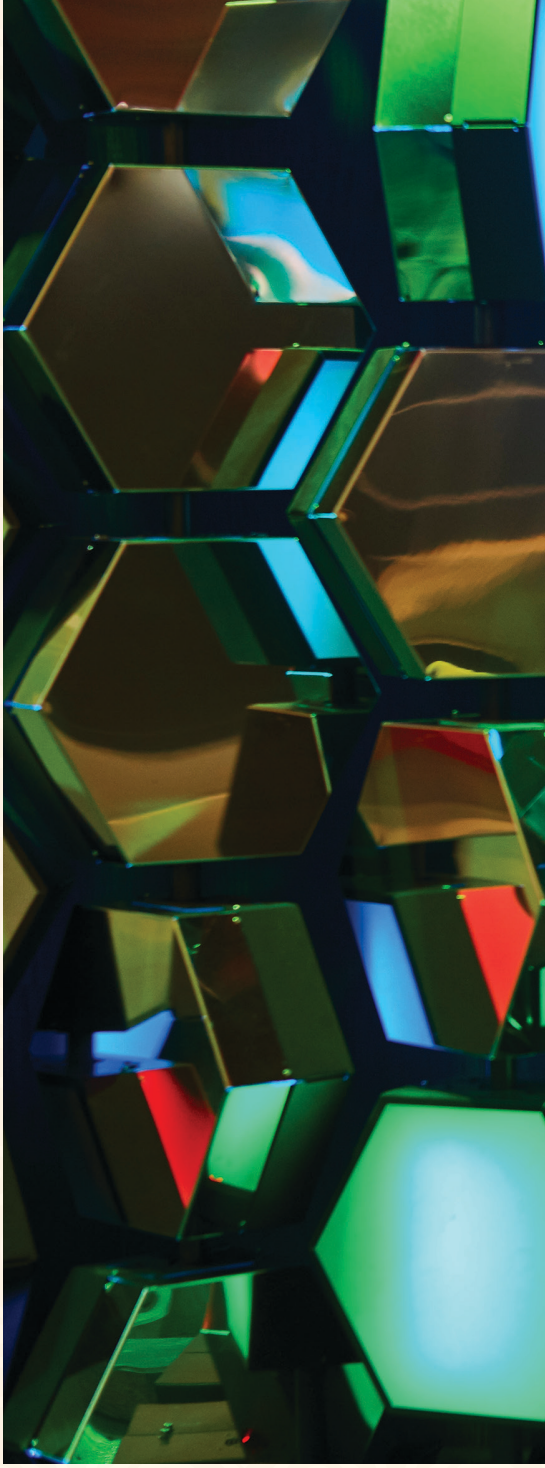
DOI:10.1145/3737450

BY AZZA ABOUZIED, FIROJ ALAM, RAIAN ALI, AND PAOLO PAPOTTI

Combating Misinformation in the Arab World: Challenges and Opportunities



MISINFORMATION AND DISINFORMATION are global risks. However, the Arab region is particularly vulnerable due to its geopolitical instabilities, linguistic diversity, and other cultural nuances. Misinformation includes false or misleading content, such as rumors, satire taken as fact, or conspiracy theories, while disinformation is the intentional and targeted spread of such content to deceive or manipulate specific audiences.[16] To limit the spread and influence of misinformation, it is essential to advance research on technological methods for early detection, tracking, and mitigation, while also strengthening media literacy and promoting active citizen participation.

Each stage in the fight against misinformation presents challenges. For example, in the Arab world, information is neither produced nor consumed in a single, universal language like Modern Standard Arabic (MSA). Instead, communication takes place across a wide range of dialects and languages—including Egyptian, Franco-Arabic, Gulf, and Levantine—as well as widely spoken languages such as English

IMAGE BY ABB PHOTO

and French. This linguistic diversity adds significant complexity to the task of misinformation detection. To illustrate, the word ؤجنجم in Tunisian Arabic (*ynajjim*) simply means "he can" or "he is able," while in Egyptian Arabic (*yenaggim*) it means "he is practicing astrology," and is often used mockingly to imply that the person is fabricating claims, predicting an unknown future, or pretending to know things he does not.

The problem of misinformation detection can thus be understood as occurring within a multidimensional space, with each axis representing a key factor: dialectal variation, context (for example, news articles, tweets, or social media posts), and modality (for example, text, memes, videos, or images). In this complex space, the "curse of dimensionality" becomes evident: There is simply not enough annotated data to train robust automated detection tools. Compounding the challenge is the sparsity of authoritative information across these dimensions, which leaves the region particularly susceptible to disinformation tactics such as *data void exploits*. These arise when search engines or social media platforms return little to no credible content for a given query, creating an opening that malicious actors can exploit by flooding the space with misleading or false information. Cultural and societal dynamics add further layers of complexity. Mistrust in formal media and fact-checking institutions; reliance on informal networks such as family, tribal, or religious ties; limited media literacy; and low motivation all hinder participation in social-correction efforts such as Community Notes. Yet, despite these challenges, there

Bridging the gap between researchers and practitioners—whether independent fact-checkers or news media agencies—is a key opportunity for combating misinformation in the Arab world.

remain meaningful opportunities for intervention and future research.

Here, we explore this complex research landscape by examining the technological methods surrounding misinformation in the Arab world. First, we examine the technological methods, models, and tools for automated detection, covering a wide range of tasks such as propaganda identification, check-worthiness estimation, and multimodal misinformation detection. Then we examine data void exploits as a form of disinformation attack and outline strategies for tracking and mitigating them. Finally, we investigate avenues for social correction, emphasizing the role of users and communities in countering misinformation, and highlight approaches that encourage community participation in mitigation efforts.

Detection: Building AI Systems

Datasets and benchmarks play a vital role in building automated misinformation-detection methods. Unfortunately, there are few annotated datasets that cover the many dialectal variants of Arabic, the different modalities of communication such as images or videos, and the more nuanced misinformation-detection tasks such as identifying propaganda markers, hate speech, or claim check-worthiness. The need for data is further driven by recent research that illustrates that data-hungry, transformer-based Arabic language models such as AraBERT and MARBERT consistently outperform traditional machine learning methods in misinformation-detection tasks.[4]

Recent efforts are overcoming the data-sparsity challenge by curating annotated text (AraNews,[7] AraFacts,[20] and COVID-19 disinformation[3]) and multimodal datasets (ArMeme[2]), as well as shared task benchmarks (CheckThat! Lab,[a] ArAIEval,[b] and OSACT[c]). But scaling these annotation and ground-truthing efforts is difficult: One must not only find annotators that speak the different dialects but also provide annotation guidelines[d] that reflect the linguistic and cultural norms of the annotators themselves. For example, an image depicting women in revealing attire may be considered inappropriate or offensive in certain Arab cultures, highlighting the need for culturally sensitive annotation guidelines.

In addition to these research-driven efforts to create annotated datasets, grassroots initiatives such as Misbar[e] and Fatabyyano,[f] where human fact-checkers identify misleading content, verify it, and provide evidence, can also support the work of creating rich repositories necessary for the further development and validation of automatic detection methods. Automated misinformation detection can, in turn, scale the efforts of human fact-checkers by bringing check-worthy content to their attention. Bridging the gap between researchers and practitioners—whether independent fact-checkers or news media agencies—is a key opportunity for combating misinformation in the Arab world. Initiatives such as Tanbih[g]—a research platform that provides the public with tools for detecting propaganda,[h] factuality, media bias, and framing—represent promising initial efforts in this direction.[11,22]

How to build robust and reliable automated misinformation-detection tools remains an open problem. New techniques continually emerge, often surpassing existing approaches. These include leveraging pre-trained language models such as AraBERT and MARBERT vs. classical deep learning architectures like BiLSTM and CNN, using general-purpose large language models (LLMs) vs. Arabic-only transformers, applying

a https://checkthat.gitlab.io
b https://araieval.gitlab.io
c https://osact5-lrec.github.io

d https://www.digitqr.net/critical-digital-literacy/
e https://misbar.com
f https://fatabyyano.net
g https://tanbih.qcri.org/
h Propaganda is the strategic use of true or false information to promote a particular agenda, often blending facts with manipulation or emotionally charged language to shape opinion.

model-level fusion and attention mechanisms for multimodal data vs. feature-level fusion, and incorporating metadata such as user engagement features vs. relying solely on base content.[4]

Recent benchmarks demonstrate considerable variability in state-of-the-art performance across tasks such as factuality assessment, propaganda detection, and claim verification, with accuracy scores ranging from 0.55 to 0.95.[1,13,21] Comparative evaluations between LLMs and task-specific models further suggest that, despite recent advances, there is still significant room for improvement in effectively addressing these tasks. In this rapidly evolving field, ongoing research and empirical validation are essential to ensure that emerging methods can be effectively adopted and applied to the persistent and context-specific challenges of detecting misinformation in the Arab world.

Tracking and Mitigation: Combating Data Void Exploits

Disinformation is misinformation's motivated, coordinated, and targeted counterpart. Disinformation is characterized by strategic agents, such as state-sponsored actors or regional interest groups, with access to resources and information-dissemination assets, and a target-victim demographic whom they wish to influence.[16] By examining disinformation through a cybersecurity lens, we can better identify and categorize threats and vulnerabilities within the Arab region and build tools to track these threats and effectively mitigate them through precise and cost-effective responses. One such threat to which the region is particularly susceptible is a *data void exploit*.

A data void is a gap in an information ecosystem. A search begins with keywords or questions. When there is a dearth of information online that is relevant to the keywords, we are in a data void. Data voids are not inherently problematic. Random search strings lead us into voids. Disinformers, however, capitalize on the presence of data voids with respect to certain keywords or queries and the operation of search engines to drive information seekers to their narratives,[8] filling the voids with their content before the emergence of authoritative information. Information seekers then discover the content by actively searching and deeming it authentic, as it was "found" rather than passively shared with them. The exploitation of data voids can profoundly and permanently influence people's beliefs.

Data voids present unique challenges in the Arab region, especially amid frequent breaking news and emerging crises. Disinformers are quick to exploit these gaps during times of geopolitical instability, seizing the opportunity to spread fabricated narratives before credible sources can deliver verified information.

The linguistic diversity within the Arab region complicates efforts to track and mitigate disinformation. In this post-colonial context, communities communicate in multiple languages, including hybrids such as Franco-Arab, Arablish, and Arabizi. This complexity is further pronounced in cosmopolitan cities like Dubai and Doha, where diverse populations rely on a variety of global news sources and social media platforms. As each group maintains its affinity to a different linguistic or trusted news source, the lack of a universal, authoritative information source further exacerbates the potential for data voids and their exploits.

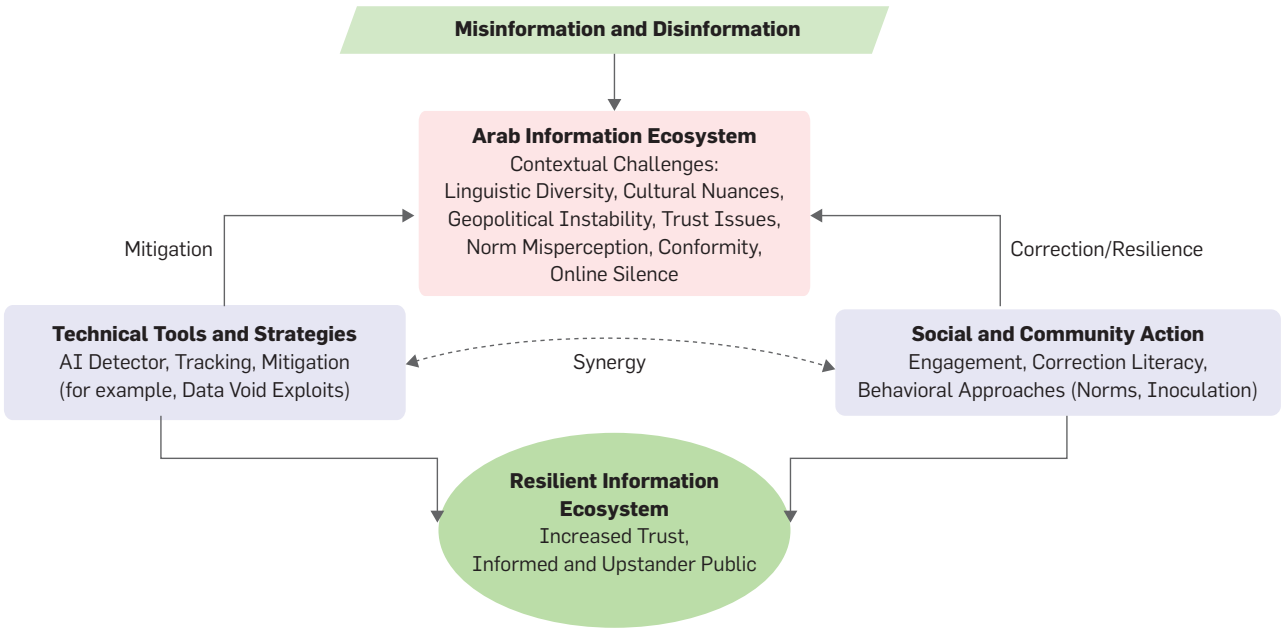
Golebiewski and boyd argue that data void exploits are largely intractable without systematic, intentional, and thoughtful management by the media and search platforms that host and index content.[8] Recently, as platforms struggle with balancing the individual's right to free speech with society's need for trusted information, they are moving further away from this kind of systematic management of disinformation.[12]

Despite the challenges posed by data voids, promising opportunities are emerging for the Arab region. Our research into these exploits has led to the development of a language-agnostic tool capable of tracking the efficacy of an ongoing exploit. We show that *search result rank* can determine the effectiveness and progress of both disinformation or mitigation efforts with respect to a set of data void keywords in any language. We validated this tracking tool across both historical and contemporary data void case studies.[15] By leveraging language models to automatically identify and label disinforming and mitigating narratives, we can effectively assess their influence in search queries. We use adversarial game-theoretic simulations on a proxy model of the Web to explore how mitigators can counter disinformers more effectively. In this setup, both sides act as competing agents with different resources, each trying to boost the visibility of their content in search results. Our results show that successful mitigation requires timely and strategic content placement. Crucially, we find that establishing high-influence information networks is core to a cost-effective response to an ongoing exploit by strategic disinformers. This amounts to creating a networked and coordinated coalition of fact-checking and credible, multilingual information sources. Therefore, even if search and media platforms were to disengage from disinformation management, it is possible for independent mitigators to build sufficient online information-dissemination assets and boost their influence through linking them to each other, allowing for an immediate, effective response to data void exploits.

Community Engagement: Promoting Social Correction

User correction relies on individuals flagging, debunking, and confronting those who post misinformation. Social correction, including crowdsourced fact-checking initiatives such as Community Notes, harnesses the collective intelligence of users to identify and rectify misinformation.[19] These collaborative efforts can be particularly effective in regions where diverse linguistic and

Figure. A sociotechnical framework for addressing misinformation in the Arab world. Misinformation enters a complex ecosystem with distinct challenges. Technical tools and community efforts work together to reduce harm and encourage correction, helping to make the ecosystem more resilient.



cultural contexts require localized understanding. But despite their potential, people often hesitate to participate in these activities. Olson and LaPoe argue that the "spiral of silence" theory applies in the context of user correction of misinformation on social media.[5] The theory suggests that the reluctance to engage in acts requiring confrontation or opposing what appears to be the majority opinion stems from a human need to belong, leading to fear of isolation and then conformity.

The question then arises: Why do people remain silent, and is there a misperception in how they think about it? Reasons for avoiding the correction of misinformation can be grouped into four categories:[10] relationship consequences (for example, "Will correcting others harm my relationship with them?"); negative impact on the person being challenged (for example, "Will they feel offended or seem less trustworthy?"); the perceived futility of the act (for example, "Is correcting misinformation even useful?"); and injunctive social norms (for example, "Is this behavior socially acceptable?"). Participants from two distinct cultural contexts, Arab and U.K. populations, exhibited misperceptions across these reasons. For instance, individuals overestimated the potential harm a correction might cause to relationships and the futility of challenging misinformation compared to reality. These misperceptions, while largely similar across cultures, significantly influence the willingness of people to engage in user correction.[18]

The identification of misperceptions creates an opportunity to leverage the social norms approach, which involves challenging individuals' assumptions and encouraging behavioral change.[10] Successfully applied in other domains, such as correcting misperceptions about smoking and alcohol consumption, this approach also holds promise for addressing misperceptions about the utility and appropriateness of misinformation correction. For it to be most effective, it requires redesigning social media platforms; however, it can also be implemented through digital literacy campaigns and framing it as a pro-social act. Such a framing resonates with cultures like the Arab culture, which values group benefit, acts of donation, and altruism. Digital nudging can also promote social correction, but more research is required to determine which nudges are more impactful. Question stickers, for example that tag content with questions such as "Is this source credible?" can be more effective than default comment boxes in encouraging social correction.[17]

A co-design study[9] revealed that users prefer social media and online forums to be enhanced by fostering a sense of secure online environments, easing confrontations, facilitating access to reliable debunking information, and leveraging social recognition and social proof. However, translating this into actual designs that are both usable and effective remains a research challenge.

Attitudinal inoculation is a psychological technique used to build resistance to persuasion, including that found in misinformation posts and news. It works by exposing individuals to weak counterarguments or small doses of opposing views, triggering critical thinking and inoculating them, similar to a vaccine, by activating their defenses and preparing them to counter stronger future challenges to their beliefs.[14] Incorporating attitudinal inoculation

into platform design has shown potential in enhancing resilience to persuasive online platforms, such as gambling.[6] This inoculation-based strategy, combined with social norms messaging, could provide additional opportunities for behavior change if integrated into social media platforms to address misperceptions about the usefulness and relevance of the pro-social act of user correction. For example, it could occasionally prompt users to identify misleading and persuasive elements in a simulated post and define ways they can challenge majority silence and perceived social norms.

Charting the Path Ahead

Effectively countering misinformation in the Arab world requires a coordinated sociotechnical approach that brings together automated tools, networked mitigation strategies, and culturally grounded user engagement (see the figure). AI-powered automated systems are essential for detecting and flagging misleading content at scale. But these systems must be trained with sufficient regional data and be guided by human oversight to ensure interventions are accurate, trusted, and contextually appropriate.

Beyond detection, our research shows that cost-effective responses to disinformation attacks such as data void exploits require active monitoring and the development of high-influence networks of credible, multilingual sources. Linking these sources creates a resilient infrastructure that can respond rapidly, even when platforms step back from moderation. Fact-checking organizations, news media, and civil society can all invest in and amplify these networks.

Finally, engaging users in the correction process is critically important. In Arab societies, where public correction may be discouraged by social norms, social norm messaging (for example, public campaigns that emphasize "It's a good thing to correct misinformation online!") can help shift behavior from silence to action. Platforms can support this shift by designing features that frame correction as a positive, community-oriented, and altruistic act. Governments can amplify these strategies through targeted, paid campaigns on social media, especially during high-stakes moments of misinformation spread. Education and media literacy initiatives that psychologically inoculate communities are also effective, though they require long-term investment by public institutions. Together, these technical, institutional, and social efforts form a scalable, culturally attuned response to the evolving threat of misinformation.

Acknowledgments

This publication was partly supported by NPRP 14 Cluster grant # NPRP 14C-0916–210015 from the Qatar National Research Fund (a member of Qatar Foundation), the ASPIRE Award for Research Excellence (AARE-2020) grant AARE20-307, NYUAD CITIES through Tamkeen - Research Institute Award CG001, and the ANR project ATTENTION (ANR-21-CE23-0037). The findings herein reflect the work and are solely the responsibility of the authors. C

References

1. Abdelali, A. et al. LAraBench: Benchmarking Arabic AI with large language models. In *Proceedings of the 18th Conf. of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Y. Graham and M. Purver (Eds.). Association for Computational Linguistics (2024), 487–520.
2. Alam, F. et al. ArMeme: Propagandistic content in Arabic memes. In *Proceedings of the 2024 Conf. on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y. Chen (Eds.). Association for Computational Linguistics (2024), 21071–21090.
3. Alam, F. et al. Fighting the COVID-19 infodemic: Modeling the combating misinformation in the Arab world: Challenges & opportunities 7 perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, M. Moens et al. (Eds.). Association for Computational Linguistics (2021), 611–649.
4. Alotaibi, T. and Al-Dossari, H. A Review of fake news detection techniques for Arabic language. *Intern. J. of Advanced Computer Science & Applications 15*, 1 (2024).
5. Olson, C.C. and LaPoe, V. Combating the digital spiral of silence: Academic activists versus social media trolls. In *Mediating Misogyny: Gender, Technology, and Harassment*. (2018), 271–291.
6. Cemiloglu, D. et al. Explainability as a psychological inoculation: Building resistance to digital persuasion in online gambling through explainable interfaces. *Intern. J. of Human–Computer Interaction 40*, 23 (2024), 8378–8396.
7. Elmadany, A. et al. Machine generation and detection of Arabic manipulated and fake news. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*. (2020), 69–84.
8. Golebiewski, M. and boyd, d. Data voids: Where missing data can easily be exploited. Data & Society (2019); https://datasociety.net/library/data-voids/
9. Gurgun, S. et al. Motivated by design: A codesign study to promote challenging misinformation on social media. *Human Behavior and Emerging Technologies 2024*, 1 (2024), 5595339.
10. Gurgun, S. et al. How would I be perceived if I challenge individuals sharing misinformation? Exploring misperceptions in the UK and Arab samples and the potential for the social norms approach. In *Intern. Conf. on Persuasive Technology*. Springer (2024), 133–150.
11. Hasanain, M., Ahmad, F., and Alam, F. Can GPT-4 identify propaganda? Annotation and detection of propaganda spans in news articles. In *Proceedings of the 2024 Joint Intern. Conf. on Computational Linguistics, Language Resources and Evaluation*. N. Calzolari et al. (Eds.). ELRA and ICCL (2024), 2724–2744; https://aclanthology.org/2024.lrec-main.244
12. Isaac, M. and Schleifer, T. Meta says it will end its fact-checking program on social media posts. *The New York Times* (Jan. 7, 2025); https://www.nytimes.com/2025/01/07/business/meta-community-notes-x.html
13. Kmainasi, M.B. et al. LlamaLens: Specialized multilingual LLM for analyzing news and social media content. In *Findings of the Association for Computational Linguistics: NAACL 2025*, L. Chiruzzo et al. (Eds.). Association for Computational Linguistics (2025), 5627–5649; https://aclanthology.org/2025.findings-naacl.313/
14. Lewandowsky, S. and Linden, S.v.d. Countering misinformation and fake news through inoculation and prebunking. *European Rev. of Social Psychology 32*, 2 (2021), 348–384.
15. Mannino, M. et al. Data void exploits: Tracking & mitigation strategies. In *Proceedings of the 33rd ACM Intern. Conf. on Information and Knowledge Management*. ACM (2024), 1627–1637.
16. Mirza, M.S. et al. Tactics, threats & targets: Modeling disinformation and its mitigation. In *30th Annual Network and Distributed System Security Symp., NDSS 2023*. The Internet Society (2023); https://bit.ly/46bPj5K
17. Noman, M., Gurgun, S., Phalp, K., and Ali, R. Designing social media to foster user engagement in challenging misinformation: a cross-cultural comparison between the UK and Arab countries. *Humanities and Social Sciences Communications 11*, 1 (2024), 1–13.
18. Noman, M. et al. Challenging others when posting misinformation: A UK vs. Arab cross-cultural comparison on the perception of negative consequences and injunctive norms. *Behaviour & Information Technology* (2023), 1–21.
19. Saeed, M. et al. Crowdsourced fact-checking at Twitter: How does the crowd compare with experts? In *Proceedings of the 31st ACM Intern. Conf. on Information & Knowledge Management*. ACM (2022), 1736–1746.
20. Ali, Z.S. et al. AraFacts: The first large Arabic dataset of naturally occurring claims. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, N. Habash et al. (Eds.). Association for Computational Linguistics (2021), 231–236; https://aclanthology.org/2021.wanlp-1.26/
21. Yousef, M.A., ElKorany, A., and Bayomi, H. Fake-news detection: a survey of evaluation Arabic datasets. *Social Network Analysis and Mining 14*, 1 (2024), 225.
22. Zhang, Y. et al. Tanbih: Get to know what you are reading. In *Proceedings of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Intern. Joint Conf. on Natural Language Processing: System Demonstrations* (2019), 223–228.

Azza Abouzied (azza@nyu.edu) is an associate professor at New York University Abu Dhabi, Abu Dhabi, United Arab Emirates.

Firoj Alam is a senior scientist at Qatar Computing Research Institute, Doha, Qatar.

Raian Ali is a professor in the College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar.

Paolo Papotti is a professor at EURECOM, Biot, France.