Research Statement

Benjamin D. Horne

October 9, 2023

My research seeks to answer questions about **resistance to strategic disinformation and malign influence**. This work can be mapped onto what I call the "disinformation pipeline", made up of three interconnected stages of information behaviors and processing: (1) production tactics, (2) interventions, and (3) influence/consumption. This pipeline with example questions at each stage is shown in Figure 1.

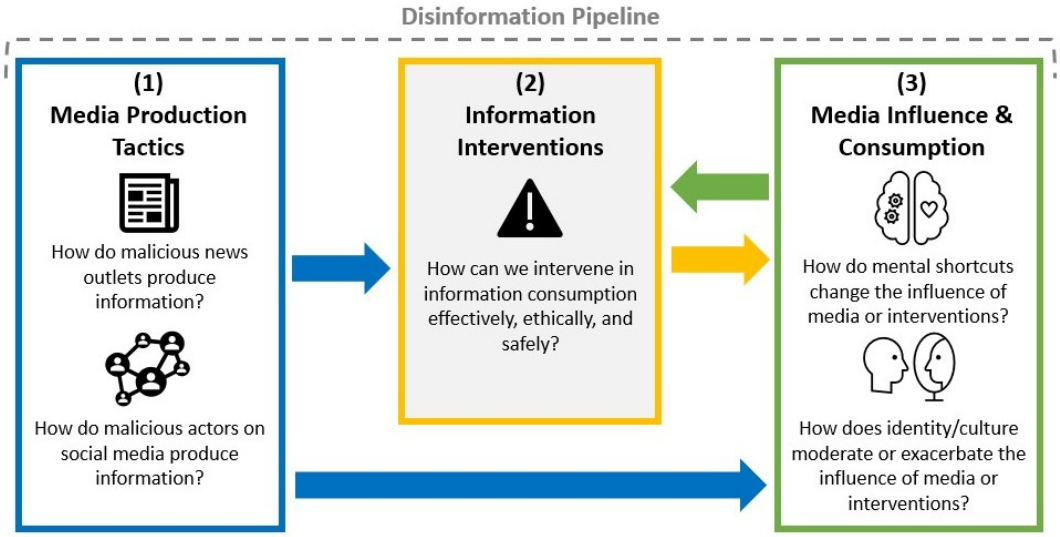

Figure 1: Three interconnected parts of my research, with broad example questions in each part. These areas connect through what could be thought of as an "disinformation pipeline": ranging from information production behaviors to information consumption behaviors. This process is often made complex through the socio-technical systems that serve information, interventions in information consumption, and the socio-cultural context of information.

1 Media Production Tactics[1]

The first step in any disinformation or influence campaign is to produce content. While on the surface this concept seems quite simple, today's media production is a complex, hybrid, multi-modality system, where content producers can range from news outlets to political elites to coordinated social media actors (Figure 2). Information within this complex system can be manipulated and spread by a range of entities, including hate groups, internet subcultures, "useful idiots [29]", political campaigns, and governments [21, 25].

My work in this subarea **describes content production, manipulation, spread, and migration across varying target contexts.** Much of this work is done with the goal of *describing*, rather than establishing causality or following traditional hypothesis-driven methods, and this work treats digital trace data as first-order objects for investigation[2]. The task of "mere description" on digital trace data is critical in hypothesis development, measuring topic importance, and generalizability. In particular, my work frequently describes settings or contexts

[1]Note, I use the term "tactics" quite loosely here, as sometimes it can be difficult, if not impossible, to establish intentions in content production across many contexts. Nonetheless, understanding content production behaviors, whether done with the intention to misinform, influence, make money, or with no intention at all, is critical to understand this complex media system.

[2]`https://epjdatascience.springeropen.com/about`

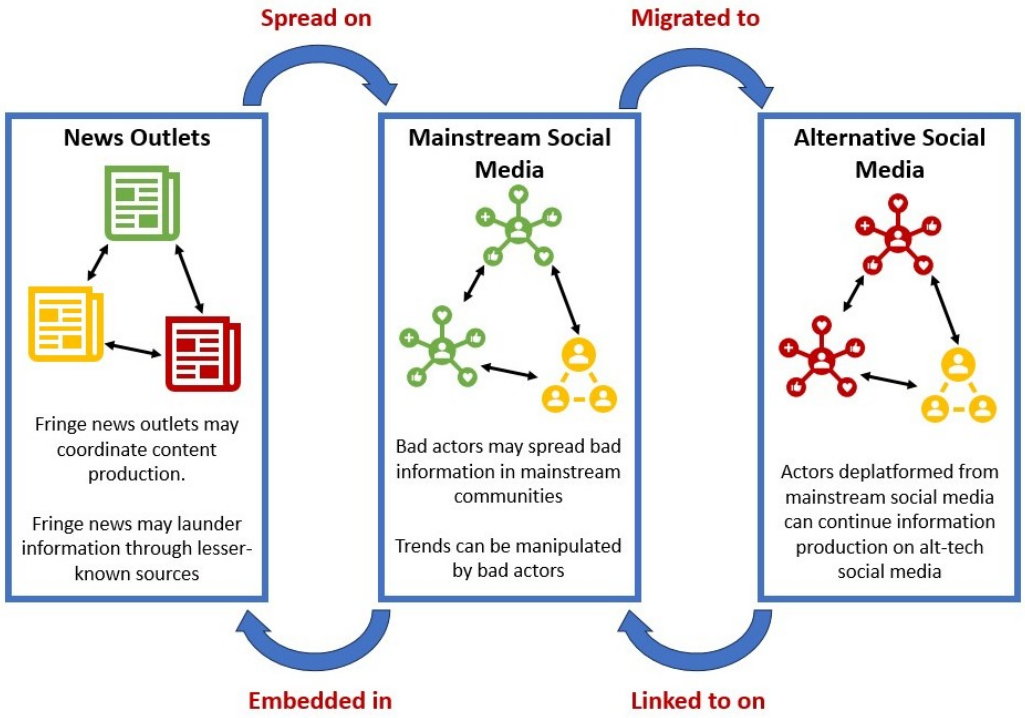

Figure 2: I use the term "media" to refer to more than just news media or social media, but rather the larger system of today's media. Disinformation and manipulation take place in a hybrid, multi-modal, networked system. In this system, news outlets can coordinate content production, spread on social networks, and embedded social media generated information in their articles. Major social media platforms, and the dynamics therein, control who sees produced information. When content is moderated on these major platforms, those measures can be subverted through alternative social media platforms, where bad content continues to be produced, sometimes appearing back on major platforms through cross-posting and links. Production and tactics are thus more than simply making content but also operating in this complex system.

that have never been explored before. So, my goal is to describe how the target context differs from or relates to known contexts.[3]

To describe and analyze large datasets of digital traces, I leverage my skills in Natural Language Processing (NLP), Network Science, and Machine Learning (ML). For example, in our 2019 ICWSM work [12], we described content sharing between news outlets in fringe and mainstream media, and we showed that content production can be coordinated among fringe news outlets. In this work, we created a novel algorithm to construct directed content sharing networks from news text data, allowing for content flow between news outlets to be established without linking data. This construction was later used in our 2021 MEDIATE workshop paper for news outlet veracity predictions through the use of attributed network embedding [6].

Another example is our WebSci 2022 work [3], in which we characterize the cross-platform mobilization of YouTube and BitChute (an alternative to YouTube) videos on Twitter during the 2020 U.S. Election fraud discussions. Through the combination of our previously published BitChute dataset [28] and the VoterFraud dataset [1], we were able to describe the prevalence of content supplied by both platforms, the mobilizers of that content, the suppliers of that content, and the content itself. This work showed that while BitChute videos promoting election fraud claims were linked to and engaged with in the Twitter discussion, they played a relatively small role compared to YouTube videos promoting fraud claims, pointing to a need for proactive, consistent, and collaborative content moderation solutions.

As a final example of work in this subarea, our 2022 ICWSM work [16] describes COVID-19 news coverage by U.S. local news outlets over time. We found that the rate of COVID coverage over time by local news outlets was primarily associated with death rates at the national level, rather than the local level. Further, we found that the volume and subtopic of COVID coverage differed depending on local politics and outlet audience size. We also

[3] `https://journalqd.org/about`

found evidence that more vulnerable populations received less pandemic-related news. A unique data collection of online local news was published out of this effort [9].

2 Information Interventions

Within this pipeline, there is an opportunity to intervene in disinformation's consumption using our knowledge about media production tactics. This moment between production and consumption is where my largest and most developed body of work lies. Specifically, my work in this subarea **evaluates the effectiveness and appropriateness of Machine Learning (ML) tools' for content moderation in the application setting** and compares those tools to non-automated alternatives.

Due to the overwhelming scale and complexity of disinformation production online, there have been many proposed technical solutions to combat it. Proposed technical solutions can filter out or automatically place warning labels on content that is of low veracity. These solutions range widely in terms of technical methods used and an extraordinary number of classifiers have been proposed. As an example of this volume, between 2016 and 2022 there were approximately 14,000 papers indexed by Google Scholar that used the phrase "fake news detection" and 210,000 papers that used the phrase "fake news"[4]. This extraordinary number of classifiers proposed demonstrates the *just make something* approach of content moderation research. While *just making something* has helped push technical advances in a variety of areas, the format often sacrifices empirical rigor, neglects the deployment context of the models, and ignores the human user. Accordingly, while many of the proposed automated approaches for content moderation have shown high accuracy in lab settings, they may be overfitting to specific data and lacking theoretical underpinnings. By simplifying or ignoring the context of a tool's deployment, we miss the potentially adverse impacts on human information consumers, creating critical gaps between research and practice.

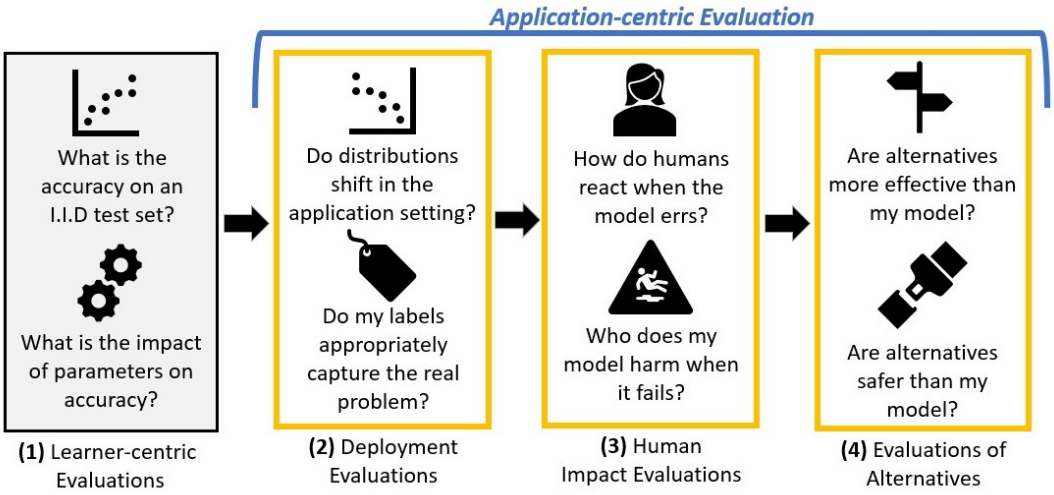

Figure 3: My framework for holistically evaluating ML tools, particularly in the context of content moderation and information interventions.

Instead, I argue that work in this area should take an *understand deeply* approach. Specifically, my current and future work seeks to move ML tool evaluations from *learner-centric* to *application-centric* (Figure 3). In the context of content moderation, I ask three broad questions: **(1) What are the limitations of machines?** Currently, tools are evaluated using decontextualized predictive behaviours [15]. By oversimplifying a complex task to be automated and reducing the systems deployment to only technical questions, it is likely that we will fail in practice. My work goes beyond traditional ML frameworks to simulate how these tools may generalize (or fail to generalize) in real-life settings. **(2) What are the costs of making mistakes?** My work evaluates the influence of ML tool interface designs on consumers' trust and information behaviors, both when those predictions are correct and when they are wrong. I pay special attention to how these costs change across the sociocultural

[4]Computed using `zenodo.org/record/1218409\#.YrHv\_NLMKrw`

differences of consumers. Who will be harmed? Who will benefit? **(3) Are there more effective and safe alternatives to automation?** Depending on the answers to the above questions, we must be open to the idea that automating content moderation may have limited use, and instead find alternatives. My work will study how effective, fair, and safe alternative interventions are compared to using automated interventions or using no interventions at all.

In the past, I focused heavily on building ML tools for news veracity classification based on the tactics used by fringe content producers [7, 13, 8]. Hence, while much of my current work in this area uses controlled experiments with methods from Human-Computer Interaction and Psychology [11, 10, 23, 20], my background provides a strong theoretical foundation in ML and predictive analytics.

3 Media Influence and Consumption

The end-goal of both disinformation campaigns and information interventions is to influence decision making. This process of changing information consumption, trust, and decision making is intricate and difficult to fully understand. In particular, previous work from multiple disciplines has demonstrated that trust in false information and its influence on decision making is dependent on multiple interconnected factors (Figure 4).

It is well known that mental shortcuts - rules of thumb for making decisions without exhaustively comparing all available options - are used when making decisions about the information we trust [19, 10, 23]. For example, confirmation bias, the notion that people tend to process information in a way that favors their previously held beliefs, plays a significant role. These prior beliefs may be influenced by media effects, generational effects, and cultural effects, each of which are difficult to disaggregate [2, 26, 27, 22, 14]. People are also more likely to believe information that is familiar, fluent, believed by others, and tells a coherent story [19, 24]. Other cognitive drivers include lack of analytical thinking and memory failures [5]. Other socio-affective drivers include source cues and the desire to be accepted in a group [4, 5]. With trust depending on numerous social, cultural, and cognitive factors, one can imagine that intervening with warning labels is also a complex process with similar dependencies.

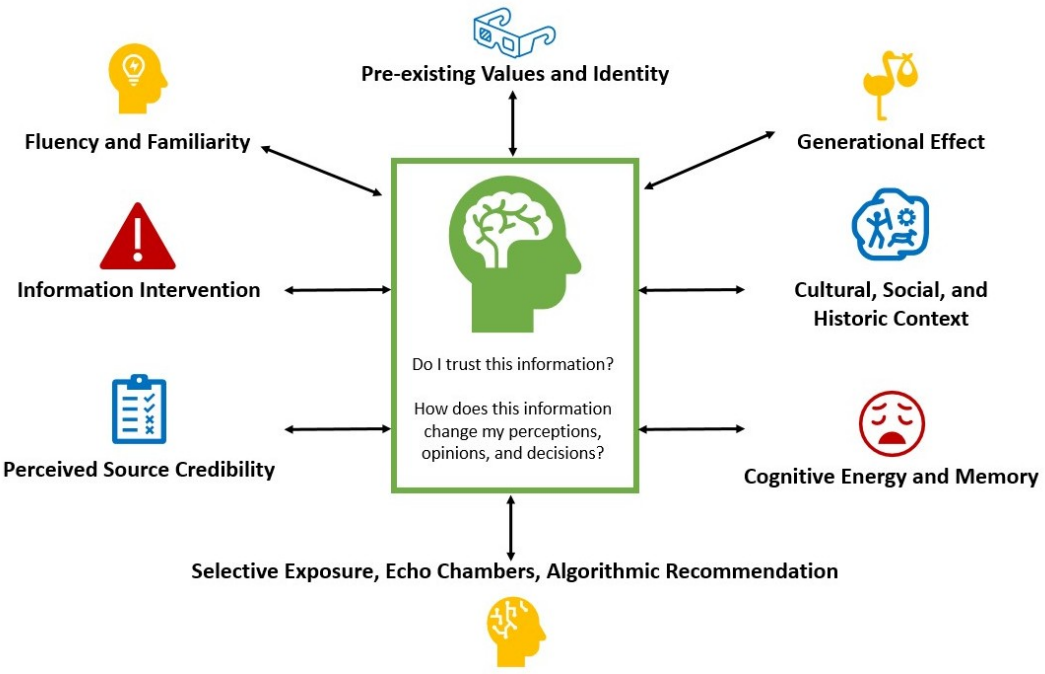

Figure 4: Information trust and influence sits within a complex set of reasoning, processes, and systems.

Within this space, I primarily try to understand media consumption and influence when consumers are intervened with (the yellow arrow in Figure 1), and how media trust decisions impact intervention effectiveness (the green arrow in Figure 1). For example, in our 2022 Communications of the ACM paper [23], we provide insight on the effectiveness of AI advice across information novelty, where novelty refers to the extent to which incoming information is similar to prior knowledge. The theoretical reasoning for this research comes from the

well-studied concept of confirmation bias (described above). In novel information situations we expect that prior beliefs are weak, while in familiar information situations, we expect prior beliefs to be strong. Hence, we used a between-subjects experiment to test if AI interventions were more effective when information was novel. Indeed, we found that interventions were significantly more effective in novel news situations, implying that the timing within the news cycle is a critical factor for intervention effectiveness.

The complex set of factors used in information trust play a role both when interventions happen and when they do not. Thus, it is also critical to develop a clear understanding of each factors role and in what context each factor matters when interventions are not used. While the influence of media is a well-studied topic across several disciplines, with several established theories of behavior, there are still many understudied gaps in our understanding, particularly as consumption contexts change online. My work in this area (when interventions are not in play) is my least developed body of work. So far, my focus here has been on the interplay of media reliance and cultural values on opinion formation. For example, in our 2023 Humanities and Social Sciences Communications paper [14], we explore generational and media-choice effects of information consumers in the Former Soviet Republics (FSRs) of Belarus, Ukraine, and Georgia. Through our analysis of representative surveys in each country, we found that media consumption does relate to opinions, such as opinions about one's country's future. However, these effects can be significantly moderated or exacerbated by generational effects (i.e., generations who grew up in the Soviet Union versus afterward).

4 Where Does My Work Fit?

4.1 Computational Social Science

I consider myself a Computational Social Scientist with expertise in Computer Science. Computational Social Science (CSS) is an emerging academic discipline that uses computational methods to analyze social science problems [18]. Work in this discipline is often done in teams (ideally of "computationally literate social scientists and socially literate computer scientists" [18]) and its researchers are housed in a variety of academic departments, including computer science, social science, cognitive science, business, information sciences, and even physics (often under the branch of physics that works on complex systems).

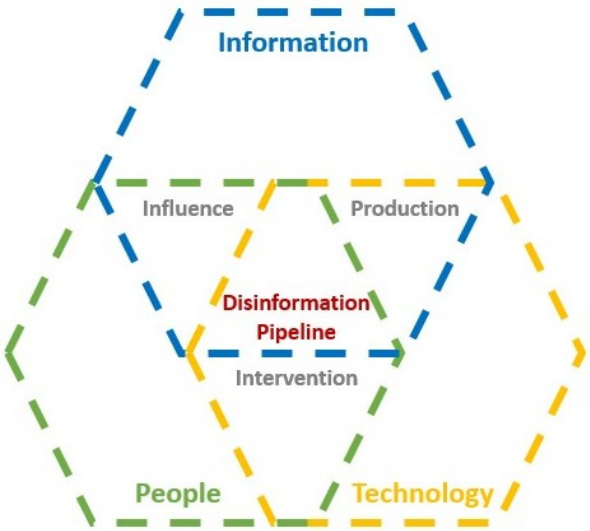

Figure 5: My work within Information Sciences.

As a Computational Social Scientist, I primarily focus on minimizing and mitigating the "wicked problem[5]" of disinformation and malign influence[6], rather than making methodological contributions to specific disciplines. However, given my problem domain, my work often makes contributions to research areas such as Human-Computer Interaction, Human-Information Interaction, Applied Machine Learning, and Social Computing. More broadly, I would argue that CSS and this particular wicked problem fit extremely well into the field of information sciences. A commonly used Venn diagram when describing the modern field of information sciences is the intersection of information, people, and technology (Figure 5). The disinformation pipeline framework used in my work maps directly to this Venn diagram, with media production tactics fitting between information and technology, interventions fitting between technology and people, and media influence fitting between people and information.

[5] https://en.wikipedia.org/wiki/Wicked_problem

[6] https://www.brookings.edu/articles/disinformation-as-a-wicked-problem-why-we-need-co-regulatory-frameworks/

4.2 An Interdisciplinary Approach

To deeply understand disinformation and to make practical contributions to society, we must cross disciplinary boundaries. As advocated in [17], disinformation research should be grounded in history, culture, and politics, all of which are fields of study with their own rich literature. Hence, I strategically collaborate and published with researchers from areas outside of the fields I was trained in, such as journalism, communication, political science, psychology, sociology, and anthropology, to learn diverse research methods and perspectives. These varying methods and perspectives are then fed back into my primary research agenda. I also try to use these diverse perspectives to make practical contributions other than publications, such as creating datasets for both computational and non-computational scientists and doing policy engagement.

Academic researchers are often not properly incentivized to do cross-discipline research, despite the terms "interdiscipline" and "transdiscipline" being heavily used in university marketing, grant proposal calls, journal scopes, and even job descriptions. In my opinion, part of this disconnect is due to communities and assessors not recognizing what interdisciplinary research outputs look like. Fruitful interdisciplinary collaborations are two-ways streets. Each discipline (and thus each person) has different top publication venues, writing conventions, terminology, and topical/theoretical priorities. Productive collaborations embrace this diversity, having each member contribute their expertise and skill's to a variety of studies and publication venues, creating what on the surface may look like an incoherent body of research from any one individual. Yet, as a whole these diverse publications create knowledge that informs each research agenda. Hence, this approach means that at times I produce work that is adjacent to my core research agenda. However, from a long-term point of view, I think these contributions and experiences will create better science for each member's discipline and help solve real world problems.

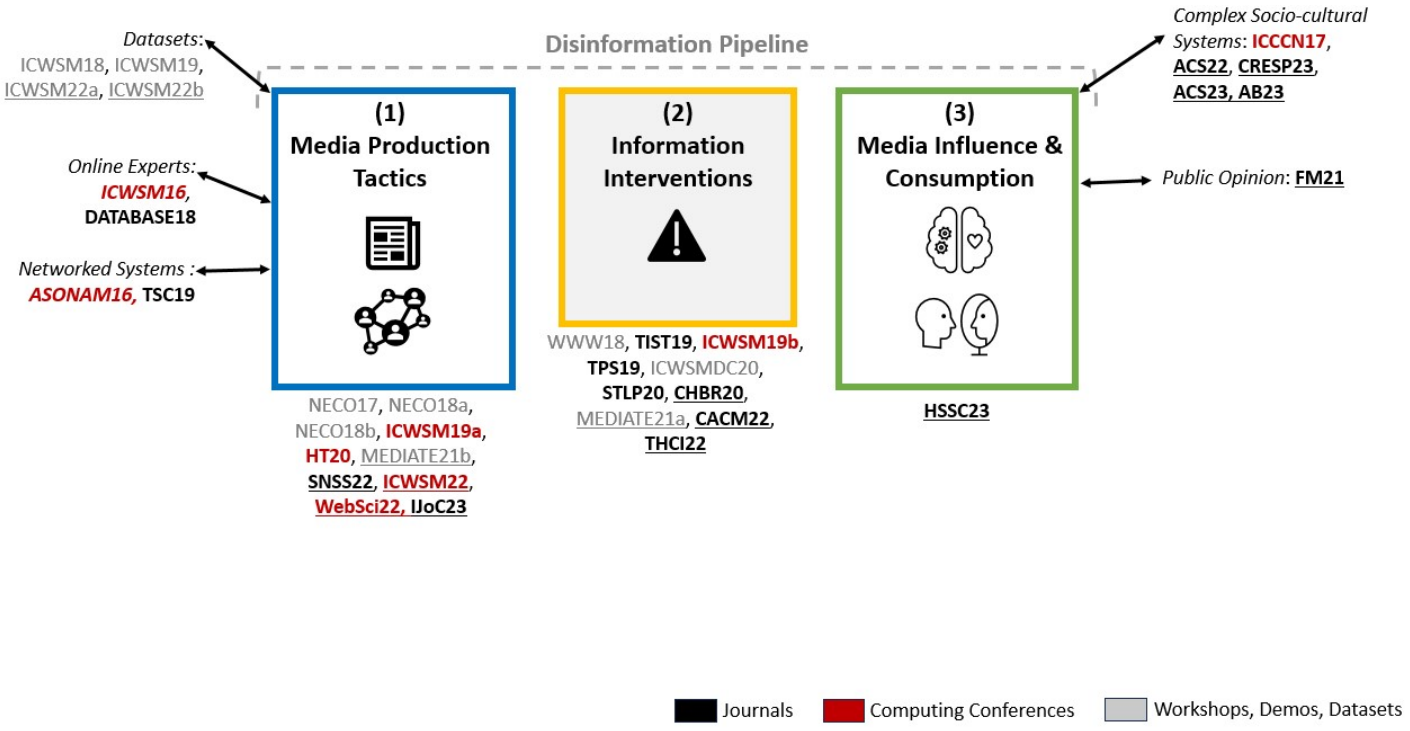

Figure 6: Outputs mapped onto research categories, where underlined works happened during my time at UTK. Sometimes papers can fit in multiple of these categories. For example. often the results from my works under information interventions overlap with media influence (as control variables often capture to social, culture, and trust factors), but I have placed them under information interventions as that is their primary focus. In addition to showing where outputs fit within my core I agenda, I show adjacent works that support parts of my core agenda through related knowledge creation.

References

[1] Abilov, A., Hua, Y., Matatov, H., Amir, O., and Naaman, M. Voterfraud2020: a multi-modal dataset of election fraud claims on twitter. *arXiv preprint arXiv:2101.08210* (2021).

[2] Aral, S., Muchnik, L., and Sundararajan, A. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences 106*, 51 (2009), 21544–21549.

[3] Childs, M. C., Buntain, C., Trujillo, M. Z., and Horne, B. D. Characterizing youtube and bitchute content and mobilizers during us election fraud discussions on twitter. *Proceedings of ACM WebSci Conference* (2022).

[4] Douglas, K. M., Sutton, R. M., and Cichocka, A. The psychology of conspiracy theories. *Current directions in psychological science 26*, 6 (2017), 538–542.

[5] Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., and Amazeen, M. A. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology 1*, 1 (2022), 13–29.

[6] Gruppi, M., Horne, B. D., and Adali, S. Tell me who your friends are: Using content sharing behavior for news source veracity detection. *MEDIATE Workshop at ICWSM* (2021). Available at: https://arxiv.org/abs/2101.10973.

[7] Horne, B. D., Dron, W., Khedr, S., and Adali, S. Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news. In *WWW Companion* (2018).

[8] Horne, B. D., Gruppi, M., and Adali, S. Do all good actors look the same? exploring news veracity detection across the us and the uk. *ICWSM Data Challenge* (2020). Available at: https://arxiv.org/abs/2006.01211.

[9] Horne, B. D., Gruppi, M., Joseph, K., Green, J., Wihbey, J. P., and Adali, S. Nela-local: A dataset of us local news articles for the study of county-level news ecosystems. In *Proceedings of the International AAAI Conference on Web and Social Media* (2022), vol. 16, pp. 1275–1284.

[10] Horne, B. D., Nevo, D., Adali, S., Manikonda, L., and Arrington, C. Tailoring heuristics and timing ai interventions for supporting news veracity assessments. *Computers in Human Behavior Reports 2* (2020), 100043.

[11] Horne, B. D., Nevo, D., O'Donovan, J., Cho, J.-H., and Adali, S. Rating reliability and bias in news articles: Does ai assistance help everyone? In *Proceedings of the International AAAI Conference on Web and Social Media* (2019), vol. 13, pp. 247–256.

[12] Horne, B. D., Nørregaard, J., and Adali, S. Different spirals of sameness: A study of content sharing in mainstream and alternative media. In *Proceedings of the International AAAI Conference on Web and Social Media* (2019), vol. 13, pp. 257–266.

[13] Horne, B. D., Nørregaard, J., and Adali, S. Robust fake news detection over time and attack. *ACM Transactions of Intelligent Systems Technology* (2019).

[14] Horne, B. D., Rice, N. M., Luther, C. A., Ruck, D. J., Borycz, J., Allard, S. L., Fitzgerald, M., Manaev, O., Prins, B. C., Taylor, M., et al. Generational effects of culture and digital media in former soviet republics. *Humanities and Social Sciences Communications 10*, 1 (2023), 1–11.

[15] Hutchinson, B., Rostamzadeh, N., Greer, C., Heller, K., and Prabhakaran, V. Evaluation gaps in machine learning practice. In *Proceedings of ACM FAccT* (2022).

[16] Joseph, K., Horne, B. D., Green, J., and Wihbey, J. P. Local news online and covid in the us: relationships among coverage, cases, deaths, and audience. In *Proceedings of the International AAAI Conference on Web and Social Media* (2022), vol. 16, pp. 441–452.

[17] Kuo, R., and Marwick, A. Critical disinformation studies: History, power, and politics. *Harvard Kennedy School Misinformation Review 2*, 4 (2021), 1–11.

[18] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al. Computational social science. *Science 323*, 5915 (2009), 721–723.

[19] Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., and Cook, J. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest 13*, 3 (2012), 106–131.

[20] Manikonda, L., Nevo, D., Horne, B. D., Arrington, C., and Adali, S. The reasoning behind fake news assessments: A linguistic analysis. *AIS Transactions on Human-Computer Interaction 14*, 2 (2022), 230–253.

[21] Marwick, A. E., and Lewis, R. Media manipulation and disinformation online.

[22] Munger, K. *Generation Gap: Why the Baby Boomers Still Dominate American Politics and Culture.* Columbia University Press, 2022.

[23] Nevo, D., and Horne, B. D. How topic novelty impacts the effectiveness of news veracity interventions. *Communications of the ACM 65*, 2 (2022), 68–75.

[24] Pennycook, G., Cannon, T. D., and Rand, D. G. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general 147*, 12 (2018), 1865.

[25] Rid, T. *Active measures: The secret history of disinformation and political warfare.* Farrar, Straus and Giroux, 2020.

[26] Ruck, D. J., Bentley, R. A., and Lawson, D. J. Religious change preceded economic change in the 20th century. *Science advances 4*, 7 (2018), eaar8680.

[27] Ruck, D. J., Matthews, L. J., Kyritsis, T., Atkinson, Q. D., and Bentley, R. A. The cultural foundations of modern democracies. *Nature human behaviour 4*, 3 (2020), 265–269.

[28] Trujillo, M., Gruppi, M., Buntain, C., and Horne, B. D. The mela bitchute dataset. *Proceedings of the International AAAI Conference on Web and Social Media 16* (2022).

[29] Zannettou, S., Sirivianos, M., Blackburn, J., and Kourtellis, N. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality (JDIQ) 11*, 3 (2019), 1–37.