Considering Contestability in Automated Fact-Checking Systems

EMMA LURIE, UC Berkeley School of Information

Using machine learning to improve and increase the output of fact-checkers is viewed as a promising way to combat misinformation. However, adding machine learning to the fact-checking pipeline increases opacity. This is problematic as reducing transparency and expliainability runs counter to the fact-checking community's norms. This paper outlines 1) the benefits of thinking about contestability in fact-checking, 2) preliminary design characteristics for contestable fact-checking ML systems, and 3) challenges raised by envisioning contestable design in the context of fact-checking systems.

1 INTRODUCTION

Increased attention and investment in fact-checking has been a key response to concerns about problematic online information [9]. Developing meaningful machine learning tools for fact-checking ("automated fact-checking") has been a subject of considerable research [4, 5, 7]. However, transparency, explainability, and interpretability are rarely mentioned with regard to these systems. This is surprising as there is a strong alignment of values of the global fact-checking community and the benefits of incorporating values like contestability.

The International Fact-checking Network (IFCN) "Fact-checkers' Code of Principles" [1] presents five criteria of high-quality fact-checking organizations: 1) nonpartisanship and fairness, 2) transparency of sources, 3) transparency of funding and organization, 4) transparency of methodology, and 5) a commitment to open and honest corrections. These values have substantial overlap with the objectives of contestable design laid out in [8] and displayed in Table 1 especially in regard to an emphasis on decision making process transparency and a commitment to having a robust corrections mechanism.

Full Fact, the fact-checking organization is largely considered the global leader of automated fact-checking, divides the fact-checking pipeline into four steps [3]:

- **Monitor:** examine large amounts of social media content, news articles, transcripts, etc.
- **Spot Claims:** identify claims worth fact-checking.
- **Check Claims:** research and evaluate claims.
- **Create and Publish:** write and disseminate fact-checks.

While there is arguably a place for contestability in every step of this pipeline, this paper focuses on two technical systems used to *spot claims*:

Checkworthiness: One aspect of spotting claims is selecting significant statements that have not been previously fact-checked. There are many statements online that could be fact-checked, but we rely on the professional expertise of fact-checkers to select important claims to be fact-checked. Recent research has cast doubt on whether current techniques of manual fact-checker claim selection are ideal [2], and technical systems that determine checkworthiness are in the early stages of development [1, 6].

Claim matching: Another step in spotting claims is identifying similar already fact-checked claims. Fact-checkers have observed that many untrue or misleading statements in speeches, debates, tweets, etc. are repeated claims that may have already been fact-checked (e.g. Obama was not born in the U.S.; vaccines cause autism). Claim matching systems link a piece of content (e.g. "New study: Vaccines Linked to Autism") to an already existing fact-check article. Claim matching

[1] https://www.poynter.org/ifcn-fact-checkers-code-of-principles/

Table 1. A comparison of the values in Hirsch et al. [8] and in the IFCN Fact-Checker's Code of Principles. There is significant overlap in commitments to methodological transparency, providing explanations for decisions, and the presence of corrective mechanisms.

Hirsch et al.[8] Objectives of Contestability	IFCN Fact-checker's Code of Principles
accuracy via iterative deployment and incentivizing feedback	Nonpartisanship and fairness
Legibility by providing explanations, confidence levels, and traces of system predictions	Transparency of sources
Training that explicitly addresses system limitations and allows experimentation to develop shared understandings	Transparency of funding and organization
Mechanisms for questioning and disagreeing with system behavior whether at the individual or aggregate scale	Transparency of methodology
	A commitment to open and honest corrections

systems are further in in development than checkworthiness systems, but their performance is still relatively poor [7, 13].

The rest of this paper considers the following two questions:

(1) How can we imagine designing for contestability in fact-checking systems?
(2) What challenges and additional questions arise when implementing contestability in fact-checking systems?

2 HOW CAN WE IMAGINE DESIGNING FOR CONTESTABILITY IN FACT-CHECKING SYSTEMS?

Most research about automated fact-checking does not include discussions of contestability and related principles. Nguyen et al. [12] design an interface with sliders to indicate their confidence in the automated system's results for a particular claim that alters the system results, but find that overall users are overly trusting of the original predictions. Beyond this experiment, little attention has been given to how to infuse ideas of contestability into automated fact-checking tools.

2.1 Contestability in Checkworthiness Assessment Systems

Fact-checkers often depend on manually reading large amounts of documents and reader suggestions to identify "checkworthy" claims that should be fact-checks.

This approach has two key drawbacks: it encompasses a tremendous amount of fact-checkers time and isn't surfacing the most checkworthy claims. Babaei et al. found that users do not recommend stories that are likely to be confusing to users, but rather recommend stories that are clearly untrue [2].

However, the most promising implementations of such systems have precision scores of 0.23 and rely on neural networks [6]. An early experiment with fact-checkers, found that a similar system overwhelmed them with unhelpful suggestions and was not a primary source of claims for fact-checkers [1].

Considering the objectives laid out by Hirsch et al. there is substantial overlap between these objectives and the needs of checkworthy ML systems: 1) improving accuracy and incentivizing

feedback is essential for the adoption of these tools, 2) as current models rely on opaque ml, centering contestability in design can emphasize the importance of interpretability, 3) checkworthiness is already somewhat of a contested concept and experimentation may help fact-checkers develop a shared understanding, and 4) as these systems improve mechanisms for disagreeing with the suggestions will be essential to fact-checkers continued adoption of checkworthiness tools.

Because the primary focus of development of checkworthiness tools has been increasing the quality of their predictions, few have focused on the ideas of contestability.

While additional research and conversations with fact-checkers are necessary, I propose a few characteristics that may be be steps towards a contestable designs:

- System traces for each prediction: What features were most influential in producing that checkworthy claim?
- Sliding scales of feature importance: Some fact-checkers may primarily fact-check politicians while others may focus on environmental issues. Allowing fact-checkers to select which features are the most valuable to them not only makes the tool more useful to them, but also helps the system better learn from fact-checkers suggestions, as one of the key ways to improve these systems is to better incorporate domain-specific knowledge.
- A simple appeals process: As further detailed in the claim matching system, making sure that fact-checkers have an easy way of disagreeing with the system will be essential to improving the accuracy of these systems.

While checkworthiness tools are a ways off from being widely used by fact-checkers, there is an opportunity to change the course of their development away from a precision no matter the cost mentality to a system that will further the norms of ideals of fact-checking.

2.2 Contestability in Claim Matching

Under the framework laid out by Kluttz et al. [10], contestability is used to give experts more of a stake in automated systems by engaging them in the algorithmic decision making process. However, when thinking about non-expert systems this framework needs to be adjusted. Even though fact-checkers are "experts," claim matching does not imitate a process done by fact-checkers as fact-checkers typically stop interacting with a fact-check article once it is published. Claim matching systems match previously published fact-check articles to problematic claims. In fact, claim matching systems often rely on the training data of novice crowd workers [13]. In claim matching systems, where the machine learning component of these systems is not in the drafting of the fact-check article, but in the assignment of a previously written fact-check article to a new iteration of the claim, it seems that contesting or appealing the assignment of a fact-check to a claim is the most in demand. While the IFCN lists having a mechanism for corrections as a principle for fact-checking (see Table 1), there is no standard for this process and some have called into question the efficacy of these processes.[2] However, as illustrated in Figure 2, there are already implementations of an appeals process for similar features.

One implementation of a claim matching system was Google's short-lived Reviewed Claims feature (see Figure 1), which appeared on a subset of news publishers' Google search result pages. Reviewed Claims matched claims from news publishers' articles to already existing fact-check articles. While some fact-checks explicitly mentioned the news publisher in the article, approximately half were claim matched.

Two months after the feature was released, conservative news outlets complained that they were being unfairly targeted by the feature. After initially removing a few matches directly mentioned

[2]https://www.realclearpolitics.com/articles/2018/07/24/is_there_recourse_when_fact_checkers_get_it_wrong_137599.html

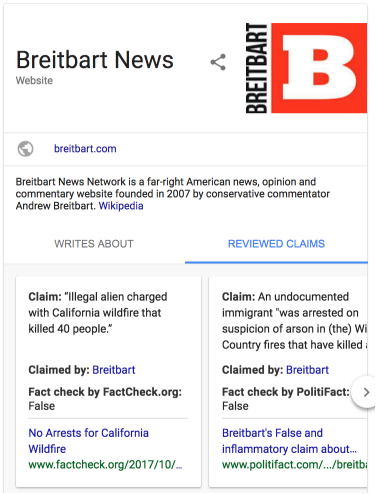

Fig. 1. Image of the Reviewed Claims feature that appeared on some news publisher search result pages from November 2017 to January 2018. Reviewed Claims was removed by Google after conservative media backlash and amid quality concerns.

by the articles, Google removed the entire feature explaining to Poynter that they had "encountered challenges in our systems that maps fact checks to publishers, and on further examination, it's clear that we are unable to deliver the quality we'd like for users."

At the time of the Reviewed Claims feature, there was no easy or meaningful mechanism for an appeal. So, several news publishers wrote angry articles that were picked up by mainstream press. I conducted [11] a content analysis of the twelve articles news publishers affected by Reviewed Claims panels wrote complaining about the feature and identified five critiques of the feature:

(1) The fact-check did not evaluate the claim to be false.
(2) The article is satire.
(3) The fact-check is assessing a different claim than the one in their article.
(4) The fact-check concerns a minor detail in the article.
(5) The article is simply reporting the facts, not making a broader political claim.

With the exception of the fifth critique, which involves epistemological examination of the role of fact-checks, it seems like a robust appeals process could resolve these controversies. While Google has not released a new iteration their claim matching tool, Google is currently experimenting with a feature that could be a model for an appeals process that makes it simple for users to report content on Knowledge Panel (the box of info in the top-right of many search results pages) as shown in Figure 2.

Surfacing fact-checks on platforms like Google dramatically increases the reach of fact-checkers, but it also brings more conflict and questions over the fact-checks themselves. This is challenging to navigate as platforms that prefer to act as a neutral player in the fact-checking ecosystem and not take any responsibility for the fact-checks that appear on their platforms.

3 WHAT CHALLENGES AND ADDITIONAL QUESTIONS ARISE WHEN IMPLEMENTING CONTESTABILITY IN FACT-CHECKING SYSTEMS?

There are several design choices that require further consideration:

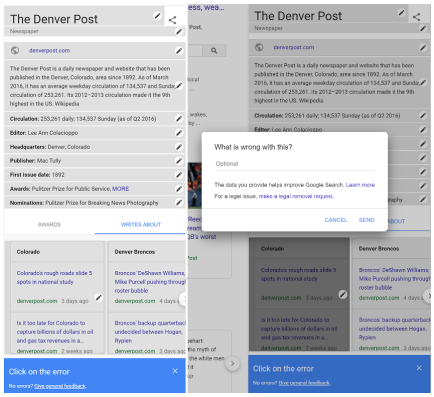

Fig. 2. These two screenshots illustrate Google's current experiment with the appeals process that makes it simple to offer feedback about specific parts of Knoweldge Panel.

Incorporating other stakeholders: The fact-checking ecosystem is a complex sociotechnical system that involves several actors including online information seekers, fact-checking organizations, platforms, and news publishers. The Kluttz et al. [10] model of contestability focuses on engaging experts, while news publishers and users may have a larger stake in the outputs of a fact-checking system. Moreover, if automated fact-checking systems are a tool to combat misinformation for users, perhaps they should be primarily accountable to users, not fact-checking organizations.

Accusations of fact-checking bias: Across the world, people in power accuse fact-checkers of being biased. Fact-checking organizations feel that they are under fire. As director of the IFCN, Alexios Mantzarlis described the state of online fact-checking at a June 2018 global summit of fact-checkers as: "a dark cloud hangs over us... the disaffection and distrust that have plagued mainstream media outlets for many years is now spilling over to fact-checkers. In Turkey, the Philippines and especially Brazil it broke out in the form of concerted campaigns aimed to vilify fact-checking as an instrument."[3]

Contestability and Trust: Further research is needed to explore this issue, but fact-checking rests on the assumption that there is a truth, and that fact-checkers are good at identifying it. While corrections are an important value of fact-checking, most people do not seem to know about that process. Could making an appeals process more robust decrease trust in the determinations of fact-checkers?

Identifying System Objectives: An automated fact-checking system may be designed to help online information seekers assess credibility (as in the Reviewed Claims example), or the purpose of the system could be designed to help a platform like Facebook identify and suppress problematic information[4]. It's hard to see how the same design could 1) display fact-checks alongside problematic information (to assist users in identifying the information as potentially problematic) and 2) limit the amount of problematic information surfaced in platform algorithms (as in the Facebook example).

[3]https://www.washingtonpost.com/news/fact-checker/wp/2018/06/25/rapidly-expanding-fact-checking-movement-faces-growing-pains

[4]https://www.facebook.com/help/1952307158131536

4 CONCLUSION

The interest and importance of automated fact-checking coupled with the fact-checking alignment of values with the ideas of contestability make it an important domain for further research.

As the fact-checking pipeline contains multiple stages, it is important to think critically about how to design for these values in the context of the objectives of each specific stage of the automated fact-checking process. Additionally, while much of the previous work around contestability has focused on engaging experts, in the case of fact-checking, some of the attention must also concern the online information seekers are the ones facing the brunt of the problematic information online.

Many of these algorithms and automated fact-checking systems are still in the early stages of development, which offers an interesting opportunity for the contestability community to get involved in designing for contestability

REFERENCES

[1] Bill Adair, Mark Stencel, Cathy Clabby, and Chengkai Li. The human touch in automated fact-checking. In *Proceedings of the Computation + Journalism Symposium*, 2019.

[2] Mahmoudreza Babaei, Abhijnan Chakraborty, Juhi Kulshrestha, Elissa M Redmiles, Meeyoung Cha, and Krishna P Gummadi. Analyzing biases in perception of truth in news stories and their implications for fact checking. In *FAT*, 2019.

[3] Mevan Babakar and Will Moy. The state of automated factchecking. *Full Fact*, 2016.

[4] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193, 2015.

[5] Lucas Graves. Understanding the promise and limits of automated fact-checking. *Factsheet*, 2:2018–02, 2018.

[6] Casper Hansen, Christian Hansen, Stephen Alstrup, Jakob Grue Simonsen, and Christina Lioma. Neural check-worthiness ranking with weak supervision: Finding sentences for fact-checking. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 994–1000. ACM, 2019.

[7] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1803–1812. ACM, 2017.

[8] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E Imel, and David C Atkins. Designing contestability: Interaction design, machine learning, and mental health. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, pages 95–99. ACM, 2017.

[9] Caroline Jack. Lexicon of lies: Terms for problematic information. *Data & Society*, 3, 2017.

[10] Daniel Kluttz, Nitin Kohli, and Deirdre K Mulligan. Contestability and professionals: From explanations to engagement with algorithmic systems. *Available at SSRN 3311894*, 2018.

[11] Emma Lurie. The challenges of algorithmically assigning fact-checks: A sociotechnical examination of google's reviewed claims. 2019.

[12] An T Nguyen, 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 *The 31st Annual ACM Symposium on User Interface Software and Technology*, pages 189–199. ACM, 2018.

[13] Xuezhi Wang, Cong Yu, Simon Baumgartner, and Flip Korn. Relevant document discovery for fact-checking articles. In *Companion Proceedings of the The Web Conference 2018*, pages 525–533. International World Wide Web Conferences Steering Committee, 2018.