Designing Assistive AI Technologies to Support Human Judging of Information Reliability

Matthew Lease
School of Information
University of Texas at Austin
ml@utexas.edu

How can we design responsible AI technologies to curb the digital spread of misinformation? While it has become clear that there are many ways to "game" news delivery and circulation systems such as Facebook, how to solve this problem is less clear. Many recent AI models proposed for automatic fact-checking largely define the problem space in terms of computational issues alone. This ignores important human factors which underlie the real-world problem and are critical to the adoption of practical solutions.

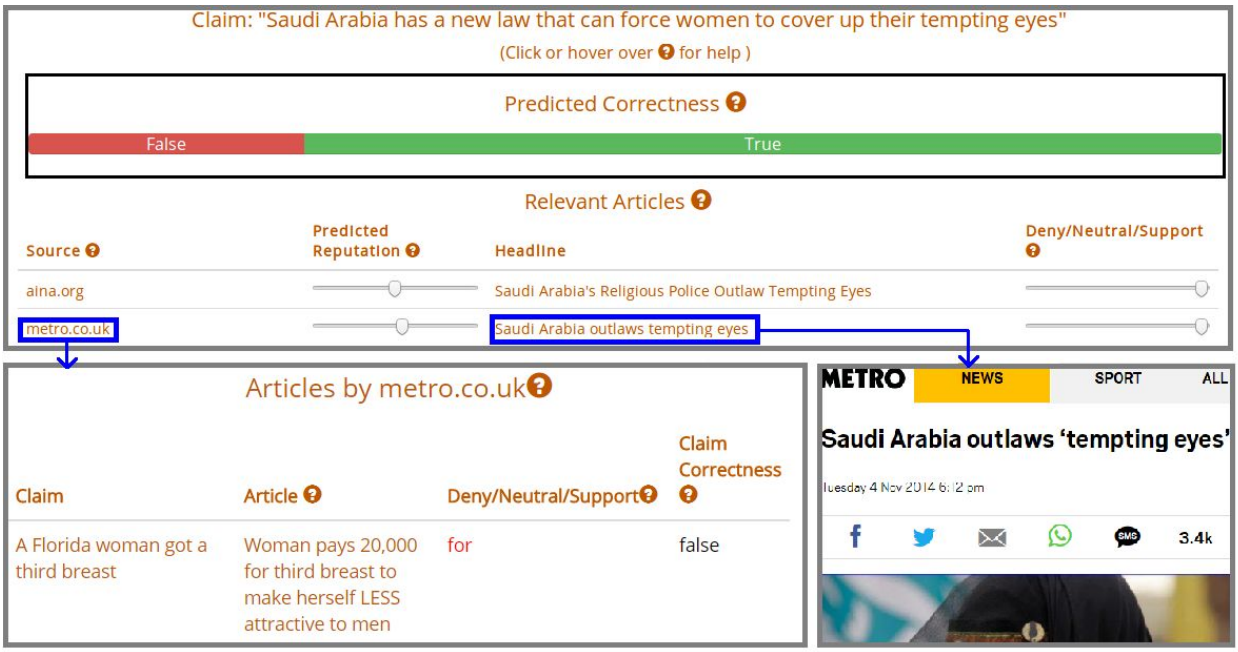
Having AI models produce accurate predictions is clearly important, but we must go further and ask why anyone would trust a black-box AI model telling them what to believe when many people distrust even well-known news outlets and fact-checking services? How do people decide what is actually misinformation? We must further recognize that AI itself is also contributing to the problem, with adversarial AI technologies being actively developed to spread misinformation, particularly using emotional and visual appeals that trigger quick responses. Consequently, AI researchers not only wrestle with issues of transparency and trust, but must further anticipate how new AI technologies could be misused and design appropriate countermeasures to prevent or mitigate harm.

Rather than frame AI as automation to make decisions for people, my lab envisions responsible AI as an *assistive technology*, designed to enhance and augment human abilities rather than supplant them. We integrate human-centered, front-end interface design with back-end language processing algorithms, assisting people in not only finding reliable, relevant online information, but in critically exploring, interpreting, and evaluating it. The goal of this synergistic combination of back-end and front-end technologies is to create a human-AI partnership based on *mixed-initiative* principles that exploits each party's respective strengths.

In 2012, we prototyped a system for searching and browsing memes underlying news: similar phrases which spread and evolved across sources. Once detected, these latent memes were revealed to users via generated hypertext, allowing memes to be recognized, interpreted, and explored in context. "Our vision [was] to complement traditional forms of critical literacy education with . . . smarter browsing technology . . . Instead of understanding online narrative through only a single source, we can instead explore how broader community discourse has shaped its development . . . [especially for] campaigns which flood social media with repeated stock phrases while obfuscating their . . . source."

While back-end AI modeling approaches to misinformation are receiving much attention, we argue that front-end human-computer interaction (HCI) and design are equally necessary for responsible AI yet have been largely neglected. Our current work builds on probabilistic graphical modeling (PGM) methods we apply as a foundation for our work. Neural (aka "deep learning") models now achieve superior predictive accuracy to PGMs on many tasks, but they are typically very difficult to interpret. What does it matter if a given prediction model achieves 85% vs. 87% accuracy if a user does not understand or trust it? We argue that model interpretability is a crucial foundation for developing transparent, responsible, and trustworthy assistive AI technologies.

Consider our existing prototype system for online fact-checking: http://exfacto.herokuapp.com.



Given a claim the user would like to check, relevant articles are retrieved. The model's predicted source reputation and stance for each retrieved article is shown to the user and can be revised via simple sliders to reflect user beliefs and/or to correct erroneousness model estimates. Whenever the user alters a given slider, the overall model prediction and confidence are re-estimated in real-time and shown to the user. Such real-time interaction is enabled by a fast variational method we developed for parameter estimation, achieving only modest degradation in model estimation vs. full-fledged Gibbs sampling.

In general, three broad criteria appear critical for AI to be adopted and useful in practice: (i) transparency, (ii) incorporation of users' knowledge; and (iii) uncertainty quantification. With misinformation, people will naturally be skeptical about any fact-checking AI. Our PGM approach is transparent because all sources of evidence are displayed, the method for aggregating

evidence is known and simple, and users can learn how the model behaves and override it by "playing" with the sliders. Because users can incorporate their knowledge, they can verify the model arrives at expected predictions based on the available evidence, and they can jointly reason alongside it. By communicating all modeling uncertainty, users can see that the model is fallible, then reason and make decisions with awareness of uncertainty about the available evidence. Moreover, it underscores that AI models are not infallible oracles, but rather make predictions only on the basis of the information provided by people.

In research to information retrieval (IR) for fact checking, we are also exploring how the relative ranking and presentation of news search results impact credibility assessment, as well as how users, unaware of personalized filtering of their news via an AI recommender system, may be more easily misled by misinformation. Which search results should we return, how should we present them, what modes of interaction should we provide, and how should we evaluate success? While assessing the authority of pages for ranking and filtering is not new, fact checking presents a different framing of authority, with ranking and filtering decisions potentially impacting user trust of the system and fears of being manipulated. Beyond topical diversification of search results, how might we diversify political (or other forms of) bias to provide diverse perspectives, especially on controversial topics? In terms of personalized search, how do we balance giving users search results matching their existing beliefs vs. challenging those beliefs with alternative viewpoints, and without such challenges prompting search engine switching behavior? Just as people follow different news outlets having different political leanings, perhaps a new class of vertical search engines will soon arise which rank and filter search results to match a given audience's views? How do we frame, measure, and address potential harm of search results including "alternative" facts, be they search result errors or intentional diversification?

References

Anubrata Das and Matthew Lease. **A Conceptual Framework for Evaluating Fairness in Search**. Technical report, University of Texas at Austin, July 2019. arXiv:1907.09328. [pdf]

Anubrata Das, Kunjan Mehta, and Matthew Lease. **CobWeb: A Research Prototype for Exploring User Bias in Political Fact-Checking**. In *ACM SIGIR Workshop on Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval* 2019. [pdf]

An Thanh Nguyen, Aditya Kharosekar, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C. Wallace, and Matthew Lease. **Believe it or not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking**. In *ACM User Interface Software and Technology Symposium*, pages 189--199, 2018. [pdf | demo | sourcecode | video | slides]

An Thanh Nguyen, Aditya Kharosekar, Matthew Lease, and Byron C. Wallace. **An Interpretable Joint Graphical Model for Fact-Checking from Crowds**. In *Proceedings of the*

Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), pages 1511--1518, 2018. [pdf | demo | sourcecode | video | slides]

Matthew Lease. **Fact Checking and Information Retrieval**. In *Proceedings of the 1st Biannual Conference on the Design of Experimental Search & Information REtrieval Systems (DESIRES)*, pages 97--98, 2018. [pdf | conference-website | slides]

Hohyon Ryu, Matthew Lease, and Nicholas Woodward. **Finding and Exploring Memes in Social Media**. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, pages 295--304. ACM, 2012. [bib | pdf | demo | sourcecode | video]