Ten Insights From Other Domains That Inform Responsible AI Frameworks

Dunstan Allison-Hope[1], Patrick Gage Kelley[2], Reena Jana[2], Angela McKay[2*], Allison Woodruff[2]

[1]Dunstan Hope Consulting
[2]Google
dunstan@dunstanhope.com, patrickgage@acm.org, rmjana@google.com, angelamckay23@outlook.com, woodruff@acm.org

Abstract

As AI rapidly evolves, so too do the guidelines, principles, best practices, standards, and regulations that seek to ensure the responsible development and use of AI systems. This article outlines ten insights from other domains that responsible AI frameworks can draw upon. We highlight the importance of using well-established international human rights standards and emphasize the value of tailoring risk assessment methodologies to suit the AI context, deploying system-wide strategies, and undertaking meaningful and effective stakeholder engagement, amongst other durable learnings. Through a mix of continuity and adaptation of frameworks in other domains—and knowing how and when to deploy each—we chart a more practical, policy-based path that supports the responsible design, development, deployment, and use of AI systems.

Introduction

The rapid growth of *AI systems*[1] is being accompanied by a proliferation of new guidelines, principles, best practices, standards, and regulations (hereafter "frameworks") that seek to ensure the responsible design, development, deployment, and use of AI systems. These frameworks are seen as one mitigating force in ensuring that the impacts AI has on society (labor and jobs, the environment, privacy and data control, etc.) are beneficial rather than harmful (Bender et al. 2021; Dwivedi et al. 2023). These frameworks include, for example:

- Government-led efforts, such as the EU AI Act (European Union 2024), the Bletchley Declaration (AI Safety Summit 2023), and the NIST AI Risk Management Framework (NIST 2023).
- Multilateral initiatives, such as the OECD AI Principles (OECD.AI Policy Observatory 2019), the UN High-Level Advisory Body on AI (United Nations AI Advisory Body 2024), the UNESCO Recommendation on the Ethics of Artificial Intelligence (UNESCO 2024), the UN Global Digital Compact (United Nations Office for Digital and Emerging Technologies 2024), and the Hiroshima Process International (G7) Code of Conduct (G7 2023).
- Collaborative initiatives, such as the WEF Responsible AI Playbook for Investors (Gillam and Jain 2024), the Partnership on AI Guidance for Safe Foundation Model Deployment (Partnership on AI 2023), and the Data and Trust Alliance Data Provenance Standards (Data and Trust Alliance 2023).
- Single company efforts, such as the responsible AI principles published by Amazon (Amazon 2023), Google (Pichai 2018), Microsoft (Microsoft 2018), OpenAI (OpenAI 2025), and others.
- Academic initiatives, such as unifying AI principles (Floridi et al. 2018), taxonomizing AI harms (Shelby et al. 2023), evaluating AI use (Peters et al. 2020), or defining AI literacy (Long and Magerko 2020).

These frameworks sometimes emphasize that AI systems bring new, different, or unique characteristics that require some tailored approaches; however, the design, development, deployment, and use of AI also share many characteristics of innovations and business transformations that have gone before and for which well-established frameworks can be used. Building on our own experience operationalizing such frameworks, our premise is that the substance, implementation, progress, and evolution of these responsible AI frameworks can be informed and improved by the practical experience of pursuing similar desired outcomes in other relevant *domains*. These pursuits include regulatory compliance, such as adhering to EU regulations; voluntary initiatives, such as operationalizing company human rights policies; and broader company efforts, such as adhering to international standards of responsible business conduct.

This article outlines ten insights responsible AI frameworks can draw upon from other domains, thereby enhancing their effectiveness and minimizing the need to invent entirely new approaches. These insights were created by the authors as experts with thousands of hours of practical experience in responsible AI; amongst these experts there is further expertise in human rights, human-computer interaction, online safety, privacy, responsible business, risk management, security, stakeholder engagement, sustainability,

[*]Angela McKay contributed to this work while at Google.

[1]Throughout the paper we will provide definitions or extended explanations of certain key terms, organizations, or documents. These are each italicized, like *AI systems*, and can be found in our glossary at the end.

transparency reporting, and related issues. To generate the insights from these domains, we used a simple "repetition test".[2] Insights passing the repetition test are those experts with extensive practical experience find themselves repeating frequently when providing advice and describing the best practices that have emerged from their domains.

We center this work on how responsible AI frameworks can and should evolve from their current state, and so we are necessarily aspirational rather than describing how any single organization or company is acting today. Instead, we aim to show how responsible AI frameworks can be enhanced by incorporating well-established, widely accepted, and vetted approaches from other domains, such as prior efforts to apply international standards of responsible business conduct, respect human rights, and achieve global sustainable development goals. While many of the ideas in this article may be familiar to experts in topics such as human rights, we hope our descriptions can provide valuable context for researchers and practitioners who may not be familiar with some of the experience and ideation in particular domains. Further, we hope this work helps articulate current weaknesses and gaps in common AI frameworks, and how existing frameworks from other domains can close these gaps.

In the insights that follow (see Table 1), we highlight the importance of using well-established international human rights standards and industry-agnostic frameworks for responsible business practices that are interoperable across jurisdictions and provide a shared terminology for both developers and deployers of AI systems. We emphasize the value of tailoring risk assessment methodologies to suit the AI context, deploying system-wide strategies, and undertaking meaningful and effective stakeholder engagement. We conclude by underscoring the benefits of holistic approaches that address the risks and opportunities of AI in combination.

Ten Insights

Insight #1 – International human rights provide a reliable and durable risk taxonomy

International human rights instruments set out the rights and freedoms that every human being is entitled to, including civil and political rights (e.g., privacy; freedom of expression; non-discrimination; equality before the law) and economic, cultural, and social rights (e.g., health; education; just and favorable conditions of work; access to culture). These instruments provide a credible, robust, and tested taxonomy for identifying and assessing risk and ensuring a complete analysis (United Nations B-Tech and Human Rights Office of the High Commissioner 2024). The value of international human rights instruments is enhanced by decades of accompanying interpretation that practitioners can draw upon, including abundant analysis from *UN treaty bodies and special procedures*, courts, and other experts (BSR 2025).

For example, Article 34(1)(b) of the *Digital Services Act* (European Union 2022) (DSA) requires companies to assess risks with the *EU Charter of Fundamental Rights* (European Parliament 2000), a regional human rights instrument, as a reference point. This provides a well-established baseline against which companies can assess risks to people and societies, helping ensure that the systemic risk assessments utilize a durable human rights taxonomy and do not miss important impacts. At the same time, Article 34 of the DSA adds unnecessary complexity by also requiring an assessment of risks to society-wide normative goals, such as illegal content, civic discourse, and public health, alongside human rights (Del Campo, Zara, and Álvarez-Ugarte 2025). However, these risks can be directly tied to specific human rights (such as the right to bodily security, the right to vote, and the right to health), causing unnecessary duplication and conceptual confusion.

Relevance for AI Frameworks: State that the realization, fulfillment, and enjoyment of human rights is a universal desired outcome to which AI can contribute and that all companies have a responsibility to assess, avoid, and address risks to human rights posed by AI. Human rights should be framed as being inclusive of (rather than separate from) other desirable outcomes, such as democracy, rule of law, and public health, and with direct reference to relevant international and regional human rights instruments.

Insight #2 – Foundations for AI risk management can be borrowed from frameworks for responsible business conduct

Industry-agnostic but well-established frameworks for responsible business conduct, especially the *UN Guiding Principles on Business and Human Rights* (UNGPs) (United Nations Office of the High Commissioner 2011) and the *OECD Guidelines for Multinational Enterprises* (OECD 2023), provide a valuable conceptual foundation for risk analysis, mitigation, and remediation in the AI domain.

The international human rights instruments described in the first point were written for governments rather than companies; however, the UNGPs and OECD Guidelines provide authoritative guidance for how companies should operate responsibly and respect human rights in practice, even while governments maintain the ultimate duty to protect human rights. Utilizing these internationally recognized frameworks can mitigate against jurisdictional fragmentation and vastly reduce the number of "similar-but-slightly-different" regulatory requirements that companies in the AI domain will need to address in the coming years.

For example, following UNGPs guidance to assess the severity of risk to people using the three criteria of scope (number of people), scale (gravity of harm), and remediability (ability to make good) allows all risk assessments to align with international best practices and other emerging industry-specific and industry-agnostic frameworks and regulations, such as anticipated EU regulation (United Nations B-Tech and Human Rights Office of the High Commissioner

[2]This repetition test could be viewed in the same light as saturation in qualitative analysis (Guest, Bunce, and Johnson 2006); we do not suggest that we outline every insight on AI Frameworks, but rather these ten cover a wide set of weaknesses we have repeatedly encountered.

	Insight	Domains Drawn Upon
1.	International human rights provide a reliable and durable risk taxonomy	Human rights, Online safety
2.	Foundations for AI risk management can be borrowed from frameworks for responsible business conduct	Human rights, Business conduct
3.	Distinguish between high-level processes versus detailed requirements, and know when each strategy is appropriate	Content provenance, Online safety,
4.	Collaboration and comparison across systems are required to address risk successfully	Violent extremism, Risk management, Content moderation
5.	Risk assessment and due diligence should be an ongoing priority, not a moment-in-time activity	Risk management
6.	Reporting and disclosure of AI risks by companies are evolving	Financial disclosure, Transparency reporting, Climate & sustainability
7.	Meaningful and effective stakeholder engagement requires significant investment in scale and specialist skills	Online safety, Climate & sustainability, Stakeholder engagement
8.	Stakeholder engagement strategies should not be underestimated or constrained by regulation	Public policy, Online safety, Stakeholder engagement
9.	Common assumptions about risk management should be challenged	Risk management, Cybersecurity
10.	The risks and opportunities of AI systems for people and society can be pursued together	Human rights

Table 1: Our list of ten insights that can inform the design, development, and application of responsible AI frameworks, highlighting domain expertise which each of these insights builds upon.

2020). This should significantly enhance compliance efficiency by enabling companies, governments, and stakeholders to focus more resources on strategic priorities and fewer on paperwork (OECD 2023).

Further, these industry-agnostic frameworks of responsible business conduct are interoperable across industries, technologies, and borders. They can provide a shared terminology and common frame of reference between (a) technology companies developing and selling AI systems and (b) "non-tech" companies deploying AI systems in industries such as retail, healthcare, and financial services. The "non-tech" companies will generally be more familiar with industry-agnostic frameworks for responsible business conduct but less familiar with the more recent ecosystem of specialist responsible AI practitioners. The common ground enabled by industry-agnostic frameworks will become more important over time as the upstream developers of AI systems increasingly collaborate with downstream deployers of AI to identify, assess, and address the impacts of AI across shared value chains and AI systems.

Relevance for AI Frameworks: Pursue risk assessment and management of AI systems in alignment with international standards of business conduct set out in the UNGPs and OECD Guidelines for Multinational Enterprises. The May 2024 revisions to the OECD AI Principles provide a good example of how to do this in practice by directly referencing OECD standards for responsible business conduct; by contrast, the *EU AI Act* does not reference the UNGPs or the OECD Guidelines, increasing the likelihood that companies use fragmented methodologies for risk assessment and management. The *NIST AI Risk Management Framework* (NIST 2023) issued by the US Department of Commerce in 2023 also did not reference the UNGPs or the OECD Guidelines, though a subsequent Risk Management Profile for Artificial Intelligence and Human Rights (Bureau of Cyberspace and Digital Policy 2024) issued by the US Department of State in 2024 (and subsequently archived) sought to bridge the gap between human rights and risk management approaches with direct reference to both the UNGPs and the OECD Guidelines.

Insight #3 – Distinguish between high-level processes versus detailed requirements, and know when each strategy is appropriate

Distinguish between (a) high-level processes designed to achieve broadly defined desired outcomes and (b) detailed requirements for technology-specific risks, and know when each strategy is appropriate.

There are times when detailed frameworks designed to address the "known-known" risks of specific technologies are an appropriate strategy. In the 2010s, significant resources were invested in identifying, labeling, and authenticating political advertising to address a particular challenge relating to electoral integrity; in the context of AI, the development of technical standards for certifying the source and history

("provenance") of media content by the *Coalition for Content Provenance and Authenticity* (C2PA)[3] is a good parallel example.

However, there are other contexts in which this approach may not be sustainable over time. For example, the drafting of the EU DSA and the EU AI Act pre-dated prominent developments in generative AI, so these regulations may struggle to address new impacts that arise from it. These regulations successfully established approaches oriented towards the highest profile concerns of the moment (such as recommender systems and emotion recognition), but it remains to be seen whether they can address the longer-term sociotechnical developments of generative AI that will emerge over time. Defining overall desired outcomes (e.g., knowing and addressing risks to people and society) and regulating general processes companies use (e.g., risk assessment, mitigation, and transparency) may be more durable than technology-specific requirements.

The best way to ensure that resources are consistently targeted on the most important risks is to establish a mix of (a) high-level processes designed to achieve broadly defined desired outcomes and (b) more detailed requirements for technology-specific risks. While these should be connected and interoperable, they have very different characteristics that should not be confused.

Relevance for AI Frameworks: Be clear about the purpose of AI frameworks and whether they exist to set high-level direction or address a specific problem such as provenance, copyright, or energy use. Setting high-level direction may be better suited for public policy, laws, and regulations because these can be difficult to refine over time; by contrast, addressing particular challenges may be better suited for voluntary industry standards because these can more easily be updated as technology, context, and use cases evolve. There are exceptions—for example, some particular use cases may be rightly outlawed by regulation, while industry standards benefit from high-level normative direction—so a smart mix of approaches is appropriate in each sphere.

Insight #4 – Collaboration and comparison across systems are required to address risk successfully

While individual company performance is important, many risks are *system*-wide and cannot be mitigated by companies acting alone (The Danish Institute for Human Rights 2019; Brodeur and Achterberg 2025). This can be especially true in the technology industry, which is inherently interconnected across industries, technologies, value chains, and the internet stack.

An example of the importance of system-wide approaches is provided by efforts to address illegal and harmful content online. Here, we have learned the importance of risks that move from one service to another and involve multiple companies, such as the way terrorist and violent extremist activity online exploits multiple platforms and types of online services (Baele, Brace, and Ging 2024; Mitts 2021) or how those engaged in the sexual exploitation and abuse of children switch between social media and private messaging services (Freed et al. 2023) as part of strategies to cause harm. Any single company can only see and address a fragment of the risk, so collaborative approaches—such as the data, insight, and signal-sharing strategies deployed by the *Global Internet Forum to Counter Terrorism*[4] and the *Tech Coalition*[5]—have become essential for company risk mitigation efforts.

Further, the nature of the internet is such that companies operating in different parts of the system have distinct roles to play when addressing illegal and harmful content (The Global Network Initiative 2023). For example, search, social media, and private messaging services all have different insights into the content being shared, distinct risk management choices available to them, and mitigation options that bring different knock-on effects for freedom of expression (United Nations Office of the High Commissioner 2016). The "right" content moderation decision in one part of the internet stack may be the "wrong" decision elsewhere in the internet stack, and it is only by collaborating with each other (e.g., signal sharing) and being aware of the unique but complementary role played by each service that they will achieve consistently their shared goals of a trustworthy content ecosystem.

As a result, when the companies regulated by the DSA published their first risk systemic risk assessment reports in November 2024, some of the most helpful analyses of these reports considered them collectively, seeking to understand how the entire industry addressed systemic risk together and reflect upon each platform's respective role. While comparing companies' reports with each other and asking "who did it best" has some merit and may inspire a race to the top, it falls short of taking a system-wide perspective. Indeed, the DSA's emphasis on around twenty of the EU's largest services ("very large online platforms" and "very large online search engines") risks generating only a partial view of systemic risk in the EU.

While the specifics differ, our premise is that similarly differentiated and collaborative approaches should be taken with AI systems. For example, when an AI system is used in the financial services industry, the company developing the AI system will have different insights and a different role in identifying and addressing risk than the company deploying the AI system; the former will have greater insight into how the AI system is initially trained and works, while the latter will have greater insight into the specific inputs and outputs of the AI system and how it is refined over time. Similarly, the developer of a generative AI model has more insight into and leverage over how the model is designed, created, or deployed by themselves than when and where it is deployed by others. As with content moderation, shared goals of safe, trustworthy, and responsible AI systems will be achieved by collaboration—and by being aware of each company's unique but complementary role.

It is important to emphasize the importance of the system as a unit of analysis when considering the impacts of respon-

[3] https://c2pa.org/

[4] https://gifct.org/tech/

[5] https://www.technologycoalition.org/

sible AI frameworks. In efforts to address climate change, one company selling a fossil fuel asset to another company doesn't address risk but simply moves it from one company to another. Similarly, one AI company choosing not to release a specific AI system or to restrict access to certain capabilities doesn't necessarily result in better outcomes if a different company steps in to fill a void.

Relevance for AI Frameworks: Pursue multi-stakeholder and multi-company strategies that take system-wide and "whole of society" approaches to the responsible design, development, deployment, and use of AI systems. Broad collaborative efforts such as the Partnership on AI[6] (a diverse community addressing the future of AI across topics such as safety, labor, and the media) and more targeted initiatives such as MLCommons[7] (to measure and improve the accuracy, safety, speed, efficiency, and safety of AI technologies), the Frontier Model Forum[8] (to advance AI safety research), and the Coalition for Content Provenance and Authenticity (to develop technical standards for certifying the source and history of media content) will all have important roles to play.

Participation in multi-stakeholder and multi-company approaches will require at least three dimensions: first, the well-resourced, scaleable, and transformative approaches that very large companies can bring; second, the involvement of companies from across whole value chains, such as those that exist in financial services, retail, healthcare, and the public sector; and third, "big tent approaches" that remove barriers to the participation of smaller and less well-resourced companies and organizations, such as cost, time zones, and location. These approaches should also prioritize the meaningful participation of stakeholders that impact and are impacted by AI, not just companies.

Collaboration will also be needed outside of formal multi-stakeholder and multi-company initiatives, such as the pre-competitive sharing of threat intelligence, actionable insights, and best practices between companies and other actors across AI value chains on priorities of shared concern.

Insight #5 – Risk assessment and due diligence should be an ongoing priority, not a moment-in-time activity

The logical step-by-step process of risk assessment (e.g., identify, prioritize, mitigate, track, report, remedy) provides a useful framework and formal record for review (OECD 2018). However, risk assessments take considerable time to work through and do not coexist well with the rapid, responsive, and more organic nature of AI model, product, and service development. Assessments can provide a useful moment-in-time baseline but risk attracting too much attention and often creating the false impression of representing a timeless and authoritative statement of the truth.

Alternative approaches include establishing a formal register of priority risks to people and society that is updated over time, reviewing risk mitigations regularly (e.g., quarterly or annually), and undertaking regular dialogue with external stakeholders with insight into the current societal context. Maintaining effective processes for people to make complaints and secure access to remedy when they have been victims of harm also provides insight into how risks may evolve over time. These ongoing approaches to risk may also be more practical, effective, and scaleable for smaller companies that are earlier in their growth trajectory but taking on global significance very quickly. Ongoing risk assessment may be especially well-suited to situations where novel products are constantly evolving, insights are continually changing, and investment decisions are being made rapidly.

There is also a human element, ensuring that experts with insight into specific impacts are embedded at key points in the product development lifecycle, and that these decisions and their documentation can complement existing process (Winecoff and Bogen 2025).

Relevance for AI Frameworks: Take an approach to due diligence (i.e., assessing, addressing, tracking, and reporting risk) that is embedded into other business processes and operations (e.g., product design, development, and launch) rather than disconnected from it, and which relies on having the right mix of skills and experience engaged at every stage. When due diligence is undertaken on major cross-cutting issues, ensure the results are widely disseminated and communicated to the relevant teams. View due diligence as a value-adding activity that enhances product quality and launch, rather than a narrow exercise in regulatory compliance.

Insight #6 – Reporting and disclosure of AI risks by companies are evolving

Companies can draw upon over 100 years of financial reporting experience and more than 20 years of sustainability reporting experience, but less than five years of reporting on responsible AI—so it is unsurprising that high-quality reporting of AI risks by companies remains a work in progress (Arize 2024; Ozdemir and Ludena 2024). The rapid evolution of AI and its associated impacts, risks, and opportunities adds to this challenge.

Priorities include the development of quantitative and qualitative AI disclosures that are decision-useful for investors, regulators, civil society organizations, academics, and users (Bommasani et al. 2025). This means the development of disclosures that are comparable longitudinally within a single company (to demonstrate progress over time), comparable across many companies (to enable benchmarking and competitive analysis), and interoperable across sectors and systems (so that companies can be analyzed as a collective whole–see point 4).

Progress is also needed to develop the right combination of "numbers and narrative" in reports, given that quantitative data alone (such as model performance, the volume of content removed, or the number of appeals received) are not necessarily indicative of good or bad performance without further insights that provide context and meaning. Today's *transparency reports* on how the policies and actions of gov-

[6]https://partnershiponai.org/

[7]https://mlcommons.org/

[8]https://www.frontiermodelforum.org/

ernments and companies affect privacy, security, and access to information are full of numbers but can lack accompanying analysis to place the numbers in context or interpret what they mean for different stakeholders.

It is also important to distinguish between disclosures about specific AI models, products, and systems (e.g., model cards) from disclosures about an entire company's approach (e.g., responsible AI reports) and between reports about a company's own actions (e.g., content moderation) and reports about actions required of a company (e.g., government demands), and where these artifacts may be in conflict (Kawakami, Wilkinson, and Chouldechova 2024).

Fortunately, reporting on responsible AI at a company level can use well-established standards for sustainability reporting that benefit from a common architecture across voluntary reporting standards (such as the Global Reporting Initiative and International Sustainability Standards Board) and regulated requirements (such as the European Sustainability Reporting Standards) (Saltman 2022). Building upon a decade of experience with the Task Force on Climate Related Financial Disclosures, these standards now deploy a four-part structure (governance, strategy, risk, metrics and targets) that can be redeployed for responsible AI. While additional AI-specific guidance may be beneficial, existing reporting standards can be used by developers and deployers of AI systems to (a) disclose the impacts of their AI systems on people, society, and the environment and (b) describe to investors the risks and opportunities of AI systems for the company's business model, strategy, and operations. These well-established standards for sustainability reporting are also well-aligned with the NIST AI Risk Management Framework.

Relevance for AI Frameworks: Develop guidance, best practices, and/or recommendations for how existing reporting standards (such as the International Sustainability Standards Board and European Sustainability Reporting Standards) can support company-level disclosure relating to the responsible design, development, deployment, and use of AI. These company-level disclosures should meet the information needs of investors, regulators, and other stakeholders and complement, rather than replace, disclosures made about specific AI models, products, and systems. They should be connected with (or embedded into) existing company reporting about impacts, risks, and opportunities relating to social and environmental matters, including mainstream public financial filings and voluntary/regulated sustainability reports.

Insight #7 – Meaningful and effective stakeholder engagement requires significant investment in scale and specialist skills

The concept of stakeholder engagement appears in many regulatory requirements, such as the DSA and the EU AI Act, and voluntary frameworks, such as the OECD AI Principles; however, it is much easier to state a commitment to stakeholder engagement in theory than it is to implement a commitment to stakeholder engagement in practice. Four prominent challenges to *meaningful and effective stakeholder engagement* are scale, skill, expertise, and experience.

The deployment of AI can touch upon any topic for any person anywhere in the world, and its impact can vary significantly according to context. The breadth of stakeholder engagement required for informed decision-making can exceed that of many other industries (Deshpande and Sharp 2022). Investment commensurate with this scale—across intersecting dimensions like geography, issue, and community—is needed for companies to develop AI services that significantly improve the lives of as many people as possible.

However, even if an investment is made in scale, considerable specialist skills are required in methodology, approach, and execution for stakeholder engagement to be conducted effectively and achieve desired outcomes. Stakeholder engagement can be messy, human, and complicated, requiring skills in areas such as relationship building, empathy, problem-solving, negotiation, conflict resolution, and strategic analysis (Taylor 2024). Various issues of research ethics need to be addressed, such as knowledge asymmetry and participant privacy, safety, and security (European Center for Not-for-Profit Law 2023). Informed dialogue in the field of AI also requires that diverse skills in the disparate fields of computer science, social science, business ethics, and public policy be combined. Embedding stakeholder engagement into a risk management process is a strategic priority that requires specialist expertise.

Finally, even if channels for meaningful stakeholder engagement exist, the complex character of AI creates an exaggerated version of the expertise and experience gap that can exist in technology sector stakeholder engagement more broadly. Specifically, effective engagement should enable meaningful participation by people with a wide range of expertise and experience with computer and data science and adjust for knowledge, power, and resource asymmetries that may impact the engagement. This mirrors similar challenges in environmental justice, where the perspective of those with real lived experience of air pollution, water scarcity, and biodiversity loss can be quite different than those writing environmental policies, laws, and regulations (Bullard et al. 2008; Beckman, Khare, and Matear 2016). This should also involve methods of engagement that go beyond direct interactions and include skills such as public opinion and scaled measurement, allowing data (including where the populace may differ from expert opinions) to be considered (Moreira et al. 2025; Kelley et al. 2021).

Relevance for AI Frameworks: Establish AI frameworks capable of engaging stakeholders meaningfully, effectively, and at scale. This will require investment (i.e., money, time, attention) in the specialist skills and capacity to implement meaningful stakeholder engagement, with sufficient emphasis placed on the perspective of users, civil society organizations, and other experts during risk assessment and management processes.

Further, stakeholder engagement should prioritize those most at risk of becoming vulnerable or marginalized, so one high priority should be the capacity to engage with stake-

holders in markets outside the US and EU, especially the Global South or locations with more linguistically diverse populations and languages spoken by fewer people.

Insight #8 – Stakeholder engagement strategies should not be underestimated or constrained by regulation

It is welcome that emerging technology regulations recognize the importance of stakeholder engagement, but regulating engagement too closely may underestimate the significance of the engagement already undertaken "to run the business" and threaten its quality, authenticity, and impact. There is a risk that regulated stakeholder engagement becomes too distant and atomized from real-life decision-making and the value that meaningful and effective stakeholder engagement generates for both business and society.

Companies are not starting from zero on stakeholder engagement, and there have been many teams—such as those in functions such as human rights, user experience research, content policy, responsible innovation, and trust and safety—that have accumulated years of experience developing relationships with stakeholders and integrating signals back into the company, e.g., (Meta 2024). It is more important for regulation to utilize and build upon this existing practice already undertaken to "run the business" than it is to establish entirely new processes undertaken solely for reasons of compliance; indeed, the limited time, resources, and availability of stakeholders themselves demands this approach (Anderson and Allison-Hope 2024).

Further, meaningful and effective stakeholder engagement requires a multi-track approach, such as closed-door discussions, public processes, and expert dialogue. There is a risk that the documented discipline required for compliance may undermine more creative and free-flowing approaches to dialogue; there is little more constraining for effective dialogue than all stakeholders knowing that the discussion will be recorded, scrutinized, and debated for years to come. Many best practices appropriately apply to meaningful and effective stakeholder engagement (such as using the Chatham House Rule, confidentiality agreements, and informed consent during user research) that prevent the details of the engagement from being shared with third parties.

Finally, it is noteworthy that the jurisdictions currently regulating stakeholder engagement largely do not overlap with the Global Majority markets where the risks of AI may be most significant. This creates a possibility that stakeholder engagement undertaken to achieve compliance in the EU or other Global Minority markets will divert resources away from more important stakeholder engagement efforts elsewhere.

One solution for this conundrum may be to supplement stakeholder engagement undertaken "to run the business" with additional open, public, and community-based stakeholder engagement to support regulatory requirements, such as the recent engagements hosted by the Global Network Initiative[9] and the Digital Trust and Safety Partnership[10] in support of DSA systemic risk assessments. However, it is essential to emphasize that all engagement—both engagement undertaken "to run the business" and engagement undertaken for the narrow purpose of a regulated risk assessment—should contribute to regulatory compliance and "count" as evidence. This holistic approach will require becoming comfortable with the notion that anonymized summaries of engagement and a demonstration of accumulated stakeholder insight over time (rather than engagement undertaken during a single assessment period) is valuable and for companies to be more open about how the outputs of stakeholder engagement are used in practice.

Relevance for AI Frameworks: Provide space for multi-track approaches to stakeholder engagement, such as closed-door discussions, public processes, and expert dialogue, that take place in various settings and deploy a diverse range of stakeholder engagement methods. Take a strategic approach to meaningful and effective stakeholder engagement that connects and combines engagement undertaken for the purpose of AI development and deployment (e.g., user experience research, participatory methods) with engagement undertaken for the purpose of systemic change (e.g., multi-stakeholder initiatives, civil society organizations) and regulation (e.g., compliance requirements), while recognizing the distinct and different purpose of each. Deploy global approaches to stakeholder engagement that prioritize people and locations where adverse impacts may be most severe, rather than locations that are most highly regulated, and seek to remove barriers to participation (e.g., expertise, resources, time).

Insight #9 – Common assumptions about risk management should be challenged

So far, we have made the case that risk assessment methodologies based upon well-established frameworks of responsible business conduct provide a good foundation for responsible AI frameworks. However, there are also some important ways that well-established methodologies can be revised, challenged, or stretched to more effectively address the complex environment of risks arising from downstream technology use (such as the use of AI systems in the criminal legal system or financial services industry, or to generate content), which can be more unpredictable, more complex, and less controllable than upstream risks.

Traditional risk management approaches tend to identify a specific risk, pinpoint the appropriate mitigation for that risk, and then track the effectiveness of that mitigation over time. However, mitigations in the context of risks arising from downstream technology use may also require a more nimble approach.

First, in the context of rapidly evolving technology and the ability of nefarious actors to change tactics, the appropriate mitigations for any specific risk may quickly change over time. In the "arms race" relating to content provenance, product misuse, and AI safety, it may be more important to be agile, responsive, and reactive in risk mitigation than to stick to a more traditional annual cycle of risk and compliance review. Here, approaches drawn from the field of cybersecurity may have a lot to offer the field of responsible

[9] https://globalnetworkinitiative.org/

[10] https://dtspartnership.org//

AI, such as deploying real-time threat intelligence, incident response, and scenario planning as part of strategies founded upon adaptive risk management and continuous monitoring and assessment.

Second, many technology sector characteristics commonly perceived to be *risk drivers* may also be essential *risk controls or mitigations*, such as recommender systems, data analytics, and LLMs. For example, while the DSA rightly highlights recommender systems as a driver of risk (such as by enabling harmful content to go viral), recommender systems are also an essential tool for raising the profile of high-quality content and thereby support a more proportionate approach to addressing harmful content risk than removing content altogether. The traditional risk management framework—that risk drivers and mitigations are distinct from each other—may simply not apply, and last year's risk driver may become next year's mitigation strategy.

Third, in the context of downstream technology use, there are many complex ethical dilemmas to navigate, such as the circumstances under which it may be appropriate to restrict access to general-purpose technology or how to address tensions that exist between rights to freedom of expression, privacy, civic discourse, and safety. The notion that there is a single correct answer to mitigating the risks of AI systems—a case of identifying a known effective mitigation strategy and diligently tracking progress over time—might not always apply. The "right" mitigation strategy may vary significantly by location or context.

Finally, these three factors point to another aspect of risk assessment methodology that may need to be revised in an AI context, namely moving away from a culture of solely "getting to yes" where a new product, service, or feature will be launched and towards a culture where some AI products, services, and technologies should be profoundly rethought or rejected entirely. In the context of AI frameworks, the process should also allow "getting to no."

Relevance for AI Frameworks: Specialists in the field of AI should collaborate with specialists in the field of risk management to develop methodologies, tools, and guidance that are tailored to the context of AI systems, deviate from traditional risk management methodologies where it is appropriate to do so, and evolve based on practical experience. The NIST AI Risk Management Framework provides one example where non-traditional risk management methods could be tested with integrity by experimenting within the guardrails provided by a credible framework. These risk management methods should combine traditional risk management features (e.g., identification of controls and mitigations) with new features that offer not launching or deeply re-thinking a product, service, or feature as a reasonable outcome of the process. Further, these risk management methods and their results should be described in company reports (see point 6) at a level of detail that is decision-useful for both investors and other stakeholders and as evidence of effective governance, strategy, and management.

Insight #10 – The risks and opportunities of AI systems for people and society can be pursued together

There are many ways in which companies have become increasingly well-versed in assessing and addressing risks to people and society, as distinct from risks to the enterprise. However, while methodologies are advancing well, significant progress, and even new approaches and lenses (Seymour et al. 2022), may be needed to establish and maintain a culture of addressing both risks and opportunities for people and society holistically.

In the field of responsible business, it is understood that benefits should not be used to offset harms–this is set out in UNGPs Principle 11, for example. While concluding that benefits significantly outweigh harms can support a decision to proceed, the societal risks arising from new products, services, and technologies should still be appropriately addressed. However, the principle of not using benefits to offset harms (which we believe to be correct) can result in an unfortunate disconnect between efforts to pursue the opportunities of technology to support social benefit and efforts to identify and address risk (Vaughn and Allison-Hope 2021).

On the one hand, there are efforts to develop products, services, and technologies "for good" and address some of society's biggest challenges. However, this can happen without risk professionals at the table, and efforts to explore opportunities for technology to support the realization of human rights can fail to identify risks.

On the other hand, there are increasingly sophisticated risk assessment processes to identify and address some of the most significant adverse impacts of technology. However, this can be disconnected from model, product, and service design and development, with risk professionals playing a game of "catch up" after significant decisions have already been made.

The field of responsible AI can avoid this challenge by adopting more holistic approaches. In the same way that disclosing enterprise risk does not constrain business strategy, so too we need to find ways of combining risk and compliance (i.e., the disciplined achievement of regulatory requirements for AI and mitigation of risk) with technology ambition (i.e., the pursuit of innovation and normative goals aligned with the company mission) in the realm of responsible AI.

Relevance for AI Frameworks: The developers and implementers of AI frameworks should pay attention to both their letter and their spirit. The "spirit" of AI frameworks should ultimately be about managing the impacts, risks, and opportunities of AI systems holistically such that they support the realization of human rights and the achievement of sustainable development; at the same time, the "letter" of AI frameworks should include disciplined policies, processes, and guardrails so that risks are addressed in an effective and timely manner.

This involves focusing on what AI frameworks seek to achieve and keeping their original intent and purpose (i.e., their spirit) at the forefront; it also means improving standards of business conduct (i.e., their letter) to support trust,

safety, and compliance. The two elements—the spirit and the letter—should be viewed as two complementary and mutually reinforcing objectives that are best achieved in combination rather than in competition.

Discussion

Over the past two decades, we have experienced how industry-agnostic and tried-and-tested methods for responsible business retain significant value in the technology industry; at the same time, we have learned to make adjustments to suit the specific characteristics of the technology industry, such as taking nimble approaches to risk management and accounting for the reality that even the smallest companies can very quickly take on global significance.

With insights from the policy and regulatory landscapes, and through our extensive work in regulatory development and response, policy research, and the development and study of AI frameworks, we have distilled ten durable insights that can improve AI frameworks. These insights are valuable to people developing, deploying, and assessing the use of AI frameworks, within research, government, civil society, product development, and more.

- For the creators of AI frameworks — We encourage the review of frameworks in other domains, to borrow mature, long-considered approaches from human rights, risk management, and other areas which offer solid foundations for AI frameworks. We hope the insights here provide direction on how to consider when AI does and doesn't need a new approach.
- For those applying AI frameworks while building AI systems — These insights offer detail on how to interpret and operationalize specific AI frameworks used by one's organization. For example, product development teams may use these insights to improve and fill in underspecified practices and processes in existing frameworks.
- For those assessing AI systems — For researchers evaluating specific AI systems, these insights point to potential weak spots in AI frameworks that are often used to measure deployments or proposed systems. By going beyond what a specific AI-focused framework outlines and leveraging other frameworks and domains, it is possible to more deeply examine AI benefits and risks.

We have seen frameworks across other areas of scholarship (Cranor 2008; Gorski, Iacono, and Smith 2023; Bonneau et al. 2012; Gomez-Barrero et al. 2018; Matthews et al. 2025). These frameworks have allowed practitioners and researchers to share and synthesize their insights, making them more accessible to developers, regulators, advocates, and others. Such frameworks are particularly valuable for emerging technologies and topics, where complex real-world practical challenges may not yet have been widely articulated or discussed. As our academic community continues to share AI frameworks (Floridi et al. 2018; Shelby et al. 2023; Woodruff 2019; Peters et al. 2020; Long and Magerko 2020), we encourage continuing to build on the foundations described above.

AI is characterized by innovation, transformation, and disruption, and these qualities make it reasonable to challenge some of our common assumptions about risk management and responsible business conduct. Yet, we have also experienced how mature methods retain significant value, such as utilizing human rights taxonomies, deploying well-established standards of responsible business conduct, and using reporting standards initially created for a different purpose. We believe that this thoughtful mix of continuity and adaptation—and knowing how and when to deploy each—will support the responsible design, development, deployment, and use of AI systems.

Glossary

AI system: A computer system that uses math and logic to simulate human reasoning and perform various advanced functions. AI systems can learn from data, recognize patterns, make predictions, generate content, or act based on those patterns. The OECD defines an AI system as "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. Different AI systems vary in their levels of autonomy and adaptiveness after deployment." (OECD.AI Policy Observatory 2019)

Coalition for Content Provenance and Authenticity: The Coalition for Content Provenance and Authenticity (C2PA) addresses the prevalence of misleading information online through the development of technical standards for certifying the source and history (or provenance) of media content. C2PA is a Joint Development Foundation project, formed through an alliance between Adobe, Arm, BBC, Intel, Microsoft, and Truepic.

Digital Services Act: A regulation introduced by the European Union (EU) in 2023 to regulate large online services and intermediaries doing business in the EU. Its provisions include assessing and mitigating risk to people and society, increasing transparency, and maintaining user reporting and appeals channels.

Domain: A field or area of expertise or knowledge, for example, sustainability or risk management.

EU AI Act: The EU AI Act came into force in 2024 to regulate the use of AI in the EU and establish obligations for developers and deployers of AI depending on the level of risk from AI. The EU AI Act classifies AI applications into different risk categories of unacceptable risk (e.g., emotion recognition in schools), high risk (e.g., AI in essential public services), and minimal risk (e.g., chatbots). Unacceptable-risk applications are banned, and high-risk applications are subject to strict regulations.

EU Charter of Fundamental Rights: The EU Charter came into force in 2009 to enshrine political, social, and economic rights (i.e., human rights) for EU citizens and residents.

Global Internet Forum to Counter Terrorism: An initiative bringing together the technology industry, government, civil society, and academia to foster collaboration

and information-sharing to counter terrorist and violent extremist activity online.

International human rights instruments: The treaties and other international texts that serve as legal sources for international human rights law and the protection of human rights. For example, they include: The Universal Declaration of Human Rights; the International Covenant on Civil and Political Rights; the International Covenant on Economic, Social and Cultural Rights; the International Convention on the Elimination of All Forms of Racial Discrimination; the Convention on the Elimination of All Forms of Discrimination Against Women; the Convention on the Rights of the Child; the Convention on the Rights of Persons with Disabilities.

Meaningful and effective stakeholder engagement: Stakeholder engagement is the process of interaction between a company and stakeholders. The OECD defines meaningful stakeholder engagement as being characterized by two-way communication based on the good faith of participants on both sides (OECD 2018). Meaningful stakeholder engagement is proactive, responsive, and ongoing and is often conducted before business decisions are made. Effective stakeholder engagement occurs when both the company and the stakeholders involved feel satisfied with how a given engagement or series of engagements was carried out and what was achieved. It is inherently meaningful but also achieves desirable outcomes for both sides.

NIST AI Risk Management Framework: Issued in 2023 by the US Department of Commerce National Institute of Standards and Technology (NIST), this framework is intended for voluntary use and to improve the ability to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems.

OECD Guidelines for Multinational Enterprises: The OECD Guidelines are recommendations addressed by governments to companies on (1) positive contributions to economic, environmental, and social progress and (2) minimizing adverse impacts in areas such as human rights, labor rights, environment, and bribery and corruption. The OECD Guidelines were originally introduced in 1976; the most recent update was issued in 2023.

Risk controls or mitigations: A control focuses on preventing a risk from happening in the first place and aims to reduce the likelihood of a risk event occurring; mitigation focuses on reducing the impact of a risk if it does happen and seeks to lessen the severity of the consequences.

Risk drivers: An event, condition, or factor that increases the likelihood of a risk occurring or worsens the impact of that risk if it does happen. Risk drivers can exist inside companies (e.g., undue pressure to increase sales) or outside companies (e.g., geopolitical conflict).

System: We are using the term system in a broad sense to mean any group of interacting or interrelated elements, such as those within communications (e.g., the internet stack), an industry (e.g., healthcare), a value chain (e.g., retail), or a problem (e.g., terrorism).

Tech Coalition: An alliance of global technology companies who are working together to combat child sexual exploitation and abuse online, including independent research, collective action, and signal sharing.

Transparency reports: These reports provide data that intends to shed light on how the policies and actions of governments and companies affect rights to privacy, security, freedom of expression, and access to information, such as restricting access to content or sharing data with law enforcement agencies. These reports come in various forms and have historically been voluntary disclosures; however, there are a growing number of mandatory disclosure requirements, such as the transparency reports required by the DSA.

UN Guiding Principles on Business and Human Rights: The UN Guiding Principles for Business and Human Rights (UNGPs) set out principles and guidelines for how companies should meet their responsibility to respect human rights, including undertaking due diligence to identify, prevent, mitigate, and account for how they address their impacts on human rights, and providing remedy when they have caused or contributed to adverse human rights impacts. The UNGPs were unanimously endorsed by the UN Human Rights Council in 2011.

UN treaty bodies and special procedures: The human rights treaty bodies are committees of independent experts that monitor the implementation of the core international human rights treaties. The special procedures of the Human Rights Council are independent human rights experts who report and advise on country situations or thematic issues in all parts of the world.

Acknowledgments

We are grateful to Nick Bauer, Marsden Hanna, Melike Yetken Krilla, Joe Misseri, and Danielle Osler for valuable comments on this work. Google provided funding for the first author's work on this project.

References

AI Safety Summit. 2023. The Bletchley Declaration by Countries Attending the AI Safety Summit.

Amazon. 2023. Transform responsible AI from theory into practice.

Anderson, L.; and Allison-Hope, D. 2024. Effective Engagement with Technology Companies: A Guide for Civil Society. BSR.

Arize. 2024. The Rise of Generative AI in SEC Filings.

Baele, S.; Brace, L.; and Ging, D. 2024. A Diachronic Cross-Platforms Analysis of Violent Extremist Language in the Incel Online Ecosystem. *Terrorism and Political Violence*, 36(3): 382–405.

Beckman, T.; Khare, A.; and Matear, M. 2016. Does the theory of stakeholder identity and salience lead to corporate social responsibility? The case of environmental justice. *Social Responsibility Journal*, 12(4): 806–819.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 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*, FAccT '21, 610–623. New York, NY, USA: Association for Computing Machinery.

Bommasani, R.; Klyman, K.; Kapoor, S.; Longpre, S.; Xiong, B.; Maslej, N.; and Liang, P. 2025. The 2024 Foundation Model Transparency Index. arXiv:2407.12929.

Bonneau, J.; Herley, C.; van Oorschot, P. C.; and Stajano, F. 2012. The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes. In *2012 IEEE Symposium on Security and Privacy*, 553–567.

Brodeur, C.; and Achterberg, E. 2025. Innovative Pathways: When and how to use alternative approaches to Human Rights Impact Assessments. *Oxfam*.

BSR. 2025. Fundamentals of a Human Rights-Based Approach to Generative AI.

Bullard, R. D.; Mohai, P.; Saha, R.; and Wright, B. 2008. Toxic wastes and race at twenty: Why race still matters after all of these years. *Environmental Law*, 38(2): 371–411.

Bureau of Cyberspace and Digital Policy. 2024. Risk Management Profile for Artificial Intelligence and Human Rights. *U.S. Department of State*.

Cranor, L. F. 2008. A Framework for Reasoning About the Human in the Loop. In *Proceedings of the 1st Conference on Usability, Psychology, and Security*, UPSEC '08. USENIX Association.

Data and Trust Alliance. 2023. Data Provenance Standards.

Del Campo, A.; Zara, N.; and Álvarez-Ugarte, R. 2025. Are Risks the New Rights? The Perils of Risk-based Approaches to Speech Regulation. *Journal of Intellectual Property, Information Technology and Electronic Commerce*, 16(2).

Deshpande, A.; and Sharp, H. 2022. Responsible AI Systems: Who are the Stakeholders? In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 227–236. New York, NY, USA: Association for Computing Machinery.

Dwivedi, Y. K.; Kshetri, N.; Hughes, L.; Slade, E. L.; Jeyaraj, A.; Kar, A. K.; Baabdullah, A. M.; Koohang, A.; Raghavan, V.; Ahuja, M.; Albanna, H.; Albashrawi, M. A.; Al-Busaidi, A. S.; Balakrishnan, J.; Barlette, Y.; Basu, S.; Bose, I.; Brooks, L.; Buhalis, D.; Carter, L.; Chowdhury, S.; Crick, T.; Cunningham, S. W.; Davies, G. H.; Davison, R. M.; Dé, R.; Dennehy, D.; Duan, Y.; Dubey, R.; Dwivedi, R.; Edwards, J. S.; Flavián, C.; Gauld, R.; Grover, V.; Hu, M.-C.; Janssen, M.; Jones, P.; Junglas, I.; Khorana, S.; Kraus, S.; Larsen, K. R.; Latreille, P.; Laumer, S.; Malik, F. T.; Mardani, A.; Mariani, M.; Mithas, S.; Mogaji, E.; Nord, J. H.; O'Connor, S.; Okumus, F.; Pagani, M.; Pandey, N.; Papagiannidis, S.; Pappas, I. O.; Pathak, N.; Pries-Heje, J.; Raman, R.; Rana, N. P.; Rehm, S.-V.; Ribeiro-Navarrete, S.; Richter, A.; Rowe, F.; Sarker, S.; Stahl, B. C.; Tiwari, M. K.; van der Aalst, W.; Venkatesh, V.; Viglia, G.; Wade, M.; Walton, P.; Wirtz, J.; and Wright, R. 2023. Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71: 102642.

European Center for Not-for-Profit Law. 2023. Framework for Meaningful Engagement.

European Parliament. 2000. *Charter of fundamental rights of the European Union*. Office for Official Publications of the European Communities.

European Union. 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). *Official Journal of the European Union*.

European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). *Official Journal of the European Union*.

Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; Schafer, B.; Valcke, P.; and Vayena, E. 2018. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28: 689–707.

Freed, D.; Bazarova, N. N.; Consolvo, S.; Han, E. J.; Kelley, P. G.; Thomas, K.; and Cosley, D. 2023. Understanding Digital-Safety Experiences of Youth in the U.S. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery.

G7. 2023. Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems.

Gillam, C.; and Jain, D. 2024. Responsible AI Playbook for Investors. *World Economic Forum*.

Gomez-Barrero, M.; Galbally, J.; Rathgeb, C.; and Busch, C. 2018. General Framework to Evaluate Unlinkability in Biometric Template Protection Systems. *IEEE Transactions on Information Forensics and Security*, 13(6): 1406–1420.

Gorski, P. L.; Iacono, L. L.; and Smith, M. 2023. Eight Lightweight Usable Security Principles for Developers. *IEEE Security & Privacy*, 21(1): 20–26.

Guest, G.; Bunce, A.; and Johnson, L. 2006. How Many Interviews Are Enough? An Experiment with Data Saturation and Variability. *Field Methods*, 18(1): 59–82.

Kawakami, A.; Wilkinson, D.; and Chouldechova, A. 2024. Do Responsible AI Artifacts Advance Stakeholder Goals? Four Key Barriers Perceived by Legal and Civil Stakeholders. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 670–682.

Kelley, P. G.; Yang, Y.; Heldreth, C.; Moessner, C.; Sedley, A.; Kramm, A.; Newman, D. T.; and Woodruff, A. 2021. Exciting, Useful, Worrying, Futuristic: Public Perception of Artificial Intelligence in 8 Countries. In *Proceedings of*

the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21, 627–637. New York, NY, USA: Association for Computing Machinery.

Long, D.; and Magerko, B. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 1–16. New York, NY, USA: Association for Computing Machinery.

Matthews, T.; Bursztein, E.; Kelley, P. G.; Kissner, L.; Kramm, A.; Oplinger, A.; Schou, A.; Sleeper, M.; Somogyi, S.; Szostak, D.; Thomas, K.; Turner, A.; Woelfer, J. P.; You, L. L.; Zahorian, I.; and Consolvo, S. 2025. Supporting the Digital Safety of At-Risk Users: Lessons Learned from 9+ Years of Research & Training. *ACM Transactions on Computer-Human Interaction*, 32(3): 1–39.

Meta. 2024. Guide for conducting inclusive stakeholder engagement.

Microsoft. 2018. Principles and Approach.

Mitts, T. 2021. Banned: How Deplatforming Extremists Mobilizes Hate in the Dark Corners of the Internet.

Moreira, G.; Bogucka, E. P.; Constantinides, M.; and Quercia, D. 2025. The Hall of AI Fears and Hopes: Comparing the Views of AI Influencers and those of Members of the U.S. Public Through an Interactive Platform. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery.

NIST. 2023. Artificial intelligence risk management framework (AI RMF 1.0).

OECD. 2018. OECD Due Diligence Guidance for Responsible Business Conduct.

OECD. 2023. Common guideposts to promote interoperability in AI risk management. *OECD Artificial Intelligence Papers*, (5).

OECD. 2023. OECD Guidelines for Multinational Enterprises on Responsible Business Conduct.

OECD.AI Policy Observatory. 2019. OECD AI Principles overview.

OpenAI. 2025. How we think about safety and alignment.

Ozdemir, A. C.; and Ludena, M. P. 2024. The Impacts of Digitalization: Identifying emerging challenges and opportunities for sustainability reporting. *Global Reporting Initiative (GRI)*.

Partnership on AI. 2023. PAI's Guidance for Safe Foundation Model Deployment.

Peters, D.; Vold, K.; Robinson, D.; and Calvo, R. A. 2020. Responsible AI—two frameworks for ethical design practice. *IEEE Transactions on Technology and Society*, 1(1): 34–47.

Pichai, S. 2018. AI at Google: our principles. *The Keyword*.

Saltman, E. 2022. Introducing 2022 GIFCT Working Group Outputs. *Global Internet Forum to Counter Terrorism (GIFCT)*.

Seymour, W.; Van Kleek, M.; Binns, R.; and Murray-Rust, D. 2022. Respect as a Lens for the Design of AI Systems. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 641–652. New York, NY, USA: Association for Computing Machinery.

Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Rostamzadeh, N.; Nicholas, P.; Yilla-Akbari, N.; Gallegos, J.; Smart, A.; Garcia, E.; and Virk, G. 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*, AIES 2023, 723–741. New York, NY, USA: Association for Computing Machinery.

Taylor, A. 2024. *Higher Ground: How Business Can Do the Right Thing in a Turbulent World*. Harvard Business Review Press.

The Danish Institute for Human Rights. 2019. Sector-Wide Impact Assessments (SWIA).

The Global Network Initiative. 2023. Taking an Ecosystem-wide Approach to Human Rights Due Diligence in the Technology Sector.

UNESCO. 2024. Ethics of Artificial Intelligence.

United Nations AI Advisory Body. 2024. Governing AI for Humanity.

United Nations B-Tech and Human Rights Office of the High Commissioner. 2020. The UN Guiding Principles in the Age of Technology.

United Nations B-Tech and Human Rights Office of the High Commissioner. 2024. Taxonomy of Human Rights Risks Connected to Generative AI.

United Nations Office for Digital and Emerging Technologies. 2024. Global Digital Compact.

United Nations Office of the High Commissioner. 2011. Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework. *UNHCR*, 1–35.

United Nations Office of the High Commissioner. 2016. A/HRC/32/38: Report on freedom of expression, states and the private sector in the digital age.

Vaughn, J.; and Allison-Hope, D. 2021. The Shared Opportunity to Promote: A Second Decade Priority for the UNGPs. BSR.

Winecoff, A.; and Bogen, M. 2025. Improving Governance Outcomes Through AI Documentation: Bridging Theory and Practice. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery.

Woodruff, A. 2019. 10 things you should know about algorithmic fairness. *Interactions*, 26(4): 47–51.