AI RISKS AND REALIGNMENTS

IN HUMANITARIAN CRISIS CONTEXTS

30TH APRIL 2024

ISSUE #1

UCL SCHOOL OF MANAGEMENT

IMMAP - Better Data Better Decisions Better Outcomes

AUTHOR AND INSTITUTIONAL PROFILES

Shivaang Sharma is Adjunct Lecturer and PhD researcher at UCL (University College London) School of Management. His research expertise and practitioner experience is at the intersection of AI and humanitarian crisis contexts.

UCL SCHOOL OF MANAGEMENT

UCL School of Management is a world leading Management education and research school based in London, United Kingdom. The primary focus of the school is on innovation, technology management, sustainability and best practices to address pressing social and environmental challenges.

Guido Pizzini

Business Development, Impact and Partnerships Director, **IMMAP Inc**.

Bertrand Rukundo

Strategic Planning, Technology and Innovation Director, **IMMAP Inc**.

Rishi Jha

Global Communications and Outreach Coordinator, **IMMAP Inc**.

iMMAP

iMMAP Inc. is a leading international nonprofit that delivers information management services to humanitarian and development organizations. As a full member of six humanitarian clusters—Health, WASH, Nutrition, Food Security, Protection, Logistics, and Emergency Telecommunications—iMMAP Inc. supports information value chains, enables decision-making to enhance operational and strategic outcomes in emergency and development contexts.

TABLE OF CONTENTS

INTENDED AUDIENCE

Who

This research report series is primarily intended for organizations that develop, implement, and support the use of AI in humanitarian contexts. This includes inter-agency working groups, research institutes, information systems specialists, AI ethics influencers, and humanitarian funders and donors

Additional resources will be provided in the appendix to help stakeholders better navigate their AI implementation journey.

Why

Humanitarians find themselves in the wake of an active AI race - i.e. various AI technologies are being built or 'imported' - and are beginning to voice concerns over AI washing - i.e. deliberate or inadvertent ethical misalignments of AI technologies with humanitarian values and organizational objectives. This research series aspires to catalyze coordination on data and AI governance in ways that champion core humanitarian values.

What

This report explores current ethical concerns over AI, dissonances over 'the correct' standard for assessing AI risks, and paths towards alignment (human to human, and human to machine) within humanitarian contexts. It offers a multi-stakeholder perspective that bridges numerous high-level AI guidelines with the actual, practical concerns with AI facing humanitarians.

KEY TAKEAWAYS

Funders and Donors

Support initiatives that adopt a holistic approach towards risk management, demonstrate substantive stakeholder inclusion in the AI deployment process, help address (rather than exacerbate) power imbalances between to so-called global north and global south.

Humanitarian personnel

Consider a more rigorous, multifaceted approach to impact and risk while developing AI in-house or importing 'off the shelf' models. Pay particular attention to social considerations (e.g. continuing consent of beneficiaries) with technical considerations (e.g. data strategy).

Information systems specialists

Ensure the implementation of robust cybersecurity measures and data privacy protocols, including anonymization and encryption, to protect sensitive information. Engage in activities that foster AI explainability and AI literacy or upskilling when integrating AI systems.

Humanitarian researchers

Explore how the vast catalogue of concerns over AI risks is prevented and mitigated in different pockets of humanitarian practice. This will help create a coda of best practices and help better understand ethical pluralities on AI.

COMMON DICHOTOMIES OVER AI RISKS

Transparency vs. Opacity

While some stakeholders advocate for designing AI systems with maximum transparency to identify and correct errors and biases, others warn that too much transparency could introduce security risks such as hacking or data poisoning, suggesting a balance between transparency and security is necessary.

Open Source vs. Restricted Access

Some advocate for the democratization of AI through open-source models to encourage innovation and prevent monopolies, while others argue for the regulation of potentially hazardous AI technologies, drawing parallels with controlled fields like IT security and nuclear science to prevent their exploitation.

Human oversight vs. Automation

Some stakeholders emphasize the need for human oversight in AI systems to prevent overreliance, particularly because humanitarian data and insights are sensitive. Others argue that the complexity of AI systems might necessitate a level of trust in automation, despite potential transparency issues, to manage the delicate balance between operational efficiency and security concerns.

Participatory AI vs. Private Sector AI

On the one hand, some support the involvement of the private (tech for good) sector in AI for humanitarian response, arguing that its beneficial for innovation, capacity and financial continuity of humanitarian AI projects. On the other hand, others advocate for participatory AI, involving end-users in development to enhance their understanding and agency, expressing that private sector may overstate capacities, deploy untested models, and other data ownership concerns

INTRODUCTION

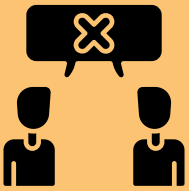
The ***AI risks and realignment*** research series is a multi-stakeholder initiative by researchers and practitioners working at the intersection of AI and humanitarian crisis contexts. This initiative bridges the glaring gap between important, high-level but somewhat abstract guidelines on AI issued by policymakers and researchers, and the ground-level, 'live', ethical and practical concerns with AI and data management experienced by crisis response teams.

This research series is also motivated to address the lack of understanding of a 'how to guide' on AI integration and interoperability with existing data workflows, data compliance, and standards in digital humanitarianism. Previous reports and reviews on AI in humanitarian work, although useful in providing a lay of the technological landscape (what AI technologies are used) or broadly cataloguing AI supercharged concerns (what AI risks exist), offer limited practical and accessible guidance on how humanitarians can seek alignment in managing AI and data risks.

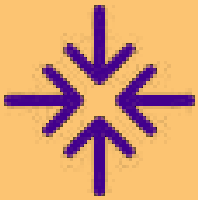
Our research, therefore, explores AI and data integration issues at the intersection of AI risks, dichotomies over AI risks, AI realignments in humanitarian crisis contexts. Broadly defined, AI risks refer to concerns over practical and existential negative effects of AI experienced by stakeholders (i.e. humanitarian field agents, participatory communities, information systems officers, data gatekeepers, etc). Dichotomies over AI risks imply polarizing disagreements amongst stakeholders over their experiences, evaluations and prioritizations of AI risks. AI realignments, in context of humanitarianism, refer to efforts of stakeholders to ensure consistent alignment of AI technologies with core humanitarian values following changes to AI features, information management protocols and AI application contexts.

FROM THE LITERATURE

AI RISKS



DICHOTOMIES OVER AI RISKS



AI REALIGNMENTS

AI RISKS

AI risks are constantly documented and defined in numerous reports, recommendations, and resolutions. In humanitarian contexts, AI risks refer to negative effects – potential or actual, existential, or procedural – on vulnerable social groups and humanitarian agents and their partners (i.e. on all 'stakeholders' of AI tools) during procurement, development, implementation and iterations of AI technologies. In crisis contexts, such negative effects can be categorized as 'high-risk' (see Europe 'AI Act') due to the profound impacts of AI on the 'fundamentals' of individual agency (e.g. AI accountability), our well-being (e.g. AI safety), and on our collective sense of responsible resource distribution (e.g. AI and equity).

In humanitarian settings AI risks include a variety of socio-technical concerns throughout the AI pipeline. These include issues about continuing consent and collating representative training data sets to concerns about lack of model transparency and ownership of AI outputs. Previous work emphasises that such a ubiquitous and ever-evolving spectrum of concerns ought to be constantly monitored, prevented and mitigated. The concerted efforts of humanitarian agencies to catalogue AI risks has led to a consensus over what are the types of significant AI risks in crisis contexts. We categorize these catalogue of AI risks into five meta categories of Safety, Humanity, Accountability, Reliability and Equity.

Figure 1: An illustration of AI risks facing humanitarians

Emerging question 1: ***How can humanitarian agents (particularly crisis response teams and data partners) prioritize AI (and data) risks?***

DICHOTOMIES OVER AI RISKS

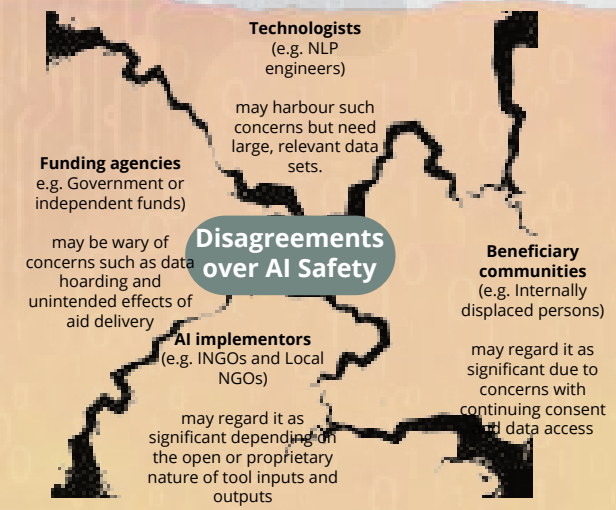
We found that some AI risks are evaluated as 'objectively' problematic (e.g. AI hallucinations, Sycophantic behaviours, various malicious uses of AI). In principle, technologists and humanitarian agents tend to agree over the 'correct ways' to minimize these so-called AI risks (e.g. use data anonymization algorithms, red teaming).

Most AI risks, however, in practice, are evaluated by humanitarian stakeholders as subjective. This often resulting in polarizing perspectives on AI risks. Our analysis of prior research and current recommendations show disagreements amongst stakeholders due to variances in experiencing and using AI. We identified four dichotomies over AI risks on transparency, access, human oversight, and stakeholder participation.

In the transparency debate, some advocate for clarity to boost AI explainability and accountability, while others favor opacity to avoid AI misuse. The open-source vs restricted use argument centers on fostering innovation and preventing monopolies versus avoiding exploitation. The human oversight vs automation discussion weighs human judgement in humanitarian missions against automation efficiencies. Finally, the participatory vs private AI debate contrasts benefits of private sector involvement with the ethical implications of user-inclusive AI development.

Figure 2: An illustration of disagreements amongst stakeholders about AI safety

Emerging question 2: ***How should humanitarian stakeholders that hove varying experiences of AI (and data) risks and over appropriate forms of risk mitigation, develop consensus over the development and use of AI in crisis contexts?***



AI REALIGNMENTS

To prevent and mitigate AI risks and dichotomies over AI, humanitarians must consider the deep interrelations between the technical and social aspects of AI. Previous research statees that disagreements over AI risks emerge when misalignments repeatedly occur between humans and machines and amongst humans. This necessitates continuous engagement amongst stakeholders – policymakers, technologists and users. The quest for AI alignment therefore, is a complex, negotiated, consistent and collective endeavor. For humanitarians, AI misalignments may occur due to the inherent flux of crisis contexts, rapid changes in AI technologies, applications of a seemingly robust tool in a new context, and changes social values and local attitudes towards an AI tool's perceived effects. Misalignments hinder procurement and uptake of AI.

To develop a roadmap towards realignments, existing literature provide various principles and frameworks. These instruments can be applied depending on the unique combination of core human values, stakeholder expectations and AI properties specific to application contexts (e.g. see UNESCO AI ethical impact assessment). Although most of these guidelines are being gradually imported into routine humanitarian AI applications, humanitarians are yet to develop a normative standard on AI and ways to document best practices for AI realignments.

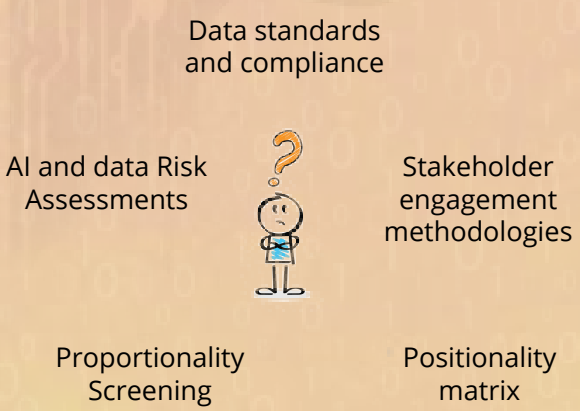
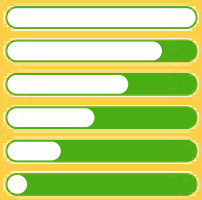


Figure 3: Humanitarian staff are inundated by a plethora of AI (and data) guidelines, standards and frameworks

Emerging question 3: ***How can humanitarians leverage several pre-existing instruments to ensure continuous alignment on AI and data ethics?***

FROM THE FIELD



PRIORITIZING AI RISKS

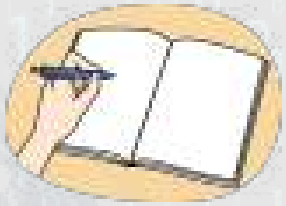
CONSENSUS ON AI RISKS

INTEGRATING GUIDES FOR AI REALIGNMENTS

LESSONS FROM iMMAP INC.

Emerging Question1: How should humanitarian agents (particularly crisis response teams and data partners) prioritize management of AI (and data) risks?

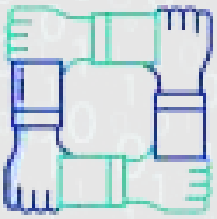
In the dynamic and often chaotic environment of emergency response, the management of AI risks is a critical concern that must not be overlooked. Humanitarian agents, particularly crisis response teams and data partners, should prioritize this by equipping themselves with **guidance notes** that provide clear instructions on the best practices and potential pitfalls of using AI during crises. These notes should be an integral part of training and onboarding to ensure that responders are well-prepared to make informed decisions. Additionally, the creation of **deployable toolboxes** containing AI tools and templates specifically designed for crisis scenarios can greatly enhance the efficiency of emergency operations. These toolboxes should be crafted to facilitate the seamless integration of AI solutions without compromising risk management protocols. Investment in **anticipatory action** is also vital, as it allows for the pre-emptive identification and mitigation of potential crises through the use of AI, such as flood detection and damage assessment. Advocating for increased funding in this area is essential to harness the full potential of AI-driven anticipatory measures. Lastly, **leadership training** is crucial to bridge the understanding gap between technical experts and decision-makers, ensuring that AI is responsibly integrated into humanitarian projects from the outset.



Equip guidance notes: Best practices and pitfalls.

Create deployable toolboxes: designed for various contexts.

Invest in anticipatory action: preparedness and awareness.



Collaborative leadership: bridge technical staff and decision makers.

LESSONS FROM iMMAP INC.

Emerging Question 2: How should humanitarian stakeholders holding dichotomic experiences of AI (and data) risks develop consensus over using AI during crisis?

Achieving consensus among humanitarian stakeholders on the development and use of AI in crisis contexts is a complex task, given the diverse experiences and perspectives on AI and data risks. To navigate this challenge, stakeholders must establish a **dedicated forum** for open dialogue, where they can share insights and resources, fostering a spirit of cooperation and mutual understanding. This collaborative space is crucial for transcending individual agendas and aligning efforts around the common goal of saving lives. By focusing on this shared objective, organizations can work collectively towards more effective crisis response. Operationalizing this approach involves **leveraging existing collaboration platforms** such as Technical Working Groups and Humanitarian Coordination Cluster systems. These established networks facilitate the integration of AI as a complementary tool within familiar frameworks and promote the development of joint policies and capacity-building initiatives that encourage the responsible use of AI across all stakeholder organizations. It is important to not view AI as a standalone solution but one that involves various **diverse voices** across the implementation pipeline. AI can be seen as an enabler for existing collaborative structures, which is essential for overcoming differences in AI risk perceptions and harnessing its potential (and related technologies) to enhance crisis response efforts.

Dedicated knowledge communities:
openly share concerns and resources on data and AI.

Coordinate with institutionalized working groups:
inter-agency clusters, associations.

Seek out various voices:
consider differential impacts of AI.

LESSONS FROM iMMAP INC.

Emerging Question 3: How can humanitarians select and combine existing instruments to ensure continuous alignment on AI and data ethics ?

Humanitarians are often confronted with a plethora of frameworks and guidelines that can guide ethical AI and data practices. To ensure continuous alignment on AI and data ethics, it is crucial to strategically leverage these existing instruments while remaining steadfastly **aligned with humanitarian principles**. The rapid evolution of technology, particularly in the realm of AI, poses a significant challenge to the traditional policy development process. To address this, humanitarians must adopt a proactive stance, continually monitoring and **adapting existing instruments** to reflect the latest ethical considerations in AI and data usage. This requires a commitment to ongoing review and revision of policies to ensure their relevance and effectiveness in safeguarding humanitarian values amidst technological advancements. Collaboration and knowledge-sharing among diverse stakeholders are paramount in this process. Engaging with experts from various sectors enriches the understanding of ethical considerations and enhances the robustness of governance frameworks. By **harnessing the collective wisdom** of stakeholders, existing instruments and remaining agile in response to emerging ethical challenges, humanitarians can maintain the highest standards of ethical conduct in their use of AI and data. This will help ensure that their actions are firmly rooted in compassion, integrity, and respect for human rights.

Alignment with humanitarian principles: prevents 'mission drifts' in partnerships with other parties.

Update existing sandards: to maintain relevance and aspire to 'future-proof' AI guidelines.

Harness collective intelligence: seek out lessons learnt from non-humanitarian sectors.

SELECT REFERENCES

'AI Act', European Commission, 13 March 2024, Accessed 13 March 2024 https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law

'Artificial Intelligence Toolkit', INTERPOL, February 2024, Accessed 20 March 2024, https://www.interpol.int/en/How-we-work/Innovation/Artificial-Intelligence-Toolkit

Beduschi, A. (2022). Harnessing the potential of artificial intelligence for humanitarian action: Opportunities and risks. International Review of the Red Cross, 104(919), 1149-1169.

Cronin, M. A., & George, E. (2023). The why and how of the integrative review. Organizational Research Methods, 26(1), 168-192.

'Digital Dehumanisation' StopKillerRobots.org, Accessed 2 January 2024 https://www.stopkillerrobots.org/stop-killer-robots/digital-dehumanisation/

'Ethical impact assessment', UNESCO, September 2023, Accessed 12 February 2024 https://unesdoc.unesco.org/ark:/48223/pf0000386276#:~:text=As%20stated%20in%20article%2050,monitoring%20measures%2C%20among%20other%20assurance

'FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence', The White House US Government, 30 October 2023, Accessed 2 January 2024 https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence

Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and machines, 30(3), 411-437.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), 627-660.

SELECT REFERENCES

'Health Equity', World Health Organization, Accessed 2 January 2024 https://www.who.int/health-topics/health-equity#tab=tab_1

Iyengar, R. 2024. 'The U.N. Gets the World to Agree on AI Safety', Foreign policy, Accessed 20 March 2024 https://foreignpolicy.com/2024/03/21/un-ai-regulation-vote-resolution-artifical-intelligence-human-rights/

'Joint Statement on AI Safety and Openness', Mozilla, 31 October 2023, Accessed 20 March 2024 https://open.mozilla.org/letter/#:~:text=We%20are%20at%20a%20critical,%2C%20transparency%2C%20and%20broad%20access

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., ... & Gao, W. (2023). Ai alignment: A comprehensive survey. arXiv preprint arXiv:2310.19852.

Kanter, B, Fine, A, Deng, P., '8 Steps Nonprofits Can Take to Adopt AI Responsibly', 7 September 2023, Accessed 27 January 2024 < https://ssir.org/articles/entry/8_steps_nonprofits_can_take_to_adopt_ai_responsibly

'MODEL CARD: DRC Foresight Model', United Nations Office for the Coordination of Humanitarian Affairs, September 2020, Accessed 25 October 2023, https://data.humdata.org/dataset/2048a947-5714-4220-905b-e662cbcd14c8/resource/be6ab2c8-f3c4-4045-9acf-529f6091c253/download/drc-model-card.pdf

Milaninia, N. (2020). Biases in machine learning models and big data analytics: The international criminal and humanitarian law implications. International Review of the Red Cross, 102(913), 199-234.

Naughton, J., 'ChatGPT exploded into public life a year ago. Now we know what went on behind the scenes', 9 December 2023, Accessed 10 March 2024, https://www.theguardian.com/commentisfree/2023/dec/09/chatgpt-ai-pearl-harbor-moment-sam-altman

Motalebi, N, Verity, A., 2023. 'Generative AI for Humanitarians', September 2024, Accessed January 2024 https://digitalhumanitarians.com/generative_ai_for_humanitarians-september_2023/

SELECT REFERENCES

Paulus, D., de Vries, G., Janssen, M., & Van de Walle, B. (2023). Reinforcing data bias in crisis information management: The case of the Yemen humanitarian response. International Journal of Information Management, 72, 102663.'

Passi, S., & Vorvoreanu, M. (2022). Overreliance on ai literature review. Microsoft Research.

Policy implications of artificial intelligence (AI)' UK Parliament Post, 9 January 2024, Accessed 20 March 2024 https://researchbriefings.files.parliament.uk/documents/POST-PN-0708/POST-PN-0708.pdf

'Recommendation on the Ethics of Artificial Intelligence', UNESCO, 16 May 2023, Accessed 20 March 2024 https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

Someh, I., Wixom, B. H., Beath, C. M., & Zutavern, A. (2022). Building an Artificial Intelligence Explanation Capability. MIS Quarterly Executive, 21(2)

'The Bletchley Declaration by Countries Attending the AI Safety Summit', UK Department for Science, Innovation and Technology, 1 November 2023, Accessed 2 January 2024, https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

Vanneste, B. S., & Puranam, P. (2024). Artificial Intelligence, Trust, and Perceptions of Agency. Academy of Management Review.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., & Chen, Y. (2023). Siren's song in the AI ocean: a survey on hallucination in large language models. arXiv preprint arXiv:2309.01219

APPENDIX

Appendix A: Relating core humanitarian principles with concerns over AI responsibility.

Principle	Definition	Related AI concerns
Leave no one Behind (LNOB)	A central, transformative promise of all humanitarian agencies that implies no social group should be ignored while providing aid	E.g. Algorithmic bias. AI prioritizes aid to certain groups over others based on biased data.
Do no Harm	Not causing harm – intended or inadvertent – to any social group during operations by humanitarians themselves.	E.g. Data Privacy. Personal data of affected groups is exploited by bad actors
Neutrality (or Independence)	Position held by all humanitarian agencies (and their representative agents) is not to favor any specific 'side' during crisis.	E.g. Biases in training data and model. AI decision-making influenced by agendas and favours.
Anonymity	To ensure the personal and social identities of vulnerable social groups is protected and not shared with potentially malicious actors.	E.g. Cybersecurity concerns. Biometric data collected by AI is hacked, exposing identities of refugees.
Transparency	To openly share and communicate all aspects of humanitarian processes with the general public and stakeholders.	E.g. Explainability. Lack of clarity about model logic induces algorithmic aversion,
Empowerment	Enable marginalized social groups to partake in society, make them more resilient and to build capacities to uplift.	E.g. Power Imbalance. AI tools may amplify and deepen existing inequalities in society.

APPENDIX

Appendix B: Resources for humanitarians on responsible AI and data practices.

Resource	Initiative by	Description
DHN publications repository	Digital Humanitarian Network (DH Network)	Informs humanitarians on topics such as generative AI, digital identity, managing remote teams, chatbots, UAVs etc.
The Centre for Humanitarian Data (Centre for humdata)	Centre for Humdata, UN-OCHA.	A UN family centric data repository on various data workstreams (learning, practice, responsibility),
IOM Global Data Institute	International Organization for Migration (IOM)	integrates data insights from DTM and GMDAC to enhance operational strategies and understand global migration trends
HDX repository	The Humanitarian Data Exchange repository by UN-OCHA services	An open data-sharing platform that simplifies the discovery and analysis of humanitarian data.
Humanitarian AI today Podcast and Community	Humanitarian AI today	Podcast series on the latest AI and technology iniatives, challenges and applications in humanitarian settings.
The Alan Turing Institute Learn and Apply Skills	The Alan Turing Institute	Open-source resources and training to foster responsible and ethical data science and AI practices

Also see: The EU AI Act development tracker, HXL Repository, Microsoft AI Hub.

APPENDIX

Appendix C: A note on literature review methodology for the research series.

The summarized literature review section of this report series (i.e. *AI Risks and Realignments*) adopts a rigorous integrative review approach (see here). The advantage of this approach over alternative review methodologies adopted in prior work (e.g. systematic or thematic reviews) is that it provides a practical and theoretically meaningful analytic framework to help practitioners and researchers orient their future work. The review section produces three questions that are important for humanitarian AI applications and have been insufficiently explored in academic and grey literatures. This framework (i.e. to assess AI risks, dichotomies over AI, realignments) is based on qualitative analysis of 112 recent (2018-2024) academic publications, reports, laws, resolutions and other grey literature. Theses qualitative data sources were crowdsourced from humanitarian personnel who are directly involved in developing and using AI technologies. The reviewed data sources include all forms of AI technologies i.e. generative AI, predictive analytics, computer vision, other forms and applications of deep learning that are being tested and scaled. These include AI technologies that are 'home grown' within humanitarian sectors or are imported into crisis contexts.

The empirical analysis (i.e. *Lessons from IMMAP Inc.*) section of the report series is based on interviews and surveys with leading humanitarian agencies and practitioners working in crisis contexts. The questions emerging from the summarized integrative literature review are posed to humanitarian practitioners involved in this research series (or are juxtaposed with real-world and up to date concerns that practitioners continue to have with AI and data ethics).

UCL SCHOOL OF MANAGEMENT | iMMAP

SCAN QR code for feedback
or to contact lead author