety of forms ranging from open-ended questionnaires to graded score-cards. However, to date that has been limited evaluation of existing AIIA instruments. We conduct a classroom study ($N = 38$) at a large research-intensive university (R1) in an elective course focused on the societal and ethical implications of AI. We assign students to different organizational roles (*e.g.,* an ML scientist or product manager) and ask participant teams to complete one of three existing AI impact assessments for one of two imagined generative AI systems. In our thematic analysis of participants' responses to pre- and post-activity questionnaires, we find preliminary evidence that impact assessments can influence participants' perceptions of the potential risks of generative AI systems, and the level of responsibility held by AI experts in addressing potential harm. We also discover a consistent set of limitations shared by several existing AIIA instruments, which we group into concerns about their *format* and *content*, as well as the *feasibility* and *effectiveness* of the activity in foreseeing and mitigating potential harms. Drawing on the findings of this study, we provide recommendations for future work on developing and validating AIIAs.

1 Introduction

One cannot understate the impact that Artificial Intelligence (AI) systems have had on society, from organizing our social media feeds [10] to influencing how we find work [27, 1], or informing critical decisions about our lives, such as whether we will receive a loan [40] or be granted parole [2]. However, as an increasing number of incidents have demonstrated AI systems' potential to cause harm [16, 34], policymakers, regulators, and corporations face a new sense of urgency to govern AI. One proposal which has gained traction among regulators [4, 21, 25] is AI Impact Assessements, or "AIIA"s. Inspired by the long history of impact assessments in other scientific or policy domains (*e.g.,* environmental impact assessments) [24], AIIAs provide structured processes for organizations to consider the implications to people and their environment of a proposed AI system. Recent efforts have proposed many diverse instantiations of impact assessments for AI tools (surveyed by [37]) which range in form from open-ended questionnaires to graded score-cards, some of which are already in-use or mandatory under existing governance regimes [22, 17, 21]. However, to date there has been limited empirical evaluation of existing AIIA instruments [18].

Contributions. In this work, we conduct a *preliminary classroom study* to explore the usability and effectiveness of existing AIIA templates. We design a role-playing activity where we assigned $N = 38$ students enrolled in an elective course focused on the societal and ethical implications of AI, to different organizational roles, *e.g.,* an ML scientist or product manager. We then asked participant

Workshop on Regulatable Machine Learning at the 37th Conference on Neural Information Processing Systems (RegML @ NeurIPS 2023).

teams to complete one of the three prominent AI impact assessment instruments for one of two hypothetical generative AI systems: a general-purpose video chatbot, and a video chatbot fine-tuned to conduct initial screening interviews with job candidates. We surveyed participants immediately before and after the group activity to understand how completing an AIIA influenced their thinking about the potential impacts of generative AI, and collected their feedback on each template.

Our thematic analysis of students' responses provides preliminary evidence that existing AIIA instruments *can* influence respondents' perceptions of the potential risks of generative AI systems. After completing an AIIA, several students reported an *increased level* of concern and perceived level of responsibility as machine learning experts, and collectively reported a more comprehensive and actionable set of potential issues with their product. Thus, existing AIIAs have the potential to promote the imagination and mitigation of potential harms. However, our analysis of students' critiques reveals a consistent set of limitations shared by prior AIIA instruments, which we group into concerns about their *format* and *content*, as well as the *feasibility* and *effectiveness* of the activity in foreseeing and mitigating potential harms. Drawing on the findings of this study, we provide several recommendations for future work to co-design new and improved AIIA instruments.

2 Background

Related Work Following [37], we adopt a broad definition of an AI impact assessment as a structured process to help stakeholders understand the implications of a proposed AI system.[1] While many related impact assessments (such as privacy [6] or data protection [9, 13] IAs) may be relevant to AI systems, we focus our study on instruments designed specifically to consider the impact of introducing AI. Existing AIIA instruments vary widely along several critical dimensions, including their articulated scope, purpose, form, and content [37]. Existing instruments propose a wide variety of different workflows, at varying levels of specificity. While all AIIA instruments outline a process for organizations to follow [30], only a subset have a clear set of expected deliverables from their processes [22, 11, 17]. Many such instruments are structured as questionnaires designed to facilitate reflection about societal impact, where the AIIA is "completed" when all questions have been answered. Some AIIA frameworks also propose processes by which the AIIA is then *reviewed*, *e.g.,* public agencies should solicit public comments on their AIIA before moving forward with the proposed system [30]. Given the limitations of conducting a classroom study, we focus our research on characterizing challenges faced by assessors when they are asked to fill out an AIIA questionnaire, and how the process of completing an AIIA affects their thinking about potential impacts of AI. We note that developing an effective questionnaire (the focus of this work) is just one important step of assembling a functional and robust AIIA regime [18, 32, 8, 19].

In response to the rising popularity of AIIAs as a potential accountability mechanism for AI, a growing body of scholarship has proposed evaluative criteria or studies to assess the efficacy of AIIA instruments [18, 20, 41, 32, 8]. An increasing number of studies have run empirical validations of AIIAs, such as OpenLoop's experiment where researchers partnered with 10 European AI companies to co-design a new AIIA instrument [20]. Our research contributes to this growing and timely body of work that aims to critically evaluate and characterize potential limitations of AIIAs.

AIIA Instruments Our classroom study was inspired by our simple observation that many of the most prominent and well-known AIIA instruments at the time which we ran our study were released without any publicly available evaluation of their usability or effectiveness.[2] We selected three different instruments pictured in Figure 2 (released by the US CIO Council, Canadian Treasury Board, and Microsoft) that past work has referenced as mature and well-known instantiations of a general, domain-agnostic AIIA [18], and that also differed in interesting ways. We provide a detailed summary of the form and content (*e.g.,* what types of impact users were prompted to consider) of each AIIA instrument in Appendix A. All three instruments were at some point actively in-use within an existing AI governance regime, *e.g.,* all Canadian executive agencies must complete the Canadian

[1] Several instruments that we consider use the term "algorithmic impact assessment" (AIA) and are developed to assess specific types of automated systems. We include these instruments because of the wide overlap in their scope and applicability with existing AI impact assessments and AI products.

[2] Of the three templates we selected, two (the Canadian ADM and Microsoft) state that the template was revised through consultation with relevant stakeholders. However, to our knowledge the details of this process are not publicly available.

AIIA for any proposed AI system [22] and all Microsoft AI products were required to be assessed by AIIA [17]. We invite the reader to explore each instrument online.[3]

3 Study Design

We conducted an in-class research activity where we assigned students enrolled in an elective course to different organizational roles, and asked them to work together in teams to complete an AIIA. The study was approved by an Institutional Review Board (IRB) process and conducted synchronously on Zoom in a single 80-minute class session in March 2023. We provide further details on our study design and text of all study materials in Appendix B, and summarize key details in this section.

Activity overview. Students were assigned at random to teams, where each student within the team role-played an assigned organizational role to complete an existing AIIA for an imagined product scenario.

Organizational roles. Each student was assigned at random to a team of 3 or 4 students. Within each team, each participant was assigned to one of three roles: (1) a *machine learning scientist* representing the team in charge of design or development, (2) a *product manager* in charge of making the project economically successful, and (3) a *user representative* meant to represent the interests of potential users of the product. Each team had at least one participant assigned to all three roles, and teams with four participants had two user representatives.

Scenarios. We designed 2 product scenarios, and assigned half of the teams to each one. Both scenarios told participants to role-play that they are employed by a private firm that has developed a *generative AI model* that is capable of powering a hyper-realistic video-chat agent. While the AI model in both scenarios was described to have the same capabilities, the two scenarios differed in the *level of specificity* of the system's intended use. In one scenario, individual *users* of the AI product could "specify the characteristics of the agent they would like to converse with", like other general-purpose products, *e.g.,* ChatGPT. In the other scenario, the company was specifically working on a product that would use the agent to conduct initial job screening interviews with candidates. We hypothesize that existing AIIA tools may be less useful for general-purpose technologies, *i.e.,* that teams tasked to complete the AIIA for a product where the *user specifies the end use* may encounter more difficulties than teams who are given a more well-specified end use of screening candidates. We provide the complete scenario text shown to participants in Appendix B.1, which we wrote to provide students with the level of detail that would typically be available during the product ideation stage, *i.e.,* before any particular model has been developed.

Study population. Our study population consisted of 38 students at a large research-intensive university (R1) enrolled in an elective course on the societal and ethical considerations of AI. All participants had self-reported intermediate knowledge of machine learning, as introductory machine learning was listed as a prerequisite for the course. Students are future technologists and practitioners who may someday be asked to complete an AIIA, and as such we believe that their feedback is valuable in creating an AIIA template that is broadly accessible. There are several important limitations when using a classroom role-playing exercise to evaluate a proposed AIIA workflow: for example, students do not have situated expertise that practitioners would have (*e.g.,* in a hiring scenario), and we did not provide them context on existing governance or accountability structures within their simulated organizations. Thus, we believe our study, while informative, is no replacement to running grounded empirical evaluations of proposed AI systems within real organizational contexts.

Study procedure & data collection. The course instructor began the study by introducing the study task, team assignments, and scenarios. Before beginning the role-playing exercise, students individually completed a *pre-questionnaire* form. Once they completed the pre-questionnaire, students entered break-out rooms with their team and begin the role-playing activity (*i.e.,* fill out the assigned impact assessment template). Students were specifically told to *think-aloud* by verbalizing all of their thoughts as they worked together to complete the impact assessment. Once they completed the impact assessment (or the course time ended) students individually completed a *post-questionnaire* form. We provide the complete text of the pre- and post-questionnaires in Appendices B.3 and B.4.

[3] Students accessed each instrument online on March 30, 2023. They accessed version 0.10.0 of the Canadian AIA (released on March 26th, source code, link), the alpha version of the US CIO AIA (link), and the June 2022 release of Microsoft's Responsible AI Impact Assessment (link).

3.1 Research Questions

Our first set of research questions compare students' answers to the same set of questions (copied verbatim from the pre- to post-questionnaire) to examine if, and how, completing an AIIA changed their thinking about both the potential implications of the imagined AI product, and more broadly about generative AI.

- **RQ1**: *Imagining potential harms*. How does completing an AIIA affect the issues and values that students believe are the most urgent to address for their product?
- **RQ2**: *Sentiment towards generative AI*. How does completing an AIIA affect students' self-reported excitement and concern about potential uses of generative AI (beyond the presented scenario)?
- **RQ3**: *Responsibility of ML experts*. How does completing an AIIA affect students' perceived relative level of responsibility held by machine learning experts in addressing potential harms?

Our final research question analyzes students' reflections about their AIIA after completing the exercise:

- **RQ4**: *Characterizing limitations of & opportunities for AIIA instruments*. What concerns or critiques do students have of existing AIIA instruments? What considerations should those designing improved AIIA instruments keep in mind?

For all of our research questions, we also explore differences in participants' responses across treatments (*i.e.,* stratified by AIIA instrument or scenario condition). For our quantitative analyses (RQ2 and RQ3), we report differences in measures (such as the average self-reported excitement score across participants) across conditions. For our thematic analyses, we highlight what we believe are particularly informative differences across instruments or scenarios.

Data analysis In our quantitative data analysis, we did not run statistical significance tests. Instead, we opted to report summary statistics only and compare them with students' responses to closed-form questions. Due to our small sample size and the nature of the study, our numerical reports should be understood as suggestive evidence (and not statistically significant claims) to contextualize the qualitative findings and inform hypothesis making for future larger-scale studies, conducted with a more representative sample of the target population. To analyze the free response questions, we adopted a thematic analysis approach where we qualitatively coded participants' responses in a shared coding session. We conducted a bottom-up affinity diagramming process [3] to identify higher-level groups. We then compared the frequency of each code across participants' experimental conditions, *i.e.,* across the template or scenario conditions, to understand if themes were shared across or unique to a particular template or scenario condition.

4 Findings

In this section, we summarize key findings from our analysis of participants' pre- and post-questionnaires, organized by research question. We contextualize our thematic analyses using quotes from participants' responses, and refer to each participant using their four-digit ID number.

4.1 Imagining Possible Impacts (RQ1)

Before and after they completed the AIIA, we asked students to report which one out of seven possible categories of values they believed "*is most urgent to address for [their] product*" (see Appendix B.5 for the definitions of each value). We visualize students' changes in beliefs before and after completing the AIIA in Figure 1. We observed that students' responses varied significantly based on which scenario they were assigned: a larger number of students tasked to complete the AIIA for the hiring product rated "*Fairness*" as most urgent, while a much larger number of students assigned to the general-purpose use scenario chose "*Safety*" or "*Human autonomy & agency*". Many students also changed their response after exposure to the AIIA: 50% of all students chose a different value in their post-questionnaire than their original choice in the pre-questionnaire.

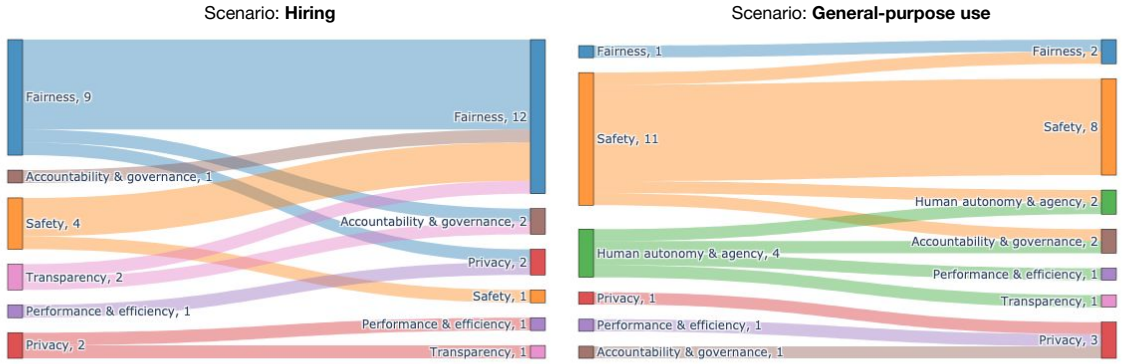

Figure 1: Sankey diagram to visualize students' changes in beliefs (*i.e.,* their answers to the question, "*Which one of the following broad categories of values do you believe is most urgent to address for your product?*") before (left) and after (right) completing the AIIA. We group students' responses by the scenario (hiring vs. general-purpose use) that their team was assigned. We observed that students responses differed depending on their scenario: students assigned to complete the AIIA for the hiring product were significantly more likely to rate "*Fairness*" as most urgent, while a much larger number of students assigned to the general-purpose use scenario chose "*Safety*" or "*Human autonomy & agency*". See Appendix B.5 for complete definitions of each value.

In addition to selecting the single most urgent value, we also asked students to "*outline [in free text] the top three issues you believe the development team of the AI product assigned to your team should carefully address*". Before completing their AIIA, students reported a wide range of potential issues for both the hiring and general-use products in their pre-questionnaires. Common responses shared by students across *both* scenarios included concerns about the accuracy, effectiveness, and functionality of the system; data privacy, security, and collection without users' consent; data quality or dataset bias; potential to be "biased" or "unfair" against protected groups; and different dimensions of transparency [39], *e.g.,* whether the AI is *explainable* and whether the user is notified that they are interacting with an AI.

Beyond these shared concerns, students also **identified concerns that were bespoke to their given scenario**. Students tasked with the hiring scenario reported concerns about gaming, as the AI product may introduce "*more opportunities for the candidate to cheat* (P1462), or may "*be gamed if candidates state relevant but uninformative key words*" (P1541). Students tasked with the general-purpose use scenario listed concerns about its potential for malicious use, including creating deepfakes (*e.g.,* "*stealing someone's identity*", P1153), or violent, explicit, or toxic content (P1384, P1063, P1816). Several students also expressed concern about users' potential over-reliance or persuasion by faulty AI: "*the AI may play a role in giving users information, so it is important that the AI is as unbiased and factual as possible, to not mislead the user with false or radical ideas*" (P1958).

When we compared individual students' responses to the same question *after* they completed the AIIA, we observed two interesting behaviors. First, **different students that were presented with the same instrument articulated a *similar* set of *new* concerns (that were not listed in their pre-questionnaires) after completing the AIIA**. For example, 5 students who completed the Microsoft template (and no other students) listed stereotyping as one of their top concerns after completing their AIIA, and several students that completed the US CIO AIIA expressed concern about how a general-purpose conversation agent would affect children. Notably, in both of these cases, the students' AIIA templates explicitly prompted them to consider these specific impacts (*e.g.,* the US CIO AIIA was the only template to include a question about the system's "*impact on children under the age of 18*"). This finding illustrates how AIIA instruments can inform students' awareness of possible harms that they may have previously not prioritized or considered. Further, that students prioritized potential issues that were *unique to their AIIA instrument* demonstrates how the *content* of each AIIA (*i.e.,* which particular harms are emphasized or excluded) has the potential to influence users.

Second, we observed that **existing AIIAs facilitated reflection on new types of impact that *no* student had listed as being important in their pre-questionnaires**. After completing their AIIA, several students listed that they believed the system's *broader impacts* to society and individuals whom are not direct users of the product, such as "*job displacement*" (P1387, Microsoft) and "*economic effects of automation (who and what will be replaced)*" (P1081, Canadian), were most important to address. Students also emphasized the importance of developing *processes* (such as audit trails or procedural fairness) to facilitate governance best practices, such as developing a "*rigorous logging and audit trail [...] to capture model failures and address potential user harm as quickly as possible*" (P1359, Canadian) and "*ensuring that the model is auditable*" (P1384, Canadian). We also observed that more students listed preliminary ideas or explicit actions to *mitigate* potential negative impacts of the AI system (rather than simply naming potential harms) after completing the AIIA: for example, one participant named "*[implementing] ways to check whether [generated] recordings are real or AI-generated*" as an important issue for the general-purpose use product.

4.2 Shifting Mindsets: Excitement, Concern, and Responsibility (RQs 2 & 3)

Excitement and concern for general-purpose technology Overall, most students entered into the AIIA exercise with high levels of both excitement and concern about potential uses of foundational generative AI models. While seemingly contradictory, being simultaneously excited and concerned is a common sentiment voiced by many politicians, experts, and regulators who share great concern about potential harms of AI systems, absent measures for accountability and governance [26, 23]. After completing their AIIA, the class' average excited rating slightly *decreased* (-0.16 points) from 4.11 to 3.94 (where a rating of 5 indicates "very excited"), and concerned rating *increased* (+0.32 points) from 4.00 to 4.32 (where a rating of 5 indicates "very concerned"). A greater proportion of students changed their concernedness rating after completing the AIIA: only 24% of students changed their excitedness rating, while 50% of students changed their concern rating. Of the students that changed their rating, 7 out of 9 students became *less* excited, and 15 out of 19 students became *more* concerned. We observed differences in how students' self-reported excitement and concern changed for different instruments, and different scenarios (complete results in Appendix C), where students who were given the general-purpose use scenario and US CIO or Microsoft AIIAs had relatively larger increases in their concern. We hypothesize that there was a slightly greater effect on participants' concern because existing impact assessments are primarily focused on articulating potential *negative* rather than positive impacts of proposed systems.

Responsibility of ML experts To survey students' perceptions of *who* should be held accountable to mitigate potential impacts, we asked them to rate "*the relative level of responsibility you imagine for ML experts (compared to other stakeholders) to address [the issues they identified]*". The class' average rating increased ($+0.26$) from a score of 4.24 to 4.50 after completing the AIIA. 20 (52% of all) students changed their response in their post-questionnaire, the majority (15) of whom reported an *increased* relative level of responsibility for ML experts. The extent to which students' ratings increased varied across AIIA instruments, where the Canadian AIIA resulted in the largest average increase (+0.42), which we speculate may be because the Canadian AIIA had the most detailed and thorough mitigation sections (summarized in Appendix A). We present extended results of how ratings varied across conditions in Appendix C.2.

4.3 Limitations of Existing Impact Assessment Instruments (RQ4)

We highlight interesting themes that emerged in students' reflections of the limitations of each AIIA, *i.e.,* their responses to the question: "*What do you think are the major limitations of the AIA you completed for your team's product?*" (RQ4). We found that students both listed a consistent set of limitations that were shared by all three AIIA instruments, and also identified limitations specific to individual instruments. We further group students' responses into concerns about instruments' *format*, *content*, *feasibility*, and *effectiveness*.

Format. While each AIIA consisted of a series of questions that users must respond to, each AIIA instrument varied in its *format* – design decisions such as whether the AIIA is administered as a static form or dynamic website, or the expected response type of each question (*e.g.,* multiple choice or free response). We summarize key findings that are either shared across instruments or bespoke to individual instruments:

- **The AIIA has open-ended questions that grant freedom, yet require creativity**. All three AIIA instruments included open-ended questions that elicited free text responses. One student pointed out that due to this open-endedness, the quality of the final completed AIIA "*depends heavily on the reliability and creativity of the person/team filling it out*" (P1709, Microsoft). Another student also noted that the instrument required them to make several subjective judgments because several terms were under-specified: "*it required a lot of qualitative assessments like those for safety and reliability and intelligibility that were not well-defined and could vary a lot depending on who fills out the AIIA and when*" (P1750, Microsoft).
- **The AIIA has close-ended questions that hide nuance**. Participants also critiqued questions that restricted responses to a specific form, such as answers to Yes/No questions about the AI system, or checklists and multiple choice questions that asked users to sort the AI into relevant categories. One participant explicitly called attention to a specific section of their instrument (Microsoft Section 5.2, "Goal Applicability"): "*the [...] section classifies the systems into only two classes (i.e., yes or no). This may not measure the systems well.*" (P1816).

  The Canadian AIIA differed from the others in that it used users' responses to closed-ended questions to calculate a numeric "raw impact score" and "risk score", which were displayed at the bottom of the AIIA webpage (shown in Figure 2). These scores were then used to sort the AI into one of four "impact levels", which determine the mitigation steps that the system owners are required to take. However, many participants stated that they did not understand how the numeric scores were calculated (P1081), or found them difficult to interpret (P1384, P1153). One participant expressed a desire to better understand the method used to calculate the risk score: "*we were confused about how points corresponded to our answers. It may be more helpful to be able to walk-through (after a first draft is completed) and see what alternative choices would result in a lower score*" (P1081).

Content. The three AIIA instruments had overlap, yet varied considerably in their *content*: the types of impacts and mitigation steps that users were prompted to consider. We provide a detailed comparison across instruments in Appendix A, and summarize participants' critiques below:

- **The AIIA is missing types of impact**. Several participants noted that they believed their AIIA was not comprehensive enough. Participants wrote that their instrument's "*categorized limitations and harms*" are "*non-extensive*" (P1063, Canadian), or that their AIIA "*doesn't address broader issues beyond privacy and non-discrimination*" (P1281, Microsoft). Another participant explicitly noted types of impact that they believed should have been included in their AIIA: "*no question had us think critically about the accessibility of the system, or which subgroups/subpopulations could be disproportionately harmed if the model was to be deployed*" (P1081, Canadian).[4]
- **The AIIA includes questions that may not be applicable to all AI systems**. Some participants commented that not all questions may be relevant to all AI systems. One participant that completed the Microsoft AIIA found that "*some questions did not strictly apply to our product*" (P1819). This calls into question whether it makes sense to assess all proposed AI systems with a common (versus more specialized) AIIA instrument, a tension that has been raised by past work [18, 20, 37].

Feasibility. While participants were instructed to complete the AIIA in a facilitated classroom exercise, several participants pointed out perceived or imagined obstacles that may prevent organizations from completing an AIIA in practice. We group these concerns as those related to the *feasibility* of the AIIA, *i.e.,* what realistically must happen to effectively complete each questionnaire.

- **Users completing an AIIA desire supervision**. Several students listed a lack of supervision (*i.e.,* "*no oversight*") as a limitation of the AIIA process (P1505, P1709). While all three instruments provided an email address that users could contact with questions or feedback on each instrument, only the Canadian AIIA listed named persons who could provide synchronous help and consultation.
- **Users completing an AIIA desire clear expectations and transparency about how their responses will be reviewed**. Some students expressed uncertainty or confusion about how their completed AIIA would be evaluated or reviewed (P1613, P1709), or were "*unsure if we answered the questions properly*" (P1456). One student expressed concern specifically about how their

[4]The Canadian AIIA has been updated to include additional questions regarding accessibility since the students participated in the study.

open-ended responses would be evaluated: "*a lot of the questions were short answer, which makes sense since there's many different types of responses, but this sometimes made it unclear what should be answered*" (P1613).

- **Users may lack information about the AI that is necessary for the AIIA at the time of its completion**. Study participants were given a short one-paragraph description of a potential AI system. However, many students reported that this description did not provide enough information about the product to complete the AIIA (P1620, P1613, P1866, P1286, P1528). One participant commented that their AIIA "*seems to be designed for more fleshed out products rather than just ideas*" (P1286, Microsoft). Another participant's team "*didn't have a lot of context about the development of the AI system itself, so when answering these questions, we made a lot of assumptions about use cases and the development process*" (P1613, US CIO). When participants opted against making restrictive assumptions, they noticed that existing instruments did not allow them to consider multiple downstream possibilities: "*some of the questions where we need to select only a single option to be hard to choose from since the product might apply to multiple options provided (for example, the role of humans in the product)*" (P1819, Microsoft).

  We note that participants' critiques may not only reflect potential limitations of the AIIA templates, but also of our study design (*e.g.,* that students lacked knowledge or context about the imagined scenario that a practitioner would have). The scenario text (pasted in Appendix B.1) includes details of the proposed user interface but excludes implementation details such as the training dataset or model architecture, and was designed to simulate the information available in an AI services' early ideation stages. Thus, we hypothesize that many of the above critiques may still be relevant for real AI systems that are at a similarly early phase of their design.

- **The AIIA requires coordinating a team of stakeholders with interdisciplinary expertise**. Several participants reported that they lacked relevant knowledge or the appropriate domain expertise necessary to complete the AIIA. One participant that completed the US CIO AIIA was confused by "*legality questions*", writing, "*as technical members of the team, we were unfamiliar with the terminology used*" (P1456). Another participant noted a "*lack of explanation on some of the terms [and] regulations*" (P1696, Canadian). This finding points to the practical need to assemble a *team* of stakeholders with appropriate interdisciplinary expertise, and identifying who is responsible for each piece. Existing AIIA interfaces presently do not include support for collaborating asynchronously across multiple devices.

Effectiveness. Participants reflected on whether the AIIA was effective at helping them imagine, and develop a plan of action to mitigate potential harmful impacts of the proposed AI.

- **Potential harms & mitigation steps are vague or under-specified**. Several participants reported that their AIIA used language that was "*not clear*" (P1145, P1505), "*vague*" (P1005, P1886), or "*difficult to understand*" (P1620). Unclear language throughout the AIIA posed challenges to helping users imagine and enumerate specific harms: "*many of the questions were vague and did not go over the harms that could be caused*" (P1886, Canadian). Some participants specifically called attention to their AIIA's suggested mitigation steps, which they felt were under-specified: participants wrote that Microsoft's AIIA "*doesn't provide specific guidelines on how to address risks*" (P1281), "*didn't give us concrete directions of actions we need to take*" (P1121), and "*while good at identifying harms, [is] not very helpful in suggestions for mitigation*" (P1286). One participant noted how this under-specification could potentially be exploited in favor of the organization completing the AIIA, thus making it less effective as an accountability mechanism: "*many of the evaluations for mitigations were concerned with the existence of a procedure to handle something, but allowed this to be an in-house procedure and did not lay down any specifications as to what the procedure should consist of*" (P1359, Canadian).
- **Existing AIIA instruments are less effective for general-purpose technologies**. Across all three instruments, participants who were assigned to complete the AIIA for the general-purpose technology scenario reported a disconnect where parts of their AIIA did not seem applicable to general-purpose technologies. Participants that were shown the Canadian AIIA were especially more likely to note this disconnect (P1359, P1384, P1153, P1866, P1273), likely because the Canadian AIIA includes specific sections where the user must report the desired end use of the system (Section 7). Participants wrote that "*the questionnaire seemed to be designed for someone both creating a model and applying it to a given application domain. It didn't work as well for our case of creating a model that could be used in many domains*" (P1359) and that "*the AIIA*

does not account for the difference in scope of the project vs the potential applications" (P1153). Beyond critiques of existing instruments, one participant found imagining the potential impacts of a general-purpose system to be more challenging: "*I think the major limitation is that the description of the product was vague so depending on which industry the product was being applied to there could be all sorts of different issues that AIIA could not really pick up*" (P1866).

5 Discussion

Overall, our study provides preliminary evidence that exposure to existing AIIA templates *can* influence respondents' perceptions about generative AI systems (RQ2, RQ3), and also affect the issues and values that they believe are most urgent to address (RQ1). We analyzed students' critiques to identify four broad categories of concerns that limit the effectiveness of existing AIIA instruments (RQ4). These gaps point to opportunities for future work on **co-designing and validating (1) new and improved AIIA instruments, and (2) processes and workflows around them** to ensure they are completed and acted on appropriately.

Our study sheds light on the several issues in existing *instruments*. First, *existing AIIA templates are incomplete*. They do not include a comprehensive set of likely harms, and the harms they identify are not adequately defined for the user. This finding points to several important directions for future work. One direction is to conduct think-alound validation studies with a representative sample of stakeholders to ensure that key terms are interpreted correctly and uniformly. Another is to devise a mechanism to encourage users to think about context-specific harms [15]. Such approaches may explore how to meaningfully consult and involve impacted communities at the time of impact assessment [36, 18]. Additionally, participants found that *the user interface of existing AIIA templates are at times structured and restrictive, or too unstructured and open-ended*, hindering effective deliberation and imagination in both cases. These challenges are not bespoke to AIIAs, but rather are common challenges that have been well-documented in related work on survey design [7]. Future work can experiment with alternative interfaces and processes to strike a better balance between structured thinking and expressiveness. We also encourage future work to engage with existing best practices from research in survey design and impact assessments across domains [5, 38].

Perhaps more importantly, our work firmly establishes the need to embed AIIAs in the appropriate *workflows and processes*. As pointed out by prior work [18, 32], many existing AIIAs don't clearly specify *who* should participate in the impact assessment activity, and how different stakeholder groups should fill out the form individually or as a group. Further, existing AIIAs don't clearly specify *when* the AIIA should be completed within the AI development lifecycle. For example, some existing AIIA templates ask the system owners to answer questions about the dataset or model architecture that will be deployed. However, in practice these details are often unknown or subject to being changed at early phases of ideation, design, or development [33]. Yet, past work has underscored the need for impact assessment and participation at the earliest possible phases of ideation, before too much investment into the system has been made [30]. Developing AIIA processes that are designed to be *iterative and continued* across the stages of AI development [28], and clarifying at what stage of development certain parts of an AIIA can be completed, are important directions for future work.

Finally, foreseeing the impacts of general-purpose AI technologies is particularly challenging due to the vast space of possible use cases and application domains [29, 35]. To ensure that risks are properly anticipated and mapped using existing templates, the intended use of the AI under assessment must be well-scoped and contextualized. Effective impact assessment for general-purpose AI may require designing new impact assessment specifically designed for general-purpose use.

While artificial intelligence impact assessments have been referenced in more and more proposals, few such proposals have been made concrete. What exactly should an AIIA entail? By eliciting students' feedback on existing AIIA templates, we outline preliminary criticisms and potential paths forward. While existing AIIA templates are imperfect, we observed that they *did* shape students' thinking on potential harms, and we believe are promising tools for imagining potential risks of AI systems. Importantly, assessors' imagination of possible harms is just one of many articulated goals of AIIAs [37], and is just one step within the larger process of creating an effective accountability regime [32, 18, 12], which involves stakeholders beyond the assessors (such as impacted communities and the general public). More broadly, we hope that our work can inspire researchers and policymakers alike to conduct and share further empirical evaluations of AIIAs that assess their effectiveness for a wider set of stakeholders and goals.

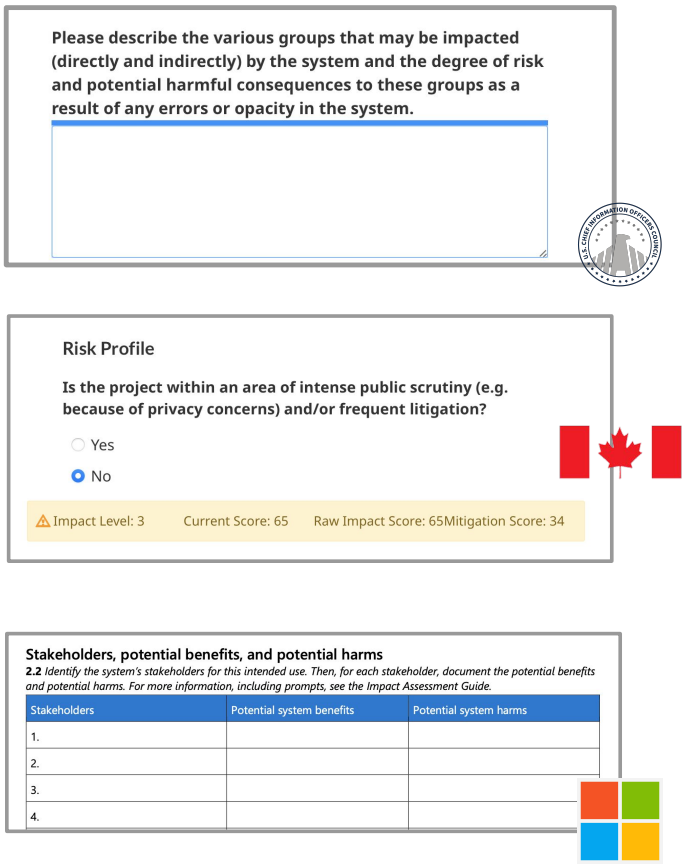

Figure 2: Screenshots of a subset of questions that prompt users to imagine potential harms from the three AIIA instruments evaluated in our study. Both the US CIO (Top) and Canadian Treasury (Middle) templates share a similar responsive web format, containing mostly closed-form questions (such as the 'Yes/No' questions pictured here) and the occasional free-response question. In contrast, the Microsoft template (Bottom), which is a static PDF, consists mostly of free-response exercises (such as filling out the pictured table). The Canadian template (Middle) calculates a series of scores pictured in the yellow bar, which are used to determine mitigation steps required for the proposed AI system [22].

Acknowledgments and Disclosure of Funding

We thank our study participants who made this research possible. We also thank the reviewers of the NeurIPS 2023 Regulatable ML workshop for their feedback. H. Heidari and N. Johnson acknowledge support from NSF (IIS2040929 and IIS2229881) and PwC (through the Digital Transformation and Innovation Center at CMU). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the National Science Foundation and other funding agencies.

References

[1] Nil-Jana Akpinar, Cyrus DiCiccio, Preetam Nandy, and Kinjal Basu. Long-term dynamics of fairness intervention in connection recommender systems, 2022.

[2] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art, 2017.

[3] Hugh Beyer and Karen Holtzblatt. Contextual design. *Interactions*, 6(1):32–42, jan 1999.

[4] Alessandra Calvi and Dimitris Kotzinos. Enhancing ai fairness through impact assessment in the european union: A legal and computer science perspective. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1229–1245, New York, NY, USA, 2023. Association for Computing Machinery.

[5] Colleen Cameron, Sebanti Ghosh, and Susan L. Eaton. Facilitating communities in designing and using their own community health impact assessment tool. *Environmental Impact Assessment Review*, 31(4):433–437, 2011. Health Impact Assessment in the Asia Pacific.

[6] The US Federal Trade Comission. Federal trade commission privacy impact assessments, 2023.

[7] Jean M. Converse and Stanley Presser. *Survey questions : handcrafting the standardized questionnaire*. Sage Publications, 1986.

[8] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. Who audits the auditors? recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1571–1583, New York, NY, USA, 2022. Association for Computing Machinery.

[9] Katerina Demetzou. Data protection impact assessment: A tool for accountability and the unclarified concept of 'high risk' in the general data protection regulation. *Computer Law & Security Review*, 35(6):105342, 2019.

[10] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. "i always assumed that i wasn't really that close to [her]": Reasoning about invisible algorithms in news feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 153–162, New York, NY, USA, 2015. Association for Computing Machinery.

[11] Lara Groves. Algorithmic impact assessment: a case study in healthcare, 2022.

[12] AI Now Institute. Algorithmic accountability: Moving beyond audits, 2023.

[13] Yordanka Ivanova. The data protection impact assessment as a tool to enforce non-discriminatory ai. *Springer Proceedings of the Annual Privacy Forum*.

[14] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. How different groups prioritize ethical values for responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, jun 2022.

[15] Nikolas Martelaro and Wendy Ju. What could go wrong? exploring the downsides of autonomous vehicles. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '20, page 99–101, New York, NY, USA, 2020. Association for Computing Machinery.

[16] Sean McGregor. Preventing repeated real world ai failures by cataloging incidents: The ai incident database, 2020.

[17] Microsoft. Microsoft responsible ai standard, v2 (general requirements), 2022.

[18] Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. Assembling accountability: algorithmic impact assessment for the public interest, 2021.

[19] Matti Mäntymäki, Matti Minkkinen, Teemu Birkstedt, and Mika Viljanen. Putting ai ethics into practice: The hourglass model of organizational ai governance, 2023.

[20] Norberto Nuno, Gomes de Andrade, and Verena Kontschieder. Ai impact assessment: A policy prototyping experiment, 2021.

[21] Serena Oduro, Emanuel Moss, and Jacob Metcalf. Obligations to assess: Recent trends in ai accountability regulations. *Patterns*, 3(11):100608, 2022.

[22] The Government of Canada. Algorithmic impact assessment tool, 2023.

[23] The White House Office of Science and Technology Policy. Blueprint for an ai bill of rights: A vision for protecting our civil rights in the algorithmic age, 2022.

[24] Leonard Ortolano and Anne Shepherd. Environmental impact assessment: Challenges and opportunities. *Impact Assessment*, 1995.

[25] Leonard Ortolano and Anne Shepherd. Vectors of ai governance - juxtaposing the u.s. algorithmic accountability act of 2022 with the eu artificial intelligence act. *Berkman Klein Center for Internet & Society*, 2023.

[26] the Czech Republic Finland France Estonia Ireland Latvia Luxembourg the Netherlands Poland Portugal Spain Position paper on behalf of Denmark, Belgium and Sweden. Innovative and trustworthy ai: Two sides of the same coin, 2020.

[27] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, jan 2020.

[28] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing, 2020.

[29] Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, and Saleema Amershi. Supporting human-ai collaboration in auditing llms with llms, 2023.

[30] Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. Algorithmic impact assessments: A practical framework for public agency, 2018.

[31] Teresa Scassa. Administrative law and the governance of automated decision-making: A critical look at canada's directive on automated decision-making. *SSRN Electronic Journal*, 01 2020.

[32] Andrew D. Selbst. An institutional view of algorithmic impact assessments. *35 Harvard Journal of Law & Technology 117*, 2021.

[33] Shreya Shankar, Rolando Garcia, Joseph M. Hellerstein, and Aditya G. Parameswaran. Operationalizing machine learning: An interview study, 2022.

[34] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction, 2023.

[35] Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim, Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul Christiano, and Allan Dafoe. Model evaluation for extreme risks, 2023.

[36] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. Participation is not a design fix for machine learning, 2020.

[37] Bernd Carsten Stahl, Josephina Antoniou, Nitika Bhalla, Laurence Brooks, Philip Jansen, Blerta Lindqvist, Alexey Kirichenko, Samuel Marchal, Rowena Rodrigues, Nicole Santiago, Zuzanna Warso, and David Wright. A systematic review of artificial intelligence impact assessments. *Artificial Intelligence Review*, 2023.

[38] USAID. Collecting and using data for impact assessment, 2006.

[39] Ramak Molavi Vasse'i, Jesse McCrosky, and Mozilla Insights. Ai transparency in practice, 2023.

[40] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E. Hines, John P. Dickerson, and Chirag Shah. Counterfactual explanations and algorithmic recourses for machine learning: A review, 2022.

[41] Elizabeth Anne Watkins, Emanuel Moss, Jacob Metcalf, Ranjit Singh, and Madeleine Clare Elish. Governing algorithmic systems with impact assessments: Six observations. AIES '21, page 1010–1022, New York, NY, USA, 2021. Association for Computing Machinery.

A Overview of AIIA Instruments

US CIO AIA. The US CIO AIA (Figure 2, Top) was developed to "*help [US] federal government agencies begin to assess risk associated with using automated decision systems*". At the time the study was conducted, the AIA's webpage states that it is in "alpha mode", and is presently not mandatory. The AIA is administered dynamically online, and contains a mix of free-response and multiple choice questions organized into 12 total sections. The AIA presently has no review process or named persons available to contact for assistance. Information in the AIA is only stored locally on one's own computer, and upon completing the assessment the user can export their responses to PDF.

The first eight sections of the AIA prompt users to report various design decisions that they have made in the development of the system and its intended use, such as the agency's motives for using automation (Section 1), the named people responsible for deploying the system (Section 2), details about the data source, data provenance, and data privacy (Sections 3, 4, 5) and model (Section 7). Section 8 of the AIA, titled "System Impact Assessment and Risk Profile", asks the user to reflect on different types of possible impacts their system may have to individual rights or freedoms, individuals' health or well-being, the environment, and the economy. The final four sections ask the user to self-report any plans that they have made to mitigate potential harms, including consulting with external stakeholders (Section 9), following data quality and provenance best-practices (Section 10), and maintaining an "audit trail" [28] (Section 12).

Canadian Treasury Board AIA. The Canadian Treasury Board AIA (Figure 2, Middle) was developed to support the Canadian Government's Directive on Automated Decision Making, which mandates that all government institutions subject to the directive are required to complete and publicly release results of the AIA (by uploading them to an Open Government portal) for any automated decision system [31]. Like the US CIO AIA, the AIA is administered dynamically online, and contains a mix of free-response and multiple choice questions organized into 13 sections.

The Canadian AIA notably differs from the others studied in that it scores participants responses to closed-form questions in real time. Specifically, the website calculates a numeric "raw impact score" and "mitigation score" (out of 51 and 34 total points, respectively), which are combined to sort the system into one of four "impact levels". The final impact level of the system determines the mitigation steps that the organization is required to take under the Directive on Automated Decision-Making [22].

While the content of the Canadian AIA is similar to that of the US CIO, the two templates only share a small handful of questions. Like the US CIO AIA, the Canadian AIA begins by asking a series of descriptive questions about the AI system, such as the agency's motives for introducing automation (Section 2), a "risk profile" that assesses the "stakes" of the system, the named persons responsible for completing the AIA and building the system (Sections 1 and 4), and basic information about the system, such as its capabilities (Sections 5 and 6).

The Canadian AIA's single "Impact Assessment" section is more thorough than that of the US CIO, which in addition to asking users to reflect on the same set of specific impacts from that template (*e.g.,* rights and freedoms, health and well-being, etc.), also includes many other questions about the larger social context in which the algorithm is embedded, *e.g.,* "*Please describe the output produced by the system and any relevant information needed to interpret it in the context of the administrative decision*" and, "*Will the system perform an assessment or other operation that would not otherwise be completed by a human?*".

Similarly, the Canadian AIA's three mitigation sections contain a larger and more comprehensive set of questions than the US CIO AIA. The "Data Quality" section asks if users plan to undertake processes to assess dataset bias or follow other best-practices (*e.g.,* releasing training data), and also asks if the user plans to "*make information [associated with these processes] publicly available*". The "Procedural Fairness" section asks about users' plans to implement best practices such as maintaining an audit trail of documentation [28], eliciting user feedback, or enabling human override of decisions. The final "Privacy" section asks if the users intend to complete a separate privacy impact assessment and lists potential mitigations to reduce risk (*e.g.,* de-identifying data before it is stored).

Microsoft Responsible AI Impact Assessment. In June 2022, Microsoft's Responsible AI team publicly released a PDF "Responsible AI Impact Assessment Template" (link), and accompanying guide (link) in a public blog post (link) that publicly announced Microsoft's Responsible AI Standard.

The Responsible AI Standard attempts to operationalize Microsoft's six AI principles to define *product development requirements* for all AI products at Microsoft. Goal A1 of v2 of the standard states that under the standard, all Microsoft AI systems are assessed by impact assessments, which are then reviewed "*according to your organization's compliance before development starts*".

One important difference that distinguishes the Microsoft template is that it targets the process of AI *development* (*e.g.,* with product teams and engineers), rather than grounded real-world use of AI (as was the case for the US CIO and Canadian templates, which are intended to be completed by public agencies who are planning to use AI). Another important difference is the form of the template: the majority of questions are free-response, and ask users to fill out tables (*e.g.,* a table where each row corresponds to a different stakeholder, and each column facilitates reflection on potential benefits and harms they may experience). In our classroom study, we presented students with the template PDF only (which does not contain links or references to the accompanying guide).

The content of the template is organized into five sections:

1. The "System Information" section asks descriptive questions about the system, such as its present stage in the "AI lifecycle", intended purpose and uses, features, and areas where the system may be deployed.
2. The "Intended uses" section asks the user to assess the system's "fitness for purpose", name potential benefits and harms to named stakeholders, and identify specific types of stakeholders (*e.g.,* name the person who will be responsible for overseeing the system post-deplyoment). The section also has a section on "fairness considerations", that prompts the user to identify any "demographic groups" that may require fairness considerations. The section also contains questions on technology readiness, task complexity, the role of humans in the system, and deployment environment.
3. The "Adverse Impact" section prompts the user to reflect on potential uses beyond intended use, such as restricted use, unsupported use, sensitive use, or intentional/un-intentional mis-use. The section also asks the user to report known limitations of the system and to imagine the impact of failure on stakeholders.
4. The "Data Requirements" section asks the user to assess existing data requirements that may apply to the system and the suitability of available training datasets.
5. The "Summary of Impact" section asks the user to "*describe initial ideas for mitigations*" for the potential harms that they identified earlier in the IA. It also asks users to check a box for whether each of Microsoft's "Responsible AI Goals" (*i.e.,* product requirements grouped into accountability, transparency, fairness, reliability & safety, privacy & security, and inclusiveness) applies to their system.

B Study Materials

We include the complete text (shown to study participants) of our study materials below.

B.1 Scenarios

Scenario #1: General-purpose technology. The private firm AI-X is developing a generative AI model capable of powering a hyper-realistic conversational video-chat agent. The system's user interface is very similar to common video conferencing softwares: the user logs in, describes the characteristics of the agent they would like to converse with, then a video chatbot exhibiting those characteristics appears on the screen and engages in an open-ended conversation with the user. You work at AI-X, and you have been tasked with completing an AIA for your team's project (i.e., the video chatbot).

Scenario #2: Hiring. The private firm AI-X has developed a generative AI model capable of powering a hyper-realistic conversational video-chat agent. The system's user interface is very similar to common video conferencing softwares: the system owner describes the characteristics of the agent they would like to create, the user logs in. then the video chatbot appears on the screen and engages in a conversation with the user according to the goals and characteristics specified by the system owner. AI-X is working on a product that fine-tune this agent to conduct initial screening interviews with job candidates.

B.2 Organizational roles

For both scenarios, we presented students with identical text for the first two roles:

1. ML scientist representing the team in charge of design and development 2. Product manager in charge of making the project economically successful

The text for the third and final role varied depending on which scenario the group was assigned to:

(Scenario #1: General purpose technology) 3. A member of the evaluation team who represents the interests of potential users.

(Scenario #2: Hiring) 3. A member of the evaluation team who represents the interests of potential job applicants.

Students were assigned to organizational roles alphabetically by last name.

B.3 Pre-Questionnaire

Before beginning the activity, students individually completed a pre-questionnaire Google Form that was designed to take 5 minutes.

1. (multiple choice) Please select your team.
2. (likert) On a scale of 1 to 5, how excited are you about the potential uses of foundational generative AI models (e.g., the product assigned to your team) in socially consequential domains? (1=not excited at all–5=very excited).
3. (likert) On a scale of 1 to 5, how concerned are you about the potential uses of foundational generative AI models (e.g., the product assigned to your team) in socially consequential domains? (1=not concerned at all–5=very concerned).
4. (free response) Please outline the top three issues you believe the development team of the AI product assigned to your team should address carefully before product release.
5. (multiple choice) Which one of the following broad categories of values do you believe is most urgent to address for your product? (A high- level description of each value–taken from prior work–is provided below. Options are ordered randomly).

- Fairness
- Safety
- Transparency
- Privacy
- Accountability & governance
- Human autonomy & agency
- Performance & efficiency

6. (likert) On a scale of 1 to 5, what is the relative level of responsibility you imagine for ML experts (compared to other stakeholders) to address the above matters? (1=very low at all–5=very high).
7. (free response) If you have any feedback for the teaching instructor or the research team about the questionnaire, please leave your comments here.

B.4 Post-Questionnaire

After completing the activity, students individually completed a post-questionnaire Google Form that was designed to take no more than 10 minutes. Note that all but two questions (Q2-3) which are new, are repeated verbatim from the pre-questionnaire.

1. (multiple choice) Please select your team.
2. (multiple choice) What was your individual role in your team?
3. (free response) What do you think are the major limitations of the AIA you completed for your team's product?
4. (likert) On a scale of 1 to 5, how excited are you about the potential uses of foundational generative AI models (e.g., the product assigned to your team) in socially consequential domains? (1=not excited at all–5=very excited).
5. (likert) On a scale of 1 to 5, how concerned are you about the potential uses of foundational generative AI models (e.g., the product assigned to your team) in socially consequential domains? (1=not concerned at all–5=very concerned).
6. (free response) Please outline the top three issues you believe the development team of the AI product assigned to your team should address carefully before product release.
7. (multiple choice) Which one of the following broad categories of values do you believe is most urgent to address for your product? (A high- level description of each value–taken from prior work–is provided below. Options are ordered randomly).
   - Fairness
   - Safety
   - Transparency
   - Privacy
   - Accountability & governance
   - Human autonomy & agency
   - Performance & efficiency
8. (likert) On a scale of 1 to 5, what is the relative level of responsibility you imagine for ML experts (compared to other stakeholders) to address the above matters? (1=very low at all–5=very high).
9. (free response) If you have any feedback for the teaching instructor or the research team about the questionnaire, please leave your comments here.

B.5 Value Category Definitions

We provided students with a brief description of each of the seven value categories taken verbatim from prior work [14].

- **Transparency**: A transparent AI system produces decisions that people can understand. Developers of transparent AI systems ensure, as far as possible, that users can get insight into why and how a system made a decision or inference.
- **Fairness**: A fair AI system treats all people equally. Developers of fair AI systems ensure, as far as possible, that the system does not reinforce biases or stereotypes. A fair system works equally well for everyone independent of their race, gender, sexual orientation, and ability.
- **Safety**: A safe AI system performs reliably and safely. Developers of safe AI systems implement strong safety measures. They anticipate and mitigate, as far as possible, physical, emotional, and psychological harms that the system might cause.
- **Accountability**: An accountable AI system has clear attributions of responsibilities and liability. Developers and operators of accountable AI systems are, as far as possible, held responsible for their impacts. An accountable system also implements mechanisms for appeal and recourse.
- **Privacy**: An AI system that respects people's privacy implements strong privacy safeguards. Developers of privacy-preserving AI systems minimize, as far as possible, the collection of sensitive data and ensure that the AI system provides notice and asks for consent.
- **Human Autonomy & Agency**: An AI system that respects people's autonomy avoids reducing their agency. Developers of autonomy-preserving AI systems ensure, as far as possible, that the system provides choices to people and preserves or increases their control over their lives.
- **Performance & Efficiency**: A high-performing AI system consistently produces good predictions, inferences or answers. Developers of high-performing AI systems ensure, as far as possible, that the system's results are useful, accurate and produced with minimal delay.

C Extended Results

In this section, we present results from additional analyses excluded from the main text due to space constraints.

C.1 Excitement & Concern

We report the average pre- and post-questionnaire excitement and concern ratings, stratified by (1) AIIA instrument and (2) scenario. We visualize the distribution of individual students' responses before and after completing the AIA in Figures 3 and 3.

AIIA Instrument - **Excitement** about potential uses of generative AI:

- **US CIO** ($n = 12$): $4.25 \rightarrow 4.25$ (-0)
  - 2 out of 12 students changed their mind
  - 1 student became less excited
- **Canadian Treasury** ($n = 12$): $4.08 \rightarrow 3.75$ (-0.33)
  - 3 out of 12 students changed their mind
  - 3 students became less excited
- **Microsoft** ($n = 14$): $4.00 \rightarrow 3.86$ (-0.14)
  - 4 out of 14 students changed their mind
  - 3 students became less excited

Scenario - **Excitement** about potential uses of generative AI:

- **Hiring** ($n = 19$): $4.21 \rightarrow 4.05$ (-0.16)
  - 5 out of 19 students changed their mind
  - 4 students became less excited
- **General use** ($n = 19$): $4.0 \rightarrow 3.84$ (-0.16)
  - 4 out of 19 students changed their mind
  - 3 students became less excited

AIIA Instrument - **Concern** about potential uses of generative AI:

- **US CIO** ($n = 12$): $4.08 \rightarrow 4.50$ ($+0.417$)
  - 6 out of 12 students changed their mind
  - 5 student became more concerned
- **Canadian Treasury** ($n = 12$): $4.08 \rightarrow 4.17$ ($+0.083$)
  - 5 out of 12 students changed their mind
  - 3 students became more concerned
- **Microsoft** ($n = 14$): $3.86 \rightarrow 4.29$ ($+0.43$)
  - 8 out of 14 students changed their mind
  - 7 students became more concerned

Scenario - **Concern** about potential uses of generative AI:

- **Hiring** ($n = 19$): $4.11 \rightarrow 4.26$ ($+0.16$)
  - 8 out of 19 students changed their mind
  - 5 students became more concerned
- **General use** ($n = 19$): $3.89 \rightarrow 4.37$ ($+0.47$)
  - 11 out of 19 students changed their mind
  - 10 students became more concerned

C.2 Level of Responsibility

We report the average pre- and post-questionnaire relative levels of responsibility stratified by condition, where a rating of 1 is very low and 5 is very high.

AIIA Instrument - Relative **responsibility** of ML experts:

- **US CIO** ($n = 12$): $4.42 \rightarrow 4.58$ ($+0.17$)
  - 6 out of 12 students changed their mind
  - 4 student reported increased responsibility
- **Canadian Treasury** ($n = 12$): $3.83 \rightarrow 4.25$ ($+0.42$)
  - 9 out of 12 students changed their mind
  - 7 students reported increased responsibility
- **Microsoft** ($n = 14$): $4.4 \rightarrow 4.64$ ($+0.21$)
  - 5 out of 14 students changed their mind
  - 4 students reported increased responsibility

Scenario - Relative **responsibility** of ML experts:

- **Hiring** ($n = 19$): $4.26 \rightarrow 4.63$ ($+0.37$)
  - 10 out of 19 students changed their mind
  - 8 students reported increased responsibility
- **General use** ($n = 19$): $4.21 \rightarrow 4.37$ ($+0.16$)
  - 10 out of 19 students changed their mind
  - 7 students reported increased responsibility

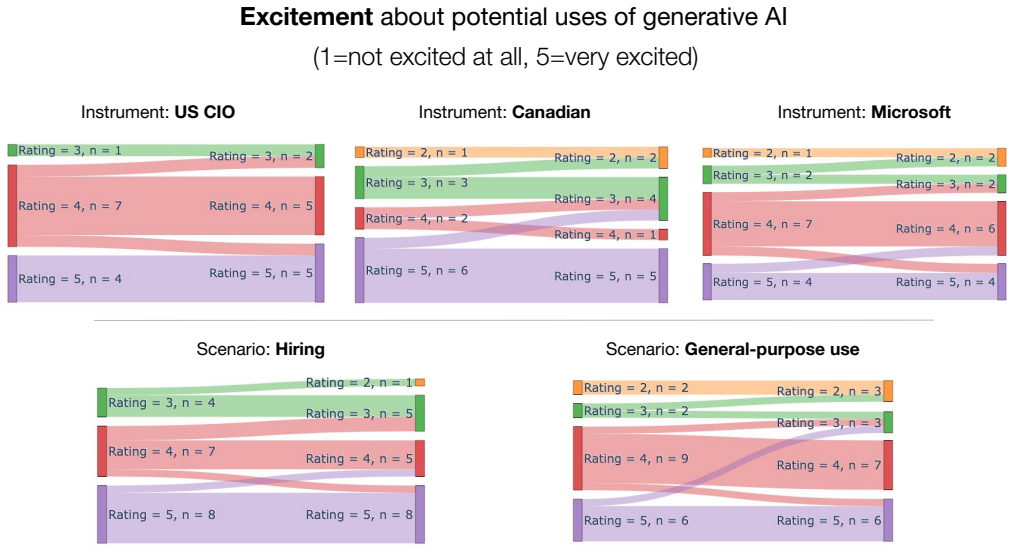

Figure 3: Sankey plots visualizing changes in students' response to the question, "*On a scale of 1 to 5, how **excited** are you about the potential uses of foundational generative AI models (e.g., the product assigned to your team) in socially consequential domains? (1=not excited at all–5=very excited).*" before (left) and after (right) completing the AIIA. We group students' responses by the AIIA instrument (Top) and scenario (Bottom) that their team was assigned.

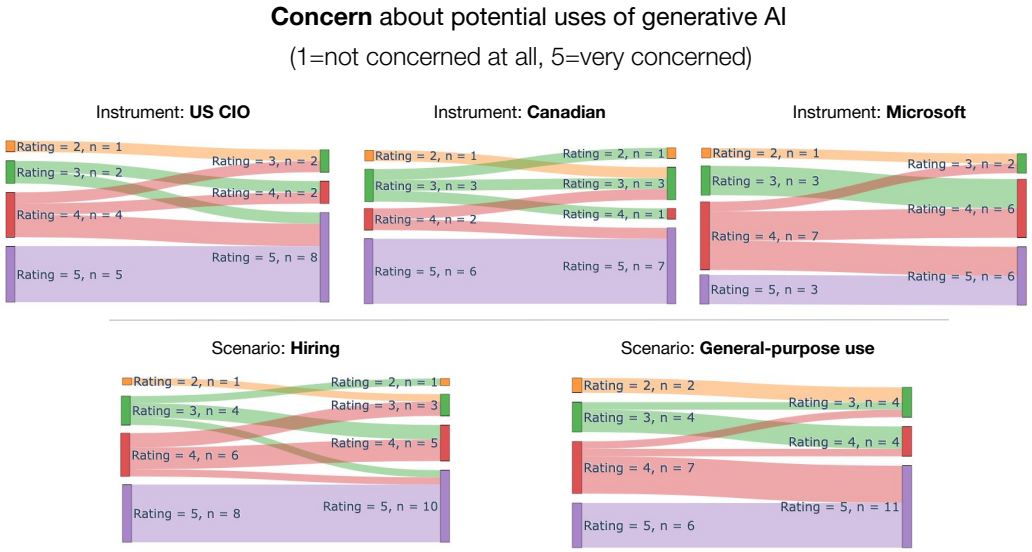

Figure 4: Sankey plots visualizing changes in students' response to the question, "*On a scale of 1 to 5, how **concerned** are you about the potential uses of foundational generative AI models (e.g., the product assigned to your team) in socially consequential domains? (1=not concerned at all–5=very concerned).*" before (left) and after (right) completing the AIIA. We group students' responses by the AIIA instrument (Top) and scenario (Bottom) that their team was assigned.