MIS Quarterly Executive

Volume 24 | Issue 2 Article 4

June 2025

How Stakeholders Operationalize Responsible AI in Data-Sensitive Contexts

Shivaang Sharma

Angela Aristidou

Follow this and additional works at: https://aisel.aisnet.org/misqe

Recommended Citation
Sharma, Shivaang and Aristidou, Angela (2025) "How Stakeholders Operationalize Responsible AI in Data-Sensitive Contexts," *MIS Quarterly Executive*: Vol. 24: Iss. 2, Article 4.
Available at: https://aisel.aisnet.org/misqe/vol24/iss2/4

This material is brought to you by the AIS Journals at AIS Electronic Library (AISeL). It has been accepted for inclusion in MIS Quarterly Executive by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

MIS Quarterly Executive

How Stakeholders Operationalize Responsible AI in Data-Sensitive Contexts

Operationalizing the responsible use of AI in data-sensitive, multi-stakeholder contexts is challenging. We studied how six AI tools were operationalized in a humanitarian crisis context, which involved aid agency decision makers, private technology firms and vulnerable populations. From the insights gained, we identify five types of "AI responsibility rifts" (AIRRs - the differences in subjective expectations, value sand perceived impacts of stakeholders when operationalizing an AI tool in data-sensitive contexts). We propose the self-assessment SHARE framework to mitigate these rifts and provide recommendations for closing the identified gaps.[1,2]

Shivaang Sharma
UCL School of Management (U.K.)

Angela Aristidou
Stanford University (U.S.) and UCL School of Management (U.K.)

The Need to Operationalize Artificial Intelligence Responsibly in Multi-Stakeholder Contexts

A characteristic of today's world, which is marked by unprecedented social polarization,[3] is that long-standing adversaries[4] are forsaking their historical animosities to address the shared challenge of mitigating risks triggered by artificial intelligence. Terms associated with AI risks that once seemed alien to organizational leaders—including training data bias, algorithmic fairness, AI black box and deepfakes—are now part of everyday business and leadership discourse.

Executives and boards of directors are increasingly inundated with advice on how to address AI risks. This guidance typically involves highly technical measures, such as red-

1 Hind Benbya is the senior accepting editor for this article.

2 We are deeply thankful to executives from leading United Nations agencies, International NGOs, private organizations (from the "tech for good" sector), and the Humanitarian AI Today community who contributed their time and valuable insights for the research that informed the insights in this article. We acknowledge the support and funding provided by UK Research Innovation (reference number withheld to maintain the anonymity of research participants). We also thank: 1) Cornell University's Reppy Institute for Peace and Conflict Studies for the input that our work received from academics and practitioners and 2) the Stanford University Centre on Philanthropy and Civil Society and the Stanford Institute for Human-centered Artificial Intelligence for feedback on earlier drafts of this article. All opinions and oversights in this article remain the responsibility of the authors, who acknowledge equal contribution.

3 For a review of the use of natural language processing in research on political polarization, see Németh, R. "A Scoping Review on the Use of Natural Language Processing in Research on Political Polarization: Trends and Research Prospects," *Journal of Computational Social Science* (6:1), April 2023, pp. 289-313.

4 Iyengar, R. *The U.N. Gets the World to Agree on AI Safety*, Foreign Policy, March 2024.

DOI: 10.17705/2msqe.00114

teaming and adversarial training,[5] as well as broader sociotechnical strategies like improving model explainability,[6] tailoring AI applications to specific tasks[7] and ensuring seamless system integrations.[8] We refer to this approach as the "AI risk lens," which consists of recommendations to mitigate AI risks deemed as objectively undesirable, such as AI hallucinations[9] and AI sycophancies.[10] Through this lens, executives are encouraged to manage AI risks by working with technologists, IT teams and technology partners to align AI tools with recommended technical safeguards.

By itself, the AI risk lens may be inadequate for ensuring the responsible operationalization of AI tools in data-sensitive, multi-stakeholder environments, which encompass both AI design and real-world deployment. This limitation became evident in our empirical study of the operationalization of six AI tools during the ongoing humanitarian crisis in Gaza. Humanitarian crises require heightened data sensitivity[11] and bring together diverse stakeholders, including AI developers, designers and end users.

Leveraging our exclusive research access in the data-sensitive context of the ongoing humanitarian crisis in Gaza, we observed that the AI risk lens is valuable for highlighting objective ethical principles and standardized technical policies. However, it frequently overlooks the subjective and varied expectations, values and perceived impacts (both benefits and risks) of the full spectrum of stakeholders in a given AI application context. This oversight revealed by our research is particularly problematic in a humanitarian crisis context where stakeholders may hold starkly different views on critical issues, such as when it is safe to use an AI tool or whether a tool is sufficiently accountable. Failing to address these subjective dimensions risks marginalizing vulnerable and affected stakeholders, particularly nontechnical end users directly impacted by AI tools. Excluding the subjective aspects hinders the operationalization of responsible AI because amplifying the voices of diverse stakeholders is increasingly recognized as a core responsibility of executives and board members in supporting ethical AI practices. Addressing both the objective and subjective aspects of AI issues is therefore essential to foster more inclusive and effective AI operationalization, particularly in data-sensitive, multi-stakeholder AI contexts.

Introducing AI Responsibility Rifts and the SHARE Framework

In this article, we set out an approach to taking account of both the objective and subjective aspects of AI operationalization. Drawing on our research on deploying AI applications during a humanitarian crisis, we propose a complementary approach to the existing AI risk lens. First, based on our empirical findings, we argue that responsibly operationalizing AI requires shifting the focus from AI risks to what we term "AI responsibility rifts." We define AI responsibility rifts (AIRRs) as the misalignments or gaps in stakeholders' subjective expectations, values and perceptions of an AI system's benefits and risks during its design and deployment in application contexts. Stakeholders encompass all individuals, organizations and social groups involved in the operationalization of AI tools, including AI developers, end users, supply

5 For an explanation of these terms, see *Planning Red Teaming for Large Language Models (LLMs) and Their Applications*, Microsoft, May 2025, available at https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/red-teaming.

6 For example, by focusing on the four challenges of AI projects identified in Someh, I., Wixom, B. H., Beath, C. M. and Zutavern, A. "Building an Artificial Intelligence Explanation Capability," *MIS Quarterly Executive* (21:2), June 2022, pp. 143-163.

7 For example, by following the recommendations for tackling generative AI impact on knowledge work in Benbya, H., Stritch F. and Tamm T. "Navigating Generative Artificial Intelligence Promises and Perils for Knowledge and Creative Work," *Journal of the Association for Information Systems* (25:1), 2024, pp. 23-36.

8 See, for example, Abdel-Karim, B. M., Pfeuffer, N., Carl, K. V. and Hinz, O. "How AI-Based Systems Can Induce Reflections: The Case of AI-Augmented Diagnostic Work," *MIS Quarterly* (47:4), December 2023, pp. 1395-1424.

9 Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y. Chen, Y., et al. "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models," *arXiv*, September 2023.

10 Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., et al "Towards Understanding Sycophancy in Language Models," *arXiv*, October 2023.

11 The EU Artificial Intelligence Act (2024) mentions that humanitarian crisis contexts have a greater need for the deployment of responsible AI systems due to heightened data concerns about vulnerable social groups. The act can be accessed at https://artificialintelligenceact.eu/the-act/.

 misqe.org | © 2025 University of Minnesota

chain partners and the broader community.[12] In our study, this includes leading humanitarian organizations—such as United Nations agencies, international nongovernmental organizations, crisis response teams and disaster management teams—together with "tech-for-good" companies and local communities. We introduce the "AI responsibility rifts lens" (AIRR lens) as a necessary complement to the AI risk lens, to ensure a more inclusive and context-sensitive approach, bridging the divides that often arise between technical, objective expertise and the lived, subjective realities of potential stakeholders.

Second, we introduce the SHARE framework—the first framework grounded in AI responsibility rifts. SHARE focuses on five areas—safety, humanity, accountability, reliability and equity—that emerged as key themes during our fieldwork, informed by interviews with multiple stakeholders across the six AI tools we studied. (Our data collection and analysis methods, based on in-depth interviews with executives and other stakeholders - including AI developers and AI end-users - are summarized in the Appendix table). Our empirically derived SHARE framework was further validated through a review of literature on existing principles and responsible AI frameworks, was evaluated by cybersecurity experts, and is being used in the operationalization of AI in the ongoing humanitarian crisis in Gaza.

Designed as a self-diagnostic tool, the SHARE framework supports organizational leaders in identifying and addressing AIRRs during AI operationalization. SHARE offers stakeholders a robust resource to resolve both existing and emerging responsibility challenges associated with AI tools. The self-diagnostic SHARE questionnaire is shown in Table 1.

Organizations can use the SHARE framework to identify gaps in each of the five AI responsibility rifts. The article concludes by providing practical recommendation actions for closing the gaps.

Shifting the Executive Lens from AI Risks to AI Responsibility Rifts

The conventional AI risks lens, which underpins numerous international resolutions on AI safety, continues to prove valuable for organizational leaders in two crucial ways. First, it highlights several objective AI risks including AI hallucinations, AI sycophancies and obvious malicious uses of AI tools through producing false content to undermine democratic processes.[13] It also provides standardized technical approaches and social governance mechanisms that are suited to monitor, assess and minimize these objective risks. However, especially in data-sensitive and multi-stakeholder contexts, the AI risks lens is insufficient to help executives better understand how to address the subjective aspects of operationalizing AI, which we argue is crucial for deploying AI applications in real-world settings (See Table 2).

Subjective Aspects of Operationalizing AI Revealed in Our Study

Our study of six AI tools deployed in the ongoing humanitarian crisis in Gaza, together with data from in-depth interviews (see the Appendix), revealed the cost of overlooking the subjective aspects of operationalizing AI. The six tools are referred to by the pseudonyms "Conflict-NLP," Climate-Predict," "Violence-Predict," "Resilience-Map," "Shelter-Map" and "Hunger-Predict" are described in Table 3.

We found that the operationalization of all six AI tools involved disagreements about subjective expectations, values and perceived impacts, and these disagreements were a significant factor in preventing the use and scaling of the tools. For example, the CEO of the company that developed the Climate-Predict AI tool informed us about disagreements in the operationalization of a computer vision tool for emergency shelter planning: "…one person's ethical violation might appear quite reasonable or maybe even desirable to someone else using the same [AI] tool." Most executives and developers interviewed identified similar points

12 Deshpande, A. and Sharp, H. "Responsible AI Systems: Who Are the Stakeholders," *Proceedings of the 2022 AAAI/ACM Conference on AI*, Ethics, and Society, July 2022, pp. 227-236.

13 Karnouskos, S. "Artificial Intelligence in Digital Media: The Era of Deepfakes," *IEEE Transactions on Technology and Society* (1:3), June 2020, pp. 138-147.

Table 1: SHARE Self-Diagnostic Questionnaire for Identifying AI Responsibility Rifts Among Stakeholders of an AI Tool

SHARE Dimensions of Responsible AI (Rows: AI Risks); (Columns: AI Rifts)	Core questions (all stakeholders except AI developers, please respond)	Comments (all stakeholders please fill in, and assign level of concern)	Core questions (for AI developers)	Comments (from AI developers and executives plus level of concern)
Safety	Do you have any concerns over the AI tool using your input data (i.e., concerns over privacy, ownership, and continuing use of data)? Level of concerns: (select one level for each SHARE dimension:) High level of concerns Concerns are being addressed No perceived concerns		Are adequate data-related safeguards (technical and governance) put in place for the AI tool? (e.g., preventing training-data poisoning, input manipulations, membership inferencing, abuses and misuses)? Level of concerns: (select one level for each SHARE dimension:) High level of concerns Concerns are being addressed No perceived concerns	
Humanity	By using the AI tool, are you experiencing an increase in "distance" from the decision-making context (i.e., concerns about reducing interpersonal relations, inability to relate or understand contexts)? Is the AI tool negatively affecting your sense of meaning and identification with your work?		Does the AI tool offer features that can help users better relate (i.e., "close the distance") with the AI application contexts? (For executives) Are existing initiatives on retraining and re-allocating personnel affected by automation (i.e., those experiencing transitional automation anxieties and aversions) sufficient?	
Accountability	Are you able to clearly assign accountability to outputs generated by the AI tool to specific stakeholders or claim ownership over decisions you made with assistance of the AI tool? (i.e., consider various "entanglements" with and "pervasiveness" of the tool across tasks leading up to the decision)		Are there adequate AI integration monitoring mechanisms in place and are they visible to core users of the AI tool? (i.e., activity logs of human and AI interactions, specific points of human intervention)	

misqe.org | © 2025 University of Minnesota

Table 1: SHARE Self-Diagnostic Questionnaire for Identifying AI Responsibility Rifts Among Stakeholders of an AI Tool (Continuation)

SHARE Dimensions of Responsible AI (Rows: AI Risks); (Columns: AI Rifts)	Core questions (all stakeholders except AI developers, please respond)	Comments (all stakeholders please fill in, and assign level of concern)	Core questions (for AI developers)	Comments (from AI developers and executives plus level of concern)
Reliability	Can you rely on the AI tool (and its outputs—recommendations, classifications) for your task and application context? (i.e., consider contextual sensitivity, false positives, response appropriateness) Does using the AI tool free up your time and resources to devote to other tasks? (i.e., consider your level of trust in AI capabilities)		Does the AI tool have fail-safe features that enable users to override automated decisions or query outputs and logic behind decisions made? (i.e., consider AI explainability mechanisms and tracing presence of biases, or inappropriate weights on variables)?	
Equity	Does the AI tool lack inclusiveness of relevant voices and values in its design and implementation? (i.e., human-centered design, inclusion of marginalized voices, checking for unintended negative effects of the tool in the long run)		Are stakeholders who are likely to be affected by the AI tool substantively involved in the design, development and use of the tool? (i.e., are their voices baked into training datasets and models)?	

Table 2: Examples of How the AI Risks and AI Responsibility Rifts Lenses Highlight Key Issues in Multi-Stakeholder, Data-Sensitive Contexts

AI risk lens: Draws attention to objective AI risks on which stakeholders readily reach consensus—i.e., there is alignment on what risks should be minimized and how to minimize them.	**AI responsibility rifts (AIRRs) lens:** Draws attention to subjective disagreements among stakeholders of an AI tool arising from their differing expectations, values and perceptions of the tool's impact—i.e., lack of alignment due to minor or major disagreements about aspects of safety, and the effects on humanity, accountability, reliability and equity.	
Examples of objective AI risks: • AI hallucinations and deceptions • Sycophantic AI behaviors • Malicious use of AI by bad actors	*Examples of minor disagreements among stakeholders:* • Extent of needed safety guardrails • Data ownership and consent protocols • How representative an AI tool is of the "real" world	*Examples of major disagreements among stakeholders:* • Extent of stakeholder inclusion in AI design • Extent of AI contextual sensitivity • Extent to which AI dehumanizes users

Table 3: The Six Artificial Intelligence Tools Studied

AI Tool (Pseudonym)	Tool Description	Example Tool Use Cases
1. Conflict-NLP	A natural language processing (NLP) platform that contains various auto-tagging and auto-summarization features to generate reports for humanitarian agents.	Humanitarian agents create routine needs assessment reports (i.e., what types of aid are needed such as food and medicines) to trace the growing needs of vulnerable populations in crisis zones.
2. Climate-Predict	A predictive analytics tool based on an early warning system that predicts non-food-related items needed by affected communities in the event of climate change disasters.	Local volunteers use the tool to assess local capacities and non-food stocks for disaster preparedness.
3. Violence-Predict	A predictive analytics-based conflict alert system or early warning system that predicts the incidence and scale of political violence across countries.	Analysis cells develop situational analysis reports to keep track of escalating conflicts and guide on-the-ground teams.
4. Resilience-Map	A computer vision service that detects features (e.g., human habitations) on maps based on open-source images from satellites and unmanned aerial vehicles (UAVs).	Humanitarian "nomads" or virtual mapping teams identify and map blank spots on maps to aid local logistics.
5. Shelter-Map	A computer vision and predictive tool that automates the process of identifying and labeling geographical features and integrates the tool with other NLP tools used by frontline crisis responders.	Government and humanitarian agencies use the tool for modeling disaster risks and human adaptations to hazards.
6. Hunger-Predict	A predictive analytics tool that leverages various quantitative data sets to trace and predict food security in near real time across more than 90 countries.	Originally deployed to trace food insecurity during COVID-19; now adapted for conflict zones.

of disagreement among stakeholders about an AI tool's operationalization, most notably that AI developers may "evangelize" a tool's features while cautious AI users may want to limit the use of the same features.

We also found that disagreements among stakeholders can range from minor disagreements about an AI tool's impact to major disagreements where stakeholders can hold nearly antithetical views of an AI tool's impacts (both risks and benefits) in a given context (see Table 2 above).

An example of a major disagreement among stakeholders was provided by a developer of the Conflict-NLP AI tool. He told us about an incident where an update to the tool's automation capabilities had worked "too well." The update meant that tasks previously performed by humanitarian analysts could now be handled in their entirety by the AI tool instead of going through human data labelers and data collectors trained in the nuances of handling sensitive humanitarian data. This update increased efficiency in performing core tasks and saved data analysts time because the tool could now repeatedly conduct secondary data reviews by following established interagency analytical frameworks. However, the update alienated other user communities. Data taggers felt it downgraded the meaning of their work and were concerned that the automated features could potentially and permanently erode crucial human-to-human relationships—widely considered to be the bedrock of humanitarian aid.[14]

14 Akingbola, A., Adeleke, O., Idris, A., Adewole, O. and Adegbesan, A. "Artificial Intelligence and the Dehumanization of Patient Care," *Journal of Medicine, Surgery, and Public Health* (3), August 2024, pp. 100-138.

Impacts of Overlooking the Subjective Aspects of Operationalizing AI

In our study, we repeatedly encountered disagreements among stakeholders due to misalignments of their subjective expectations, values and their evaluations of the perceived impact of AI tools. We observed significant resources were being dedicated to the management of such disagreements, which would suddenly emerge or become prominent following each new feature update or every novel application extension of an AI tool. Further, we noted that even when stakeholders agreed on addressing specific AI risks, they disagreed on how risks should be evaluated and on ways to develop context-appropriate solutions.

The AIRR lens highlights gaps or misalignments among stakeholders (including AI developers, end users, technologists, executives and vulnerable populations) in their subjective expectations, values and perceived impacts, including both benefits and risks, when operationalizing an AI tool. For example, adopting the AIRR lens can help identify disagreements on the appropriate level of stakeholder representation in providing training data for AI tools and the acceptable level of ethical safeguards.

The AIRR lens emphasizes the importance of paying attention to subjective expectations because, when overlooked, these rifts can have unintended negative effects on stakeholders and result in costly delays in the sustained operationalization of AI tools.[15] Ignoring the rifts can generate significant troubleshooting and retraining costs for the user organization, threaten the reputation of technology evangelists of an AI tool or executives recommending its use for data-sensitive contexts, and result in the loss of an organization's social license to operate in data sensitive contexts due to concerns about negative social impacts.

As shown in Figure 1, the emergence and persistence of AI responsibility rifts is cyclical because AI tools can have differential effects on stakeholders, which in turn can lead to a variance in expectations, values and ethical positions that stakeholders may have about any given AI tool. Such differences may cause stakeholders to either foster or constrain future development of AI based on their subjective experiences and preferences. And if such differences among stakeholders remain unchecked, they can even undermine the sustainable integration and successful scaling of AI tools.

In our study, we found that there were disagreements on responsibly operationalizing a specific AI tool due to different stakeholders having different perceptions of responsible AI.[16] However, we found no meaningful guidance for executives on how to proactively identify and navigate in real-world AI use the differing subjective values, expectations and perceptions of the impacts of AI held by stakeholders. To address this shortcoming, we propose the empirically driven SHARE framework and associated actionable strategies for managing AIRRs.

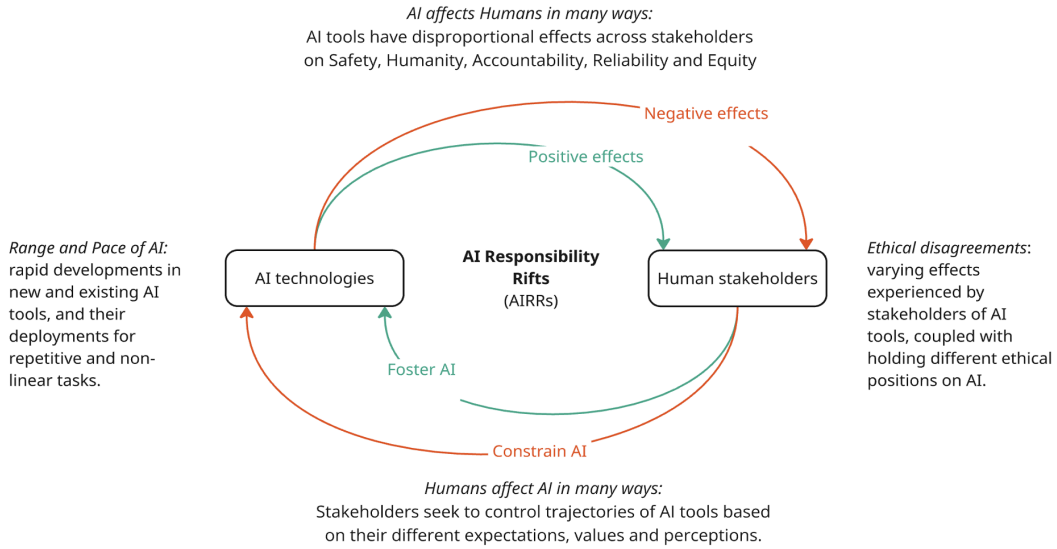
The SHARE Framework for Identifying and Managing AI Responsibility Rifts in Multi-Stakeholder Contexts

Drawing from insights gained from our interviews and further supplemented by our direct engagement with AI tools in a humanitarian crisis context, we identified five specific types of misalignments—i.e., rifts—in the subjective expectations, values and perceived impacts held by different stakeholders. These five rifts underpin the AIRR lens and fall into five areas—safety, humanity, accountability, reliability and equity (SHARE). Paying attention to these five rift areas will enable executives and stakeholders to seamlessly integrate the AIRR lens into practice. We believe that executives should consistently monitor and address these five rift areas to prevent a newly developed or preexisting AI tool from falling into disuse due to ethical concerns.

15 For more on the unintended consequences of AI, see Mikalef, P., Conboy, K., Lundström, J. E. and Popovič, A. "Thinking Responsibly About Responsible AI and 'the Dark Side' of AI," *European Journal of Information Systems* (31:3), February 2022, pp. 257-268.

16 For insights on variance among stakeholders on AI use arising from various organizational policies and standard operating procedures, see *Artificial Intelligence Toolkit*, Interpol Innovation Centre, June 2023.

Figure 1: A Cyclical Model of How AI Responsibility Rifts Emerge and Persist



Safety-Related AI Responsibility Rifts

> *"Where you choose to draw the lines in the sand or cut-off points which make sense in terms of safety really depends on who you talk to and when you talk to them."* NLP engineer, Conflict-NLP AI tool

Since 2023, a plethora of generative AI models have become publicly available, accompanied by a raft of multi-governmental declarations and legislation[17] and transnational resolutions.[18] At the industry level, there has been a burst of multi-stakeholder AI safety-related initiatives.[19] In our study of the deployment of AI tools in a data-sensitive humanitarian context, we also observed that the most attention of AI technologists and executives centers on AI safety.

We found that users and developers of AI tools tend to agree on certain aspects of AI safety, such as the intentional abuse and inadvertent misuse of AI tools, concerns about the continuing consent of training data that draws on vulnerable stakeholders, the implications of co-owning or sharing data on the human rights of specific stakeholders, and the establishment of "red lines" or safeguards to prevent discriminatory outcomes of machine learning models. However, interviews with executives and other stakeholders revealed disagreements on the "finer" operational aspects of AI safety. These disagreements were on issues such as what social considerations, technical considerations and design approaches (e.g., various human-centered design and "human in the loop" or "crowd in the loop" approaches[20]) will ensure the safe design and operationalization of AI tools. We also observed differences among stakeholders about their assessments of whether human rights considerations were being baked into the model-development phase of AI tools and what measures were effective and context-appropriate for establishing "cut-off" safety points.

17 See, for example, *AI Act Enters into Force*, European Commission, August 2024, available at https://commission.europa.eu/news/ai-act-enters-force-2024-08-01_en.

18 See, for example, Mishra, V. *General Assembly Adopts Landmark Resolution on Artificial Intelligence*, United Nations, March 2024, available at https://news.un.org/en/story/2024/03/1147831.

19 See, for example, *Prime Minister Launches New AI Safety Institute*, 2023, U.K. Government press release, November 2, 2023, available at https://www.gov.uk/government/news/prime-minister-launches-new-ai-safety-institute.

20 See Li, W., Wu, W. J., Wang, H. M., Cheng, X. Q., Chen, H. J., Zhou, Z. H. and Ding, R. "Crowd Intelligence in AI 2.0 Era," *Frontiers of Information Technology & Electronic Engineering* (18:1), February 2017, pp. 15-43

An example of a safety-related AIRR was provided by a board member of the company that developed the Conflict-NLP AI tool used for humanitarian crisis monitoring and aid logistics. This board member described the ongoing friction between developers and users (i.e., humanitarian analysts, targeted crowdsourced volunteers and refugee communities) who were troubleshooting a deep disagreement about the way training data was collected for the AI tool. On the one hand, executives, developers and some humanitarian agents pressed for collecting, curating and cleaning vast pools of highly contextual and personalized data to train the AI model because this would enable the tool to make robust recommendations. On the other hand, other user groups, particularly refugee stakeholders, voiced concerns about the "unnecessarily detailed information" that was being gathered from vulnerable stakeholders to feed the AI tool and raised concerns about continuing consent and potential misuse of data that may impact the safety of refugees.

Our study shows that this example is not an isolated incident of friction among stakeholders. Rather, it is representative of similar disagreements encountered by executives who are tasked with encouraging the uptake of AI tools by humanitarian crisis response teams. It is therefore crucial that executives anticipate and prepare for resolving disagreements among core stakeholders about AI safety priorities.

Humanity-Related AI Responsibility Rifts

> *"Our focus should be less on automation and more on humanity. … We need to be mindful if AI is empowering or reducing us to data points, to stereotypes."* Violence-Predict AI tool developer

Given the rapid speed of advancements in AI, understanding the effects of AI tools on the innate humanity of users and on human society may seem nearly impossible. And these effects are expected to become more varied across society as the AI landscape pivots from large language models (LLMs) to large multimodal models (LMMs).[21]

Our study of the early operationalization of generative AI in a crisis context revealed key disparities in how different stakeholders regard and benefit from such AI tools. For example, the first author—acting as a crisis coordinator—asked users (humanitarian analysts) to experiment with a new suite of NLP features provided by the Conflict-NLP AI tool that is designed to save time while still generating critical and timely situational analysis reports.[22] But he encountered a clear divide in the stakeholders. On the one hand, some humanitarian analysts questioned the relevance of such NLP features (sometimes even without using them) and even voiced concerns about "algorithmic enslavement," akin to attention-trapping social media algorithms that make users uncritical of other patterns and perspectives. On the other hand, some users supported the use of new NLP features (e.g., abilities to summarize vast corpora of text and develop automatic chronological events), stating that such AI features enabled them to develop a more nuanced understanding of crisis dynamics and saved them time. Similarly, we learned from our engagement with the other AI tools that such disagreements were anchored in different perceptions of "being and feeling human" and were connected to the effects of AI on users' meaning-making rituals and professional identities.

Other humanity-related AIRRs revolved around the degree to which executives and decision makers should rely on "digital" humanitarian agents, such as the use of AI tools vs. the field knowledge of locally embedded human agents. There were diverse responses to this issue because executives varied in their values and expectations for assessing the appropriateness of aid delivery and in whether the core humanitarian principle of "humanity" was being violated when AI tools were used to make life-saving decisions.[23] Other humanity-

21 Zhang, D., Yang, J., Lyu, H., Jin, Z., Yao, Y., Chen, M. and Luo, J. "CocCot: Contrastive Chain-of-Thought Prompting for Large Multimodal Models with Multiple Image Inputs," *arXiv*, January 2024.

22 For example, see *ECARO Humanitarian Situation Report, Mid-Year, 2024, UNICEF*, available at https://www.unicef.org/documents/ecaro-humanitarian-situation-report-mid-year-2024.

23 See *Humanitarian Principles*, European Commission, available at https://civil-protection-humanitarian-aid.ec.europa.eu/who/humanitarian-principles_en.

related AIRRs were concerned with whether AI tools can have an inadvertent effect of withholding life-saving humanitarian aid. Some stakeholders argued that AI tools can help prioritize specific vulnerable stakeholders in accessing essential life-saving aid such as food, medical supplies and information about safe zones. Other stakeholders claimed the opposite: They said that AI tools can unintentionally deprive specific vulnerable stakeholders from accessing aid due to biases in the training data.

Accountability-Related AI Responsibility Rifts

> *"The lines have become blurred. It is hard to know exactly, and pinpoint, how much AI is involved in helping make decision(s), which was not the case before. ... We are still working on ways to assign accountability."* Conflict-NLP AI tool developer

Issues about AI accountability are especially difficult to address in multi-stakeholder AI operationalization contexts because such contexts involve intertwined interactions among different human stakeholders and AI tools that perform several decision-making tasks. For instance, an executive of the company that developed the Conflict-NLP AI tool mentioned that even if stakeholders agree about AI safety issues (e.g. practices to ensure appropriate levels of data privacy), they may disagree about AI accountability (e.g. assigning credit or blame for decisions and outputs). We learned that these disagreements typically centered on which human or set of humans should be held liable when AI tools make errors or provide inappropriate suggestions. While some stakeholders pushed for assigning blame to the "nearest human" to an automated decision-making output, others argued that the owners of the AI tool or its developers should be held accountable.

We observed that accountability-related AIRRs arise when stakeholders have differing views on how to make decisions more traceable and what type of authority should assess the validity of AI outputs. Another factor is different perceptions about the "power" held by various human agents (and AI tools) involved in accomplishing complex humanitarian tasks (e.g., predicting what types of aid are required in a given region in the next three months). In practice, accountability-related AIRRs in the context of a humanitarian crisis are complicated because several AI tools and human agents are involved in making life-saving decisions. For example, a board member of the company that developed the Conflict-NLP AI tool told us that because humanitarian agents use several AI tools in decision-making (e.g., combining the Violence-Predict and Conflict-NLP tools) and sometimes integrate the tools with other digital repositories, it is challenging to know who (or what tool) to hold accountable for decisions that are sub-optimal or harmful. She also said that though accountability should "start and end" with AI developers and technologists responsible for operationalizing an AI tool, other board members had called for assigning accountability to humanitarian (nontechnical) decision makers who may over-rely on AI tools.

We learned that accountability-related rifts among stakeholders will likely continue due to the absence of a legal agency assigned to AI tools. The executives we interviewed said that the lack of common, updated legal standards for the governance and use of AI tools in multi-stakeholder contexts will likely continue to exacerbate accountability-related AIRRs. Because humanitarian agents are expected to become more deeply involved with using AI tools and because the tools themselves will likely become more pervasive across data-sensitive humanitarian contexts, accountability-related AIRRs will remain an area that requires the urgent attention of executives.[24]

Reliability-Related AI Responsibility Rifts

> *"We may want to scale the opportunities to use AI for the sake of efficiency and also because how resource-strapped we are. But we also have to be very careful about creating safer informational landscapes. ... Disinformation and bad outputs [of AI] will be a big issue."* Executive, Resilience-Map AI tool developer

24 See Kokina, J., Gilleran, R., Blanchette, S. and Stoddard, D. "Accountant as Digital Innovator: Roles and Competencies in the Age of Automation," *Accounting Horizons* (35:1), March 2021. pp. 153-184.

 misqe.org | © 2025 University of Minnesota

As the use of LLM AI models (in particular ChatGPT) continues to grow, we observed that humanitarian personnel were often swayed by news articles from practitioner reports and other reputable sources.[25] Such articles created a wave of surging caution about the perils of relying on AI-generated outputs and even sparked controversy about the rapid adoption of AI tools in humanitarian contexts. We noticed that some humanitarian practitioners were initially averse to experimenting with AI tools for even the simplest of tasks, such as generating daily updates about aid delivery. The extent of this skepticism is evident from a survey conducted by *The New Humanitarian*—the leading practitioner publication for humanitarian crisis response teams. Around 84% of survey respondents indicated being confused by or feeling uncomfortable in adopting AI tools for sensemaking and decision-making tasks.[26]

Our own engagements with the use of AI tools in a humanitarian crisis setting offer deep insights into reliability-related AIRRs. We observed that several stakeholders (such as local humanitarian data analysts and field agents) were cautious of relying (or claimed not to rely) on AI-generated outputs for decision-making. For example, while participating in crisis response teams that used the Hunger-Predict and Violence-Predict AI tools, we observed that humanitarian analysts found these AI tools produced accurate outputs only for very specific humanitarian tasks and in particular contexts only (e.g., in helping predict food and medicine aid requirements for a specific internally displaced refugee community). However, these tools were not deemed reliable for predicting aid requirements of other vulnerable communities in similar crises. On the other hand, other stakeholders who engaged more deeply with the same AI tools in the same context held more positive expectations and perceptions of the impacts. These stakeholders (humanitarian agents) perceived the AI tools as being objective and apolitical, with the AI-created reports or predictive algorithmic outputs from large datasets being seen as impartial depictions of reality. These stakeholders told us that AI tools such as Violence-Predict can efficiently examine vast volumes of data, remove unreliable or biased data sources, and help save time for other critical humanitarian tasks.

Reliability-related AIRRs also centered on stakeholders disagreeing about whether humanitarian agents would become over-reliant on AI due to uncritical usage of AI tools and outputs or would deliberately become under-reliant because of "AI aversion."[27] Stakeholders also had varying views on:

- Whether AI-generated outputs have sufficient cultural sensitivity
- The levels of trust, due to the black-box or unexplainable nature of AI (i.e., lack of transparency)
- Model logics and unexpected AI behaviors, especially in generative AI tools
- Whether humanitarian agents could be susceptible to AI-induced confirmation biases that provide technical legitimacy to otherwise biased or instinctive human decisions.

A relatively well-known case of an accountability-related rift is a disagreement about using AI tools to facilitate communication between humanitarian agencies and the crisis-affected communities they serve. In this specific case, we learned that AI developers pushed for introducing rigorously tested language translation services in the belief that the outputs generated would be accurate and relevant in the local language. But refugees and on-site humanitarian agents cited instances where the tool failed to understand cultural nuances (e.g., because the same local words had multiple, varied contextual meanings). These misunderstandings resulted in awkward communications between "foreign" humanitarian workers and the "native" displaced refugee stakeholders. We also observed several similar cases of reliability-related AIRRs centered on disagreements among stakeholders about the trustworthiness of an AI tool's features, outputs and logics.

25 See Hendrycks, D. "The Darwinian Argument for Worrying about AI," *Time*, June 2023, available at https://time.com/magazine/south-pacific/6284502/june-12th-2023-vol-201-no-21-asia-south-pacific/.

26 Margffoy, M. "AI for Humanitarians: A Conversation on the Hype, the Hope, the Future," *The New Humanitarian*, September 5, 2023.

27 Passi, S. and Vorvoreanu, M. *Overreliance on AI: Literature Review*, Microsoft, June 2022.

Equity-Related AI Responsibility Rifts

> *"... after surveying what other [AI] tools are being trialed during [the] crisis, we asked ourselves are [the tools] deepening pre-existing inequalities or actually helping us positively readjust power imbalances [among vulnerable stakeholders]?"* Design lead, Climate-Predict AI tool

The preceding four types of AIRRs arise at the individual stakeholder level, but equity-related AIRRs[28]—i.e., disagreements about whether AI tools are helping or hindering the quest for social justice in society—transcend all stakeholders. As such, they are challenging to observe and mitigate because the long-term effects of AI applications, at the societal level, tend to gradually unfold over time (i.e., post hoc, after stakeholders have experienced any damage or benefits from AI). Moreover, these AIRRs can only be indirectly experienced by AI developers and only when they continue to interact with stakeholders who experience unintentional negative effects of AI (e.g., not providing life-saving information about safe zones to social groups).

Our observations revealed that equity-related AIRRs tend to be about the disproportional effects of AI tools, where different stakeholders are affected disparately by the same AI tool. These AIRRs can arise from competing views of the appropriate level of stakeholder representation when AI tools are operationalized—i.e., whether an AI tool's training data and output accurately reflects on-the-ground realities, or the values and expectations of relevant stakeholders, and does not perpetuate negative stereotypes of specific stakeholders (e.g., ethnic minorities) or entirely excludes them.

Our interviews—in particular, with humanitarian analysts and refugee communities—revealed that stakeholders had varying expectations and beliefs related to how the algorithmic models used in AI tools were designed. More specifically, whether AI models privileged one set of stakeholders (e.g., a crisis-affected social group requiring food aid) over another (e.g., an overlooked nomadic ethnic minority that needs urgent medical aid but lacks a voice). Humanitarian analysts expressed concerns about whether the long-term use of AI tools would deepen preexisting structural inequalities (e.g., colonial and neocolonial legacies that accumulate power in the hands of typically Western AI developers[29]) or would address the inequalities.

We observed that stakeholders of AI tools tend to widely disagree on interpreting and measuring the equity effects of AI. These disagreements are partly due to the "fuzzy" operationalizable nature of AI equity and partly due to stakeholders having different ethical standards because of their operational context and their degree of interaction with and extent of dependency on AI tools to perform humanitarian tasks. We also learned that the political alignment of developers and end users could strongly influence their disagreements or misalignments with each other on the purpose of AI tools and their impacts on society.

A recent example of such a misalignment that resulted in a temporary hiatus in the use of an AI tool is Gemini AI[30]—Google's generative AI multimodel tool. In this widely publicized "scandal," developers, users and even Google's CEO openly disagreed about the appropriate ways to encode equity into the generative algorithms and where the line should be drawn between equitable representations and historical accuracy.

In another example, the first author was privy to a closed-door panel meeting of board members of the Conflict-Predict AI tool developer. In this meeting, board members disagreed about how their firm's tool was predicting and portraying some countries in terms of their political violence in ways that entrenched neocolonial perceptions of African and Arab nations as unsafe relative to European nations. Technical stakeholders such as AI developers took the view that because the tool is trained on textual data based on specific journalism sources, it could mask political biases against specific stakeholders. Other stakeholders voiced concerns about definitions of fundamental concepts and the weights ascribed to specific

28 See Lin, Y. T., Hung, T. W. and Huang, L. T. L. "Engineering Equity: How AI Can Help Reduce the Harm of Implicit Bias," Philosophy & Technology (34), July 2021, pp. 65-90.

29 See Arora, A., Barrett, M., Lee, E., Oborn, E. and Prince, K. "Risk and the Future of AI: Algorithmic Bias, Data Colonialism, and Marginalization," *Information and Organization* (33:3), September 2023, pp. 1-7.

30 See Pequeño, A. IV. "Google's Gemini Controversy Explained: AI Model Criticized by Musk and Others over Alleged Bias," *Forbes*, February 26, 2024.

misqe.org | © 2025 University of Minnesota

prediction variables in the model driving the AI tool.

By aggregating these examples of equity-related AIRRs, we conclude that these rifts arise from a lack of consensus among stakeholders regarding the long-term, disproportionate effects (risks and benefits) of AI tools on stakeholders over time.

Recommended Actions for Closing AI Responsibility Rifts

To ensure that AI is responsibly operationalized for all stakeholders, it is essential that executives and decision makers consider the five dimensions of the SHARE framework—safety, humanity, accountability, reliability and equity. Above, in Table 1, we have provided a self-assessment questionnaire to enable organizations to identify the emerging or most urgent AI responsibility rifts among stakeholders, which voices should be included in navigating the rifts, and where they arise in the AI pipeline (e.g., in creating training datasets, defining the model architecture driving an AI tool, assigning ownership of AI tools and assessing impacts of AI tools). We now provide recommended actions for closing the identified rifts.

For each type of AIRR, we provide a core recommendation and a caveat that forewarns executives about potential challenges that may hinder the operationalization of the recommendations to close the AIRRs.

Closing Safety-Related AIRRs

Safety-related AIRRs were regarded as the most critical aspect by nearly all of our stakeholder interviewees. AI developers and technologists tended to prioritize swift deployment of AI technologies that measured up to technical benchmarks, while humanitarian analysts, ethics officers in humanitarian agencies and some users pushed for comprehensive safety evaluations and clarifications regarding data usage. Our study identified the following core recommendation and caveat.

Core recommendation: Conduct comprehensive safety evaluations that include nontechnical stakeholders—including end users, beneficiaries and vulnerable stakeholders—who could be affected by the design and operationalization of an AI tool. This approach embeds cultural sensitivities from the start and minimizes the risk of unintended negative effects of AI tools.

Caveat: A significant challenge is managing the inevitable tensions and competing interests of corporate or for-profit groups (the "tech for good" sector) and users who espouse humanitarian values. Private businesses that sell to or assist humanitarian agencies may overstate AI capabilities or have limited understanding of the multilevel and multifaceted concerns of the effects of safely deploying AI technologies in crisis contexts.

Closing Humanity-Related AIRRs

Humanity-related AIRRs—particularly when deploying generative AI applications—can be contentious. Our research indicates that while some experts view AI technologies as a means to enhance human interactions, others are cautious about the perceived loss of human empathy when critical knowledge tasks are delegated to AI technologies. Our discussions with stakeholders of AI tools revealed the following ways to address this polarizing challenge.

Core recommendation: Facilitate participatory AI design[31] processes by involving nontechnical local community members and ethicists in defining data collection and usage protocols. This helps develop AI tools that preserve personal identity, provide a sense of "meaningfulness" in doing work, and ensure that local customs and values are respected.

Caveat: A challenge facing executives is to recognize that the current phase of AI is polarizing—i.e., the politicization of AI and lack of AI literacy implies that humanity-related AIRRs cannot be "permanently" closed by top-down solutions. Executives and AI technologists must therefore ensure that the AI systems "imported" into their sector or organizational workflows do not compromise the preexisting cultural values of users or employees who integrate AI tools in their day-to-day workflows.

31 For an example of participatory AI in humanitarian contexts, see Berditchevskaia, A., Peach, K. and Malliaraki, E. *Participatory AI for Humanitarian Innovation: A Briefing Paper*, Nesta, September 15, 2021, available at https://www.nesta.org.uk/report/participatory-ai-humanitarian-innovation-briefing-paper/.

Closing Accountability-Related AIRRs

Our research indicates that accountability-related AIRRs center on who should be credited (or blamed) for an AI tool's performance. While some stakeholders advocate assigning responsibility to AI developers, others believe accountability should lie with those closest to an AI-related failure. Although no clear solution exists to tackle the AI accountability problem, our engagements with stakeholders revealed the following recommendation and caveat.

Core recommendation: Establish a multi-stakeholder oversight committee that is tasked with enhancing transparency about an AI tool's logic, quality, the contextual relevance of its output and the provenance of training data. By constantly monitoring an AI tool's deployment and future developments, this committee can balance different perceptions of accountability among stakeholders.

Caveat: Executives will face the challenge of striking the right balance between the need for transparency (typically demanded by regulators, civil society organizations and user communities) and the need to protect proprietary information (typically the concern of shareholders, executives and investors). As a result, executives need to find the optimum trade-off between transparency and security. Finding the "Goldilocks zone" between these two crucial considerations needs to be done in conjunction with relevant core stakeholders.

Closing Reliability-Related AIRRs

Our research reveals that stakeholder reliance on AI outputs varies based on their experience with AI tools. For example, we found that first-time users who favorably evaluated AI outputs may become over-reliant, whereas those witnessing early AI failures may develop "algorithmic aversion,"[32] leading to under-reliance despite the tool's proven performance. Based on our empirical observations about reliability-related AIRRs, we recommend the following to mitigate these issues.

Core recommendation: Develop AI auditing practices (externally conducted) and internal features in AI tools that allow users to dynamically adjust the level of AI assistance they receive. These practices and features can be created by keeping an open, real-time, feedback channel between AI developers and a core group of AI tool users.

Caveat: Previous studies have advocated that organizations develop explanation capabilities to unpack the black-box nature of AI technologies, such as developing an AIX (AI explanation capability). In practice, however, due to differing user experiences of AI tools, variance in AI literacy and the inherent complexity of deep learning technologies, the logic behind AI-generated decisions cannot be made fully clear to all stakeholders (proprietary and security concerns may also be a factor here). Executives should therefore consider including bias-awareness exercises in AI training modules for staff and users, which will encourage potential and actual users to reflect on their perceptions of specific AI tools and contextual operationalizations.

Closing Equity-Related AIRRs

Equity-related AIRRs are concerned with the long-term effects of AI technologies on information flows, access to essential social services and amplification of existing inequalities. We identified a dichotomy between stakeholders prioritizing the efficiency-enhancing effects of AI tools and those advocating that AI tools should foster democratic representation. We propose the following recommendation and caveat for closing equity-related AIRRs.

Core recommendation: Create working groups or communities of practice for an AI tool that has historically marginalized stakeholders affected by unintended negative effects of using the tool over time. The insights from these groups or communities can provide crucial long-term foresight to executives; they cannot gain these insights from internal committees.

Caveat: A key challenge facing executives is assessing which stakeholder values and worldviews to prioritize when designing and deploying an AI tool. Moreover, the urgency to deploy AI tools must be weighed against the potential long-term unintended and unexpected societal impacts. The "eternal" ethical dilemma of pitting short-term and organization-focused efficiency against long-term and broader societal implications needs to be managed

32 See Dietvorst, B. J., Simmons, J. P. and Massey, C. "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them," *Management Science* (64:3), March 2018, pp. 1155-1170.

collaboratively by users, developers, regulators and beneficiaries. Executives should be mindful of the often hidden and at times politically motivated, nature of disagreements among stakeholders when managing equity-related AIRRs, especially in data-sensitive contexts. They can expect resistance from stakeholders who may perceive solutions—such as inclusive practices of incorporating voices from "the other side of the sociopolitical spectrum"—as threatening, which may hinder the genuine collaboration needed to close equity-based AIRRs.

Concluding Comments

Operationalizing responsible AI technologies is becoming the new "holy grail" for executives, ethicists, technologists, public policy practitioners and end users, especially in data-sensitive and multi-stakeholder contexts. The continuous flow of reports and guides that adopt an AI risk lens has proved helpful for developing high-level consensus on the common, objective AI risks (such as AI hallucinations and sycophantic behaviors) and how they can be addressed. However, operationalizing AI responsibly remains challenging.

Our empirical study revealed that a continued focus on AI risks masks deep-rooted disagreements among stakeholders regarding their expectations, values and perceived impacts of AI technologies in a given context. We call these subjective aspects of responsible AI "AI responsibility rifts" (AIRRs). Focusing on existing and ever-evolving AIRRs will enable executives to anticipate and identify misalignments among stakeholders of an AI tool. Given the speed of changes in AI capabilities and potential changes in stakeholder expectations toward AI tools, these disagreements will become more prevalent. We believe that by collaboratively working across the user and developer divide on the five types of AIRRs we have identified—safety, humanity, accountability, reliability and equity, which together form the basis of the SHARE framework—organizations can ensure that their AI integrations not only remain scandal-proof but also become paragons of responsible AI.

Finally, rather than adding to the crowded space of ad hoc recommendations on responsible AI, we draw on successful cases of responsible AI deployments in a data-sensitive humanitarian crisis context to identify the best practices that were leveraged to close the AIRR gaps. We provide executives with a core recommendation (and a caveat) for each type of AIRR to guide them in implementing the solutions. We also provide a self-assessment tool (see Table 1) that executives and AI tool stakeholders can use to identify the AIRRs for each SHARE dimension and thus operationalize responsible AI in an inclusive and sustained manner.

Appendix: Data Collection and Analysis Methods

The data in the next table was collected and analyzed during 2023 and 2024.

About the Authors

Shivaang Sharma

Shivaang Sharma (shivaang.sharma.19@ucl.ac.uk) is an adjunct lecturer and Ph.D. candidate at UCL School of Management, England. His research focuses on human-AI collaboration and AI ethics in extreme contexts such as humanitarian crises. As an engaged scholar, he actively collaborates with United Nations agencies and nonprofit organizations to translate research findings into practice and directly participates in the implementation of AI technologies in crisis contexts across Asia and Africa.

Angela Aristidou

Professor Aristidou (a.aristidou@ucl.ac.uk; aaristid@stanford.edu) speaks, writes and advises about the real-life deployment of emerging digital technologies for the public good. Her research spans healthcare, higher education, nonprofit organizations and humanitarian aid in the U.K., U.S., Canada and several Asian countries. Angela's current work on the deployment of AI tools has been honored with a Stanford CASBS Award and a U.K. Research Innovation Award. She specializes in strategy and entrepreneurship at UCL School of Management, is a Fellow at the Stanford Digital Economy Lab and the Stanford Institute for Human-Centered AI, and holds degrees from Cambridge and Harvard.

Data Type	Data Source	Data Collection and Analysis
17 interviews	Six preliminary Interviews with developers for each AI tool. 11 semi-structured Interviews with AI developers and end users, including crisis response staff and executives of firms developing AI tools.	Authors scanned interview excerpts for evidence of disagreements among an AI tool's stakeholders.
Observations (50+ hours)	AI tool training sessions where humanitarian organizations are trained to integrate AI tools into workflows and decision-making. Workshops on AI ethics run by AI developers to explain safeguards and limitations of AI tools.	Authors attended workshops and panel discussions to learn how AI developers and end users discuss whether core humanitarian principles and ethical standards were baked into AI tools—e.g., anonymizing data uploaded about vulnerable social groups.
Participation (20 hours)	Pilot-testing AI tools with humanitarian staff in simulations and a real-world, humanitarian crisis.	Authors joined crisis response teams as a volunteer or coordinator (e.g., using AI tools to estimate vulnerable groups at risk and what types of aid they require).
Archival data (500+ pages)	AI tool websites, blogs, email updates that contain information about new features (e.g., auto-labeling crisis data) and case studies of AI tools deployed during humanitarian crises. Standards and reports on AI ethics that are used by AI developers and end users in humanitarian crisis response.	Authors looked for evidence of how concerns raised by stakeholders on the safety and reliability of AI tools were addressed by AI developers. Authors used existing standards on AI ethics as supplemental data to help triangulate the core dimensions of AI risks.

 misqe.org | © 2025 University of Minnesota