From Silos to Systems: Process-Oriented Hazard Analysis for AI Systems

Shalaleh Rismani*
shalaleh.rismani@mail.mcgill.ca
McGill University
Montreal, Quebec, Canada

Roel Dobbe
R.I.J.Dobbe@tudelft.nl
Delft University of Technology
Delft, The Netherlands

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

Abstract

To effectively address potential harms from Artificial Intelligence (AI) systems, it is essential to identify and mitigate system-level hazards. Current analysis approaches focus on individual components of an AI system, like training data or models, in isolation, overlooking hazards from component interactions or how they are situated within a company's development process. To this end, we draw from the established field of system safety, which considers safety as an emergent property of the entire system, not just its components. In this work, we translate System Theoretic Process Analysis (STPA) – a recognized system safety framework for analyzing AI development and operation processes. We focus on systems that rely on machine learning algorithms and conduct STPA on three case studies involving linear regression, reinforcement learning, and transformer-based generative models. Our analysis explored how STPA's control and system-theoretic perspectives apply to AI systems and whether unique AI traits—such as model opacity, capability uncertainty, and output complexity—necessitate significant modifications to the framework. We find that the key concepts and steps of conducting an STPA apply to AI systems but require targeted adaptations to address AI-specific challenges—such as model opacity, autonomy, and output complexity—that arise to differing degrees across AI paradigms such as linear models, reinforcement learning, and generative models. We present the Process-oriented Hazard Analysis for AI Systems (PHASE) as a guideline that adapts STPA concepts for AI. Applying and interpreting STPA using the PHASE guidelines enables four key affordances for analysts responsible for managing AI system harms: 1) detection of hazards at the systems level, including those from accumulation of disparate issues; 2) explicit acknowledgment of social factors contributing to experiences of algorithmic harms; 3) creation of traceable accountability chains between harms and those who can mitigate the harm; and 4) ongoing monitoring and mitigation of new hazards.

CCS Concepts

• **Social and professional topics → Computing / technology policy**; • **Human-centered computing → Empirical studies in HCI**.

Keywords

System safety, Sociotechnical AI Safety, Hazard Analysis, Responsible AI Development, System Theory, Control Theory, System-level Analysis, System Theoretic Process Analysis

*Corresponding author.

1 Introduction

When it comes to safety and risk management of Artificial Intelligence (AI)-based technologies, there is an emerging recognition that hazards - potential sources of harm - from such technologies need to be viewed and analyzed at the (sociotechnical) system level [26, 80]. Regulatory efforts such as the European Union AI Act explicitly distinguish between the terms AI model and AI system [27]. In parallel, major AI developers are releasing system cards in addition to model and data cards, recognizing that the performance of many recently released Application Programming Interfaces (APIs) is the result of the amalgamation of multiple models and software components[4, 49]. Additionally, the role of the algorithmic supply chain, which includes the actors involved in developing, integrating, and using an AI system, is increasingly recognized as central to understanding how harms emerge from AI systems [5, 23]. Finally, there is growing recognition of the importance of defense-in-depth approaches to mitigating risks from AI systems, which advocate for multiple mitigation mechanisms applied across the AI development lifecycle [11]. Despite growing views of AI-based technologies as sociotechnical systems, there has been surprisingly minimal empirical effort to develop and examine frameworks that facilitate hazard analysis at the system level for AI systems. In this work, we draw on a well-established hazard analysis framework from the field of system safety and examine how it could be applied to analyze the processes involved in developing and integrating an AI system.

Hazard analysis has a long-standing history in the field of safety engineering, and a track record of being an essential component of developing and deploying complex safety-critical technologies such as nuclear power plants and passenger planes [20, 36, 40]. Yet, the AI community has been slow at embracing the fundamental know-how and practices well established in safety engineering [20, 36, 40]. We specifically draw on the field of system safety, one of the most recent and significant developments in safety engineering[39]. System safety emerged from the observation that safe systems cannot be achieved by merely analyzing technical components in isolation. Instead, system safety emphasizes the necessity of considering the broader social and technological context. System safety frameworks provide the means of analyzing systems as a whole rather than conducting multiple, separate analyses of their composing subsystems (e.g., machine learning (ML) model, development team, user) [43, 51].

In this work we examine to what degree a well-established hazard analysis framework can be readily applied to assess AI systems and what it affords to those conducting the assessment. In particular, we focus on STPA, a framework well-known for its utility in analyzing complex technical systems [41]. STPA offers two powerful affordances essential for examining hazards at the system level:

1) It provides the means to map out and analyze complex social and technical systems, leveraging concepts from system theory and control theory. 2) It centers on the potential loss of stakeholder values, which facilitates harm analysis beyond safety-critical harms (i.e., loss of life, damage to property, and environmental harm). However, many of the current guidelines for conducting STPA primarily focus on traditionally safety-critical systems such as nuclear power plants.There are currently only a handful of examples of its application to AI systems and algorithmic harms.[33, 60]. When applying STPA to AI systems, we build on existing work while focusing on characteristics and challenges unique to conducting hazard analysis for AI, such as system opacity, capability uncertainty, and output complexity, as well as the diverse range of potential harms that could result from their adoption. Crucially, these characteristics are not uniform across AI systems: a linear regression model, a reinforcement learning agent, and a transformer-based generative model differ substantially in their opacity, autonomy, and output complexity, and therefore place different demands on each stage of the analysis. We treat this heterogeneity as central rather than incidental to our study, and our case studies are chosen to expose it. Specifically, we ask the following research questions:

- RQ1: How can the steps of the STPA framework be translated to analyze potential hazards from AI systems?
- RQ2: What does STPA afford an analyst in identifying, assessing and mitigating potential harms?

We conducted three case studies spanning three AI systems (linear regression, reinforcement learning, transformer-based text-to-image models) across two domains of application (creative practice and medicine). Following the existing handbook by Leveson and Thomas, we conducted STPA on the three case studies [41] and documented our interpretive and reflexive experience in the process. We find that STPA-based analysis of AI provides the means for detecting potential sources of harm along both the development and operation of AI systems. We present a guideline, Process-oriented Hazard Analysis for AI Systems (PHASE), documenting our translation of STPA for AI, considering the aforementioned unique characteristics. The guideline introduces key considerations that need to be accounted for when applying STPA for AI systems. This work contributes to the growing body of responsible and safe AI practices discourse by: 1) producing PHASE guidelines that adapt STPA to the specific challenges of AI systems and enable system-level hazard analysis for AI; 2) providing three diverse case studies to illustrate the outcome of the analyses, and 3) presenting four affordances as shown by empirical investigation of these case studies.

2 Background

In this section, we situate our work in today's AI safety discourse, Responsible AI (RAI), and system safety literature. We then briefly describe the key tenets and the steps of the STPA framework in Section 2.3.

Note that in this work, we use the Organisation for Economic Co-operation and Development (OECD) definition of an AI system: *"a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments."* [54]. An AI system consists of multiple components, including the input dataset, model, and resulting output.

2.1 Examining Potential Sources of Algorithmic Harm

Alongside the increased public awareness of how AI systems can introduce societal disruptions and harm to individuals, RAI scholars have been actively studying the various types of algorithmic harms and means of harm mitigation. Shelby et al. [66], for example, present at least 170 articles related to sociotechnical harms of algorithmic systems – defined as "adverse lived experiences of individuals and communities." The rapid advancements and adoption of large-scale, multi-modal, and general-purpose ML models in the last five years have further intensified the discussions about AI-related harms [10, 15, 69]. The emerging epistemic community of AI safety is also drawing more attention to concerns about AI-driven existential and catastrophic risks to humanity [3, 18, 61]. To date, the types of harm that can be expected from current and near-future AI systems remain a subject of heated debate.

Alongside these discussions are multitudes of theoretical frameworks, tools, and processes proposed to analyze and mitigate potential sources of such harms. Qualitative and normative frameworks are commonly used to identify societal, ethical, and legal factors, helping analysts reason about how AI systems may produce unfair, exclusionary, or otherwise harmful outcomes in specific contexts of use [47, 58, 79]. Complementing these approaches, technical tools, often implemented as software libraries or code bases, are employed to examine potential sources of harm within datasets and models, such as data quality issues, bias, and performance disparities[7, 53]. Safety and security engineering tools – inspired by processes such as red teaming [30] or Failure Mode and Effects Analysis (FMEA) – aim to anticipate how AI models may behave under adversarial, unexpected, or edge-case conditions, thereby identifying model-level failure modes that could lead to downstream harms [31, 44, 56, 59]; Additionally, participatory and community-based research methods have been adopted to uncover harms that emerge through real-world interactions between AI systems, individuals, and communities, particularly those that may be overlooked by purely technical evaluations [6, 16, 45, 82].

Existing methods are designed to be effective in analyzing hazards either by focusing on a single component of an AI system (e.g., representation in a dataset) or examining interaction instances (e.g., how much a decision maker relies on an AI assistant) in silos without examining their connections [75, 84]. As a result, these approaches do not support the kind of integrated, end-to-end reasoning about hazards that is central to system safety analysis.

2.2 Safety Engineering and System Safety for AI

Safety engineering frameworks have been used to establish the safety of technical systems as early as the 1940s with the use of FMEA and Fault Tree Analysis (FTA) in military technology and the Apollo mission [19, 20, 71]. These early frameworks follow an analytically reductive approach where component failures and human errors can be analyzed as a chain of events that lead to an incident. With the increase in complexity of technological systems

and specifically the introduction of software, the analytically reductive approach was deemed ineffective in capturing new types of hazards. In the early 2000s, Nancy Leveson proposed an alternative accident model called Systems-Theoretic Accident Model and Processes (STAMP) [39]. Consistent with system theory [76], STAMP frames safety as an emergent property of a system, resulting from the interactions and control structures among people, technical artifacts/agents, and the environment [40]. This paradigm of safety motivated the safety engineering community to look beyond the analysis of units and components, and consider analyzing the interaction between software programs, physical parts, and people needed for establishing a safe system [36, 42, 43]. Subsequently, Leveson and Thomas developed System Theoretic Process Analysis (STPA) as a *hazard analysis* method based on STAMP [41], which was promptly adapted across a number of safety-critical application domains [8, 22, 55].

Identifying and mitigating harms from AI systems requires a systematic approach that explores and maps out the distribution of existing product development practices beyond an ML model and across individuals, automated systems, collectives, and institutional mechanisms [26, 33, 37, 38]. Narrowly examining and designing for safety at the ML model level is important but not sufficient for ensuring the safety of products that only partially rely on these models [17, 35, 47]. The system theoretic perspective recognizes that AI models do not exist in silos in practical product environments. They are often integrated in a complex network of API calls, other AI models, a range of datasets and model outputs, and various user interfaces [24, 51, 83]. In addition, an AI system design is influenced by the decisions of a wide range of actors, including the data curation team, model development team, responsible AI analysts, users, company executives, product managers, and regulators [28, 48, 67].

Traditionally, STPA and system safety have been applied to technologies where there is consensus on what is safe or unsafe (i.e., an airplane that could crash is viewed as unsafe). In contrast, as discussed above, harms arising from AI systems are diverse, contested, and often context-dependent, with no single agreed-upon notion of acceptable or unacceptable outcomes. Accordingly, we treat potential harms from AI as emergent properties of the system that arise through interactions among technical components, human actors, and institutional processes across development and operation. Furthermore, AI systems have unique differentiating characteristics compared to other complex cyber-physical systems: they can have varying levels of outcome complexity, and the system capabilities [85] can be probabilistic and distributed. Importantly, these properties are not shared uniformly by all AI: they vary markedly across AI paradigms—for instance, between an interpretable linear model, an adaptive reinforcement learning agent, and an opaque generative model—so that the same analytical step can demand different considerations depending on the type of AI under study. These properties complicate hazard identification and prediction, and must be explicitly considered when applying system-theoretic hazard analysis. These considerations motivate the need to adapt STPA for AI systems in a way that supports system-level reasoning about harms across the AI development and deployment processes.

2.3 System Theoretic Process Analysis (STPA)

Below, we provide a brief description of the key steps for conducting STPA outlined in the STPA handbook [41]. These steps are meant to be performed iteratively and cyclically across the lifecycle of technology development.

2.3.1 Identify purpose of the analysis. The STPA process starts by identifying losses, system boundaries, and hazards. By definition, a ***loss*** occurs when a stakeholder's value is disregarded or violated. *Hazards* are a *system state or conditions that could be a potential source of harm*. Hazards are defined per ***system boundary***, which outlines parts of the system where designers, developers, engineers, or automated agents could exercise some level of control. To provide a simplistic example, a construction company repairing a roof of a house may point to safety considerations (loss of human life) to be important for the project; identify the possibility of workers falling from the steep and bare edges of the rooftop as a fall hazard; and is only able to install safety equipment such as fences within the boundaries of the roof being renovated and not that of the neighbouring house (system boundary).

2.3.2 Create control diagrams. The next step involves creating control diagrams to visually capture the model of the system within its system boundary. It represents the elements and interactions between different elements within the system in the form of a feedback loop called *control feedback loop*. The two key elements of a control feedback loop are the ***controller*** – an algorithm, a person, or an organization that can make decisions or perform actions to control something – and the ***controlled process*** – the component, process, or another controller being controlled by the controller. A controller is represented with a *control algorithm* – a description of how the controller makes decisions – and *process model* – the controller's internal/mental model the environment used to make decisions.

The relationship between controllers and controlled processes are represented with two types of interactions: ***control action***, which is a constraint that the controller puts on the controlled process, and ***feedback***, which is the response/feedback the controlled process provides to the controller. Typically, the various controller, controlled processes, and their interactions are arranged visually in a hierarchical manner, such that those with the highest level of control/authority appear on top.

2.3.3 Identify unsafe control actions. Once a control diagram is drawn, the next step is to identify actions that could be taken by the controller that lead to any of the hazards identified earlier. These are called ***Unsafe Control Action (UCA)***. An UCA is fully described using five parts: the controller (*source*) that provides the problematic action (*control action*), the *type* of action/inaction that could lead to hazards (e.g., is it the execution, lack of execution, untimely execution of the control action that can lead to the hazard?), specific system condition or state context (*context*), and *link to hazard*.

2.3.4 Identify loss scenarios. Finally, the circumstances that could lead to a UCA – ***loss scenarios*** – are identified. To identify loss scenarios, the analysts need to consider all potential issues with the elements and interactions, such as missing feedback loops, incorrect mental or process models, and improper control algorithms. At the

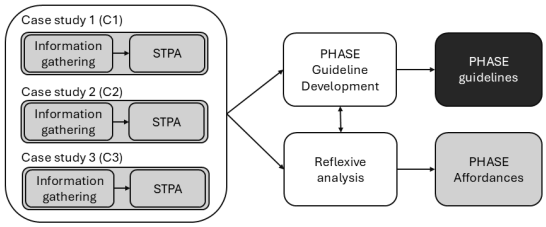


Figure 1: Methodology overview: STPA was conducted for each case study after we gathered the necessary information. The information gathering process was different for each case study. Iterative reflexive analysis led to the generation of the PHASE guideline and identification of affordances.

end of this process, the analysts develop a list of safety requirements based on the loss scenarios to provide the necessary constraints and to minimize or eliminate cases of inadequate control.

3 Methodology

Figure 1 presents an overview of our research method. As Njie and Asimirian outline, case studies as a research method enable in-depth and first-hand investigation of events and processes [52]. Therefore, we used case studies to investigate how STPA needs to be interpreted when it is used to analyze AI systems. Through an iterative, reflexive analysis of the case study applications, we developed PHASE, a guideline for conducting STPA on AI systems, and identified the key affordances it provides. We further refined the guidelines' language based on feedback from a pilot study, described in the supplementary material. Here, we describe the case studies, the details of applying the STPA framework, and the process of developing reflexively analyzing PHASE.

3.1 Case studies

To illustrate the applicability of STPA we chose the following real-world case studies based on the diverse range of ML model they represent, application context, and the nature of interaction between the AI system and a human user. Specifically, the three cases span distinct ML paradigms—supervised linear modelling, reinforcement learning, and transformer-based generative modelling—and distinct interaction modes—human-in-the-loop, human-out-of-the-loop, and human-as-tool-user—so that we could examine whether and how different types of AI place different demands on the analysis rather than treating AI as a single, homogeneous technology. An additional selection criterion was our ability to work closely with the teams developing these systems, enabling access to design decisions and development processes that are essential for conducting such analysis.

- **Case 1 (C1) - Early warning system:** A linear regression algorithm is being designed in the Netherlands to predict the likelihood of late-onset sepsis (LOS) in pre-term infants before a clinical suspicion [74]. It is to be used as a proactive warning system alerting clinicians to start early diagnostic testing for LOS and, ultimately, provide timely treatment. As a human-in-the-loop system, a clinician makes final, clinical decisions.
- **Case 2 (C2) - Insulin Injection:** An Reinforcement Learning (RL)-based, automated and personalized insulin injection system is being developed in Canada for patients with diabetes [9]. Some diabetic patients use an artificial pancreas that facilitates automatic insulin injection. The RL algorithm is being developed for use in the artificial pancreas to allow tailored insulin doses for each individual in a human-out-of-the-loop manner.
- **Case 3 (C3) - Storyboarding :** Some ML artists have adapted Text-to-image (T2I) models as part of their creative practice, such as storyboarding for video creation [29, 44, 63, 86]. In this case study, we focus on T2I demo platforms generally. While we do not focus on a specific platform, many such platforms are produced by companies in the United States, and rely on some combination of transformer-based or diffusion architecture T2I generative models.

3.2 Applying the STPA framework

To apply STPA, we started by gathering information about the AI system, its stakeholders, and how the system was developed and integrated in the given context. To do so, we connected with the relevant stakeholders of each of the projects and conducted secondary research. Considering time limitations, data availability, and access to subject-matter experts, the information-gathering process was customized for each case. For C1, the lead author held a preliminary meeting with the ML developer and data scientist to understand the project's objectives and used a peer-reviewed academic publication on the project as the main reference for the analysis. For C2, the lead author held monthly meetings for eight months with two RL researchers – one of whom developed RL algorithms for insulin injection. For C3, the lead author conducted eight interviews and one workshop (15 participants) involving experts along the T2I model development process and ML artists who use these models in their creative practice.

After information gathering, the lead author followed the STPA steps outlined earlier (Section 2.3) and applied them for each case study. While the procedural steps are those of standard STPA, the interpretive work at each step—deciding what counts as a loss, where to draw system boundaries, which components act as controllers, and what makes a control action unsafe—is where AI-specific considerations entered and where the case studies differed from one another; Section 4 details these adaptations. The lead author then cross-checked the hazards identified in the analysis against relevant literature and subject-matter experts to sanity-check the results from the STPA. The co-authors of the paper also reviewed each of the analyses and provided feedback on the application of STPA.

3.3 Developing the PHASE guideline and examining its affordances

Through reflexive analysis of the case studies and the process of applying STPA, the lead author developed Process-oriented Hazard Analysis for AI Systems (PHASE), a step-by-step guideline for conducting system-level hazard analysis of algorithmic harms in

AI systems. This work was informed by iterative discussions and feedback with the co-authors, and the joint reflexive examination of both the case study findings and the analysis process enabled the identification of the key affordances that PHASE provides for analysts. The guideline was further refined based on feedback from a pilot study, described in the supplementary material.

4 PHASE Guideline: Interpreting STPA for AI System Case Studies

Next, we present observations from applying STPA across the three case studies and highlight features unique to PHASE. Before turning to the individual steps, we briefly summarize how PHASE adapts each of the four STPA steps for AI systems.

In Step 1 (identify the purpose of the analysis), PHASE broadens how losses are conceived by drawing on value-sensitive design and by surfacing the value tensions among stakeholders, so that sociotechnical losses are considered alongside safety-critical and performance-related ones; it also draws system boundaries around the development and use processes of ML-based systems rather than around the model itself, since developers exercise control over data and design choices but not over what a model learns. In Step 2 (create a control diagram), PHASE accommodates multiple and dynamic control structures, because the way ML-based systems are used is fluid—a component can act as a controlled process in one context and as a controller in another—and it introduces components specific to AI systems, such as datasets, models, safety filters, and learning agents, as controllers or controlled processes. In Step 3 (identify unsafe control actions), PHASE summarizes the unsafe control actions characteristic of AI systems—functional failures, poor design decisions or misuse, and coordination or communication breakdowns; what distinguishes these from unsafe control actions in traditional STPA is that their unsafeness frequently depends on sociotechnical context rather than on the technical timing or sequencing of a control action. Finally, in Step 4 (identify loss scenarios), PHASE captures the loss scenarios also found in conventional STPA—missing feedback loops, incorrect mental or process models, and technical malfunctions—but adds a class unique to AI: loss scenarios arising from emerging and evolving system capabilities, where changes such as retraining, updates, or expanded use outpace existing safeguards, a source of hazard that traditional STPA, applied to systems with fixed capabilities, does not typically consider.

The PHASE guideline, along with a detailed summary of the case studies, is provided in the supplemental material.

4.1 Step 1: Identify purpose of analysis

Applying the first step of STPA - *identify purpose of analysis* - to AI systems required a substantive reinterpretation. In traditional applications of STPA, this step typically centers on defining safety-critical losses and establishing system boundaries around controllable technical components. Our case studies showed that, for AI systems, this framing is too narrow: both the notion of loss and the selection of system boundaries are deeply shaped by sociotechnical context, value tensions, and the processes through which AI systems are developed and operated. In response, PHASE adapts the first step of STPA by broadening the interpretation of loss to incorporate insights from value-sensitive design and by emphasizing process-oriented system boundaries aligned with the algorithmic supply chain. The sections below describe our observations from the case studies and how they informed these adaptations.

Identifying losses. The losses identified for each case study go beyond the traditional definition of safety (e.g., loss of life, damage to property or environment). As shown in Table 1, the identified losses consider a broader set of emergent harms from AI systems. Our reflexive analysis reveals that there are three different categories of losses pertinent in an STPA of an ML system: 1) traditional safety-critical losses, 2) performance-related losses, 3) sociotechnical losses. *Safety-critical losses* represent losses in the traditional definition of safety, such as loss of human life, property damage, and the environment (*C1.L1*). *Performance-related losses*, on the other hand, emerge when the AI system does not meet performance expectations (*C1.L2*) or when the individuals, groups, or organizations experience loss of reputation or economic loss (*C3.L6*, *C2.L4*). Lastly, *sociotechnical losses* cover a broad range of sociotechnical harms that individuals and groups can experience that may not be safety-critical or performance-related (*C1.L4*, *C2.L2*, *C3.L2*). Considering the growing body of research in algorithmic harms and our findings from the case studies, we found it much more appropriate to use existing taxonomies [13, 66, 81], industry-specific regulatory requirements [1], and community-based and participatory research methods [25, 78] for loss identification of AI systems.

Defining system boundaries. Our selection of system boundaries was guided by the structure of the algorithmic supply chain and by the information available about each case, as well as by where stakeholders indicated that they could meaningfully exercise control. Rather than centering boundaries on individual models, we drew them around development and use processes, which enabled a more thorough analysis of how hazards emerge across the lifecycle of an AI system. The number of system boundaries is not prescribed by PHASE; analysts define as many as are needed to cover the loci of control relevant to the identified losses. In our cases, three boundaries per system happened to capture the accessible points of control, but this is illustrative rather than a requirement, and prior work has shown that the ML development lifecycle can be decomposed into many more stages. The boundaries also differ across cases—for example, integration or a distinct operational boundary appears in some systems but not others—because the systems, their paradigms, and the points at which developers, users, or automated agents reported having decision-making authority differ. Importantly, drawing boundaries around sub-processes does not discard the whole-system view: because boundaries are connected through shared control actions and feedback, interactions across them remain visible rather than being lost, and a hazard localized in one boundary can propagate through these links to losses defined in another.

Across all cases, we identified system boundaries within data collection and processing (e.g., *C1.SB1*, *C3.SB1*), where developers made decisions about data sources, feature inclusion, and labeling; and within model development, evaluation, and integration (e.g., *C1.SB2*, *C2.SB1*, *C3.SB2*), where choices about learning algorithms,

Table 1: The losses identified across the three case studies. The three types of losses (safety-critical, performance-related, and sociotechnical harms) are identified with #.

Case 1 (C1) -Early alert system	L1. Loss of life or injury to the preterm infants #safety-critical L2. Loss of time (in terms of efficiency) #performance-related L3. Loss of reputation/credibility for the physician in charge of a case #performance-related L4. Loss of patient privacy (regulation about medical data) #sociotechnicalharm
Case 2 (C2) - Insulin injection	L1. Loss of life or injury to the diabetic patient #safetycritical L2. Loss of quality of life (i.e., comfort and autonomy) #sociotechnicalharm L3. Loss of efficiency #performance-related L4. Loss of money #performance-related
Case 3 (C3) - Story boarding	L1. Loss of creativity #sociotechnicalharm L2. Loss of diversity #sociotechnicalharm L3. Loss of accessibility #sociotechnicalharm L4. Loss of efficiency #performance-related L5. Loss of quality #performance-related L6. Loss of reputation #performance-related

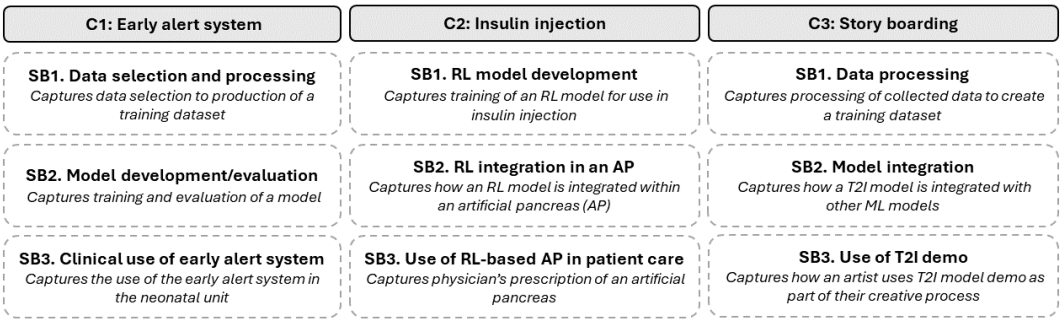


Figure 2: Three *system boundaries* (SB) defined for each of the case studies (C). In practice, analysts are free to define SBs following Section 2.3.1. The items under the Case title describe the scope of the SB we defined for the Case Study.

architectures, metrics, and outputs were exercised. However, despite this apparent control, stakeholders consistently noted limited visibility into—and control over—the learned latent representations of ML models, which constrained the usefulness of drawing system boundaries narrowly around the model itself.

The contrast between the reinforcement learning (RL) case and systems built around opaque, black-box generative models further illustrates this point. In the RL case, we identified an operational system boundary (*C2.SB2*) where control shifted from human operators to the algorithmic agent, which autonomously adjusted the amount of insulin delivered to a diabetic patient. This clear locus of control during operation made it possible to reason explicitly about hazards arising from the system's autonomous decision-making. In contrast, for opaque generative systems such as text-to-image models, stakeholders reported limited ability to interpret or directly control model behavior at inference time. For a general-purpose system such as a T2I platform, the space of possible uses is open-ended, which makes boundary definition especially consequential: a boundary drawn loosely around "use of the demo" is too vague to determine whether a given hazard is controllable within it. We therefore drew the downstream boundary narrowly around the generation and dissemination of a specific output (e.g., a storyboard used for video creation, *C3.SB3*), where a hazard such as the dissemination of false or harmful content becomes controllable—through the prompts entered, the safety filters applied, and the decision to publish—rather than around the platform's use in the abstract. Consequently, system boundaries for these systems were more meaningfully drawn around upstream development processes and downstream use contexts—such as content generation and dissemination (e.g., *C3.SB3*) - where human actors retained control over how outputs were produced, interpreted, and shared.

Outlining Hazards. Hazard identification largely followed established STPA practice and was driven by the system boundaries and loss definitions established in the first step of the analysis. The hazards identified across the case studies aligned with the three loss categories described above and often spanned multiple losses. For example, the hazard "The system creates a model or output that does not meet performance requirements" for *C1.SB2* could degrade an early-warning system and, in extreme cases, contribute to loss of life (*C1.L1*). Other hazards reflected well-documented sociotechnical concerns in the Fairness, Accountability, Transparency and Ethics (FATE) and RAI literature, such as lack of interpretability (*C1.SB2*; *L1*, *L3*) [12] and the dissemination of false or harmful

content (*C3.SB3*; *L5*, *L6*) [14]. Here, the input and output safety filters referenced throughout the T2I case are classifiers that screen prompts and generated images to block disallowed content; their performance refers to how accurately they do so, that is, their false-positive rate (blocking benign content) and false-negative rate (allowing harmful content through), and a poorly designed filter is one whose threshold admits too many harmful outputs or suppresses too much legitimate use.

4.2 Step 2: Create a control diagram

A key challenge in applying STPA to AI systems is that some components do not occupy fixed roles as controllers or controlled processes. For example, an AI assistant may function as a controlled process in one context (e.g., supporting human decision-making) and effectively act as a controller in another when users are organizationally required to follow its outputs. In PHASE, we explicitly account for this role fluidity by encouraging analysts to model control relationships as context-dependent rather than static. Capturing this role fluidity at the control-diagram stage is essential, as hazards depend not only on system behavior but also on how control shifts between humans and AI systems in practice.

Using the core concepts of STPA, we mapped controllers, controlled processes, control actions, feedback, inputs, outputs, and control hierarchies within each selected system boundary. Across the case studies, we were able to translate these concepts to AI systems and interpret how they manifest across both development and operational contexts. We constructed a control diagram for each system boundary, yielding a total of 9 diagrams; the full set is provided in the supplementary material, and we show one boundary per case study in Figure 3 to illustrate the range of control structures across the three paradigms. To read such a diagram, one traces each control action downward—the constraints a controller imposes on the process below it—and each feedback signal upward—the information returned to the controller—so that, together, they reveal where control is adequate and where it may break down. Figure 3 presents one representative control diagram from each case study. The elements represented as controllers or controlled processes ranged from individuals and teams to technical artifacts. In many cases, controllers were individuals or groups responsible for decision-making during development or use. For example, in *C1.SB1*, controllers included the data collection team, data processing team, and hospital database manager, while the collected dataset functioned as a controlled process. Similarly, in *C3.SB2*, technical artifacts such as input and output safety filters and text-to-image models were modeled as controlled processes. However, individuals and teams were not always the primary controllers. In some contexts, control was exercised by technical artifacts themselves. For instance, in *C2.SB3*, the reinforcement learning agent and the insulin pump acted as controllers by autonomously determining insulin delivery. In other cases, control relationships depended on organizational and use-context constraints. For example, in *C1.SB3*, when use of an early alert system was mandated for physicians, the AI system constrained clinical decision-making, effectively positioning it as a controller despite its nominal role as a decision-support tool. These examples illustrate why treating AI systems as having fixed roles obscures important sources of risk.

The top-to-bottom representation of the control hierarchy (with higher levels indicating greater control) was generally effective for visualizing these relationships, but the ordering of the hierarchy was often context-dependent and subject to change based on system design choices, organizational policies, or patterns of use. Across the case studies, control actions ranged from development-time decisions (e.g., dataset processing and model design choices), to use-time actions (e.g., prompts entered by ML artists interacting with a text-to-image system), to technical constraints (e.g., threshold settings for safety filters). Feedback took the form of evaluation metrics, sensor outputs, user feedback, or other forms of human communication. Inputs and outputs captured both technical artifacts (e.g., training datasets) and coordination among individuals and teams, such as collaboration between physicians and developers to ensure clinically appropriate model behavior (*C1*).

4.3 Step 3: Identify unsafe control actions

Overall, we found that Step 3 of STPA—identifying unsafe control actions (UCAs)—could be applied to AI systems with minimal modification. The core STPA logic of examining control actions in relation to system-level hazards was transferred directly; however, careful interpretation of the four conditions under which a control action becomes unsafe was essential in AI contexts. In particular, whether a control action is provided, not provided, provided too early or too late, or applied too long or too briefly often depended on sociotechnical context rather than purely technical timing or sequencing. From our analysis, we identified three main trends in UCAs applicable to AI systems.

Functional UCAs The most common type involved *functional UCAs*, where inadequate control resulted from a malfunction or dysfunction of the controller. For example, in *C2.SB2*, the insulin pump in the artificial pancreas failed to release the required amount of insulin at a given time. Similarly, in *C1.SB3*, an incorrect alarm from the early alert system could negatively influence a physician's decision-making. These UCAs were primarily associated with performance-related and safety-critical losses.

UCAs from poor design decision-making or misuse A second type of UCAs stemmed from *poor design decisions or misuse* by experts, system developers, or users. For instance, in *C3.SB2*, model developers could select an inappropriate safety filter threshold, enabling harmful prompts or image outputs. In *C1.SB1*, the data processing team could select incorrect clinical variables as features, leading to misleading model behavior. These UCAs highlight how design-time decisions can render otherwise correct control actions unsafe in context.

UCAs from communication/coordination The final type of UCAs involved *inadequate organizational coordination or miscommunication*. For example, model developers might receive incomplete or incorrect safety requirements, or lack sufficient time to address them before deployment deadlines (*C3.SB2*). In these cases, unsafe control actions emerged not from technical failure, but from misaligned processes and constraints across teams.

4.4 Step 4: Identify loss scenarios

The original STPA process was highly effective in the process of understanding loss scenarios for AI systems. We brainstormed a

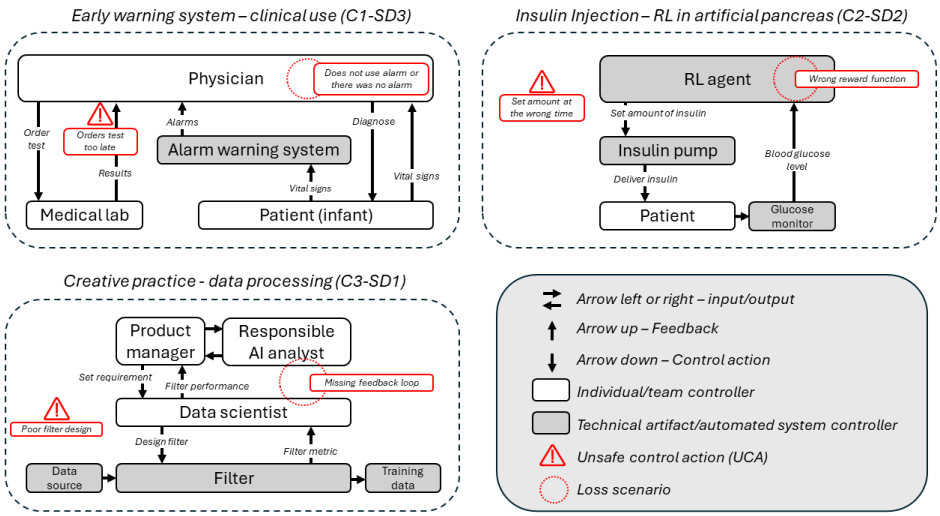


Figure 3: This diagram illustrates the simplified control structure. Each of the System Diagrams (SD) corresponds to one System Boundary (SB). We have included only three of the system boundaries in this diagram, one from each case study.

list of loss scenarios by thinking through all the elements of the control diagram. A diverse range of loss scenarios were identified, which included organizational/institutional causal scenarios (i.e., poor understanding or communication), interaction issues (e.g., incorrect mental model) and technical challenges (e.g., missing metrics/evaluations of a system). Across the case studies, however, we observed a recurring class of loss scenarios unique to AI systems: losses arising from emerging and evolving system capabilities. These scenarios did not necessarily stem from immediate component failure, but from changes in model behavior over time—such as updates, retraining, or expanded use—that outpaced existing assumptions, safeguards, or organizational processes. In addition, we observed two other prevalent forms of loss scenarios. A common institutional causal scenario across our case studies involved missing feedback loops between different teams that are in charge of various parts of the development process. For instance, often model development teams use pre-existing datasets that have been collected and processed by another group and at a different time frame than when the model (e.g., T2I model) is being developed. In such cases, there is a missing feedback loop between the developers of the model and creators of the dataset (e.g., *C3-SB2*). Another type of causal scenario is when the user has an incorrect mental model of the AI system, meaning that they do not understand how to use it correctly or how it works. This scenario is characteristic of AI systems deployed to non-expert users: unlike traditional safety-critical software operated by trained specialists, AI tools are increasingly used by people without an accurate model of the system's competence or limits, and PHASE captures this through a mismatch between the controller's process model and the actual behavior of the controlled process. For instance, for *C1.SB3*, if the physician does not understand the reason for the alarm, they could make the wrong lab order or diagnosis. The third type of loss scenario is *technical challenges* which occurs when a component (i.e., a pump, ML model, safety classifier) is functioning erroneously. This could lead to inadequate control when making decisions about a patient's health. The type of loss scenarios we identified is not exhaustive. However, they provide an understanding of the type of causal loss scenarios for inadequate control in AI systems.

5 Affordances of System Theoretic Perspective for algorithmic harm from AI systems

Based on our reflexive analysis, we identified four affordances observed across our case studies that build on STPA when applied to AI systems using our PHASE guideline. We stress that the value here is not system-level analysis in itself—which STPA already provides—but the specific adaptations PHASE makes for AI: an expanded loss taxonomy that includes sociotechnical harms, the treatment of AI components as context-dependent controllers or controlled processes rather than fixed-role artifacts, and loss scenarios that account for evolving model capabilities. The affordances described below follow from these adaptations, and it is these adaptations, rather than the general system-centric stance, that we claim as PHASE's contribution.

5.1 Affordance 1: Analyze AI at the systems level

Since STPA is designed to analyze a process or an artifact at the systems level, PHASE inherits this affordance. Prior work suggests that algorithmic harms from AI arise from system-wide issues [64, 68, 73]. For instance, representational harms are system-level issues that "occur when algorithmic systems reinforce the subordination of social groups along the lines of identity, such as disability, gender, race and ethnicity, religion, and sexuality" [34, 66]. They arise from a combination of ad-hoc factors such as inaccurate ML model performance for minority groups, misuse of the system in a given application, or exclusionary social norms. Instead of narrowly focusing on "fixing" an unbalanced training dataset, system-level analysis prompts us to examine how this dataset will interact with

other parts of the system, such dataset analysts, other ML models, or the user interface [7, 77].

STPA is a systematic method that establishes possible causal chains between undesirable outcomes of the system and inadequate control. As such, the *losses* and *harms* identified in the first step of PHASE (Section 4.1) define the negative impact of the system an analyst aims to prevent or mitigate. Subsequently, UCAs (Section 4.3) explicitly defines all possible cases of inadequate control in the chosen system boundary that lead to a hazard. Loss scenarios (Section 4.4) finally establish the causal link connecting the undesirable outcomes to specific unsafe states of the system. For example, we identified the hazard *"H: Clinician misdiagnoses the patient."* to stem from *inadequate control* over the proper use of the early alarm system and timing for ordering a blood test, which can lead to *L: loss of life*. In this case, the control action of *"clinical use of the early alert system"* could be unsafe and in several loss scenarios, such as when the clinician does not understand how to use the early alert system. Notably, this hazard would not surface from a component-level analysis of the model or dataset alone: a fairness or accuracy audit of the alert model in isolation could report acceptable performance, yet the loss still arises from the interaction between an imperfect alert and a clinician who overtrusts or misreads it. It is only by analyzing the dataset, model, and user together that this system-level hazard becomes visible.

5.2 Affordance 2: Accounting for social factors and sociotechnical harms

Formalizing losses at the beginning of STPA mobilizes a discussion and instantiates the need for consensus amongst stakeholders about harms that must be prioritized. Amongst practitioners and scholars, there is a divergence of perspectives on the types of harm that need to be assessed and mitigated for AI systems [32, 38, 62]. Currently, company policies and ad-hoc conversations amongst limited, invited individuals shape practitioners' choices on what harms need to be minimized or fully prevented [2]. However, following PHASE structures the analysis so that a broad set of sociotechnical harms is considered at the outset. It cannot compel an analyst to be exhaustive; rather, by making the enumeration of losses the explicit first step, it renders the omission of such harms visible and open to challenge rather than leaving it implicit. Furthermore, in defining the system boundary, PHASE connects the losses to a network of individuals, teams, technical artifacts, automated systems, and institutional mechanisms which form a system boundary along the AI development cycle as opposed to just examining a loss by focusing on one of these elements (i.e. blaming a developer or assuming ML models cannot be explained). Lastly, the control diagram and its components in PHASE provide the conceptual flexibility to map social and technical elements within a system boundary. For example, a controller could be a human (e.g., an artist) who has a specific mental model of how to use a T2I platform. It could also be an RL agent with a map of the environment (*process model*) and a reward model (*control algorithm*) that guides the suggested dosage of insulin (*control actions*). This conceptual flexibility allows for the integration and prioritization of sociotechnical factors in AI harms analysis.

5.3 Affordance 3: Identify accountability chain by establishing traceability

Traceability in harm reduction is a pressing concern in the AI community. Recent studies indicate that responsible AI practitioners are frustrated by the lack of systematic methods that establish a connection between a specific set of harms and the means to mitigate the harm [46, 57, 67]. This problem is exacerbated by the fact that many systems are developed by a multitude of individuals and groups across institutions that hardly interact directly with one another [51, 83]. That is, many unsafe system states can accumulate between individual decision-makers, the technical system, and the development process without the situation being detected or managed [37, 83]. As described earlier, the STPA framework allows us to connect harms to inadequate control within a system. Furthermore, it allows an analyst to make explicit who has control over what, and make the harms and hazardous states of a system traceable for everyone. Specifically, the concept of hierarchical control structure as interpreted in PHASE engages the analyst to map who or what has control in relation to other actors or artifacts by mapping their interaction using control action and feedback. For instance, by drawing the control structure, we needed to examine whether the clinician had control over the early alert system meaning that they could choose to use the warning upon reception (i.e., feedback) or that they needed to use the warning from the early alert system (i.e., the control action) by thinking through how the early warning system is integrated into clinical practice. Understanding the control hierarchy fosters an understanding of how power, control, and responsibility are distributed in a system, which ultimately results in a more managed chain of accountability for responsible AI deployment. Concretely, this is what lets an analyst localize a harm to its source. If the early alert system contributes to a harmful misdiagnosis, tracing the loss scenario back through the control structure distinguishes whether the inadequate control originated in data collection (for example, a context-specific perturbation missing from the training data, surfaced as a missing feedback loop between the data and model teams), in the model (a training deficiency, surfaced as an inadequate control algorithm), or in use (a clinician who does not understand the alert, surfaced as an incorrect process model). Because a single loss scenario can implicate more than one control action or feedback link at once, the same structure also makes explicit when a harm is combinatorial—arising from the joint effect of, say, a data gap and a user's misunderstanding—rather than attributable to any one component in isolation.

5.4 Affordance 4: Monitor and address emergent hazards over time

STPA framework enables an analyst to monitor and account for changes in individuals, teams, technical artifacts, and institutional mechanisms over time. By emergent hazards over time, we mean hazards that were absent or not analyzable at initial design but arise as the model, its use, or its organizational context changes—for example, as a reinforcement learning agent adapts its policy during operation or as a generative model is retrained or updated after deployment. As C2 illustrates, this includes the changes that stem from the adaptivity of AI systems such as RL. These adaptations can be reflected in various components of a control diagram, such

as control actions, feedback loops, and elements of a controller (e.g., decision-making protocol, process/mental model of the world). For instance, the change in the output of an AI model could be captured in a control action as constraints/information a clinician or an artist receives. Furthermore, changes in an AI model's latent representation could be captured in the process model of the controller or controlled process (depending on whether a model is acting as a controller or is a controlled process). Changes in how individuals, teams, and organizations perceive and operate, such as changes in how an AI model is evaluated, could be captured as part of the feedback loop between the model and the developer team. Such concrete means of accounting for changes in the system allow analysts to monitor the dynamics systematically and iteratively, prompting them to examine emergent hazards as AI systems change in how they are integrated or interact with their sociotechnical context. This is a critical affordance given the rapid pace of AI development and increased interest in adaptive and agentic AI systems [65, 80].

6 Discussion

In this paper, we employed a case study approach to translate STPA for the analysis of AI systems. We illustrated that STPA framework can not only be translated to analyze AI systems and identify potential sources of algorithmic harm, but that the PHASE implementation of the analysis enables four affordances critically missing in existing AI assessment frameworks. STPA framework allows analysts to map out sociotechnical system dynamics, and instantiate accountability by examining harms at the system level and tracing them to control methods for harm reduction. Moreover, the framework allows for accounting of changes in individuals, collectives, AI models and artifacts, and institutional processes that could happen over time. Here, we discuss the broader implications of adopting a system-theoretic perspective for AI governance and building a responsible AI culture in organizations. We also elaborate on limitations and avenues for future work.

6.1 Implications for policy and AI governance

From self-governance mechanisms in corporations to national regulations on AI, there has been a growing effort to practice due diligence and establish accountability mechanisms across the AI industry [1, 72]. As illustrated in our findings (Section 5.3), establishing traceability between harms and causal scenarios can be a powerful tool toward this end. It can enable an analyst to define a division of accountability where specific roles/groups are held responsible for maintaining adequate/safe control of certain subsystems [40]. Although our system boundaries did not explicitly focus on internal or external governance mechanisms, some of the controls added by internal governance mechanisms were observed in our case studies. For example, for *C3-SB1* as captured in Figure 3, the responsible AI analyst provides input on safety requirements to product managers about datasets, which are then passed onto the data scientists. These safety requirements would typically be developed based on existing regulations or internal company policies [2]. Given that STPA has already been used in other domains to examine and improve governance mechanisms; we expect PHASE to be able to afford the same function for AI systems [39].

For example, such a process could point out the missing feedback loop between the data scientist and the responsible AI analyst in *C3-SB1*, which hints at the lack of quality control of the training dataset and the potential need for a data quality approval mechanism. In addition to improving internal accountability protocols for a company, external actors such as standard organizations could leverage these affordances to establish best practices for conducting hazard analysis, and mobilize policymakers to refer to these standards in the regulation for certain AI products [70].

6.2 Building a safety culture: systems lens for responsible and safe AI

Given the focus of our three case studies in the product development context, we also observed how existing culture within the organization and management practices can contribute to sociotechnical harm. For example, in *C3-SB1*, the loss scenarios for the UCA of poor safety filter design could stem from a responsible AI analyst not having enough time to review the dataset and provide safety requirements to the product manager before the product needs to be launched. The presence of toxic or profit-driven culture within the tech industry and its relationship to the quality of products deployed in the market is not new [50]. PHASE helps account for this social factor and its relationship to a specific set of harms. Our hope is that such explicit acknowledgment of harms that can stem from social contexts can help organizations prioritize the implementation of a safety culture. Even though STPA could be a tool for building a safety culture, adoption of STPA demands a company culture that values safety and is willing to allocate resources to it. Lack of incentives and financial/human resources limits the applicability and effectiveness of many of the existing responsible AI/AI ethics frameworks [46, 57]. The persistence of these challenges for STPA would hinder adoption even with the development of guidelines that translate and illustrate the framework for RAI practice.

6.3 Limitations and future work

Recent discussion about the possible risks of developing agentic systems suggests that agentic AI will be inherently dangerous considering that lack of any meaningful methods to control these systems and comprehensively understand their process model of the world [21, 61, 65]. While none of our case studies involve an agentic AI, it was clear that the PHASE effectively maps where and what humans and technical systems/processes *have or do not have* control. We believe that this mapping can offer valuable insights in contextually examining increasingly agentic systems, especially in investigating areas of inadequate or lack of control to examine systems that are inherently unsafe or uncontrollable.

There are many known limitations of STPA that also transfer to PHASE. First, STPA frames interaction between processes or agents as control and feedback. Such framing limits how nuances in human relationships can be captured, such as collaborative decision-making and social influence between agents. Furthermore, similar to any analytical process, analysts must make normative choices, assumptions, and simplifications in conducting PHASE. This means that the views, values, and perspectives of the analysts affect the final analysis. Relatedly, the conceptual flexibility that lets analysts

map social and technical elements within a boundary is also a limitation: analysts may equally choose not to include such elements, so the same flexibility that is a strength for diverse use cases can weaken an analysis in less careful hands. To improve robustness against this analyst-dependence, PHASE can be conducted by multiple or deliberately diverse analysts and the resulting hazard and loss-scenario lists cross-validated against one another. Therefore, it is imperative to have well-established industry standards on what a quality PHASE should look like. We provide a running example in the supplementary material as a starting point. A further limitation concerns generality: our findings rest on three case studies, and although these were chosen to span distinct ML paradigms and interaction modes, they cannot represent the full AI landscape, so broader application is needed to test how far these adaptations generalize. Lastly, PHASE is a comprehensive hazard analysis that requires adequate expertise and time; to lower this barrier, the supplementary guideline provides step-by-step instructions, templates, and a worked example, though effective use still assumes basic familiarity with the system being analyzed and with control-based reasoning.

7 Conclusion

In this work, we interpret and translate STPA, a well-established system safety hazard analysis framework. We find that system safety models and methods offer a complementary perspective to existing responsible AI efforts. It helps us to defining safety as an emergent property of the system. Through three real-life case studies, we conclude that the concepts and steps of STPA apply readily to the analysis of AI systems. We developed the PHASE guideline that translates these concepts for the analysis of AI systems. Pilot testing the guidelines with students in a graduate-level responsible AI course helped us validate and improve the final PHASE guideline provided in the supplementary material. We find that PHASE affords an analyst the ability to adopt a sociotechnical lens, establish traceability, and account for changes in a system over time. We invite the AI community to leverage these affordances to address growing safety concerns with rapidly changing AI systems, establish traceable AI governance practices, and improve safety culture in the technology industry for the safety of all.

8 Generative AI Usage Statement

The first author used ChatGPT (version 5.2) sparingly to obtain suggestions related to grammar, writing style, and conciseness. All suggestions were reviewed critically, and none were incorporated verbatim. Grammarly was additionally used for grammar and spelling edits. The remaining authors primarily provided feedback through verbal discussions and written comments on the manuscript and did not use generative AI tools for providing feedback.

9 Ethical Considerations Statement

In this study, we collected information at multiple stages of our methodology as described in Section 3 of the paper. In the information gathering stage for Case Study 1 and Case Study 2, we primarily relied on publicly available resources and consulted subject-matter experts as needed to clarify our understanding of the projects. As such, and following our institution's research ethics board policy, research ethics approval for the two cases was not needed. Case Study 3, on the other hand, involved a qualitative study where we recruited and conducted semi-structured interviews with relevant stakeholders along the development process. As this pertains to designed interventions with collection from human participants, we followed and received the necessary institutional approvals prior to conducting the study. All the participants (15 for the workshop and 8 for interviews) for Case Study 3 provided informed consent.

References

[1] 2023. Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI. https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai. Accessed: 2024-1-22.

[2] 2023. Safety standards. https://openai.com/safety-standards. Accessed: 2024-1-23.

[3] Shazeda Ahmed, Klaudia Jaźwińska, Archana Ahlawat, Amy Winecoff, and Mona Wang. 2023. Building the Epistemic Community of AI Safety. (Nov. 2023).

[4] Open AI. 2025. GPT-5 System Card. https://openai.com/index/gpt-5-system-card/. Accessed: 2025-8-10.

[5] Blair Attard-Frost and David Gray Widder. 2025. The ethics of AI value chains. *Big Data Soc.* 12, 2 (June 2025).

[6] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. (Dec. 2022). arXiv:2212.08073 [cs.CL]

[7] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W Duncan Wadsworth, and Hanna Wallach. 2021. Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. 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, 368–378.

[8] Esra Bas. 2020. STPA methodology in a socio-technical system of monitoring and tracking diabetes mellitus. *Appl. Ergon.* 89 (Nov. 2020), 103190.

[9] Sumana Basu, Marc-André Legault, Adriana Romero-Soriano, and Doina Precup. 2023. On the Challenges of using Reinforcement Learning in Precision Drug Dosing: Delay and Prolongedness of Action Effects. (Jan. 2023). arXiv:2301.00512 [cs.LG]

[10] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. 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, 610–623.

[11] Yoshua Bengio, Sören Mindermann, Daniele Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Peter Fox, Ben Garfinkel, David Goldfarb, Hoda Heidari, Andrew Ho, Saurabh Kapoor, Leila Khalatbari, Shayne Longpre, Christopher D. Manning, Vasilios Mavroudis, Mantas Mazeika, Julian Michael, Jessica Newman, Kelvin Y. Ng, Chinasa T. Okolo, Inioluwa Deborah Raji, Girish Sastry, Elizabeth Seger, Theodoros Skeadas, Thomas South, Emma Strubell, Florian Tramèr, Luis Velasco, Nathan Wheeler, Daron Acemoglu, Olufunke Adekanmbi, David Dalrymple, Thomas G. Dietterich, Pascale Fung, Pierre-Olivier Gourinchas, Fredrik Heintz, Geoffrey Hinton, Nicholas Jennings, Andreas Krause, Susan Leavy, Percy Liang, Teresa Ludermir, Vidushi Marda, Helen Margetts, John McDermid, John Munga, Arvind Narayanan, Alondra Nelson, Christopher Neppel, Alice Oh, Sarvapali Ramchurn, Stuart Russell, Marietje Schaake, Bernhard Schölkopf, Dawn Song, Andres Soto, Lukas Tiedrich, Gaël Varoquaux, Edward W. Felten, Andrew Yao, Yong-Qiang Zhang, Olugbenga Ajala, Fahad Albalawi, Mayada Alserkal, Gabriel Avrin, Christoph Busch, Antonio C. P. L. F. de Carvalho, Benjamin Fox, Amanpreet S. Gill, Ahmet H. Hatip, Jussi Heikkilä, Colin Johnson, Gavin Jolly, Zohar Katzir, Shakir M. Khan, Hiroaki Kitano, Antonio Krüger, Kwan Min Lee, Daniel V. Ligot, José R. López Portillo, Oleksii Molchanovskyi, Andrea Monti, Nothando Mwamanzi, Myriam Nemer, Nuria Oliver, Rodrigo Pezoa Rivera, Balaraman Ravindran, Hasan Riza, Celestin Rugege, Cathal Seoighe, Huma Sheikh, Jack Sheehan, David Wong, and Yi Zeng. 2025. *International AI Safety Report.* Technical Report DSIT 2025/001. UK Department for Science, Innovation and Technology. https:

//www.gov.uk/government/publications/international-ai-safety-report-2025 Accessed: 2025-01-13.

[12] Bhatt Surabhi, Cohon Adam, Rose Jenna, Majerczyk Natalia, Cozzi Brian, Crenshaw Drew, and Myers Griffin. 2021. Interpretable Machine Learning Models for Clinical Decision-Making in a High-Need, Value-Based Primary Care Setting. *NEJM Catalyst* 2, 4 (2021).

[13] Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. 2023. Typology of Risks of Generative Text-to-Image Models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (<conf-loc>, <city>Montréal</city>, <state>QC</state>, <country>Canada</country>, </conf-loc>) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 396–410. doi:10.1145/3600211.3604722

[14] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul Republic of Korea). ACM, New York, NY, USA.

[15] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. (Oct. 2021). arXiv:2110.01963 [cs.CY]

[16] 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.

[17] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, J B Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. 2020. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. (April 2020). arXiv:2004.07213 [cs.CY]

[18] Benjamin S Bucknall and Shiri Dori-Hacohen. 2022. Current and Near-Term AI as a Potential Existential Risk Factor. 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, 119–129.

[19] John Calder. 1899. *The Prevention Of Factory Accidents: Being An Account Of Manufacturing Industry And Accident.*

[20] Carl Carlson. 2012. *Effective FMEAs: Achieving Safe, Reliable, and Economical Products and Processes Using Failure Mode and Effects Analysis.* Wiley & Sons, Limited, John.

[21] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 651–666.

[22] Shufeng Chen, Siddartha Khastgir, Islam Babaev, and Paul Jennings. 2020. Identifying Accident Causes of Driver-Vehicle Interactions Using System Theoretic Process Analysis (STPA). In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. ieeexplore.ieee.org, 3247–3253.

[23] Jennifer Cobbe, Michael Veale, and Jatinder Singh. 2023. Understanding accountability in algorithmic supply chains. In *2023 ACM Conference on Fairness Accountability and Transparency (FAccT '23)*. ACM, New York, NY, USA, 1186–1197.

[24] Kate Crawford. 2022. *Atlas of AI Power, Politics, and the Planetary Costs of Artificial Intelligence.* Yale University Press.

[25] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. dl.acm.org, 1–19.

[26] 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.

[27] European Union. 2024. Article 3: Definitions, including definition of "AI system". https://artificialintelligenceact.eu/article/3/. Accessed: 2025-01-13. According to Article 3(1), an "AI system" means "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.".

[28] Clàudia Figueras, Harko Verhagen, and Teresa Cerratto Pargman. 2022. Exploring tensions in Responsible AI in practice. An interview study on AI practices in and for Swedish public organizations. *Scandinavian Journal of Information Systems* 34, 2 (2022), 6.

[29] Giorgio Franceschelli and Mirco Musolesi. 2021. Creativity and Machine Learning: A Survey. (April 2021). arXiv:2104.02726 [cs.LG]

[30] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. (Aug. 2022). arXiv:2209.07858 [cs.CL]

[31] Bálint Gyevnár and Atoosa Kasirzadeh. 2025. AI safety for everyone. *Nat. Mach. Intell.* 7, 4 (April 2025), 531–542.

[32] Alex Hanna and Emily M Bender. 2023. AI Causes Real Harm. Let's Focus on That over the End-of-Humanity Hype. *Scientific American* (2023).

[33] Edgar W Jatho, Logan O Mailloux, Eugene D Williams, Patrick McClure, and Joshua A Kroll. 2023. Concrete Safety for ML Problems: System Safety for ML Development and Assessment. (Feb. 2023). arXiv:2302.02972 [cs.LG]

[34] Nadia Karizat, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi. 2021. Algorithmic Folk Theories and Identity: How TikTok Users Co-Produce Knowledge of Identity and Engage in Algorithmic Resistance. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021), 1–44.

[35] Heidy Khlaaf. 2023. *Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems.* Technical Report. Trail of Bits.

[36] J C Knight. 2002. Safety critical systems: challenges and directions. In *Proceedings of the 24th International Conference on Software Engineering. ICSE 2002.* IEEE, 547–550.

[37] Joshua A Kroll. 2021. Outlining Traceability: A Principle for Operationalizing Accountability in Computing Systems. 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, 758–771.

[38] Seth Lazar and Alondra Nelson. 2023. AI safety on whose terms? *Science* 381, 6654 (July 2023), 138.

[39] Nancy Leveson. 2004. A new accident model for engineering safer systems. *Saf. Sci.* 42, 4 (April 2004), 237–270.

[40] Nancy Leveson. 2011. *Engineering a Safer World: Systems Thinking Applied to Safety.* MIT Press.

[41] Nancy Leveson and John Thomas. 2018. *STPA Handbook.* Technical Report.

[42] Nancy G Leveson. 1986. Software safety: why, what, and how. *ACM Comput. Surv.* 18, 2 (June 1986), 125–163.

[43] Nancy G Leveson. 2017. Rasmussen's legacy: A paradigm change in engineering for safety. *Appl. Ergon.* 59, Pt B (March 2017), 581–591.

[44] 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.

[45] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. 2023. Blaming Humans and Machines: What Shapes People's Reactions to Algorithmic Harm. (April 2023). arXiv:2304.02176 [cs.CY]

[46] 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. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1 (April 2022), 1–26.

[47] Alessandro Mantelero and Maria Samantha Esposito. 2021. An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems. *Computer Law & Security Review* 41 (July 2021), 105561.

[48] Matti Mäntymäki, Matti Minkkinen, Teemu Birkstedt, and Mika Viljanen. 2022. Putting AI Ethics into Practice: The Hourglass Model of Organizational AI Governance. (June 2022). arXiv:2206.00335 [cs.AI]

[49] Meta. [n. d.]. System Cards, a new resource for understanding how AI systems work. https://ai.meta.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/. Accessed: 2024-8-8.

[50] Emanuel Moss and Jacob Metcalf. 2020. *Ethics Owners A New Model of Organizational Responsibility in Data-Driven Technology Companies.* Technical Report. Data and Society.

[51] Ehsan Nabavi and Chris Browne. 2023. Leverage zones in Responsible AI: towards a systems thinking conceptualization. *Humanities and Social Sciences Communications* 10, 1 (March 2023), 1–9.

[52] Baboucarr Njie and Soaib Asimiran. 2014. Case study as a choice in qualitative methodology. *IOSR J. Res. Method Educ. (IOSRJRME)* 4, 3 (2014), 35–40.

[53] ODSC-Open Data Science. 2022. 15 Open Source Responsible AI Toolkits and Projects to Use Today. https://odsc.medium.com/15-open-source-responsible-

ai-toolkits-and-projects-to-use-today-fbc1c2ea2815. Accessed: 2023-12-27.

[54] OECD. 2024. OECD AI Principles overview. https://oecd.ai/en/ai-principles. Accessed: 2024-1-21.

[55] Jongwoo Park, Myeongkyu Lee, Jooyoung Maeng, Changnam Ahn, and Ji Hyun Yang. 2022. A Study for STPA-based Identification of Safety Requirements from the Perspective of Drivers in Take-Over Request. In *2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS)*. ieeexplore.ieee.org, 1–3.

[56] 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.

[57] 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.

[58] Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. 2018. Algorithmic Impact Assessments: A Practical Framework for Public Agency. *AI Now* (2018).

[59] S Rismani and A Moon. 2021. How do AI systems fail socially?: an engineering risk analysis approach. *on Ethics in Engineering, Science and …* (2021).

[60] Shalaleh Rismani, Renee Shelby, Andrew Smart, Edgar Jatho, Joshua 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* (Hamburg, Germany) *(CHI '23, Article 2)*. Association for Computing Machinery, New York, NY, USA, 1–18.

[61] Stuart Russell. 2022. Artificial Intelligence and the Problem of Control. In *Perspectives on Digital Humanism*, Hannes Werthner, Erich Prem, Edward A Lee, and Carlo Ghezzi (Eds.). Springer International Publishing, Cham, 19–24.

[62] Henrik Skaug Sætra and John Danaher. 2023. Resolving the battle of short- vs. long-term AI risks. *AI and Ethics* (Sept. 2023).

[63] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. (May 2022). arXiv:2205.11487 [cs.CV]

[64] Andrew D Selbst. 2021. An Institutional View of Algorithmic Assessments. https://jolt.law.harvard.edu/assets/articlePDFs/v35/Selbst-An-Institutional-View-of-Algorithmic-Impact-Assessments.pdf. Accessed: 2023-2-10.

[65] Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler Cullen O'keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, Katarina Slama, Lama Ahmad, Paul Mcmillan, Alex Beutel, and David G Robinson. 2023. Practices for governing agentic AI systems. https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf. Accessed: 2024-1-4.

[66] Renee Shelby, Shalaleh Rismani, Kathryn Henne, Ajung Moon, Negar Rostamzadeh, Paul Nicholas, N'mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montr\'{e}al, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 723–741.

[67] 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.

[68] Nathalie A Smuha. 2021. Beyond the individual: governing AI's societal harm. *Internet Pol. Rev.* 10, 3 (Sept. 2021).

[69] Irene Solaiman. 2023. The Gradient of Generative AI Release: Methods and Considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 111–122.

[70] Dajiang Suo, Sarra Yako, Mathew Boesch, and Kyle Post. 2017. *Integrating STPA into ISO 26262 process for requirement development.* Technical Report. SAE Technical Paper.

[71] Paul Swuste. 2022. *From Safety to Safety Science: The Evolution of Thinking and Practice.* Routledge.

[72] The White House. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/. Accessed: 2023-12-27.

[73] Kristen Thomasen. 2023. SAFETY IN ARTIFICIAL INTELLIGENCE & ROBOTICS GOVERNANCE IN CANADA. *Bar Review* 101, 1 (May 2023).

[74] Merel A M van den Berg, O'jay O A G Medina, Ingmar I P Loohuis, Michiel M van der Flier, Jeroen J Dudink, Manon M J N L Benders, Richard R T Bartels, and Daniel D C Vijlbrief. 2023. Development and clinical impact assessment of a machine-learning model for early prediction of late-onset sepsis. *Comput. Biol. Med.* 163 (June 2023), 107156.

[75] Q Vera Liao and Ziang Xiao. 2023. Rethinking Model Evaluation as Narrowing the Socio-Technical Gap. (June 2023). arXiv:2306.03100 [cs.HC]

[76] Ludwig Von Bertalanffy. 1972. The History and Status of General Systems Theory. *Acad. Manage. J.* 15, 4 (1972), 407–426.

[77] Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. 2022. Measuring Representational Harms in Image Captioning. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 324–335.

[78] Qiaosi Wang, Ida Camacho, Shan Jing, and Ashok K Goel. 2022. Understanding the Design Space of AI-Mediated Social Interaction in Online Learning: Challenges and Opportunities. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1 (April 2022), 1–26.

[79] Elizabeth Anne Watkins, Emanuel Moss, Jacob Metcalf, Ranjit Singh, and Madeleine Clare Elish. 2021. Governing Algorithmic Systems with Impact Assessments: Six Observations. 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, 1010–1022.

[80] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. (Oct. 2023). arXiv:2310.11986 [cs.AI]

[81] 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.

[82] Kimi Wenzel, Nitya Devireddy, Cam Davidson, and Geoff Kaufman. 2023. Can Voice Assistants Be Microaggressors? Cross-Race Psychological Responses to Failures of Automatic Speech Recognition. (Feb. 2023). arXiv:2302.12326 [cs.HC]

[83] David Gray Widder and Dawn Nafus. 2022. Dislocated Accountabilities in the AI Supply Chain: Modularity and Developers' Notions of Responsibility. (Sept. 2022). arXiv:2209.09780 [cs.CY]

[84] Richmond Y Wong, Michael A Madaio, and Nick Merrill. 2023. Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1 (April 2023), 1–27.

[85] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. 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, 1–13.

[86] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. (June 2022). arXiv:2206.10789 [cs.CV]