AI Risk Profiles: A Standards Proposal for Pre-deployment AI Risk Disclosures

Eli Sherman, Ian W. Eisenberg

Credo AI
{esherman, ian}@credo.ai

Abstract

As AI systems' sophistication and proliferation have increased, awareness of the risks has grown proportionally. The AI industry is increasingly emphasizing the need for transparency, with proposals ranging from standardizing use of technical disclosures, like model cards, to regulatory licensing regimes. Since the AI value chain is complicated, with actors bringing varied expertise, perspectives, and values, it is crucial that consumers of transparency disclosures be able to understand the risks of the AI system in question. In this paper, we propose a risk profiling standard which can guide downstream decision-making, including triaging further risk assessment, informing procurement and deployment, and directing regulatory frameworks. The standard is built on our proposed taxonomy of AI risks, which distills the wide variety of risks proposed in the literature into a high-level categorization. We outline the myriad data sources needed to construct informative Risk Profiles and propose a template and methodology for collating risk information into a standard, yet flexible, structure. We apply this methodology to a number of prominent AI systems using publicly available information. To conclude, we discuss design decisions for the profiles and future work.

Introduction

While AI capability improvements (OpenAI 2023; Anthropic 2023) create significant opportunities for societal benefit, AI systems pose numerous risks and potential harms. These risks take many forms, including social harms (Solaiman et al. 2023), labor market disruption (Eloundou et al. 2023), threats to creativity (Franceschelli et al. 2023), erosion of democracy (Jungherr 2023), and extreme risks (Shevlane et al. 2023). It is crucial for principled approaches to making AI deployment decisions to be grounded in accurate, practical, and clear assessments of risk.

The range of stakeholders tasked with AI deployment decision-making is incredibly diverse, necessitating shared context. We hold that risk is an ideal abstraction for establishing that context among actors along the value chain. Non-technical decision-makers, for instance, readily consider the risk of adverse events during cost-benefit analyses (Špačková et al. 2015). Regulators have found risk-framing similarly useful: the draft EU AI Act (Commission 2021) defines AI developer obligations in relation to risk. Risk also aligns with technical frameworks, corresponding to the expected value of possible adverse outcomes. In essence, risk should be the *lingua franca* for AI system assessment.

Recent work has sought to clarify how risk evaluation and disclosure could serve this diversity of decision-makers. Shevlane et al. (2023), for instance, outlined a "transparency layer" within the AI development stack. This proposal makes the key observation that decision-makers will benefit from a standardized framework for pre-deployment evaluation. Such a convention should clearly outline risks in a unified taxonomy and be flexible enough to enable information distillation at several levels of technical depth.

Unfortunately, current conventions for AI risk disclosure do not satisfy these diverse needs. One popular disclosure artifact, Model Cards (Mitchell et al. 2019), aims to communicate the intended uses of a system, technical details, and ethical considerations to relatively technical audiences. However, Model Cards don't standardize risk disclosures and are less relevant for multi-purpose AI systems, where *unintended* and *unanticipated* uses are critical concerns. A second approach is releasing detailed technical reports along with powerful AI systems, as exemplified by OpenAI (2023) and Anthropic (2023). While highlighting modeling decisions and evaluations, this approach does not adequately support non-technical stakeholders, and could be improved to better serve technical consumers by expanding discussions of data sources, safety evaluations, and third-party assessments. Both approaches lack a standard risk taxonomy and fail to emphasize risk in characterizing AI systems.

In this paper, we propose a *risk-centric* approach for assessing AI systems and a standardized reporting paradigm, which we call *Risk Profiles*. Inspired by other work on risk categories (Solaiman et al. 2023; Shevlane et al. 2023; Barrett et al. 2023; Commission 2021), we propose a high-level taxonomy of risks which can guide risk assessment and disclosure. We establish a template for reporting the risks posed by an AI system and the mitigation measures provided by the system's developer, and outline a methodology for collecting and distilling system information to generate Risk Profiles. We then apply this methodology to create Risk Profiles on several popular AI foundation models using publicly available information. We conclude with a discussion of al-

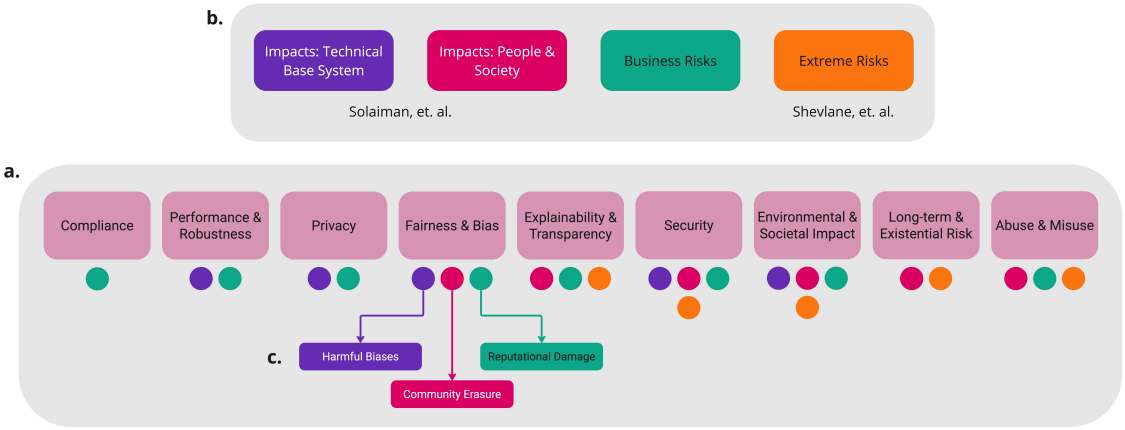

Figure 1: Illustration of how the Risk Taxonomy (a) subsumes other risk categorization frameworks (b). The Risk Taxonomy is expressive enough to capture multiple concerns, from corporate compliance interests to societal harms. Multiple risks exist under each category (c). While a 1-1 mapping is not always possible (e.g., many risks simultaneously impact privacy, security, and society), the taxonomy's primary role is to be a standardized, high-level schema for risk identification and communication.

ternative design decisions, and how Risk Profiles can help establish regulatory and industry standards for disclosure.

A High-level Risk Taxonomy

In this section, we detail our proposed taxonomy of AI risk categories. We center our taxonomy on high-level risk categories that subsume known risks, exhibited in Fig. 1. This taxonomy is both comprehensive and flexible, avoiding omission of any one risk class (e.g., social risks per Shevlane et al. (2023)) and ensuring adaptability to specific evaluator needs. The taxonomy can be used to anticipate both user-oriented harms, corporate incentives, and societal externalities. This is critical, since many risk *scenarios* reflect multiple *risks* simultaneously. For instance, a discriminatory credit prediction system poses fairness, compliance, and societal impacts. Our taxonomy follows the inspiration of recent papers in using generative and general purpose systems (genAI) as the frame of reference. Because the risks posed by genAI systems encompass those posed by other AI systems, our taxonomy is applicable in both settings.

These features allow the taxonomy to be used within risk-management processes like those articulated by the NIST AI Risk Management Framework ((Tabassi 2023); specifically MAP 2.1), which direct AI developers or deployers to identify potential negative impacts, but offer little practical guidance on how to discover or identify these risks. To maximize effectiveness, the risk taxonomy should be used in conjunction with "Trustworthy Characteristics" (Tabassi 2023). The complementary use of these tools enables comprehensive risk/benefit analysis of AI systems.

Abuse & Misuse: The potential for AI systems to be used maliciously or irresponsibly, including for creating deepfakes, automated cyber attacks, or invasive surveillance systems. Specifically denotes *intentional* use of AI for harm.

Compliance: The potential for AI systems to violate laws, regulations, and ethical guidelines (including copyrights). Non-compliance can lead to legal penalties, reputation damage, and loss of trust. While other risks in our taxonomy apply to system developers, users, and broader society, this risk is generally restricted to the former two groups.

Environmental & Societal Impact: Addresses AI's broader societal effects, including labor displacement, mental health impacts, and issues from manipulative technologies like deepfakes. Additionally, it considers AI's environmental footprint, balancing resource strain and training-related carbon emissions against AI's potential to help address environmental problems.

Explainability & Transparency: The feasibility of understanding and interpreting an AI system's decisions and actions, and the openness of the developer about the data used, algorithms employed, and decisions made. Lack of these elements can create risks of misuse, misinterpretation, and lack of accountability.

Fairness & Bias: The potential for AI systems to make decisions that systematically disadvantage certain groups or individuals. Bias can stem from training data, algorithmic design, or deployment practices, leading to unfair outcomes and possible legal ramifications.

Long-term & Existential Risk: The speculative potential for future advanced AI systems to harm human civilization, either through misuse or due to challenges in aligning AI objectives with human values.

Performance & Robustness: The AI system's ability to fulfill its intended purpose and its resilience to perturbations, and unusual or adverse inputs. Failures of performance are fundamental to the AI system's correct functioning. Failures of robustness can lead to severe consequences.

Privacy: The potential for the AI system to infringe upon individuals' rights to privacy, through the data it collects, how it processes that data, or the conclusions it draws.

Security: Encompasses vulnerabilities in AI systems that compromise their integrity, availability, or confidentiality. Security breaches could result in significant harm, ranging from flawed decision-making to data leaks. Of special con-

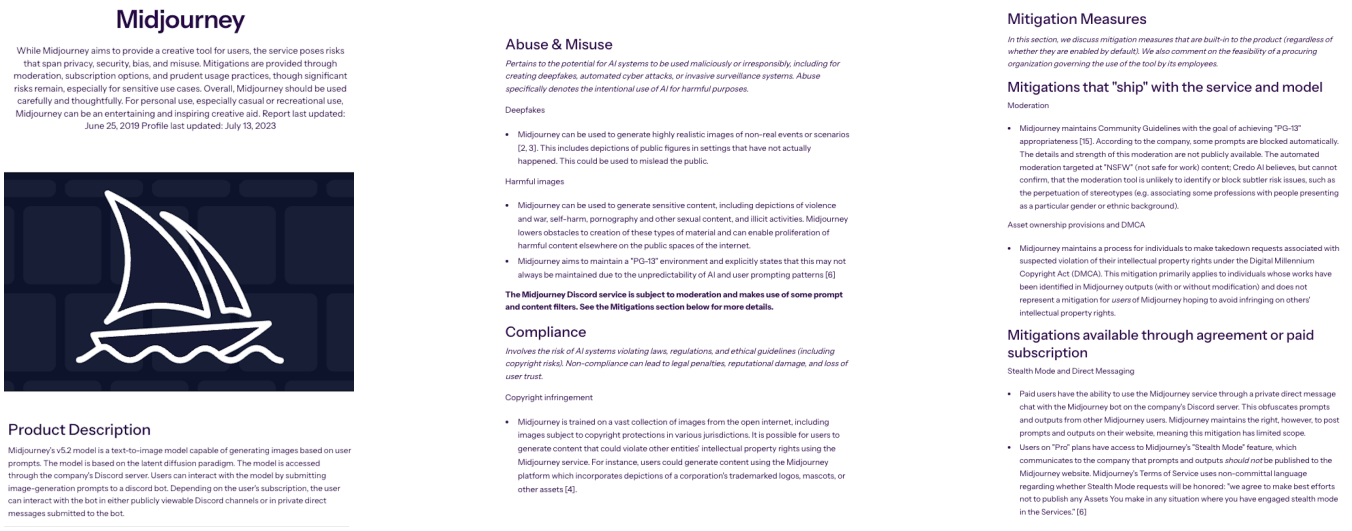

Figure 2: Example realization of our proposed reporting template featuring information about risk assessment, mitigations, evaluations, and certifications & compliance.

cern is leakage of AI model weights, which could exacerbate other risk areas. Evaluation of attack vectors is well-studied, with myriad frameworks for articulating potential vulnerabilities, such as the ATT&CK Framework (Roy et al. 2023).

A New Risk Disclosure Paradigm

Here we detail our methodology for generating AI Risk Profiles. For illustration, we compiled profiles that exemplify our risk taxonomy and report template for five popular AI systems: Anthropic's Claude, OpenAI's GPT APIs, Microsoft Copilot, GitHub Copilot, and Midjourney. These can be found at https://credo.ai/ai-vendor-directory.

Identifying Risk Scenarios

To apply the risk taxonomy to a particular AI use case, report creators must identify applicable *risk scenarios*, which cover adverse outcomes that arise from any use of the system (not simply intended use). The risk profiler must assess the AI system against a set of possible scenarios and infer whether each scenario applies based on its implicit probability and magnitude of impact. Presently, risk scenario discovery relies on subject-matter expertise with respect to AI's potential harms, with significant latitude for subjectivity. We expect that as the AI risk assessment field matures, practitioners will settle on common risk scenarios, develop tools to support risk-scenario discovery, and establish conventions for determining the thresholds of relevance for each scenario.

It is critical that identified risk scenarios are relevant to the Risk Profile's target audience. Profiles compiled by independent third parties can approximate the risk scenarios relevant to a downstream user, but comprehensive profiling requires incorporating all of the expertise and information that exists for the system.

Surfacing Risk-Relevant Information

In practice, risk scenario identification relies on inference from existing AI system documentation, which can be drawn from numerous sources including:

- academic articles detailing system design, training data composition, fitting procedures, and evaluations;
- marketing materials about the system or the developer's general research and design practices;
- independent evaluations of the system, including both replication studies and use case-specific evaluations; and
- evaluations of analogous systems.

The availability of each source will depend on the report creator's relationship with the system developer, as many developers are reluctant to disclose design and training information under pretenses of competition or potential for misuse. This disclosure paradigm will bring clarity to the risk evaluation process by highlighting *where* risk information comes from and the causes of information gaps.

Report Creation

Figure 2 illustrates our proposed template for AI risk and mitigation reporting. Key sections include an analysis of risks and mitigations, a discussion of formal evaluations of the system, and enumeration of which regulations and standards the system conforms to. We wrote our reports without partnering with the AI vendors, and so our Risk Profiles represent a particular instantiation of this template. The general methodology for profile creation is summarized by Fig. 3.

Our instantiation of Risk Profiles begins with an executive summary which lists key risks users and deployers should be aware of. The introduction summarizes the AI system, its inputs and outputs, and intended uses according to the developer. We also comment on the developer's reputation, highlighting the role of any key stakeholders, like investors.

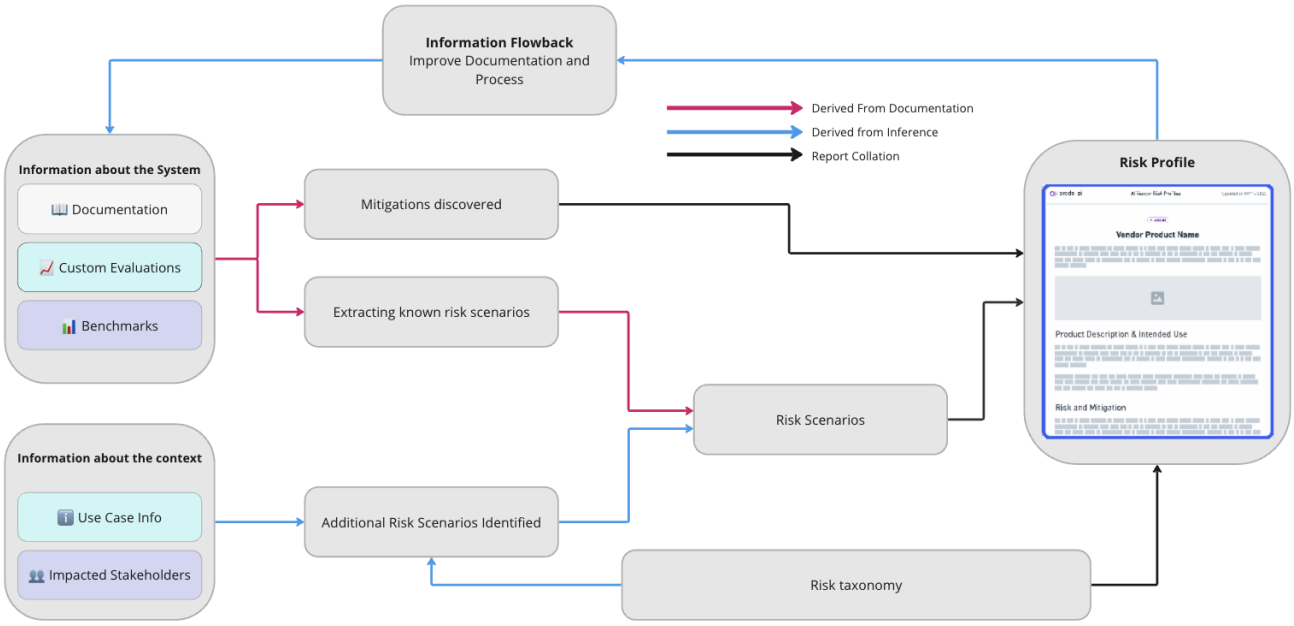

Figure 3: Our methodology for synthesizing system information into risk profiles. Arrows represent information flow, via inference or directly from documentation. Insights distilled through profile creation can inform base documentation improvements.

The Risk and Mitigation analysis serves as the central section of the profile. We first provide a risk and mitigation summary table, as in Fig. 4. The table identifies which of the 9 risks from our taxonomy apply to the system and whether the developer has implemented mitigation measures, here defined as any measure that reduces the likelihood or severity of a risk. We then describe the results of the risk scenario identification process for each risk category. Given our target audience of corporate procurement decision-makers and AI-informed regulatory bodies, we adopted an inclusive mindset for this step. We surfaced risk scenarios that were, in our view, plausible and relevant to those audiences, emphasizing scenarios that have received attention in public discourse in recent months. Likewise, while we discussed all developer-provided mitigations, we were cautious; given the difficulty of evaluating the efficacy of most mitigations, we stressed that risk *reduction* is not synonymous with *elimination*.

The evaluations section focuses on quantitative benchmarks, relying primarily on the developers' published evaluations. Where possible we highlight academic benchmarks (e.g. MMLU for large language models), but cite closed evaluations for model behaviors for which no public evaluation exists. We also cite third-party qualitative metrics, such as chatbot rankings (Zheng et al. 2023). We expect this section to expand to include further model-specific and qualitative evaluations (e.g., ARC's evaluation of GPT4, (OpenAI 2023)), as evaluation methods mature (Perez et al. 2022a).

We check system compliance with 8 standards and regulations, including the EU's General Data Privacy Regulation, California's Consumer Privacy Act, and SOC II. These are among the most relevant standards and laws for enterprises.

Discussion

Risk Taxonomy

Our risk taxonomy is designed to cover all known risk categories. This leads to potential overlap between categories, which could complicate mitigation if focusing on a single category. (Suppose regulation mandates mitigating 'Privacy' risks. Are closely related 'Security' risks also covered?) We see this as a non-issue. Risk Profiles and our taxonomy are meant to ease and enhance communication. Capturing all risks is preferable to overlooking risk scenarios by over-indexing to a non-comprehensive taxonomy.

Design Decisions

Generating our five example Risk Profiles necessitated several design decisions informed by our independent position and target audience. These decisions point to areas where report creators can make their profiles fit for purpose.

Profiler Identity Perhaps the most influential factor in Risk Profile compilation is the relationship between the Risk Profile creator and the AI system developer. Our example profiles reflect our independent synthesis of public information about each system. This avoids some biases. For instance, we can include reputationally-damaging information about the developer that would be omitted if the profiles were first-party generated. On the other hand, developer involvement has benefits: they are likely knowledgeable about system-specific risk scenarios and privy to additional evaluation details that make a profile comprehensive.

More broadly, we expect a profiler's background to have an impact on the final report. Profilers should be, at a minimum, knowledgeable about AI and the application domain.

Risk	Present	Built-in Mitigation
Abuse & Misuse	⚠️	❌
Compliance	⚠️	✅
Environmental & Societal Impact	⚠️	✅
Explainability & Transparency	⚠️	❌
Fairness & Bias	⚠️	✅
Long-term & Existential Risk	-	N/A
Performance & Robustness	⚠️	❌
Privacy	⚠️	✅
Security	⚠️	✅

Figure 4: A summary of the risks present in GitHub Copilot and accounting of which risks GitHub has built mitigations for.

Profilers need not be technical experts. In fact, when feasible, it is advisable that Risk Profiles be compiled with input from multiple individuals with differing perspectives on risk to ensure the identification of applicable risk scenarios is sufficiently comprehensive.

Risk and Mitigation Summary As a multi-page report, the Risk Profiles benefit from summary information – such as executive summaries and the risk distillation in Fig. 4 – which enables readers to efficiently glean the most important details in the transparency report. To be valuable, this summarization needs to be grounded with a reference point, e.g., addressing the question 'mitigated relative to what?' In our embodiment of Risk Profiles, we see risks arising as a function of either modeling decisions or structural decisions like the computing infrastructure. Thus, mitigation (i.e. the green checkmarks in Fig. 4) is defined relative to a model trained with performance (accuracy, perplexity, etc.) as the sole objective, and to an infrastructure with high risk tolerance, such as public cloud with no emissions reductions and permissive data use policies, respectively. The green checkmarks in our embodiment signify that the developer has taken on some responsibility towards addressing risk relative to these reference points.

Of course, numerous alternative risk distillation approaches are possible, including quantitative ones. For instance, one could provide counts of which risks (scenarios) apply to the system – either in absolute terms or as a fraction of the total number of risks considered. Further sophistication is possible by positing system-level or per-risk scores which account for the probability of the risk (scenario) and the impact were that risk to be realized, or analogously the degree of mitigation achieved through a particular strategy. In such cases, establishing a reference point for the understanding degree of mitigation is also necessary.

All of these approaches have limitations. Binarizing and counting implicitly weigh all risks equally in terms of both probability and impact. This simplification could mislead report consumers. For instance, a distant, low-likelihood risk like a large language model-based (LLM) AI developing resource acquisition capabilities may be over-weighed relative to an immediate, high-likelihood risk like that same LLM encoding racial stereotypes into employment-related decisions. Risk scoring is also problematic. Absent actual adverse outcomes data, scoring relies on subjective weighting of competing risks, meaning the risk summarization will over-index to the risk profiler's biases. We opted for the former option because it is more conservative. Nevertheless, there is clearly space to further develop these approaches – we envision a future scoring method which serves as a middle-ground by enabling transparent application of subjective weightings. This is an area of ongoing work.

Evaluations The nascent state of AI evaluations, particularly for genAI systems, makes profiling challenging. Model developers actively establishing conventions around which benchmarks are relevant to each model type (e.g. MMLU (Hendrycks et al. 2020) for LLMs) but more work is necessary. First, performance on traditional benchmarks is increasingly saturated (Maslej et al. 2023), requiring the development of more challenging and comprehensive benchmarks. Second, current benchmarks are not comprehensive due to the difficulty in creating them (though automated evaluations could help (Perez et al. 2022a)). Given the limitations of current benchmarks, profilers should supplement benchmarks with bespoke and qualitative evaluations like red-teaming results (Perez et al. 2022b), alignment-oriented evaluations (Perez et al. 2022a), and contextualized evaluations on use-case relevant datasets. Lastly, existing evaluation approaches poorly anticipate downstream impacts on stakeholders and society writ-large. Recent proposals for 'soft' deployments, such as 'regulatory sandboxes' (Truby et al. 2022), could prove useful for improving impact-forecasting, as could explicit measures of societal impact (Solaiman et al. 2023), both of which can be incorporated into Risk Profiles as supporting evaluation information.

Regardless of which evaluations are cited in the Risk Profile, the profiler has the responsibility to discuss their relevance and interpretation. Readers should be able to understand the relationship between each metric and the system's suitability for the deployment setting.

Implications for Mandated Reporting

Increasing public awareness of AI's risks has led to calls for mandated risk disclosures by AI developers (Commission 2021; Kang 2023). Our Risk Profile template and methodology address many of the concerns that motivate these proposals. For instance, the draft EU AI Act lays out trans-

parency requirements for high risk AI systems and foundation models. Obligations cover details like a 'Description of the capabilities and limitations of the foundation model' or 'Description of the model's performance, including on public benchmarks or state of the art industry benchmarks'. While capabilities and performance are partially covered by artifacts like Model Cards, the Risk Profile expands this artifact with a detailing of mitigations, a comprehensive articulation of limitations, and the added benefit that our methodology provides a guide for *contextualizing* the required risk information. Two recent related research efforts – Algorithmic impact assessments and regulatory scorecards – point to immediate steps we can take to strengthen our proposals in service of a convention for regulation-mandated risk reporting. Algorithmic impact assessments (Reisman et al. 2018) were conceived as a tool to inform regulators on algorithmic harms. In particular, they establish a principled process for identifying and evaluating likely harms to stakeholders and externalities. Our approach can incorporate this process to harden our evaluations section and fill gaps left by the present inadequacy of technical evaluations. Regulatory scorecards, like the one studied in Bommasani et al. (2023) for the EU AI Act, represent an alternative approach to risk summarization. They amount to a binary checklist, not of risk application, but of satisfaction of legal requirements. When a Risk Profile is explicitly compiled to show conformity with a specific regulation, adopting a scorecard can be hugely beneficial for clarifying and communicating how the AI system matches up against the requirements.

Monitoring & Ongoing Decision-Making

Ultimately, the full benefits of Risk Profiles will be realized when they can inform *ongoing* decision-making, following the initial deployment. As a system sees in-production use, the pre-deployment Risk Profile should be used as a guide for the perpetual process of governing the system. It can indicate the need for additional evaluations, further mitigation measures, or even changes to the system itself, like model re-training. It also can inform monitoring strategies which lead to regular revisions of the Risk Profile.

One variety of resource that can support ongoing risk assessment and risk management is AI incident databases (Pittaras et al. 2022; McGregor 2021). We expect incident databases to be directly applicable to post-deployment monitoring. Incidents indicate risk scenario realization and allow stakeholders to update their priors about the need for additional, stronger mitigation measures. This approach supplements the notion of Risk Profiles being 'living documents'. Incident databases may also be indirectly applicable to Risk Profiles: as future work, we plan to develop a principled approach to identifying similar systems to the one being evaluated at pre-deployment time. As with post-deployment same-system incidents, like-systems experiencing incidents can indicate a need for alternative mitigation approaches.

References

Anthropic. 2023. Model Card and Evaluations for Claude Models. anthropic.com. Accessed: 2023-11-30.

Barrett, A.; et al. 2023. Response to NTIA Request for Comments on AI Accountability Policy. *Berkeley CLTC*.

Bommasani, R.; et al. 2023. Do Foundation Model Providers Comply with the EU AI Act? *Stanford CRFM*.

Commission, E. 2021. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Ammending Certain Union Legislative Acts. *EUR-Lex*.

Eloundou, T.; et al. 2023. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv:2303.10130*.

Franceschelli, G.; et al. 2023. On the creativity of large language models. *arXiv:2304.00008*.

Hendrycks, D.; et al. 2020. Measuring massive multitask language understanding. *arXiv:2009.03300*.

Jungherr, A. 2023. Artificial Intelligence and Democracy: A Conceptual Framework. *Social Media+ Society*, 9(3): 20563051231186353.

Kang, C. 2023. OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing. *The New York Times*.

Maslej, N.; et al. 2023. Artificial intelligence index report 2023. *arXiv:2310.03715*.

McGregor, S. 2021. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 15458–15463.

Mitchell, M.; et al. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.

OpenAI. 2023. GPT-4 technical report. *arXiv:2303.08774*.

Perez, E.; et al. 2022a. Discovering language model behaviors with model-written evaluations. *arXiv:2212.09251*.

Perez, E.; et al. 2022b. Red teaming language models with language models. *arXiv:2202.03286*.

Pittaras, N.; et al. 2022. A taxonomic system for failure cause analysis of open source AI incidents. *arXiv:2211.07280*.

Reisman, D.; et al. 2018. Algorithmic Impact Assessments: A Practical Framework for Public Agency. *AI Now*.

Roy, S.; et al. 2023. SoK: The MITRE ATT&CK Framework in Research and Practice. *arXiv:2304.07411*.

Shevlane, T.; et al. 2023. Model evaluation for extreme risks. *arXiv:2305.15324*.

Solaiman, I.; et al. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. *arXiv:2306.05949*.

Špačková, O.; et al. 2015. Cost-benefit analysis for optimization of risk protection under budget constraints. *Risk Analysis*, 35(5): 941–959.

Tabassi, E. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). *NIST*, (NIST.AI.100-1).

Truby, J.; et al. 2022. A sandbox approach to regulating high-risk artificial intelligence applications. *European Journal of Risk Regulation*, 13(2): 270–294.

Zheng, L.; et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv:2306.05685*.