From plane crashes to algorithmic harm: applicability of safety engineering frameworks for responsible ML

Shalaleh Rismani
Google Research
McGill University
Montreal, Canada
shalaleh.rismani@mail.mcgill.ca

Renee Shelby
Google
San Francisco, USA
reneeshelby@google.com

Andrew Smart
Google Research
San Francisco, USA
andrewsmart@google.com

Edgar Jatho
Naval Postgraduate School
Monterey, USA
edgar.jatho@nps.edu

Joshua Kroll
Naval Postgraduate School
Monterey, USA
jkroll@nps.edu

AJung Moon
McGill University
Montreal, Canada
ajung.moon@mcgill.ca

Negar Rostamzadeh
Google Research
Montreal, Canada
nrostamzadeh@google.com

Abstract

Inappropriate design and deployment of machine learning (ML) systems leads to negative downstream social and ethical impact – described here as social and ethical risks – for users, society and the environment. Despite the growing need to regulate ML systems, current processes for assessing and mitigating risks are disjointed and inconsistent. We interviewed 30 industry practitioners on their current social and ethical risk management practices, and collected their first reactions on adapting safety engineering frameworks into their practice – namely, System Theoretic Process Analysis (STPA) and Failure Mode and Effects Analysis (FMEA). Our findings suggest STPA/FMEA can provide appropriate structure toward social and ethical risk assessment and mitigation processes. However, we also find nontrivial challenges in integrating such frameworks in the fast-paced culture of the ML industry. We call on the ML research community to strengthen existing frameworks and assess their efficacy, ensuring that ML systems are safer for all people.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

1 Introduction

Safety practices in the ML community often focus on ML-centered areas, such as Robustness, Monitoring, Alignment and Systemic Safety. In this paper, we propose the use of safety engineering frameworks in machine learning research and practice, and study its impact on responsible ML practices. While recent work argues that there are unique aspects to ML for which safety engineering may not generalize [8], we propose that there have not been serious efforts to apply these frameworks in ML. We examine the dialogue between safety engineering frameworks and understandings of social and ethical risks of ML systems. First, we report on ethical and social risk management practices currently used in the industry. Second, we take a developmental approach to examine how safety engineering frameworks can improve existing practices. We chose two of the most successful safety engineering frameworks used in other sociotechnical domains [27, 19, 4]: Failure Mode and Effect Analysis (FMEA) [5] and System Theoretic Process Analysis (STPA) [12, 18].

We conducted 30 semi-structured in-depth interviews with industry practitioners who shared their current practices used to assess and mitigate social and ethical risks. We introduced the two safety engineering frameworks, inviting them to envision how they might employ them to assess ethical and social risk of ML systems. The results of our study address the following research questions:

- **RQ1**: Which practices do ML practitioners use to manage social and ethical risks today?
- **RQ2**: What challenges do practitioners face in managing social and ethical risks?
- **RQ3**: How could safety engineering frameworks such as FMEA and STPA inform and improve current practices? What advantages and disadvantages of each method do ML practitioners identify?

Our findings illustrate safety engineering frameworks provide valuable structure for investigating how social and ethical risks emerge from ML system design and integration in a given context. However, successful adaptation of these frameworks requires solutions to existing organizational challenges for operationalizing risk management practices. Moreover, results of our work motivate further theoretical and applied research on adaptation of such frameworks. We start by providing an overview of current discourse in responsible ML development and contextualize the relevance of the safety engineering frameworks (Section 2). We discuss the strengths and limitations of applying safety engineering frameworks in light of current practice and call on the research community to further examine and strengthen these frameworks for ethical and social risk management of ML systems (Section 3). A full version of this paper is submitted for another archivable venue and it is currently under review.

2 Introducing safety engineering approaches to failure and hazard analysis

Safety engineering is a generic term for an assemblage of engineering analyses and management practices designed to control dangerous situations arising in sociotechnical systems [1, 7, 13]. These analyses and practices identify potential hazards or system failures, understand their impact on users or the public, investigate causes, develop appropriate controls to mitigate the potential harms, and monitor systems [26]. Safety engineering crystallized as a discipline around WWII, when military operators recognized losses and accidents were often the result of avoidable design flaws in technology and human factors [28]. Since then, implementation of safety engineering in sociotechnical domains, such as medical devices and aerospace, has significantly reduced accidents and failures [24].

We motivate use of safety engineering for social and ethical risk management given its strength in drawing attention to the relationships between risks, system design, and deployment [6, 20]. As ML systems introduce interdependencies between the ML artifact, its operational environments, and society at large [22], safety frameworks can provide strong analytical grounding for risk management [7]. Moreover, harms from ML systems are often recognized after they have occurred [21] at which point mitigating them is significantly more challenging and costly [5], and safety frameworks aid proactive harms surfacing. In this study, we focus on two safety engineering techniques designed to identify and address undesired outcomes early in development [1, 7, 13]: a failure analysis technique for improving reliability (FMEA) and a hazard analysis technique for identifying unsafe system states (STPA).

2.1 Failure Mode and Effects Analysis (FMEA)

FMEA, a long-standing reliability framework, takes an analytic reduction (i.e. divide and conquer) approach to identifying and evaluating likelihood of risk for potential failure modes (i.e. the mechanism of failure) for a technological system or process [5]. FMEA has been used in high consequence projects, such as space shuttle [11] and U.S. nuclear power plant safety [15]. The FMEA framework helps uncover potential failure modes, identify the likelihood of risk, and address higher risk failure modes for a system (i.e. bicycle), component (i.e. bicycle's tire), or process (i.e. bicycle assembly) [5]. FMEA is a multi-step framework, through which steps are iteratively performed by FMEA and system experts over the development life cycle [5]. See Figure 1 for breakdown of the steps and refer to Appendix for the details.

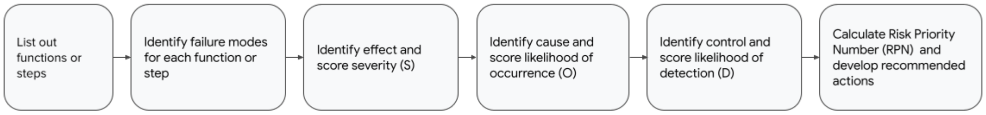
List out functions or steps
Identify failure modes for each function or step
Identify effect and score severity (S)
Identify cause and score likelihood of occurrence (O)
Identify control and score likelihood of detection (D)
Calculate Risk Priority Number (RPN) and develop recommended actions

Figure 1: Steps for conducting an FMEA [5]

2.2 System Theoretic Process Analysis (STPA)

The hazard analysis method, STPA, is a relatively new technique taking a system theoretic perspective towards safety [13]. It maps elements of a system, their interactions, and examines potential hazards (i.e. sources of harm). While analytic reduction requires a user of the tool to imagine interactions between components, modeling at the system level is meant to capture *emergent* phenomena that are well-described only by component interactions rather than individual component behavior. STPA has been employed in NASA's space program [10], the nuclear power industry [25], and the aviation industry [9].

In contrast to FMEA, the STPA process does not focus on reliability, failures, or risk likelihood. Instead, STPA models the sociotechnical system, focusing on the structure between components as well as control and feedback loops. Broadly, STPA encompasses the following steps, which are meant to be iterative (across the model of a system) and cyclic (across a system's lifecycle). See Figure 2 for breakdown of the steps and refer to Appendix for the details.

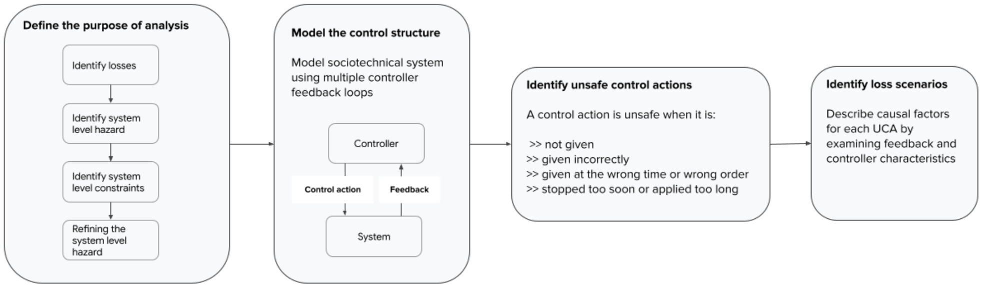
Define the purpose of analysis
Identify losses
Identify system level hazard
Identify system level constraints
Refining the system level hazard
Model the control structure
Model sociotechnical system using multiple controller feedback loops
Controller
Control action
Feedback
System
Identify unsafe control actions
A control action is unsafe when it is:
>> not given
>> given incorrectly
>> given at the wrong time or wrong order
>> stopped too soon or applied too long
Identify loss scenarios
Describe causal factors for each UCA by examining feedback and controller characteristics

Figure 2: Steps for conducting an STPA[12]

In sum, FMEA and STPA frameworks pose complementary analytical perspectives from safety engineering. Prior work suggests these techniques could strengthen the identification and mitigation of social and ethical risks of ML systems [14, 6, 23, 20]. Scholars have discussed the overall benefits of FMEA for internal ML auditing [20], illustrating how it could uncover ML fairness related failures [14], and have used it to propose an analysis of "social failure modes" for ML systems [23]. Yet, we could not locate any studies investigating ML practitioner's perspectives towards use of FMEA for social and ethical risk management. Similarly, several works suggest the value of a system theoretic framework for eliminating or mitigating social and ethical risks of ML systems [6, 17]. These works illustrate the theoretical application and benefit; however, little work to date explores industry ML practitioner's perspectives towards these techniques and how they could address perceived gaps in current risk management practices [16].

3 Findings

We interviewed 30 participants with expertise in foundational ML research; ethics review and advising; program and product management; and engineering. Interviews focused on participants' (a) current social and ethical risk management practices and (b) first impressions of safety engineering frameworks.[1]

In terms of current practices, participants described increased formalization of risk management, yet noted key aspects of their work - including defining and assessing for social and ethical risks - were characterized by an interpretive flexibility through which practitioners navigate peers with multiple and sometimes conflicting understandings of risk management work. While this flexibility accommodates the wide range of ML systems and contexts of deployment these practitioners are responsibilized to assess, it also fosters friction in multidisciplinary environments. Our findings show the emergent and chaotic nature of existing practices. We identified four key challenges that ML practitioners face when developing and implementing these practices: (a) Organizational structure: Incentive conflicts, fractured adoption, and unclear responsibilization; (b) resource constraints (e.g., time, capacity and data); (c) communication limitation of people with diverse perspectives and forms of expertise; and (d) Uncertainty and knowledge gaps for assessing ML systems.

Participant first impressions of FMEA and STPA underscored how safety engineering could bring greater structure to social and ethical risk management practice. Specifically, we find participants strongly endorsed the soundness these structures offer for social and ethical risk management. However, practitioners also emphasized understanding the context of use and implementation of the ML system remains a critical aspect of assessment, and raised concerns about employing safety engineering when context is not yet known. Lastly, implementation of FMEA- and STPA-like processes require internal capacity building and organizational shifts.

3.1 Observations on step-by-steps application of FMEA and STPA

When asked to walk through the FMEA and STPA processes as illustrated in Figures 1 and 2, participants discussed how they would apply each step on an ML system. Many participants remarked they would benefit from using FMEA and STPA processes *"earlier in development"* (R25) when they are tasked with assessing potential social and ethical risk.

Observations of the FMEA steps

- **Identify function or steps:** Practitioners expressed that there is value in outlining functions or steps for an ML system. They suggested either breaking down functions based on intended uses of an ML-based feature or product (i.e. *"to log food, to inform users and to provide the act of tracking."* (R10)) OR identifying steps for sub-processes along the ML development pipeline (i.e. *"dataset development, annotation, training, evaluation or deployment"* (R25)). Some remarked that it would be challenging to break down functions for the ML model itself. According to a research scientist *"key features/functions [of an ML system] are often embedded in some distributed representation in the model. This is especially true for larger models, and it is very hard to assess because the boundaries are not there anymore"* (R3).
- **Identify failure mode:** Participants were able and skilled in articulating failure modes (i.e. *"unfavorable chatbot responses to zip codes that are identified as lower socioeconomic status"* (R17)) but expressed uncertainty about their ability to comprehensively identify all potential failure modes due to the emerging nature of ML technologies. Moreover, they stated ethical and social risk often emerge from complex and non-tangible failures which are hard to identify and often need in-depth analysis of the ML system in context of use.

[1] Details on the study design, and interview findings can be found in the Appendix.

- **Identify effect:** Participants emphasized the importance of identifying the effect on whom or what, and expressed this should be incorporated into the FMEA process. They raised questions and seeked guidance about the extent to which practitioners should be responsible for protecting the interest of the company deploying/developing an ML system as opposed to interests of directly impacted users or the society at large.
- **Identify cause and control:** These two steps were seen as valuable, and participants noted ML technology companies have some control over changing design/implementation based on identified causes and controls. Participants noted that causal analysis and developing controls for social and ethical risk are active areas of research and suffer from challenges about uncertainties and knowledge gaps in assessing ML systems.

Observations of the STPA steps

- **Identify purpose of analysis:** Practitioners in different roles appreciated the start from stakeholder, values and losses. A participant explains *"I like the idea of starting with the negative outcomes, it's much more user oriented at the beginning, in terms of how it impacts them"* (R1). Social scientists and ethicists noted that in-depth value analysis and normative guidance is required in this step.
- **Create control structure:** Participants identified two scopes of analysis for drawing a control structure: internal company processes OR human-ML product interactions. Some noted that it is difficult to set meaningful boundaries and questioned how one could create a control structures for sociotechnical harms such as ecological harms. A participant noted that *"control structure would be very helpful in terms of limiting rather than constantly overextending where all of the potential problems or risks can come from"* (R10). Guidance is needed for where the system needs to be bounded and future work can use lessons from current STPA literature/practice. [12] Similar to reasons expressed for an FMEA, participants stated it is challenging to create a control structure for an ML model.
- **Unsafe control actions:** Participants appreciate *"the quadrant logic"* (R10) (i.e. four conditions outlined by the STPA process) for identifying how control actions could be unsafe. Some participants remarked that unsafe control actions could be mapped to *"design choices"* (R25) and stated that it would be valuable to have this type of analysis earlier on in the development process when creators can critically think about unsafe control actions to inform system requirements and corresponding design choices.
- **Loss scenarios:** Similar to FMEA, participants noted it would be hard to set boundaries for identifying causal scenarios for unsafe control actions. Considering the time and design limitations in our interview, participants did not have the time to provide adequate feedback on this step. Further investigation is necessary to understand how this step could be operationalized for ML systems.

4 Conclusion

In this work, we implicitly argue for an expanded definition of safety for ML systems - one which considers harms beyond physical harms - namely social and ethical harms. We posit that ML safety discussion need to include future work on how lessons and tools from safety engineering frameworks can offer a valuable insight for design of safer ML systems even though these frameworks were traditionally user for different technical systems. ML systems present new challenges; however, principles of safety engineering is transferrable across different technological systems. Challenges with organizational structure, resources constraints, representing diverse perspectives, and uncertainty of assessing ML systems present fertile ground for innovating social and ethical risk management tools. Quantitative, qualitative and reflexive investigative processes are emerging for defining, assessing and mitigating social and ethical risks. We examined existing practices and posit tools from safety engineering could provide value for creating more appropriate frameworks. Our preliminary discussions with ML practitioners about safety engineering frameworks, such as STPA and FMEA, showed these approaches could be adapted to provide the necessary guidance for systematically conducting failure and hazard analysis for social and ethical risks of ML systems. We discussed the strength and limitations of two processes from safety engineering and highlighted need for further research.

References

[1] Nicholas J Bahr. *System safety engineering and risk assessment: a practical approach*. CRC press, 2014.

[2] Virginia Braun and Victoria Clarke. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4): 589–597, 2019.

[3] Virginia Braun and Victoria Clarke. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, 18(3):1–25, August 2020.

[4] Nikhil Bugalia, Surjyatapa R Choudhury, Yu Maemura, and K E Seetharam. A systems theoretic process analysis (STPA) approach for analyzing the governance structure of fecal sludge management in japan. *Environment and Planning B: Urban Analytics and City Science*, page 23998083221075639, March 2022.

[5] Carl Carlson. *Effective FMEAs: achieving safe, reliable, and economical products and processes using failure mode and effects analysis*. Wiley, Hoboken, N.J, 2012.

[6] Roel I J Dobbe. System safety and artificial intelligence. February 2022.

[7] Clifton A Ericson et al. *Hazard analysis techniques for system safety*. John Wiley & Sons, 2015.

[8] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

[9] Takuto Ishimatsu, Nancy G Leveson, John Thomas, Masa Katahira, Yuko Miyamoto, and Haruka Nakao. Modeling and hazard analysis using stpa. 2010.

[10] Takuto Ishimatsu, Nancy G Leveson, John P Thomas, Cody H Fleming, Masafumi Katahira, Yuko Miyamoto, Ryo Ujiie, Haruka Nakao, and Nobuyuki Hoshino. Hazard analysis of complex spacecraft using Systems-Theoretic process analysis. *J. Spacecr. Rockets*, 51(2): 509–522, March 2014.

[11] Kouroush Jenab and Joseph Pineau. Failure mode and effect analysis on safety critical components of space travel. *Manag. Sci. Lett.*, 5(7): 669–678, 2015.

[12] Nancy Leveson and John Thomas. STPA_Handbook, March 2018.

[13] Nancy G Leveson. *Engineering a safer world: Systems thinking applied to safety*. The MIT Press, 2016.

[14] Jamy Li and Mark Chignell. FMEA-AI: AI fairness impact assessment using failure mode and effects analysis. *AI and Ethics*, March 2022.

[15] Huai-Wei Lo, James J H Liou, Jen-Jen Yang, Chun-Nen Huang, and Yu-Hsuan Lu. An extended fmea model for exploring the potential failure modes: A case study of a steam turbine for a nuclear power plant. *Hindawi*, 2021.

[16] Nikolas Martelaro, Carol J. Smith, and Tamara Zilovic. Exploring opportunities in usable hazard analysis processes for ai engineering, 2022. URL `https://arxiv.org/abs/2203.15628`.

[17] Donald Martin, Jr, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S Isaac. Participatory problem formulation for fairer machine learning through community based system dynamics. May 2020.

[18] Riccardo Patriarca, Mikela Chatzimichailidou, Nektarios Karanikas, and Giulio Di Gravio. The past and present of System-Theoretic accident model and processes (STAMP) and its associated techniques: A scoping review. *Saf. Sci.*, 146:105566, February 2022.

[19] Todd Pawlicki, Aubrey Samost, Derek W Brown, Ryan P Manger, Gwe-Ya Kim, and Nancy G Leveson. Application of systems and control theory-based hazard analysis to radiation oncology. *Med. Phys.*, 43(3):1514–1530, March 2016.

[20] 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. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 33–44, New York, NY, USA, January 2020. Association for Computing Machinery.

[21] 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. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 33–44, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372873. URL `https://doi.org/10.1145/3351095.3372873`.

[22] Lydia Reader, Pegah Nokhiz, Cathleen Power, Neal Patwari, Suresh Venkatasubramanian, and Sorelle Friedler. Models for understanding and quantifying feedback in societal systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1765–1775, 2022.

[23] Shalaleh Rismani and AJung Moon. How do AI systems fail socially?: an engineering risk analysis approach. In *2021 IEEE International Symposium on Ethics in Engineering, Science and Technology (ETHICS)*, pages 1–8, October 2021.

[24] Clarence C Rodrigues, Stephen K Cusick, et al. *Commercial aviation safety*. McGraw-Hill Education, 2012.

[25] Sung-Min Shin, Sang Hun Lee, Seung K I Shin, Inseok Jang, and Jinkyun Park. STPA-Based hazard and importance analysis on NPP safety I&C systems focusing on Human–System interactions. *Reliab. Eng. Syst. Saf.*, 213:107698, September 2021.

[26] Kristin Sharon Shrader-Frechette. *Risk and rationality: Philosophical foundations for populist reforms*. Univ of California Press, 1991.

[27] Sardar Muhammad Sulaman, Armin Beer, Michael Felderer, and Martin Höst. Comparison of the FMEA and STPA safety analysis methods–a case study. *Software Quality Journal*, 27(1):349–387, March 2019.

[28] Diane Vaughan. *The Challenger launch decision: Risky technology, culture, and deviance at NASA*. University of Chicago press, 1996.

A Methodology

We conducted 30 semi-structured interviews with ML industry practitioners specializing in assessing and mitigating ML ethics risks, from six companies. The research proposal, the interview protocol, and consent forms were reviewed and approved within one of the institutions represented in this study. Here, we describe the participants, recruiting, data collected, analysis, and study limitations.

A.1 Participants and recruiting

We used purposive and snowball sampling to recruit participants. Recruitment inclusion criteria specified participants be 18 years old or older, and currently work in an industry position conducting, managing, or researching social and ethical risks of ML system. As our primary research question is to understand industry adoption of reliability engineering tools, we excluded practitioners in academic, governmental or not-for-profit organizations. While we did not establish specific quotas for each professional position, we sought a balance of roles and backgrounds.

Four of the authors brainstormed an initial list of interview candidates based on knowledge about their existing work profile (via networking and publication or presentation track record at major conferences) and sent emails inviting their participation. Once a candidate accepted an invitation to participate, the interview was scheduled and the interviewer sent the consent form. At the conclusion of each interview session, we invited participants to recommend other candidates. The lead author conducted all interviews, which lasted approximately 60 minutes except for two 90-minute interviews.

In total, 30 practitioners from a diverse range of industry roles and educational backgrounds took part in the study **(Table1).** Participants held a range of roles including management (e.g., product, technical program, research and executive) (*n*=8), research (*n*=11), analyst/advisory roles (*n*=9), and software engineers (*n*=2). All participants worked in their current role for at least one year, and had experience assessing multiple ML systems for social and ethical risks, including classifiers, recommendation systems, large language models and text to image models. We conducted interviews between June and August 2022. All participants gave informed consent prior to participating in the study; interviews were recorded with permission. Participants were not financially compensated for their participation.

Table 1: Participant's roles and reference ID

Job Title	Description	n (%)	ID
Research (i.e. research scientist, principal researcher)	Primarily conduct interdisciplinary research in responsible ML	11 (37)	R3, R4, R5, R12, R18, R19, R21, R25, R22, R28, R29
Analyst/advisory (i.e. ethics reviewer, ethics and policy advisor, sociotechnical analyst, user researcher, research associate)	Advise project teams and review ML systems according to internal review processes	9 (30)	R6, R7, R8, R13, R14, R16, R17, R24, R26
Management (i.e. product manager, technical program manager, research manager, chief executive officer)	Manage products, programs, companies, and research projects	8 (27)	R1, R2, R9. R10, R15, R20, R23, R27
Engineer (i.e. research/software engineer)	Design and develop ML systems	2 (6)	R11, R30

A.2 Interview design

The interview protocol consisted of two parts: a) current practices and challenges, and b) first impressions of FMEA and STPA applicability for ML systems. Following confirmation of consent, we asked participants to describe their role and the type of technologies they focus on. We then asked participants how they define, assess, and mitigate social and ethical risk, broadly conceived. Moreover, we asked participants to discuss challenges they face when assessing and mitigating social and ethical risks in their current role. In the second part of the interview, we introduced the two processes using non-ML examples: FMEA was described with an example of a car tire; STPA was introduced using an example of a new surgical technique. The introduction of each process (including the example) took approximately 5 minutes. We introduced each technique one at a time and then discussed it for 10 minutes each. During this discussion, we asked participants to share their first impressions (pros, cons) while considering their potential use as a social and ethical risks assessment tool for ML systems. We invited them to talk through how they would apply such a process on an ML system they have assessed previously. To avoid order bias, the interviewer alternated between the processes for each interview. All interviews were conducted online using a video conferencing platform. Participants discussed both techniques in all interviews except in two interviews where due to time restraints one of the techniques was not discussed. This occurred once for each of the technique.

A.3 Data analysis

We used reflexive thematic analysis [3, 2] to understand the main themes in the interview data. We used an automatic transcription software for transcribing the interview recordings and then manually cleaned the transcripts. The primary author removed identifying information (e.g., current employer, specific products/projects mentioned) from the transcripts to protect the anonymity of the participants. Four of authors coded the data, first taking familiarization notes to highlight key ideas emerging early in the analysis. We then conducted open coding of the interview data, using the QSR NVivo 12 qualitative analysis software. The lead author coded all of the interviews and three other authors collectively coded 15 interviews. Every interview was coded by two researchers. The authors responsible for coding met iteratively to discuss codes, data interpretations, and progress from codes to thematic discussions. During these discussion session, researchers resolved disagreements, and generated new codes as relevant concepts emerged. In the final session, these authors convened to organize codes thematically and discussed emerging themes. The lead author compiled all the coding documents and synthesized the themes from the group discussions. Next, thematic findings were shared with broader research team for confirmation and collaborative discussion.

A.4 Author reflexivity

As with all research, our positionality and lived experiences inform our approach to designing, conducting, and analyzing this research study. All authors are researchers living in Canada and the United States. Our collective disciplinary backgrounds informing our research perspectives include ML research and engineering, mechanical engineering, robotics, human-robot interaction, sociology/ science and technology studies, cognitive sciences, and cybersecurity.

A.5 Study limitations

Our study examines how ML practitioners engage in social and ethical risk management practices, what challenges they face, and how failure and hazard analysis frameworks could inform and improve their practice. As an exploratory study, further work is needed to deepen understanding and develop ML model or other contextually-specific insights on the applicability of FMEA and STPA. Moreover, the ML practitioners interviewed for the study did not have expertise in safety and reliability engineering, and had limited time and exposure to the techniques. This study reflects first impressions of these frameworks based on their experience. In addition, our participants primarily come from larger, multinational technology organizations (4 of 6 companies represented) and reside in North America. As industry practitioners, there are limitations on what some participants could disclose due to confidentiality commitments. Thus, further work could examine views from a wider range of practitioners, which could provide deeper insights.

B STPA and FMEA Steps

B.1 Failure Mode and Effects Analysis

1. List out the *functions* of a component/system OR steps of a process.
2. Identify potential *failure modes*, or mechanisms by which each function or step can go wrong.
3. Identify the *effect*, or impact of a failure, and score its *severity* on a scale of 1 – 10 (least to most severe).
4. Identify the *cause*, or why the failure mode occurs, and score its *likelihood of occurrence* on a scale of 1 – 10 (least to most likely).
5. Identify *controls*, or how a failure mode could be detected, and score *likelihood of detection* on a scale of 1 – 10 (least to most likely).
6. Calculate *Risk Priority Number* (RPN) by multiplying the three scores; higher RPN indicates higher risk level. Develop *recommended actions* for each failure mode and prioritize based on RPN.

B.2 System Theoretic Process Analysis

1. Define the *purpose of the analysis* by identifying losses via outlining stakeholders, and their values. System specific hazards and controls are then highlighted based on the loss.
2. Model the *control structure* of the full sociotechnical system using control feedback loops which consists of a controller which sends *control actions* to a system that is being controlled while receiving *feedback* from the same system.
3. Identify *unsafe control actions* (UCA) by going through each control action and thinking about unsafe modes of (no) action, incorrect action and untimely action.
4. Identify potential *loss scenarios* by outlining potential casual scenarios for each UCA.