From Plane Crashes to Algorithmic Harm: Applicability of Safety Engineering Frameworks for Responsible ML

Shalaleh Rismani*
McGill University
Montreal, QC, Canada

Renee Shelby
Google Research
San Francisco, CA, USA

Andrew Smart
Google Research
San Francisco, CA, USA

Edgar Jatho
Naval Postgraduate School
Monterey, CA, USA

Josh A. Kroll
Naval Postgraduate School
Monterey, CA, USA

AJung Moon
McGill University
Montreal, QC, Canada

Negar Rostamzadeh
Google Research
Montreal, QC, Canada

ABSTRACT

Inappropriate design and deployment of machine learning (ML) systems lead to negative downstream social and ethical impacts – 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 an appropriate structure for 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 CHI community to strengthen existing frameworks and assess their efficacy, ensuring that ML systems are safer for all people.

CCS CONCEPTS

• **General and reference → Evaluation**.

KEYWORDS

Empirical Study, Safety Engineering, Machine Learning, Social and Ethical Risk

ACM Reference Format:
Shalaleh Rismani, Renee Shelby, Andrew Smart, Edgar Jatho, Josh A. Kroll, AJung Moon, and Negar Rostamzadeh. 2023. From Plane Crashes to Algorithmic Harm: Applicability of Safety Engineering Frameworks for Responsible ML. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3544548.3581407

1 INTRODUCTION

During a panel at the 1994 ACM Conference on Human Factors in Computing Systems (CHI), prominent scholars from different disciplines convened to discuss "what makes a good computer system good." Panelists highlighted considerations for safety, ethics, user perspectives, and societal structures as critical elements for making a *good* system [40]. Almost 28 years later, we posit that these epistemological perspectives need to be in a deeper conversation for designing and assessing machine learning (ML) systems that challenge the conventional understanding of safety and harm.

The development and use of ML systems can adversely impact people, communities, and society at large [12, 33, 82, 88, 120, 127], including inequitable resource allocation [3, 21, 107], perpetuating normative narratives about people and social groups [54, 122], and the entrenchment of social inequalities [1, 69, 75]. We frame these adverse impacts broadly as *social and ethical risks.* To manage such risks, quantitative [38, 65], qualitative [43, 70, 81, 98], and epistemological frameworks [32, 45, 82] have been proposed. Recently, scholars in the responsible ML community have advocated for use of safety engineering frameworks for managing social and ethical risks [30, 98]. Safety engineering frameworks offer an essential perspective for managing social and ethical risks for ML systems for two main reasons. Firstly, these frameworks provide the necessary analytical structure to connect harms to potential failures and hazards for existing design choices and streamline appropriate mitigation development. Secondly, these frameworks could inform active regulatory and standards activities taking place internationally [4, 36, 44, 61, 90]. Despite the growing body of empirical work on the operationalization of responsible ML practices [24, 73], there is a minimal understanding of whether proposed safety engineering frameworks are adopted in industry and how/if practitioners, tasked with managing social and ethical risks, perceive or use these methods.

Recognizing the potential advantages of safety engineering frameworks and inspired by the 1994 panelists, we examine the dialogue between these 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

[*] This work was completed during lead author's internship at Google Research.

CHI '23, April 23–28, 2023, Hamburg, Germany
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581407

two of the most successful safety engineering frameworks used in other sociotechnical domains [17, 93, 119]: Failure Mode and Effect Analysis (FMEA) [20] and System Theoretic Process Analysis (STPA) [66, 92], which we describe in detail in Section 2.

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 the ethical and social risks 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? What challenges do practitioners face in their attempts to manage social and ethical risks?
- **RQ2**: What are ML practitioners' perspectives towards using FMEA and STPA-like processes for social and ethical risk management? How could safety engineering frameworks such as FMEA and STPA inform and improve current practices?

We contribute to the emerging research on managing the social and ethical risk of ML systems in human-computing scholarship and responsible ML communities by offering:

- An overview of how practitioners define, assess, and mitigate social and ethical risks;
- A set of insights on how FMEA and STPA could inform existing practices along with their perceived advantages and disadvantages;
- Future research directions and calls to action for HCI and responsible ML scholars.

Our findings illustrate safety engineering frameworks provide valuable structure for investigating how social and ethical risks emerge from ML systems' design and integration in a given context. However, successfully adapting these frameworks requires solutions to existing organizational challenges for operationalizing formal risk management practices. Moreover, the results of our work motivate further theoretical and applied research on the adaptation of such frameworks. The remainder of this paper is organized as follows. We start by providing an overview of the current discourse in responsible ML development and contextualize the relevance of the safety engineering frameworks (Section 2). We outline our interview protocol and analysis methods in Section 3 and highlight key findings in Section 4. We discuss the value and shortcomings of applying safety engineering frameworks in light of current practices and call on the research community to further examine and strengthen these frameworks for ethical and social risk management of ML systems in Section 5.

2 BACKGROUND

Analyzing social and ethical implications of algorithmic systems is not new to computing researchers and practitioners [9, 29, 41, 91]. In the literature, terms such as harm [120], failure [97], and risk [61, 127] are often used to describe adverse impacts of ML systems. While there is currently no agreed upon definition of these terms and their relationships, we use the phrase *social and ethical risk* to frame broadly the adverse social and ethical implications ML systems can have on users, society, and the environment. This working definition provides conceptual consistency in this paper and is not meant to be normative. In the remainder of this section, we contextualize current discourses on social and ethical risks in ML to situate our study design, findings, and discussion. We highlight current epistemological perspectives and tools for responsible ML development and detail the safety engineering frameworks (FMEA and STPA).

2.1 Epistemological perspectives for anticipating and mitigating harms of ML systems

Scholars have proposed various methods for anticipating social and ethical impacts [24, 37, 109]. Anticipating harm involves thinking about the values [87, 112] and affordances of ML systems [15], with specific attention to how social norms and power dynamics constitutively shape adverse impacts of ML systems [10, 11]. The process of anticipation is aided by critical epistemologies that center the needs and standpoints of socially oppressed groups, including critical race theory [10, 45, 56, 88], post-colonial theories [82], and queer [116], and feminist HCI [7].

As social and ethical impacts are co-constituted through the interplay of technical system components, and the social world [51], design methodologies attentive to these dynamics support more meaningful harm anticipation and mitigation. For instance, Value Sensitive Design that examines what value tensions ML systems create or resolve [39, 126], supports increased stakeholder coordination [124] and consideration of technology from different social standpoints and perspectives [6]. Similarly, participatory design methods can center the needs of users, communities, and other stakeholders often excluded from the design process [132], or algorithmic governance [63, 64], especially when incorporating feminist epistemologies [7, 48]. Speculative design can also help designers imagine more socially just and racially equitable technological futures [46].

While these critical epistemological perspectives and design methodologies do not explicitly assess risk, they provide theoretical grounds for examining and mitigating social and ethical risk. We examine whether and how such epistemological perspectives inform current social and ethical risk management practices and discuss the possibilities of formally integrating them with safety engineering frameworks based on our findings.

2.2 Responsible ML tools, processes, and emerging regulations

With increased deployment of ML systems and reported harms [12, 88, 127], there is a movement towards formalizing quantitative and qualitative tools for responsible ML development. Traditionally, ML system evaluations [49, 105] prioritized assessing and optimizing for a narrow set of performance metrics, mistakenly treating these measurements (e.g., the accuracy of a test set) as a target rather than a proxy for certain risks [71]. Recognizing these shortcomings [53], ML scholars proposed alternative methods to enable more comprehensive evaluation. These methods include assessing computational fairness with alternative statistical definitions [19, 23, 25, 26], quantifying model interpretability based on statistical properties [83, 95], evaluating robustness to distribution shift [22, 57, 118] and examining model performance when exposed to adversarial examples [35, 106, 130, 131]. In parallel, significant effort has also focused on developing mixed-method (qualitative and quantitative) processes to increase accountability and assess ML systems contextually. Scholars have proposed model cards [81], datasheets [43] and auditing tools [16, 98, 111] to improve the transparency and quality of model and data practices. Human rights and algorithmic impact assessments aid the identification of potential societal level harms by examining model deployment in a given context. As part of this movement towards responsible ML development, a few scholars have also proposed the use of safety engineering frameworks for assessing and mitigating potential risks of ML systems [30, 98]. Parallel to tool and process development, there is a rapidly emerging set of international standards [90], policies [123], and regulatory frameworks [36] that examine ML systems from a risk-based perspective. [62, 76, 84]. Considering that existing responsible ML tools are not framed explicitly as risk management frameworks, we examine which tools practitioners self-report as a component of their social and ethical risk management practice. Furthermore, we investigate what pain points remain as they execute such practices and examine if using safety engineering frameworks could address some of these concerns and ease the upcoming efforts to meet regulations.

2.2.1 *Empirical studies of responsible ML practices.*

HCI scholarship examining the perceptions and needs of responsible ML practitioners has identified key challenges [47, 100], including limited definitional consensus on key terms [58] and the underlying need to translate principles into actionable guidance to catalyze transformative organizational change [28, 74]. Practitioners often work in multidisciplinary environments, where technical and non-technical stakeholders draw on different epistemologies, and perspectives [85], posing challenges to cohesive anticipation and identification of harms and risks [129]. In terms of risk assessment specifically, Raji et al. [98] underscore how the often-rapid pace and piecemeal implementation of risk assessment inhibit holistic forecasting of potential risks and their relationships to technical system components.

While there is a growing literature on practitioner needs, limited work has focused on identifying existing social and ethical risk management practices and ML practitioners' perspectives towards safety engineering frameworks. Martelaro et al.'s [77] study of the applicability of hazard analysis, and the needs of practitioners is a notable exception. From an exploratory interview study with eight participants, Martelaro et al. conclude existing hazard analysis tools from safety engineering cannot readily support ML systems and highlight how lack of team incentives, the pace of industry development, and underestimating the effort needed to create robust ML systems challenge the implementation of these tools. Nonetheless, Martelaro et al. emphasize frameworks are necessary to support risk management for responsible ML practice.

2.3 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 [5, 34, 67]. 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 [114]. 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 [125]. Since then, implementation of safety engineering in domains such as medical devices and aerospace has significantly reduced accidents and failures [104].

We motivate the 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 [30, 98]. As ML systems introduce interdependencies between the ML artifact, its operational environments, and society at large [102], safety frameworks can provide a strong analytical grounding for risk management [34]. Moreover, harms from ML systems are often recognized after they have occurred [99] at which point mitigating them is significantly more challenging and costly [20]. In this study, we focus on two safety engineering techniques designed to identify and address undesired outcomes early in development [5, 34, 67]: a failure analysis technique for improving reliability (FMEA) and a hazard analysis technique for identifying unsafe system states (STPA).

2.3.1 *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 the likelihood of risk for potential failure modes (i.e., the mechanism of failure) for a technological system or process [20]. FMEA has been used in high consequence projects, such as space shuttle [52] and U.S. nuclear power plant safety [72]. 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) [20]. FMEA is a multi-step framework through which steps are iteratively performed by FMEA and system experts over the development life cycle [20] (refer to Figure 1):

(1) List out the *functions* of a component/system OR steps of a process (e.g., everything the system/process needs to perform).
(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 (most likely to least likely).
(6) Calculate *Risk Priority Number* (RPN) by multiplying the three scores; a higher RPN indicates a higher risk level and develop *recommended actions* for each failure mode and prioritize based on RPN.

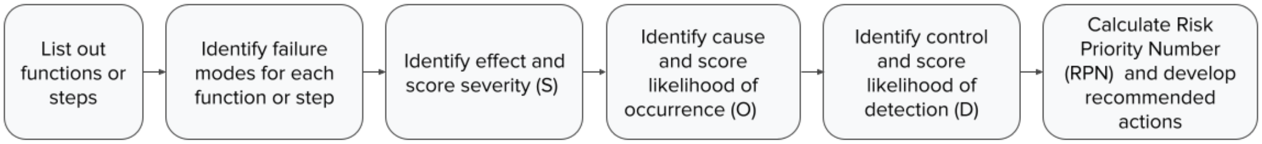

List out functions or steps
Identify failure modes for each function or step
Identify effect and score severity (S)
Identify cause and score likelihood of occurrence (O)
Identify control and score likelihood of detection (D)
Calculate Risk Priority Number (RPN) and develop recommended actions

Figure 1: Steps for conducting an FMEA [20]

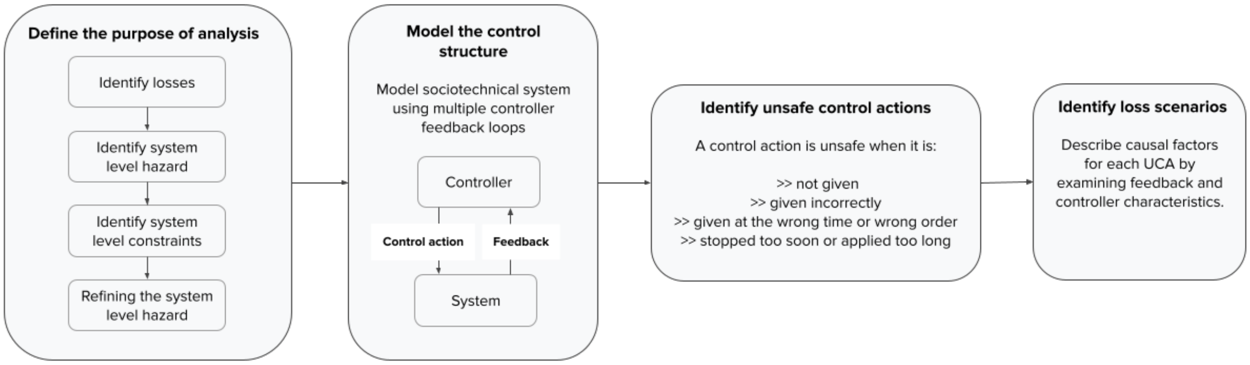

Define the purpose of analysis
Identify losses
Identify system level hazard
Identify system level constraints
Refining the system level hazard
Model the control structure
Model sociotechnical system using multiple controller feedback loops
Controller
Control action
Feedback
System
Identify unsafe control 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 [66]

2.3.2 *System Theoretic Process Analysis (STPA).*

The hazard analysis method, STPA, is a relatively new technique taking a system theoretic perspective towards safety [67]. 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 [50], the nuclear power industry [113], and the aviation industry [121].

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 (as illustrated in Figure 2) encompasses the following steps, which are meant to be iterative (across the model of a system) and cyclic (across a system's lifecycle):

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 specified 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 causal scenarios for each UCA.

These steps can be applied to positive effect at any stage in development and be used to develop requirements that must be enforced to ensure a safe sociotechnical system, such as new design decisions, requirements, procedures, operator training, test cases, or even periodic audits.

In sum, FMEA and STPA frameworks pose complementary analytical perspectives from safety engineering. Prior work suggests these techniques could strengthen identifying and mitigating social and ethical risks of ML systems [30, 68, 98, 103]. Scholars have discussed the overall benefits of FMEA for internal ML auditing [98], illustrating how it could uncover ML fairness-related failures [68], and have used it to propose an analysis of "social failure modes" for ML systems [103]. Yet, we could not locate any studies investigating ML practitioner's perspectives towards the 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 [30, 78]. These works illustrate the theoretical application and benefit; however, little work to date explores industry ML practitioners' perspectives towards these techniques and how they could address perceived gaps in current risk management practices [77].

3 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, the recruitment material, and the consent forms were reviewed and approved in accordance with the privacy and ethics guidelines of the hosting institution. The data was collected, stored, and analyzed only by the researchers working in this organization. Here, we describe the participants, recruiting, data collection, analysis, and study limitations. [1]

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

3.1 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 systems. 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.

3.2 Interview design

The interview protocol, as illustrated in Figure 3, 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 the challenges they face when assessing and mitigating social and ethical risks in their current role. Recognizing that some of the existing empirical works have already explored similar topics [73, 79], we ask these questions to: 1) understand what practitioners self-report as ethical and social risk management practices specifically, and 2) prime the conversation about safety engineering frameworks on their existing approaches to risk management. 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 risk assessment tool for ML systems. We invited them to talk through how they would apply such a process to 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 techniques.

3.3 Data analysis

We used reflexive thematic analysis [13, 14] to understand the main themes in the interview data. We used 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 the 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. The authors responsible for coding met iteratively to discuss codes, data interpretations, and progress from codes to thematic discussions. During these discussion sessions, 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

[1] We included the interview protocol and material in the supplemental material.

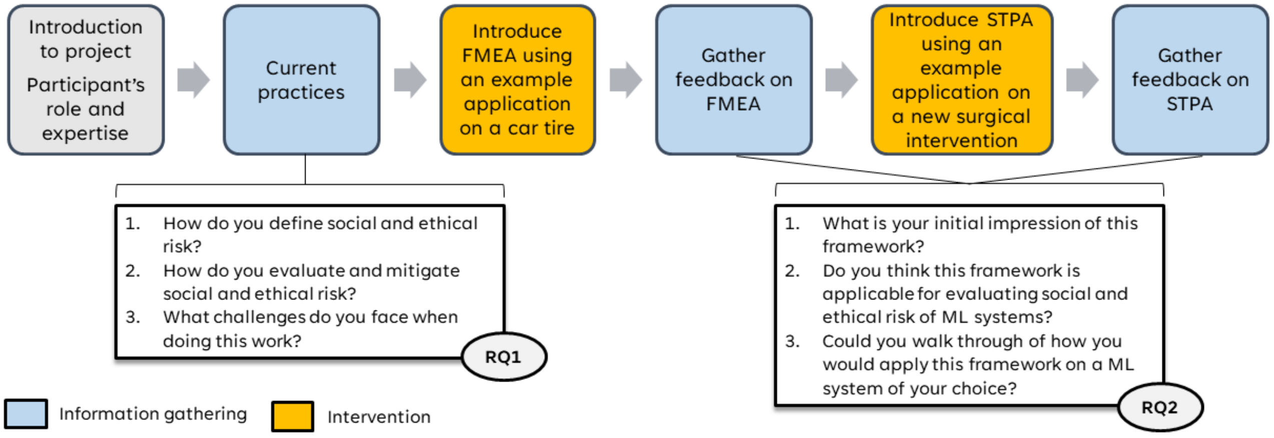

Introduction to project
Participant's role and expertise
Current practices
Introduce FMEA using an example application on a car tire
Gather feedback on FMEA
Introduce STPA using an example application on a new surgical intervention
Gather feedback on STPA
1. How do you define social and ethical risk?
2. How do you evaluate and mitigate social and ethical risk?
3. What challenges do you face when doing this work?
RQ1
1. What is your initial impression of this framework?
2. Do you think this framework is applicable for evaluating social and ethical risk of ML systems?
3. Could you walk through of how you would apply this framework on a ML system of your choice?
RQ2
Information gathering
Intervention

Figure 3: Interview Protocol Steps

synthesized the themes from the group discussions. Next, thematic findings were shared with the broader research team for confirmation and collaborative discussion.

3.4 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.

3.5 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 an 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.

4 FINDINGS

Our study examines how failure and hazard analysis frameworks could inform ML risk management practices. We present our findings in two parts to reflect our two research questions (Figure3) and start by highlighting what practitioners identify as existing social and ethical risk management practices and discuss the shortcomings and challenges of these existing practices (Section 4.1). In Section 4.2, we build on this understanding of current practices and challenges and discuss ML practitioner's perspectives on using FMEA- and STPA-like processes for social and ethical risk management.

4.1 RQ1: How are ML practitioners currently identifying and assessing social and ethical risk? What are the existing challenges they face?

Participants described increased formalization of risk management practices, 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 with multiple and sometimes conflicting understandings of risk management. 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. Organizational culture and resource constraints are power dynamics influencing these challenges.

4.1.1 Variable definitions of risk. Defining social and ethical risks sets the bounds of which system (mis)behaviors or downstream effects are acceptable or concerning. Rather than anchoring to a canonical definition, we find participants employ multiple definitions of social and ethical risks, explicitly noting that there is no widely-accepted definition in the ML community. This echoes the results described in Krafft et al. [58] where they found limited consensus on key definitions in AI ethics policy and practice. Similarly, the Ethics Owner Report by Metcalf et al. and Constanza-Chock et al.'s study investigating auditing practices for ML systems outlines the use of inconsistent and custom definitions by "ethics owners" and "auditors" of ML systems poses a challenge for conducting consistent and reliable assessments [27, 79]. Despite the lack of common definitions, there were points of convergence, each underpinned by concerns with adverse, material impacts on people.

Foremost, participants described social and ethical risks as *user and societal harms of ML systems*. Here, participants described *“harms"* broadly, without specifying uniform methods for surfacing harms to whom or what. While some participants noted a general *“user-centric [harms] framing works well [...] for a product and engineering organization”* (R15), others centered harms to “underrepresented” (R13) and “historically marginalized” (R4) communities. Participants' use of the harms perspective is echoed heavily in the critical epistemological perspectives [7, 45] toward anticipating harms for ML systems. Beyond harms, participants also described how *transgressions to a company's public AI ethics principles* offer a *“jumping off point”* (R15) to identify social and ethical risks. While these commitments provided a clear north star for identifying risks, participants also noted that the abstract nature of these principles limit their usefulness in practice. For instance, they do not help in identifying which stakeholder groups to prioritize, nor help to grapple with the constitutive role that the context-of-use and system affordances play (R15) in generating risks. These remarks echo similar limitations identified by other scholars [74, 100, 128]. Namely, Whittlestone et al. translated lessons from bioethics and suggested focusing on tensions between these principles to assess an ML system in practice[128]. Lastly, participants described social and ethical risks as *human rights violations* that could be surfaced through a human rights impact assessment. The human rights frame was less common than other definitions, though many participants recognized its value in evaluating systems deployed in cross-cultural domains. This emerging framing is echoed in recent scholarly conversation [96].

Variable definitions foster frustration, misunderstanding, and inefficiency in multidisciplinary environments [58, 74, 100]. Supporting findings from current scholarly work, participants all stated clear definitions and "formalized frameworks" are necessary for productive conversation about social and ethical risk management (R1, R10). For instance, as (R1), a product manager described: *“sociologists come from the harm perspective, whereas engineers often think of it in the failure perspective.”* Without a common language, identification and assessment of risks can be slowed. Despite the desire for a standard definition, many participants noted the value of definitional flexibility, particularly for assessing novel technologies in which strict definitions may not accommodate possible harms.

4.1.2 Multiple methods for assessing social and ethical risks. Whereas definitions of social and ethical risk constitute the boundaries of (un)acceptable ML system behavior, assessment methods shape the situated assumptions, guide the questions asked, and format how social and ethical risks are communicated. Participants described employing various risk assessment methods including qualitative, quantitative, and "reflexive investigatory" approaches, through which the methods and motivation of fellow practitioners are probed for alignment with organization principles and best practices. Participants from four of the six companies indicated that they have formal ethical review teams or programs, through which structured risk assessment occurs. These types of programs and teams mostly exist in larger technology companies, and they are not an industry norm.

Consistently, many of the participants noted that they begin social and ethical risk assessment by **qualitatively mapping potential harms of an ML system** individually or when possible, in teams with interdisciplinary expertise (i.e. product managers, AI ethicists, software engineers, and researchers working collaboratively). Mapping harms focuses attention on the adverse material impacts on people, including how ML systems can change work practices, socialization patterns, and other dimensions of social life (R18). Our participants' process of mapping harms involve *surveying existing literature*, with a focus on known impacts of *“related technologies”* and social contexts (R25). Mapping harms also involve *foresight exercises* to hypothesize potential worst/best case scenarios (R16) through free-form brainstorming and by working through structured questions created internally. Participants also referred to using more formalized assessment process such as human rights and impact assessment processes [84, 89] at this stage. A third and highly desired approach by participants, when resources permit, is *participatory methods*, where community-based stakeholders are engaged to co-identify social and ethical risks [48, 132].

Engineers and computer scientists also described **quantitatively testing ML system properties**. Such assessments begin with functional tests, where *“ML components are treated much like a piece of software”* (R11) and are subject to routine code reviews and performance tests measuring accuracy, recall, and precision. Functional tests do not explicitly measure social and ethical risks. Assessments for such risks are additive and *“bespoke for every project”* (R11), which may include disaggregated analysis [2, 8], counterfactual and causal analysis [42], and adversarial testing [35, 106, 130, 131]. These assessments are conducted pre- and post-launch, and aim to identify allocative, representational, and quality-of-service harms based on identity characteristics. They cannot, as participants note, capture non-computational harms, particularly the diffused and long-term impacts of ML systems in the world [110]. Moreover, participants described limitations in post-launch assessment, as there are no rigorous ways of identifying such risks unless reported by users, media outlets, or external auditors as echoed by scholars who currently are working on developing tools to enable large-scale auditing by everyday users [111].

Lastly, participants in management roles described **interrogating product and research development processes**. This approach is motivated by participants' recognition that technologies are influenced by the norms, intentions, and common practices of researchers and developers. Participants described reviewing product team documentation and methodologies and making recommendations to improve practices to minimize harm to marginalized communities, such as assessing how a product team evaluates ML models and identifying whether they are operationalizing any responsible ML metrics [101]. Moreover, R24, an ethicist, described *“assess[ing] the intentions of teams and . . . predict[ing] . . . the[ir] impacts.”* While participants did not detail how they conduct such epistemological assessments to maintain confidentiality, they did reference that scholarly work such as value analysis [39] informs their current processes.

Overall, practitioners noted two tensions in risk assessment. Similar tensions have been noted in empirical research investigating fairness tools, and auditing processes [27, 74]. First, assessment is most effective when there is a commitment to multidisciplinary collaboration between product teams and "subject matter experts" (i.e., non-engineering practitioners, such as ethicists, sociologists, or people with contextual expertise), although such collaboration is *"hard"* given different epistemological background (R19). Second, assessments are *"very product dependent"* and require meaningful *"conversations with the product team to understand"* the product, its use case, and where harms may arise (R6). Overall, the participants emphasized the need to better standardize current methods for a systematic evaluation of social and ethical risks.

4.1.3 Emerging approaches for mitigating social and ethical risks. Social and ethical risks are often surfaced by ethicists and social scientists who sit outside of research and product teams, and are not subject to product launch incentive structures. As such, mitigating identified risks requires significant work to build cross-functional "partnerships" (R16) and gain buy-in from teams with relevant technical expertise and control to adjust models and product design.

Product managers may take charge of mapping a mitigation strategy, however, deciding on an approach also requires collaboration, as preferred strategies vary by disciplinary training. Engineers often gravitated towards algorithmic solutions, such as fine-tuning model parameters, creating new training datasets, and implementing blocklists or filters [86] to prevent harmful model inputs or outputs. In contrast, ethicists, social scientists, and designers emphasized UX solutions, policy development, explainability and transparency artifacts [43, 81], and education. Yet, practitioners all recognized need for multiple interventions, as one computer scientist elaborated:

> *"...I tend to gravitate towards algorithmic solutions [...but] want to qualify this is not the only way to solve things [and] there are some things … not mitigatable by algorithmic techniques. In which case, essentially, I defer my expertise to somebody else because maybe the solution in that case, is more on the policy side or participatory design methods outside the scope of what I'm familiar with." (R3)*

Prioritizing mitigations is also a challenge, as some recommendations may *"take months, maybe even years to fully fix."* (R8). There are no clear guidelines on what mitigations need to happen and which ones can be put on hold; though some noted movement towards formalizing mitigation frameworks (R16). As such, resource availability and the product team's priorities dictate which mitigations will be pursued. This approach to mitigating risks is well-aligned with the tendency for reactive decisions in prevalent responsible AI practices as described by Rakova et al [100].

4.1.4 Challenges in current social and risk assessment approaches. We find that the challenges present in existing social and ethical risk assessment practices largely echo four issues that have previously been documented and particularly aligned well with findings reported by Martelaro et al. [77]. First, many of our participants expressed that the organizational incentive structures counter the mandate and the purpose of social and ethical risk assessments and mitigation strategies (R1, R3, R17). R17 articulates it as such:

> *"I think inherently, what we do is not aligned with a corporation … it's not revenue-generating work. It's work that can inhibit the bottom line and a product launch. … it can be hard to get product teams to mitigate … because they just want to launch the product."*

This systemic issue has been expressed by other scholars who have studied and designed responsible AI tools. For instance, Madaio et al. highlight practitioner's need for organizational support when assessing the fairness of AI systems based on their co-creative design workshops [73, 74]. Moreover, Rakova et al. emphasize this point by highlighting how alignment between incentives and org-level mission statements allows for the flourishing of responsible AI practices that are anticipatory as opposed to reactive [100].

Second, participants also expressed frustration about the limited time, capacity, and other resources (e.g., data) to meaningfully address ethics risks encountered. A number of examples of such resource constraints were expressed, such as the need to rush and produce a new dataset for adversarial product testing under strict time constraints (R24). This challenge is in line with the well-recognized tendency for rapid ML development cycles in the industry which are in tension with the extended time needed to adequately assess and mitigate potential social and ethical issues as revealed by Madaio et al.'s investigation of developer's use of fairness checklists and Costanza-Chock et al.'s study of auditors' practices [27, 74].

Relatedly, R17 and R25 voiced concerns over the lack of diversity in perspectives and forms of expertise involved in a risk assessment process. They expressed that the efforts to address the issue of diversity can be perceived as a resource-demanding activity that goes against the organization's underlying incentives. The concerns over the need for more diversity have been at the heart of the fairness in ML movement, as described in Sambasivan et al.'s canonical work outlining the importance and difficulty of creating quality and representative datasets [107]. Our findings suggest that significant process improvements are yet to be made to address a similar issue of diversity when it comes to identifying social and ethical risks. Lastly, as documented by The Ethics Owner's report [79] on the newly shaping practice of responsible ML, the lack of established practices and knowledge about concepts such as social and ethical risks, as well as the opaque nature of some of the complex ML systems [18] were seen as a hindrance to effective assessment and mitigation of social and ethical risks (R7, R8).

As outlined throughout Section 4.1, reported practices for social and ethical risk management reflect a wide range of epistemological perspectives and responsible ML tools that are applied inconsistently and reactively. Our findings illustrate that current practices do not follow any well-established risk management processes and the practitioners tasked with this work develop their own methods or use a mix of the publicly available tools. The lack of a systematic risk management framework can exacerbate existing challenges that these practitioners are facing. Specifically, identifying appropriate definitions, evaluations, and mitigation strategies, while justifying the need to do this type of work requires a significant amount of resources for every ML system. Furthermore, an ad-hoc approach to risk management makes it more challenging to deal with existing uncertainties and gaps in knowledge for ML systems. Building towards an "aspirational future" [100] for responsible ML development, in Section 4.2 we discuss how safety engineering frameworks could provide the necessary systematic structure for social and ethical risk assessment and help build anticipatory practices by streamlining current practices.

4.2 RQ2: What are ML practitioners' perspectives towards using safety engineering frameworks for social and ethical risk management?

Participants' first impressions of FMEA and STPA underscored how safety engineering could bring greater structure to social and ethical risk management practices (Section 4.2.1): such a structured approach could streamline disjointed definitions of ethical and social risks; it could also provide a coherent structure to the myriad assessment and mitigation methods that are currently being developed and implemented on an ad-hoc basis. Participants walked through applying both of the processes to an ML system, identified varying ways of translating these frameworks for an ML system and discussed specific gaps in knowledge. Furthermore, practitioners emphasized that understanding the context of use and implementation remains a critical aspect of the assessment. They raised concerns about the difficulty of employing safety engineering frameworks when the context is not yet known (Section 4.2.3). Others identified the existing industry norms and limited existing capacity within organizations as potential hindrances to implementing these frameworks (Section 4.2.4).

4.2.1 *FMEA- and STPA-like processes provide sound structure.*

Building on the need for a formalized and structured approach to the current social and ethical risk management practices (Section 4.1.4), many of our participants strongly agreed that FMEA- and STPA-like processes could provide such a structure. For instance, one researcher describes, *"there are definitely cases where explicitly defining failure modes (using an FMEA), trying to have a sense of what the potential causes are and how to mitigate them is a really useful framework for machine learning systems and one that has not been terribly well formalized until now."* (R12). On the other hand, an analyst (R17) noted the system theoretic approach is valuable because it frames the analysis of an ML system in relation *to both co-existing ML systems and societal power structures* which "provides a really sound structure to a process that we need in machine learning systems, especially from the ethical analysis perspective".

These structured frameworks can streamline current approaches to evaluating and mitigating potential social and ethical risks in different ways. R1 and R14 noted that FMEA links ML system's functionality to potential failure modes and their corresponding effects, causes, and detection methods. They elaborated that this connection promotes accountability in the social and ethical risk management process. On the other hand, R21, a researcher elaborates that *"a systems theory approach [as used in STPA] is very useful. It helps [an evaluator] understand relations between new pathways of harm and allows [them] to think about the multiple points of intervention"* (R21). Specifically, when introduced to the example application of STPA for new surgical interventions, R10 and R21 appreciated how the STPA process started from understanding potential losses and then linked them to specific unsafe control actions by identifying how different stakeholders interact with each other and medical devices. Furthermore, participants noted STPA and FMEA provide complementary and different analytical perspectives for examining failures and hazards. Particularly, R26 and R28 remarked that the two techniques provide two directions of analysis. They envisioned that an analyst would perform an FMEA for an ML system when they are interested in investigating a particular ML process (i.e., training data curation process) or component of a system (i.e. ML model). On the other hand, STPA could be used to examine specific losses and hazards by seeing how an ML system "integrates within a larger system", such as a product. R2 noted that there would be a value in applying both on the same ML system.

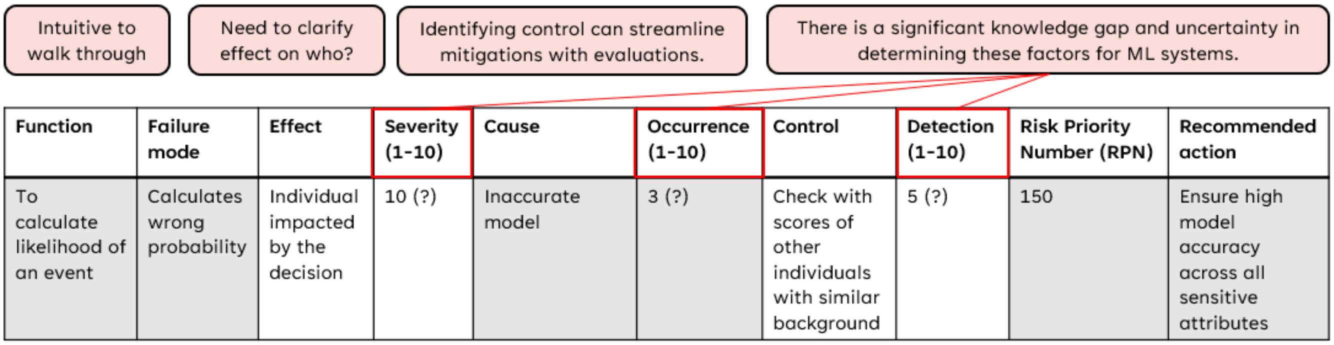
A few of our participants employed in large technology companies found commonalities between their existing practices and FMEA/STPA. For instance, R16 stated that *"we have already adopted components of the FMEA" namely the identification of failure modes, effect, and cause in our current risk assessment"*. However, they noted that FMEA provides a more comprehensive analysis by incorporating functional decomposition. R1 and R24, on the other hand, indicate that STPA is similar to current practices in that the identification of losses in STPA is analogous to the process of mapping out harms of an ML system. Apart from these elements, the majority of participants found the two frameworks novel. (i.e. a product) could create these potential hazards.

4.2.2 *Perspectives on step-by-steps application of FMEA and STPA.*

When asked to walk through the FMEA or STPA processes as illustrated in Figures 1 and 2, participants discussed how they would apply each step on an ML system. Overall, many of the participants were able to successfully walk through most of the steps for both of the processes. However, due to limitations of time, only a few participants got to provide detailed feedback on the last step of each process. Furthermore, in conducting the walk-through, the participants offered varying ways of translating FMEA and STPA steps for an ML application which can serve as important considerations for future applications of these frameworks. Reflecting on the challenges identified in Section 4.1.4, participants highlighted specific knowledge gaps and points of uncertainty when conducting each step. The following summarizes the key observations the participants made about each one of the processes. We illustrate these observations using a mock example in the illustrated walk-through of the frameworks in Figures 4 and 5.

Observations of the FMEA steps (as illustrated in Figure 4:)

- **List out function or steps:** Overall, practitioners expressed that it is valuable to explicitly outline functions or steps for an ML system when thinking about potential failures. When asked to break down functions, participants primarily selected functions based on the intended uses of an ML-based feature or product. For instance, R10, a product manager, described that a possible breakdown of functions of a nutrition tracking feature could be *"to log food, to inform users and to provide the act of tracking."* Some remarked that it would be challenging to break down functions for the ML model itself. According to a researcher *"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). Alternatively, few participants preferred to focus



Intuitive to walk through

Need to clarify effect on who?

Identifying control can streamline mitigations with evaluations.

There is a significant knowledge gap and uncertainty in determining these factors for ML systems.

Function	Failure mode	Effect	Severity (1-10)	Cause	Occurrence (1-10)	Control	Detection (1-10)	Risk Priority Number (RPN)	Recommended action
To calculate likelihood of an event	Calculates wrong probability	Individual impacted by the decision	10 (?)	Inaccurate model	3 (?)	Check with scores of other individuals with similar background	5 (?)	150	Ensure high model accuracy across all sensitive attributes

Figure 4: Perspectives on applying FMEA for ML systems as illustrated by a mock walk-through for a generic predictive ML algorithm used as a decision support tool. The illustrated example is analogous to the type of responses provided by participants in their own walk-through. Red color-filled boxes highlight the key takeaways.

on identifying steps for sub-processes along the ML development pipeline when walking through the FMEA process. For example, R25, a researcher described how they would break down the steps for *"dataset development, annotation, training, evaluation or deployment"* for an ML system and think about what could go wrong in each of those steps.

- **Identify failure modes:** After listing out potential functions or steps for an ML system, participants could comfortably articulate known and foreseeable failure modes. An analyst listed *"unfavorable chatbot responses to zip codes that are identified as lower socioeconomic status"* as one of the failure modes for the function of *"respond to use"* for a chatbot (R17)). Despite participants' expertise in identifying potential failure modes, many expressed uncertainty about their ability to comprehensively identify all potential failure modes due to the emerging nature of ML technologies. A researcher frames these *"unknown unknowns"* as *"foundational research challenges"* that require further investigation. Moreover, another product manager (R1) elaborated that ethical and social risks often emerge from complex and non-tangible failures which are hard to identify and often need in-depth analysis of the ML system in the context of use.
- **Identify effect and score severity:** Participants emphasized the importance of identifying the effect of a failure mode on *whom or what*, and expressed this should be incorporated into the FMEA process. Participants noted that failures have varying levels of impact on different stakeholders. For example, as seen earlier, R17 identified that an *"unfavorable response"* from a chatbot could have a different impact on a user from a *"lower socioeconomic status"* compared to someone with an average socioeconomic status. Similarly, R10 noted that it is important to consider how a certain product might fail for marginalized groups as opposed to an *"average user"*. Furthermore, participants noted that companies themselves could be affected by potential failure modes and raised questions about the extent to which practitioners should be responsible for protecting the interest of the company deploying/developing an ML system as opposed to the interests of directly impacted users or the society at large. Recognizing that current ways of identifying and assessing effect are ad-hoc and inconsistent, they sought guidance for who should be considered when identifying the effect of a potential failure mode.

  When asked to score the severity of an effect on a scale of 1 to 10, many participants raised questions on the possibility of defining a meaningful scale for social and ethical failures. R22, a researcher, describes: *"Uncertainty is a really big [challenge]. Sometimes [ethical and social] concerns are serious. But, it's not easy to parameterize things in the way a risk suggests. So you don't know if it's 10% likely, 90% likely. You just have deep uncertainty about whether a future risk will transpire. Particularly when we talk about [compounding] effects and complex systems."* This knowledge gap is especially salient when conducting assessments that require scoring severity or likelihood, which often require forecasting and hypothesizing. Many participants called for further work on developing validated methods and guidelines for assessing the severity of social and ethical failures. Similar challenges were noted regarding the scoring of the likelihood of occurrence and detection.
- **Identify cause score likelihood of occurrence:** When asked to identify potential causes, some practitioners mentioned examples such as *"misrepresentation in training datasets"* (R9, a product manager). However, participants noted that identifying causes could be challenging because components of an ML system are developed by different groups. R12, a researcher, notes *"One thing from my experience that I found is that it's very difficult at least in ML systems to get people to write down a model of the entire system."* Moreover, sometimes ML researchers cannot adequately understand the behaviour of an ML system. Identifying potential causes is very challenging considering these gaps in knowledge. Notably, participants highlighted the importance of further research and guidelines for identifying causes of potential failures.
- **Identify control and score likelihood of detection:** Identifying control was a more novel step for many of the participants compared to the last two steps. Many noted that they

do not explicitly identify controls in their current processes and expressed that they see value in the exercise of identifying potential controls for a failure mode. For instance, R12 and R1 listed conducting disaggregate analysis and monitoring system's performance in early releases as potential controls. However, as R21, a researcher, states it is difficult to identify a way of detecting some social and ethical failures as they have a *"secondary impact or their impacts is observed after a period of time"*. R15 also echoes the difficulty of *"meaningfully monitoring"* ML systems as they are deployed in a large scale. Overall, participants emphasized that further research on identifying appropriate control would be valuable.

- **Calculate Risk Priority Number (RPN) and develop recommended actions:** Most of the data collected from the interviews focused on discussion of the earlier steps. Mainly, participants observed that it is useful to have a way of prioritizing different failure modes. However, they questioned the efficacy of using RPN considering the issues that they raised about scoring severity, occurrence and detection.

As illustrated in the observations from the walk-through, participants see a clear need for further development and research on ways of identifying and scoring effect, cause, and control methods at a large scale for ML systems. Existing assessment and mitigation methods described in Section 4.1 could inform and be integrated into future FMEA analysis for social and ethical risks of ML systems. The following section describes the key themes of when participants completed an STPA walk-through.

Observations of the STPA steps (as illustrated in Figure 5:)

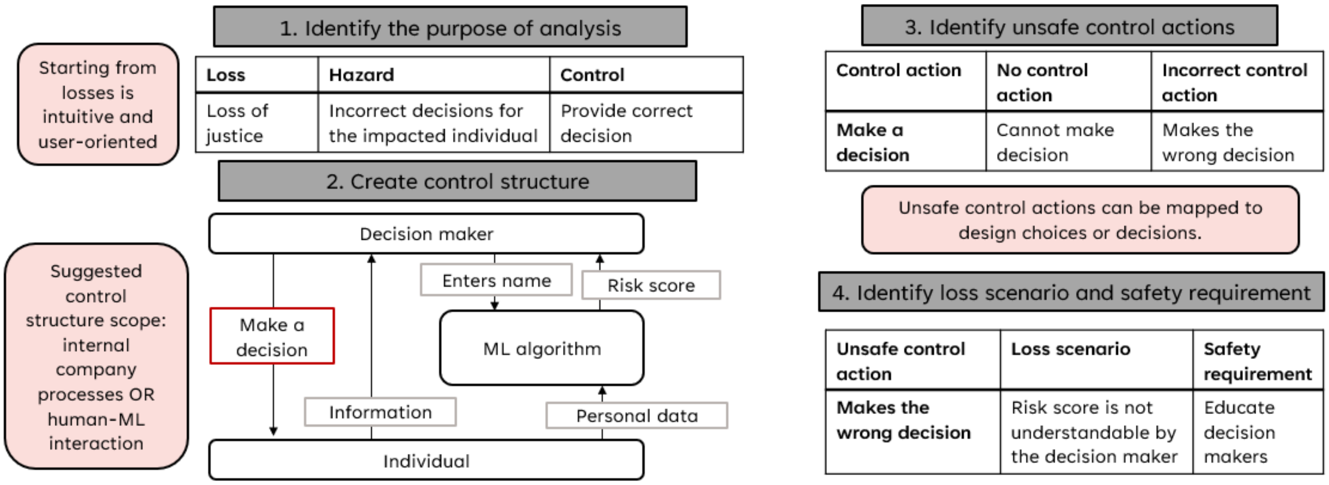
- **Identify the purpose of analysis:** Practitioners in different roles appreciated the start from stakeholder, values, and losses. Some examples of losses covered in the walk-throughs include *"loss of justice"* (R9), *"loss of dignity"* (R10) and *"loss of reputation for a company"* (R3). A product manager 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 [users]"* (R1). They elaborated that identifying losses is analogous to current practices of mapping out potential harms to a user. However, social scientists and ethicists noted that in-depth value analysis and normative guidance are required in this step. When it came to mapping out hazards, participants found it challenging to distinguish them from failure modes. Once a hazard was identified, it was intuitive to think of a constraint.
- **Create control structure:** Overall, participants identified two scopes of analysis for drawing a control structure: internal company processes OR human-ML product interactions. R19, a researcher, explained that a control structure for internal company processes would include elements such as the *"model development team"* and how they interact with *"product managers"* and *"policy advisors"*. On the other hand, a control structure for human-ML product interactions would include (but is not limited to) controllers such as the users, the distributor of the ML product, and the product itself. A product manager 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). However, participants raised a few challenges in selecting a control structure boundary for the social and ethical risks of ML systems. An ethicist noted that it is difficult to set meaningful boundaries and questioned how one could create a control structure for losses such as *"ecological harm"* (R21). Furthermore, a researcher stated it is challenging to create a control structure for an ML model. Further research and guidance are needed for where the system needs to be bounded, and future work can use lessons from current STPA literature and practices [66].
- **Unsafe control actions:** An example of a control action identified by a product manager is *"provide suggestion"* between *"a health mobile application"* and *"a user"* (R10). Once a control action is identified, participants appreciate *"the quadrant logic"* (R10, a product manager) for identifying how control actions could be unsafe. Some participants remarked that unsafe control actions could be mapped to *"design choices"* (R25). Some potential unsafe control actions could be that the app provides "the wrong suggestion" or that it *"gives the suggestion at the wrong time"* (R10). Many 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:** Identifying loss scenarios requires in-depth details about the control structure including the feedback and the process/mental model for each controller. Considering the time and design limitations in our interview, participants did not have the time to provide adequate feedback on this step as it was not possible to create a detailed enough control structure in the given time. Further investigation is necessary to understand how this step could be operationalized for ML systems.

Many of the participants were less familiar with the language around control feedback loops compared to the FMEA terminology. Participants asked for clear examples and instructions on how control structures should be drawn for different ML systems. Successful implementation of FMEA and STPA- frameworks faces two challenges identified in the next two sections.

4.2.3 *Understanding context is critical for social and ethical risk management: FMEA and STPA have limitations.*

When asked to walk through an FMEA and STPA for an ML system, participants emphasized the need for having a specific context of use. R7, an analyst, attempted to walk through the two processes for a reinforcement learning algorithm that did not have a specific use case and quickly ran into challenges with identifying failure modes for FMEA and hazards for STPA. This finding is in line with the fact that both of these processes are often applied for well-defined systems and use cases [20, 67].

Assuming that these frameworks are applied to a specific use case of an ML system, participants noticed FMEA only provides a framework for translating functions and steps to failure modes, effect, cause, and control. The FMEA does not explicitly facilitate structured thinking about social and ethical issues. To address this shortcoming, many participants suggested an FMEA analysis needs to be accompanied by a deep understanding of social issues relevant


1. Identify the purpose of analysis
Starting from losses is intuitive and user-oriented
2. Create control structure
Suggested control structure scope: internal company processes OR human-ML interaction
Decision maker
Make a decision
Enters name
Risk score
ML algorithm
Information
Personal data
Individual
3. Identify unsafe control actions
Unsafe control actions can be mapped to design choices or decisions.
4. Identify loss scenario and safety requirement

Loss	Hazard	Control
Loss of justice	Incorrect decisions for the impacted individual	Provide correct decision

Control action	No control action	Incorrect control action
Make a decision	Cannot make decision	Makes the wrong decision

Unsafe control action	Loss scenario	Safety requirement
Makes the wrong decision	Risk score is not understandable by the decision maker	Educate decision makers

Figure 5: Perspectives on applying STPA for ML systems as illustrated by a mock walk-through for a generic predictive ML algorithm used as a decision support tool. The mock walk-through represents the type of responses provided by participants. Red color-filled boxes highlight key takeaways.

to a given ML system. As a researcher explains, FMEA is *"very agnostic to the socio-technical context at first glance and it will be important to outline the use case when thinking about each component or process"* (R21). Similarly, a product manager remarked if the ethics analysts understand the*"context of deployment"* (i.e., where a model will be deployed in a product or feature), they can think of failures that are not just*"component specific"* (R1). The contextual understanding can allow for the identification of failure modes that have negative social and ethical effects. However, this would still require an analyst skilled at critically examining functions or steps of an ML system/process for social and ethical issues. Once appropriate and relevant failure modes are mapped, participants stated that thinking through the remainder of the FMEA process (i.e., the effect, cause, and control) can help practitioners gain *"foresight"* (R19) on social and ethical risks.

STPA, on the other hand, was perceived as a process that structurally considers an ML system in relation to stakeholders and other automated systems that interact with it. A researcher noted that STPA would be useful to analyze how an ML system *"fits into a larger decision-making process"* (R12). Many participants appreciated that STPA starts from understanding stakeholders, values, and losses and provides a framework for mapping relations between humans and the ML systems. A program manager expressed that STPA *"magnifies the fact that when you think of harms you have to have both the technical and then the social and ethical lenses"* and that it is valuable that STPA provides a framework to represent *"all of that"* (R2). Although STPA incorporates an understanding of context via system theoretic perspectives, participants expressed that mapping multidimensional interactions between an ML system and various stakeholders using bidirectional control feedback loops will at times lead to an inaccurate depiction of a sociotechnical system (R27). As R24, an ethicist elaborates *"once you start system theoretic analysis, you're going to abstract away and choosing proxies for social phenomena and they're going to be insufficient"* and noted that STPA-like processes should *"go hand-in-hand with expertise"* of understanding *"the limitations of systems processes analysis".* These limitations create opportunities for further theoretical development of these safety frameworks for social and ethical risk management. We discuss some potential avenues of improvement in Section 5.2.

4.2.4 Implementation of FMEA- and STPA-like processes require internal capacity building and organizational shifts. Many of our participants' first reactions to the FMEA and STPA processes was that the current industry culture and lack of internal capacity within technology companies will be hindrances to their adoption. Without a clear demonstration of their usefulness, industry adoption of similar processes will be slow. This sentiment is in line with the current challenges expressed about existing practices. R15, a product manager, states:

> *My biggest reaction is that [these processes] are so far from where our engineering culture is at. It feels like you would need to hire an entirely new type of person into these companies and over time completely change roles [...]. If [we] want engineering teams to do this themselves or be directly involved in the risk assessment [we] need to dramatically change the incentive structure."*

The need for an organizational culture shift was expressed with respect to both FMEA and STPA. However, a program manager explains STPA *"will require greater organization across teams and subject matter experts"* (R2) considering its focus on the interaction between different systems. Recognizing the challenge of creating these organizational shifts, participants were interested in exploring STPA only if they could see some concrete evidence of how it worked and what it delivered. STPA, in particular, was seen to require heavier internal capacity building to implement successfully since many participants have some familiarity with FMEA-like processes already while elements of STPA remain foreign. A researcher explains:

> *"People often think very linearly, and there's a challenge of trying a systems approach. [Practitioners] want to know X causes Y, causes Z, and they want to mitigate right at one of those points [similar to an FMEA] rather than thinking about all the connections between X, Y, and Z and what pathway is causing the most harm... [T]he first step is to get people to realize there are multiple relationships between X, Y, and Z."*

As building capacity requires time and buy-in from teams with different incentives, participants emphasized it is *"important for these processes to be simple"* (R9) so a diverse group of people can engage with them. Practitioners will need to learn new concepts, and it will be important to translate terminology used in FMEA and STPA for ML applications.

5 DISCUSSION

Policymakers and critics have been calling for establishing strong accountability practices in the ML industry [24, 60]. Strong accountability enables flexibility and experimentation while providing assurance and potential recourse to affected people [59]. A common requirement to establish a chain of accountability is whether a given harm or problem was adequately *foreseeable*. Although failure is often viewed as inevitable [31], or even desirable in ML [94, 108], safety engineering frameworks, such as FMEA and STPA, provide systematic processes to better anticipate risks [20, 67]. Our findings suggest that the previously documented challenges and limitations are still the main hindrances to today's ethical and social risk management practices: much of the definition of these risks remains flexible and variable; the organizational incentive to deliver ML products quickly conflicts with the mandate of the practitioners to assess and mitigate risks carefully; and resource limitations do not allow certain mitigation action to take place. Within this context, we find our participants welcome the systemic and structured nature of FMEA and STPA while remaining skeptical about the feasibility of implementing such frameworks in the current organizational environment. We discuss challenges and opportunities for future work to improve existing social and ethical risk management practices for ML systems.

5.1 Adapting safety engineering frameworks for responsible ML development: benefits and limitations

Our findings illustrate that practitioners today use some of the existing epistemological perspectives and responsible ML tools in identifying, assessing, and mitigating social and ethical risks. For example, participants highlighted drawing upon critical theory [45] to assess ML systems and their effect on marginalized communities and elaborated on using participatory design methodology [132] to understand the impact of an ML system on a group of users. Moreover, they referred to using various quantitative and qualitative tools such as adversarial testing protocols, transparency tools such as model cards [81], and algorithmic impact assessment [84]. Despite the increased formalization of social and ethical risk management, current practices remain disjointed. Without a systematic process, it is difficult for responsible entities to systematically identify risks [98]. In support of Raji et al. [98], and Dobbe's [30] proposals, our findings illustrate that safety engineering frameworks could inform and improve current practices in three main ways. Firstly, these frameworks provide a structured and well-defined way of thinking about key risk management components, including failure, hazard, harm, cause, and control. Similar to AI fairness checklists [74], FMEA and STPA-like processes could systematize ad-hoc social and ethical risk management practices within an organization. Moreover, these formalized analytic processes can be leveraged as a communicative tool and create consistency between how different practitioners (i.e., social scientists, computer scientists, and product managers) approach responsible ML development. This includes co-creating agreed-upon definitions of failure, hazard, and harm with product, policy, and responsible ML experts [58, 85]. Lastly, the lack of guidelines for assessing and mitigating social and ethical risk results in risk ownership without strong accountability, which creates uncertainty and frustration among practitioners [101]. Uncertainty about appropriate risk management practices and extant organizational challenges prevent the creation of enforcement mechanisms that might foster trust among potentially harmed persons and groups [79]. Safety engineering frameworks and perspectives can enable the creation of assessments and mitigations strategies of risk into organizational decision points [30] which can support the implementation of emerging policies and regulations [36, 61, 123].

FMEA and STPA-like frameworks also have two key limitations in their scope of analysis. Firstly, it is critical to recognize that safety engineering frameworks have a limited scope of analysis. STPA is designed to map out potential hazards of a system to a specific interaction between different elements of that system [66]. FMEA analyzes functions of a system for potential failures and deduces the likelihood of risk for that specific function failure [20]. They can be used to proactively think about identifying potential hazards or failures; however, these processes cannot adequately answer normative questions such as "is this a good technology for society?" For example, applying FMEA or STPA on an ML system to automatically classify an individual's gender could help mitigate ethical and social risks stemming from certain *design* decisions, but it will not be able to address the underlying ethical concerns and experienced harms related to *deploying* such a technology [55].

Furthermore, traditionally, FMEA and STPA are applied when there is a sufficient understanding of the deployment context (i.e., geographic location, typical user base, etc.). FMEA- and STPA-like analysts make assumptions about the use and development of a system, which could be wrong and not hold true in different contexts. Therefore, the outcome of an FMEA and STPA is not always valid across different contexts of application [20, 66]. According to our findings, when applying STPA- and FMEA-like frameworks, ML practitioners need to carefully outline the specific use case of analysis and recognize that their findings are valid for the given context. This is also true for other responsible ML tools [47, 73]. Integrating existing social and ethical risk evaluation and mitigation techniques with the systematic safety engineering frameworks could provide comprehensive methods for social and ethical risk management for specified scopes of our analysis.

5.2 Opportunities for improvement and challenges for adapting safety engineering frameworks

Our findings illustrated that FMEA and STPA processes are limited in identifying the social context of use and deployment. We see this as an opportunity for the theoretical development of these processes. STPA frameworks could be strengthened by deeper theoretical development around what stakeholders, values, and losses must be considered for a given system. For example, values from feminist HCI [7] such as equity, diversity, and social justice, could inform what losses need to be prioritized (e.g., the losses disproportionately experienced by communities traditionally at the margins of safety analyses) when conducting a social and ethical risk assessment. Centering feminist values in applying STPA would allow the analysts to conduct safety engineering analysis from an equity-oriented perspective, enabling ML design and implementation that better meets the needs of different users and communities. Similarly, critical, sociotechnical perspectives attentive to how ML systems are situated within social systems shaped by intersecting power dynamics [56, 88] could further inform which values should be prioritized when applying safety engineering for social and ethical analysis of a given system. Furthermore, proper identification of losses and hazards and an appropriate understanding of a control structure requires an in-depth understanding of how an ML system is used and integrated within a social system. Methodologies such as participatory design, value-sensitive design, and speculative design could guide appropriate stakeholder identification and engagement practices at the beginning of an STPA.

On the other hand, an FMEA process needs further theoretical guidance on how to think about failure modes that have social and ethical implications. Currently, failure modes are identified based on how a desired function is not met. However, social and ethical failures occur even when a product functions as intended [80]. Theoretical developments such as the concept of social failure modes presented by Millar [80] could inform FMEA processes that are more suitable for analyzing social and ethical risks [103]. Similarly, some of the feminist and critical epistemological perspectives and design methodologies mentioned above could help define equity-oriented social and ethical failures for ML systems and identify potential effects, causes and controls for marginalized stakeholders.

However, the desired impacts of safety frameworks are mediated by organizational culture. Even with a stronger theoretical grounding and echoing the findings by Martelaro et al. [77], implementation of such frameworks could suffer if organizational challenges, such as insufficient organizational incentives, homogeneous standpoints, and perspectives, and lack of resources, persist. The successful adoption of safety engineering in medical and automotive industries was accompanied by regulatory and organizational transformations [117]. Currently, the ML industry is observing some of these regulatory shifts with the newly introduced acts and standards [4, 36, 61]. For safety engineering frameworks to support emerging regulatory requirements, organizational shifts in safety culture and practices will need to follow. Necessary shifts may depend on existing organizational dynamics. As illustrated by Sloane et al.'s study of start-ups and their operationalization of AI ethics practices [115], approaches taken by small and medium-sized companies might differ significantly compared to those of larger technology companies as they will not be able to hire in-house experts for conducting safety engineering analysis. With the movement towards standardization and increased regulation of the ML industry, addressing organizational challenges will be necessary - regardless of the company size. However, practitioners, company leaders and government officials can learn from the growing research on operationalizing AI ethics. For example, Rakova et al. provide a mapping of "prevalent practices", "emerging practices" and "aspirational future", describing trends in responsible AI perspectives, and this mapping can act as a planning tool for leaders who want to identify their organization's position and goals [100]. Safety engineering frameworks are anticipatory practices that enable proactive identification and mitigation of social and ethical risks. According to findings from Rakova et al.'s study, the movement towards implementing such anticipatory practices needs to be accompanied by creating data-informed efforts to understand the impact, integrating responsible AI processes throughout all business processes, and aligning organizational values with the individual, team, and responsible AI incentives [100].

5.3 Future work and research challenges for the CHI community

This study is part of a multi-stage research project investigating the applicability of safety engineering frameworks for social and ethical risk management. We will examine how FMEA and STPA frameworks could be applied to ML applications by analyzing multiple case studies. The findings from this study will inform the case study application and shape the future research direction.

Industry practices for addressing the social and ethical risks of machine learning are rapidly emerging. Recognizing that safety frameworks are primarily designed to manage technological failures and hazards, we call on the CHI community to engage, study, critique, and improve the existing social and ethical risk management practices. Specifically, we have identified three research foci. First, there is a need for clarifying existing conceptualization of social and ethical risks (i.e. harms to a user, AI ethics principle transgression, and human rights violation), and we posit that developing taxonomies of harm, failure, and hazard - distinct concepts in safety engineering - for adverse social and ethical impacts of ML systems will provide valuable guidance in social and ethical risk management. Existing critical epistemological perspectives [7, 45, 82] on defining the harms of ML systems should inform such taxonomies. Second, theoretical framing and analytical processes for examining failures and hazards in safety frameworks such as STPA and FMEA can benefit from deeper engagement with critical and feminist epistemological perspectives for examining failures and hazards considering social and ethical implications. Lastly, many empirical studies of responsible ML practices have focused on fairness-related methods [74], transparency artifacts [81], and general AI ethics operationalization issues [101]. There is a lack of empirical studies on how practitioners are using, adapting, and developing social and ethical risk management techniques. More empirical studies are required to validate the applicability, usability, and capability for identifying and managing risks of emerging frameworks across different ML applications and organizational cultures, and use cases.

6 CONCLUSION

Challenges with organizational structure, resource 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 study existing practices and explore how tools from safety engineering could provide insights for creating more appropriate social and ethical risk management 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. In this work, we discussed the strength and limitations of these two processes and highlighted the need for further research.

ACKNOWLEDGMENTS

We thank our study participants for taking the time to share their experiences, expertise, and feedback. We also thank our anonymous reviewers for their deep engagement and valuable feedback on this paper. Finally, we thank our mentors and colleagues Roel Dobbe, Renelito Delos Santos, Sunny Consolvo, Michael Madaio, and Lauren Wilcox for their continued support and feedback during the writing process.

REFERENCES

[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 298–306.

[2] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 249–260. https://doi.org/10.1145/3442188.3445888

[3] Julia Angwin and Terry Parris, Jr. 2016. Facebook Lets Advertisers Exclude Users by Race. https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race. Accessed: 2022-9-3.

[4] IEEE Standards Association. 2022. IEEE portfolio of AIS technology and impact standards and standards projects. https://standards.ieee.org/initiatives/artificial-intelligence-systems/standards/. Accessed: 2022-9-10.

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

[6] Stephanie Ballard, Karen M. Chappell, and Kristen Kennedy. 2019. Judgment Call the Game: Using Value Sensitive Design and Design Fiction to Surface Ethical Concerns Related to Technology. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (San Diego, CA, USA) *(DIS '19)*. Association for Computing Machinery, New York, NY, USA, 421–433. https://doi.org/10.1145/3322276.3323697

[7] Shaowen Bardzell. 2010. Feminist HCI: taking stock and outlining an agenda for design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1301–1310.

[8] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, Duncan Wadsworth, and Hanna Wallach. 2021. Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. https://doi.org/10.48550/ARXIV.2103.06076

[9] Solon Barocas, Sophie Hood, and Malte Ziewitz. 2013. Governing algorithms: A provocation piece.

[10] Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code* (1 ed.). Polity.

[11] Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns (N Y)* 2, 2 (Feb. 2021), 100205.

[12] Su Lin Blodgett, Q Vera Liao, Alexandra Olteanu, Rada Mihalcea, Michael Muller, Morgan Klaus Scheuerman, Chenhao Tan, and Qian Yang. 2022. Responsible Language Technologies: Foreseeing and Mitigating Harms. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22, Article 152)*. Association for Computing Machinery, New York, NY, USA, 1–3.

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

[14] Virginia Braun and Victoria Clarke. 2020. One Size Fits All? What Counts as Quality Practice in (Reflexive) Thematic Analysis? *Qualitative Research in Psychology* 18, 3 (Aug. 2020), 1–25.

[15] Philip A E Brey. 2012. Anticipatory Ethics for Emerging Technologies. *Nanoethics* 6, 1 (April 2012), 1–13.

[16] Shea Brown, Jovana Davidovic, and Ali Hasan. 2021. The algorithm audit: Scoring the algorithms that score us. *Big Data & Society* 8, 1 (Jan. 2021), 2053951720983865.

[17] Nikhil Bugalia, Surjyatapa R Choudhury, Yu Maemura, and K E Seetharam. 2022. 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* 49, 8 (Oct. 2022), 2168–2194.

[18] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 2053951715622512. https://doi.org/10.1177/2053951715622512 arXiv:https://doi.org/10.1177/2053951715622512

[19] Dallas Card and Noah A Smith. 2020. On Consequentialism and Fairness. *Frontiers in Artificial Intelligence* 3 (2020), 34.

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

[21] Le Chen, Alan Mislove, and Christo Wilson. 2016. An Empirical Analysis of Algorithmic Pricing on Amazon Marketplace. In *Proceedings of the 25th International Conference on World Wide Web* (Montréal, Québec, Canada) *(WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1339–1349.

[22] Mayee Chen, Karan Goel, Nimit S Sohoni, Fait Poms, Kayvon Fatahalian, and Christopher Re. 2021. Mandoline: Model Evaluation under Distribution Shift. *Proceedings of Machine Learning Research* 139 (2021), 1617–1629.

[23] Alex Chohlas-Wood, Madison Coots, Henry Zhu, Emma Brunskill, and Sharad Goel. 2021. Learning to be Fair: A Consequentialist Approach to Equitable Decision-Making. https://doi.org/10.48550/ARXIV.2109.08792

[24] A Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. 2022. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 864–876.

[25] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning.

[26] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) *(KDD '17)*. Association for Computing Machinery, New York, NY, USA, 797–806.

[27] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1571–1583.

[28] Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, and Sravana Reddy. 2018. Assessing and Addressing Algorithmic Bias in Practice. *Interactions* 25, 6 (oct 2018), 58–63. https://doi.org/10.1145/3278156

[29] Catherine D'Ignazio and Lauren F Klein. 2021. *Data Feminism.* Tantor Audio.

[30] Roel Dobbe. 2022. System Safety and Artificial Intelligence. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1584.

[31] John Downer. 2011. "737-Cabriolet": The Limits of Knowledge and the Sociology of Inevitable Failure. *Amer. J. Sociology* 117, 3 (2011), 725–762.

[32] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21, Article 82)*. Association for Computing Machinery, New York, NY, USA, 1–19.

[33] Upol Ehsan, Ranjit Singh, Jacob Metcalf, and Mark Riedl. 2022. The Algorithmic Imprint. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1305–1317.

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

[35] Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. 2017. Towards Linguistically Generalizable NLP Systems: A Workshop and Shared Task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*. Association for Computational Linguistics, Copenhagen, Denmark, 1–10. https://doi.org/10.18653/v1/W17-5401

[36] European Commission. 2021. Proposal for a Regulation laying down harmonised rules on artificial intelligence. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence. Accessed: 2022-9-10.

[37] Luciano Floridi and Andrew Strait. 2020. Ethical Foresight Analysis: What it is and Why it is Needed? *Minds Mach.* 30, 1 (March 2020), 77–97.

[38] Jade S Franklin, Karan Bhanot, Mohamed Ghalwash, Kristin P Bennett, Jamie McCusker, and Deborah L McGuinness. 2022. An Ontology for Fairness Metrics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 265–275.

[39] Batya Friedman and David G Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination.* MIT Press.

[40] Batya Friedman, Nancy Levenson, Ben Shneiderman, Lucy Suchman, and Terry Winograd. 1994. Beyond accuracy, reliability, and efficiency: criteria for a good computer system. In *Conference Companion on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '94)*. Association for Computing Machinery, New York, NY, USA, 195–198.

[41] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Trans. Inf. Syst. Secur.* 14, 3 (July 1996), 330–347.

[42] Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual Fairness in Text Classification through Robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 219–226.

[43] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (Nov. 2021), 86–92.

[44] Government of Canada. 2022. Bill C-27 summary: Digital Charter Implementation Act, 2022. https://ised-isde.canada.ca/site/innovation-better-canada/en/canadas-digital-charter/bill-summary-digital-charter-implementation-act-2020. Accessed: 2022-9-10.

[45] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 501–512.

[46] Christina N. Harrington, Shamika Klassen, and Yolanda A. Rankin. 2022. "All That You Touch, You Change": Expanding the Canon of Speculative Design Towards Black Futuring. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 450, 10 pages. https://doi.org/10.1145/3491102.3502118

[47] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3290605.3300830

[48] Alexis Hope, Catherine D'Ignazio, Josephine Hoy, Rebecca Michelson, Jennifer Roberts, Kate Krontiris, and Ethan Zuckerman. 2019. Hackathons as Participatory Design: Iterating Feminist Utopias. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300291

[49] Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. 2022. Evaluation Gaps in Machine Learning Practice. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1859–1876.

[50] Takuto Ishimatsu, Nancy G Leveson, John P Thomas, Cody H Fleming, Masafumi Katahira, Yuko Miyamoto, Ryo Ujiie, Haruka Nakao, and Nobuyuki Hoshino. 2014. Hazard Analysis of Complex Spacecraft Using Systems-Theoretic Process Analysis. *J. Spacecr. Rockets* 51, 2 (March 2014), 509–522.

[51] Sheila Jasanoff. 2004. *States of Knowledge: The Co-Production of Science and Social Order.* Routledge.

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

[53] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (Sept. 2019), 389–399.

[54] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3819–3828.

[55] Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 88 (nov 2018), 22 pages. https://doi.org/10.1145/3274357

[56] Goda Klumbytė, Claude Draude, and Alex S Taylor. 2022. Critical Tools for Machine Learning: Working with Intersectional Critical Concepts in Machine Learning Systems Design. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1528–1541.

[57] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 5637–5664.

[58] P. M. Krafft, Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo. 2020. Defining AI in Policy versus Practice. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 72–78. https://doi.org/10.1145/3375627.3375835

[59] Joshua A Kroll. 2018. The fallacy of inscrutability. *Phil. Trans. R. Soc. A* 376, 2133 (2018), 14 pages.

[60] Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2017. Accountable Algorithms. *University of Pennsylvania Law Review* 165 (2017), 633–705. Issue 3.

[61] Information Technology Laboratory. 2021. AI Risk Management Framework | NIST. https://www.nist.gov/itl/ai-risk-management-framework. Accessed: 2022-9-10.

[62] Mark Latonero and Aaina Agarwal. 2021. *Human Rights Impact Assessments for AI: Learning from Facebook's Failure in Myanmar.* Technical Report. Carr Center for Human Rights Policy Harvard Kennedy School, Harvard University.

[63] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 182 (nov 2019), 26 pages. https://doi.org/10.1145/3359284

[64] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 181 (nov 2019), 35 pages. https://doi.org/10.1145/3359283

[65] Michelle Seng Ah Lee and Jat Singh. 2021. The Landscape and Gaps in Open Source Fairness Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21, Article 699)*. Association for Computing Machinery, New York, NY, USA, 1–13.

[66] Nancy Leveson and John Thomas. 2018. STPA_Handbook. https://psas.scripts.mit.edu/home/get_file.php?name=STPA_handbook.pdf.

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

[68] Jamy Li and Mark Chignell. 2022. FMEA-AI: AI fairness impact assessment using failure mode and effects analysis. *AI and Ethics* 2, 4 (Nov. 2022), 837–850.

[69] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards Debiasing Sentence Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5502–5515.

[70] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15.

[71] Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. 2021. Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. https://openreview.net/forum?id=mPducS1MsEK

[72] Huai-Wei Lo, James J H Liou, Jen-Jen Yang, Chun-Nen Huang, Yu-Hsuan Lu, and Alireza Amirteimoori. 2021. An Extended FMEA Model for Exploring the Potential Failure Modes: A Case Study of a Steam Turbine for a Nuclear Power Plant. *Complexity* 2021 (Jan. 2021).

[73] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. , 26 pages.

[74] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and

Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376445

[75] Monique Mann and Tobias Matzner. 2019. Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination. *Big Data & Society* 6, 2 (July 2019), 2053951719895805.

[76] Alessandro Mantelero. 2022. Human Rights Impact Assessment and AI. In *Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI*, Alessandro Mantelero (Ed.). T.M.C. Asser Press, The Hague, 45–91.

[77] Nikolas Martelaro, Carol J. Smith, and Tamara Zilovic. 2022. Exploring Opportunities in Usable Hazard Analysis Processes for AI Engineering. https://doi.org/10.48550/ARXIV.2203.15628

[78] Donald Martin, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S. Isaac. 2020. Participatory Problem Formulation for Fairer Machine Learning Through Community Based System Dynamics. https://doi.org/10.48550/ARXIV.2005.07572

[79] Jacob Metcalf, Emanuel Moss, and Danah Boyd. 2019. Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics. *Social Research: An International Quarterly* 86, 2 (2019), 449–476.

[80] Jason Millar. 2020. Social Failure Modes in Technology and the Ethics of AI: An Engineering Perspective. In *The Oxford Handbook of Ethics of AI*. Oxford University Press, Online. https://doi.org/10.1093/oxfordhb/9780190067397.013.28

[81] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229.

[82] Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philos. Technol.* 33, 4 (Dec. 2020), 659–684.

[83] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. 2020. Quantifying Model Complexity via Functional Decomposition for Better Post-hoc Interpretability. In *Machine Learning and Knowledge Discovery in Databases*. Springer International Publishing, 193–204.

[84] Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. 2021. Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. (June 2021).

[85] Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kästner. 2022. Collaboration Challenges in Building ML-Enabled Systems: Communication, Documentation, Engineering, and Process. In *Proceedings of the 44th International Conference on Software Engineering* (Pittsburgh, Pennsylvania) *(ICSE '22)*. Association for Computing Machinery, New York, NY, USA, 413–425. https://doi.org/10.1145/3510003.3510209

[86] Helen Ngo, Cooper Raterink, João G. M. Araújo, Ivan Zhang, Carol Chen, Adrien Morisot, and Nicholas Frosst. 2021. Mitigating harm in language models with conditional-likelihood filtration. https://doi.org/10.48550/ARXIV.2108.07790

[87] H Nissenbaum. 2001. How computer systems embody values. *Computer* 34, 3 (March 2001), 120–119.

[88] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, New York, NY.

[89] Brandie Nonnecke and Philip Dawson. 2022. *Human rights impact assessments for AI: analysis and recommendations*. Technical Report. accessnow.

[90] International Standards Organization. 2022. ISO - ISO/IEC JTC 1/SC 42 - Artificial intelligence. https://www.iso.org/committee/6794475/x/catalogue/. Accessed: 2022-9-10.

[91] Wanda J Orlikowski. 2000. Using Technology and Constituting Structures: A Practice Lens for Studying Technology in Organizations. *Organization Science* 11, 4 (2000), 404–428.

[92] Riccardo Patriarca, Mikela Chatzimichailidou, Nektarios Karanikas, and Giulio Di Gravio. 2022. The past and present of System-Theoretic Accident Model And Processes (STAMP) and its associated techniques: A scoping review. *Saf. Sci.* 146 (Feb. 2022), 105566.

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

[94] Charles Perrow. 1984. *Normal accidents: Living with high risk technologies*. Basic Books, New York.

[95] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21, Article 237)*. Association for Computing Machinery, New York, NY, USA, 1–52.

[96] Vinodkumar Prabhakaran, Margaret Mitchell, Timnit Gebru, and Iason Gabriel. 2022. A Human Rights-Based Approach to Responsible AI. https://doi.org/10.48550/ARXIV.2210.02667

[97] Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 959–972.

[98] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. 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* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 33–44.

[99] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. 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* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 33–44. https://doi.org/10.1145/3351095.3372873

[100] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where Responsible AI Meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 7 (apr 2021), 23 pages. https://doi.org/10.1145/3449081

[101] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (April 2021), 1–23.

[102] Lydia Reader, Pegah Nokhiz, Cathleen Power, Neal Patwari, Suresh Venkatasubramanian, and Sorelle Friedler. 2022. Models for understanding and quantifying feedback in societal systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1765–1775.

[103] Shalaleh Rismani and Ajung Moon. 2021. How do AI systems fail socially?: an engineering risk analysis approach. In *2021 IEEE International Symposium on Ethics in Engineering, Science and Technology (ETHICS)*. Online, 1–8. https://doi.org/10.1109/ETHICS53270.2021.9632769

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

[105] Negar Rostamzadeh, Ben Hutchinson, Christina Greer, and Vinodkumar Prabhakaran. 2021. Thinking Beyond Distributions in Testing Machine Learned Models. https://doi.org/10.48550/ARXIV.2112.03057

[106] Nataniel Ruiz, Adam Kortylewski, Weichao Qiu, Cihang Xie, Sarah Adel Bargal, Alan Yuille, and Stan Sclaroff. 2022. Simulated Adversarial Testing of Face Recognition Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4135–4145.

[107] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-Imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 315–328. https://doi.org/10.1145/3442188.3445896

[108] Andrew D Selbst. 2020. Negligence and AI's human users. *BUL Rev.* 100 (2020), 1315.

[109] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68.

[110] Renee Shelby, Shalaleh Rismani, Kathryn Henne, Ajung Moon, Negar Rostamzadeh, Paul Nicholas, N'mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2022. Identifying Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. (Oct. 2022). arXiv:2210.05791 [cs.HC]

[111] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021), 1–29.

[112] Katie Shilton. 2013. Values Levers: Building Ethics into Design. *Sci. Technol. Human Values* 38, 3 (May 2013), 374–397.

[113] Sung-Min Shin, Sang Hun Lee, Seung K I Shin, Inseok Jang, and Jinkyun Park. 2021. STPA-Based Hazard and Importance Analysis on NPP Safety I&C Systems Focusing on Human–System Interactions. *Reliab. Eng. Syst. Saf.* 213 (Sept. 2021), 107698.

[114] Kristin Sharon Shrader-Frechette. 1991. *Risk and rationality: Philosophical foundations for populist reforms*. University of California Press, California, U.S.

[115] Mona Sloane and Janina Zakrzewski. 2022. German AI Start-Ups and "AI Ethics": Using A Social Practice Lens for Assessing and Implementing Socio-Technical Innovation. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 935–947. https://doi.org/10.1145/3531146.3533156

[116] Katta Spiel, Os Keyes, Ashley Marie Walker, Michael A DeVito, Jeremy Birnholtz, Emeline Brulé, Ann Light, Pınar Barlas, Jean Hardy, Alex Ahmed, Jennifer A Rode, Jed R Brubaker, and Gopinaath Kannabiran. 2019. Queer(ing) HCI: Moving Forward in Theory and Practice. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19, Paper SIG11)*. Association for Computing Machinery, New York, NY, USA, 1–4.

[117] Alexander Styhre. 2018. *The Unfinished Business of Governance: Monitoring and Regulating Industries and Organizations*. Edward Elgar Publishing.

[118] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. 2007. Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation. In *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis (Eds.), Vol. 20. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2007/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper.pdf

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

[120] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization* (–, NY, USA) *(EAAMO '21, Article 17)*. Association for Computing Machinery, New York, NY, USA, 1–9.

[121] Takuto Ishimatsu, Nancy Leveson, John Thomas, Masa Katahira, Yuko Miyamoto, Haruka Nakao. 2010. Proceedings of the 4th IAASS Conference. In *Making Safety Matter* (Huntsville, Alabama). International Association for the Advancement of Space Safety (IAASS).

[122] Rachael Tatman. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain, 53–59.

[123] Treasury Board of Canada Secretariat. 2019. Directive on Automated Decision-Making. https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592. Accessed: 2022-9-15.

[124] Steven Umbrello. 2019. Beneficial Artificial Intelligence Coordination by Means of a Value Sensitive Design Approach. *Big Data and Cognitive Computing* 3, 1 (Jan. 2019), 5.

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

[126] Thiemo Wambsganss, Anne Höch, Naim Zierau, and Matthias Söllner. 2021. Ethical Design of Conversational Agents: Towards Principles for a Value-Sensitive Design. In *Innovation Through Information Systems*, Frederik Ahlemann, Reinhard Schütte, and Stefan Stieglitz (Eds.). Springer International Publishing, Cham, 539–557.

[127] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 214–229.

[128] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, and Stephen Cave. 2019. The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 195–200.

[129] Richmond Y. Wong, Michael A. Madaio, and Nick Merrill. 2022. Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics. https://doi.org/10.48550/ARXIV.2202.08792

[130] Guoyang Zeng, Fanchao Qi, Qianrui Zhou, Tingji Zhang, Zixian Ma, Bairu Hou, Yuan Zang, Zhiyuan Liu, and Maosong Sun. 2021. OpenAttack: An Open-source Textual Adversarial Attack Toolkit. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 363–371.

[131] Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11, 3 (2020), 1–41.

[132] Douglas Zytko, Pamela J. Wisniewski, Shion Guha, Eric P. S. Baumer, and Min Kyung Lee. 2022. Participatory Design of AI Systems: Opportunities and Challenges Across Diverse Users, Relationships, and Application Domains. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 154, 4 pages. https://doi.org/10.1145/3491101.3516506