Language Models and Online Bot Revolution

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Maxwell Patek
Spring, 2020

On my honor as a University Student, I have neither given nor received
unauthorized aid on this assignment as defined by the Honor Guidelines
for Thesis-Related Assignments

Introduction

> *Recycling is good for the world.*
> *NO! YOU COULD NOT BE MORE WRONG!!*
> *Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I'm not kidding….* (Radford et al. 2019)

The above text was generated by a computer. It was fed the first sentence as a prompt, and it outputted the rest. Computers can now talk like humans. State-of-the-art language modeling algorithms demonstrated human-realistic text during the summer of 2019. Given prompts about unicorns and news events that had not actually occurred, the model generated text that mimicked scientific writing and journalism that is indistinguishable from human text. These developments in language models have arrived when online text-based communication is more important than ever. Online forums have fundamentally changed society by connecting physically remote groups and fostering communication and dialogue. Entire new industries have developed as content creators, advertisers, and platforms cooperate to vie for online customers. Politics has transformed into a battleground of online impressions. Generation Z is now growing up getting their news from faceless tweets, posts, forums, and articles. The proliferation of advanced language models will have a profound effect on this ecosystem. Online bots that impersonate real people are able to disrupt, skew, and manipulate individuals and groups in the real world. Content creators, advertisers, platforms, engineers and research scientists, customers, and regulators comprise the status quo actor network; however, the addition of indistinguishable online bots and their deployers will disrupt that status quo. Additionally, a paradigm shift into an era where computers can pass the Turing Test will affect the way engineers and scientists adapt to the emerging technology.

Goal and Approaches

How can the software engineers and research scientists that support online forums adapt to the proliferation of bots that are indistinguishable from humans without infringing on free speech and privacy? Documentary research provides a foundation for analyzing both technical and social problems. Network analysis is an effective way to synthesize the literature and gain new insight into the intricacies of the problem space. Network analysis also provides a unique perspective on bots, since it treats both human and non-human actors as identical. Wicked problem framing is effective at extrapolating the analysis to the future of bot development and providing insights on how to adapt in the context of emerging technology. Together, documentary research, network analysis, and wicked problem framing are the methods of analysis for the problem of how to adapt to advanced language models.

Online Platforms and Bots

Online platforms have discovered a new power. Machine Learning (ML) yields an endless supply of business. ML models analyze our patterns and preferences, which then determine the content we see in our news, social media feeds, search results, and advertisements. This technology makes the business model of platforms like Google, Facebook, and Twitter profitable, facilitating growth. The general public now considers these platforms indispensable. A constant stream of text-based information keeps people informed and allows groups to communicate and share. For example, many Americans use social media to stay informed on current events, since state officials, like the President of the United States, use Twitter extensively for sharing updates on public policy. Additionally, Online forums have been used to give a voice to otherwise voiceless groups of people, allowing the oppressed to rally and enact change.

As society's dependence on the platforms increases so does our need to validate and verify the information we consume. In an in-person forum, the information that people share is not necessarily true, and participants doubt and question each other to deduce truth. However, online forums complicate this due to anonymity and distance. Users are removed from the source of the information that they are consuming, making it easy to forget that the other party may not be accurately representing themselves. Online, participant identity becomes other information that may be false and needs to be questioned.

Online identity is further complicated by the presence of online bots. People, referred to as adversaries, who seek to gain from the inability to verify online identity create algorithms to spoof human content, and technologists work to write algorithms to detect the spoofed content, creating a competition that has waged since the dawn of the world wide web (Lazer, et. al. 2018). These algorithms, referred to as bots, cause many problems. They can skew demographics, making groups feel alienated and growing hurtful groups, distorting views by spreading messages and creating a chorus of agreement or a false discourse. Often these manipulations can cause real people to change their views and act. For example, bots were used to influence voters in the 2016 US Presidential Election (Bessi, Ferrara, 2016). Bots have an amplifying effect on existing problems with online platforms. For example, the rise of fake news, which also had an effect in the 2016 election, is exasperated by bots that can interact with and promote the false content (Allcott, Gentzkow, 2017). The effects of bots can be economic as well. They can disrupt the online advertisement market by generating false interactions with ads and can generate false reviews, misleading customers on ecommerce sites (Adelani, et. al., 2019). They can also steal and duplicate original online content, siphoning revenue from content creators.

Advances in bot technology are exasperating these problems. The underlying technology of text-based bots are Language Models. Bellegarda explains the fundamental goal of language modeling in 2004 before many recent advancements in machine learning: “Regularities in natural language are governed by an underlying (unknown) probability distribution on word sequences. The ideal outcome of language modeling, then, would be to derive a good estimate of this distribution” (Bellegarda, 2004). Historically, largely due to computational limits on hardware, language models had been limited to domains where the set of possible word sequences is highly constrained. However, with modern GPUs, which enabled a renaissance of machine learning, especially deep learning, language models are able to predict increasingly large corpuses with growing accuracy. The effect is that language models are now able to take some input prompt and generate text that would likely follow that prompt. For example, the first sentence of the introduction of this research paper is a prompt fed to the GPT-2 model, and the rest of the paragraph was generated by the model (Radford et al. 2019). While not completely sensible, the paragraph does read like human text. Already, it is difficult to differentiate between bot-generated text and human text. By being convincingly human, bots are now able to elicit interaction from real people and infiltrate real online communities.

Paradigm Shift and Actor Network

Science Technology and Society frameworks are an effective way to contextualize the problems just outlined. Paradigm Shift Theory is applicable here. American Physicist and Philosopher, Thomas Kuhn describes paradigm shifts as time periods when the accepted rules and norms of a scientific discipline change (Kuhn, 1964). Such a paradigm shift is necessary in order to adapt to the latest language model technology. The Turing Test is an experiment to see if bots can imitate humans. If a human is unable to differentiate between a bot and a real person, the bot is said to have passed the test. Research scientists that work to detect online bots have assumed that humans can differentiate between human text and algorithmically generated text, i.e. that the Turing Test is not being passed. Because bots are now infiltrating groups of real humans online, that assumption is starting to fail. Even with the most sophisticated deep learning models, if humans cannot differentiate between bot text and human text, there is no ground truth to train a differentiating model, meaning that it may be impossible to algorithmically detect bots. This shift has huge implications on scientists' approach to finding truth online.

As with any framework, there are shortcomings and criticisms. While Paradigm Shift is an effective way to analyze the recent developments in language model research, British Philosopher Martin Cohen claims that paradigm shift is inadequate, since it does not acknowledge the greater shortcomings of scientific research in general (Cohen, 2015). He notes that scientific research is transitory anyway and that identifying shifts is almost redundant. This criticism is certainly applicable since machine learning is a domain where the rules are constantly being broken; rapid development of deep learning in the past decade has proven that.

Actor-Network Theory, invented by sociologists Michel Callon, Bruno Latour, and John Law, is a method of describing the interactions between human and non-human actors over time (Cressman, 2009). Because of the blurred lines between artificial and genuine intelligence, and the ability for non-human actors to interact so intimately with human actors in this case, Actor-Network Theory is especially relevant. However, the theory is criticized for being too general and incorrectly applied. An actor network can be constructed to fit almost any given case, even if the network is not descriptive of reality. Designed networks can easily punctualize important subnetworks that are actually crucial to gaining correct insight. In such a complex problem space as online bots, this criticism is certainly relevant.

Analysis of Actor Network

Despite being a highly connected actor network, there are several cornerstone actors with the ability to adapt to the paradigm shift at hand: human users, forum creators, and regulators. Users can verify themselves to make identifying bots easier, forum creators can facilitate transparency and verification on a spectrum, respecting privacy, and regulators can deter bot creators. However, each of these adaptations has fundamental conflicts with the interests and expectations of other actors. Thus, additional adaptation beyond the current actor network is necessary. Either privacy can be dramatically sacrificed, or bots can be reevaluated as legitimate actors in an online forum. These adaptations would have long-lasting effects on online forums, but would likely be more effective than other approaches. However, the problem with online bots can be reframed into larger wicked problems inherent in online text-based forums at large scale. Solving those problems is the truest approach to adapting to indistinguishable online bots.

Housing association expert Dankert recommends starting Actor Network analysis with the central actant, then following connections outward (Dankert, 2011). The essential actant of the research question is the bots, which occupy an interesting middle ground of human and non-human, since they attempt to appear human and are designed by humans, but are not human themselves. The two most immediately connected actants are the bot's creator and the bot's audience, which it is trying to deceive. In order for the bot to operate, it needs a medium to communicate through. For the purposes of this research, the medium is online, text-based forums, including Facebook, Twitter, Reddit, and the comments on any given video or other user content. Herein lies the third actor, the forum. Like the bots, the forum has two connected actants, the forum creators and the forum users. The bot audience is a subset of the forum users. The creators of the forums are the engineers, research scientists, managagers, and other members of the organizations that build forum websites. For the most popular forums, these organizations are for-profit corporations like Facebook, Inc. Twitter, Inc. Google, LLC (via YouTube and hosted blogs etc), Amazon, Inc (in product reviews), etc. There are also many less popular forums that are created by individuals and smaller for/non-profit entities; however, because they are not created by engineers and research scientists and are generally less influential, they will not be part of this research. The forum creators have several connected actants, including the regulators that govern their practices and the third parties that advertise on the forum. In addition to being connected via the forum, the users and the forum creators are connected directly via a conceptual actor, trust. Similarly, users trust regulators to govern the creators in their best interest. Together, users, advertisers, and bots create the content that populates the forum. This content may or may not be the focus of the forum. For example, Amazon product reviews are a forum, but the product itself is content that is not part of the user, advertiser, bot system. This extra content is not directly relevant to the discussion, but the content in the forum is. An overview of the actor network can be seen in Figure 1.

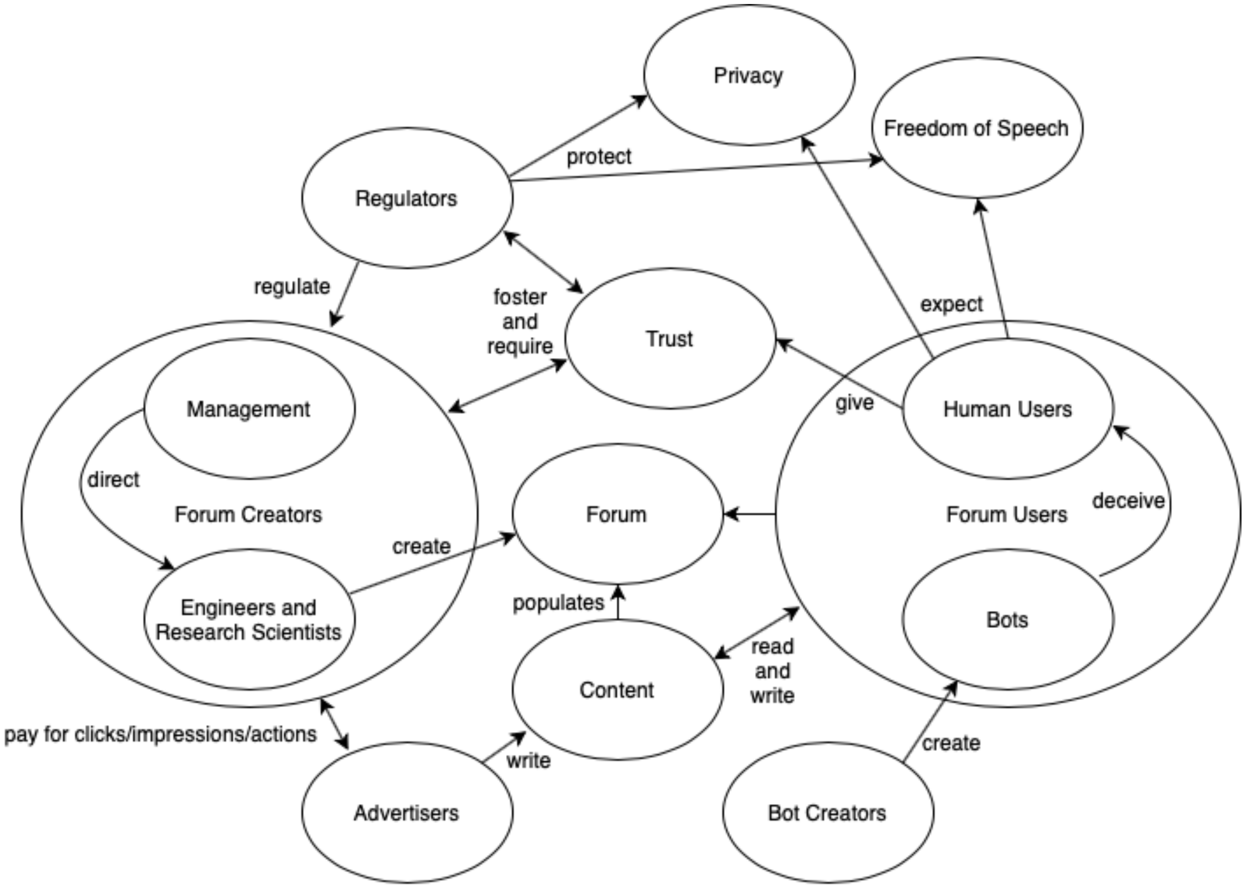

Figure 1: Actant Network Originating from Bots

To start analyzing this network, the problem needs to be contextualized in its terms, which is done by temporarily removing all bot-related actants from the graph. Without bots, information flows from the human users and advertisers to the forum as content, and the content is then fed back to the human users as the content they see when they read the forum. Trust flows from the human users to the forum creators and regulators. The regulators maintain trust by introducing and enforcing policies that guarantee privacy and freedom of speech, which users expect. However, the way that forum creators maintain trust is unclear, since they are more directly incentivized by the advertisers that pay for user attention.

Reintroducing bots affects the behavior of forum users. With bots, the content that users see is created by advertisers, human users, *and* bots. If the content is recognizable as bot-generated, then the human user can disregard or qualify it, but if not, it will appear as genuine human content. Knowing that content is created by a bot introduces a layer of indirection in the way the content is interpreted. Most humans on online forums want to interact with other humans, so if the bot is acting independently, there is no perceived value in the content, and users react strongly (Aiello et. al. 2014). The only way the bot represents a human is through its creator. However, users know that the bot creator likely has many such bots, and so that human's voice is being multiplied. To avoid bias, many human users will disregard such content to avoid becoming biased towards bot-multiplied voices.

If the bot is indistinguishable, however, human users in the current societal culture tend to assume that a real human generated the content, which is partly due to limited awareness of the presence of bots that cannot be distinguished from humans. However, even if awareness were high, by definition, the indistinguishable bot cannot be disqualified by any criteria inherent in its content. The best an aware human user could do would be to guess on the humanity of the content or to assign some partial value or partially disregard the content, which is not ideal.

The consequences of this deception are traced outward from the human users. If humans believe the bot's content, they can be manipulated into acting in someone else's best interest. This can be seen in the US election in 2016 (Bessi, Ferrara, 2016). Aware humans, knowing that bots are generating at least some of the content they see, and knowing that that content misrepresents the world, begin to distrust the content, and therefore the forum and the forum creator (Aiello et. al. 2014). Thus, the first place users might expect a fix for this problem is from the forum creators.

The forum creator could have the power to disambiguate bots from humans by requiring proof of the identity of the user associated with the content. By requiring some information that is not easily forged, forums could have a variety of verification levels, ranging from "unverified, possibly a bot" to "completely verified, we know exactly who this person is." Content could then be tagged with the amount of verification the content creator has, allowing users to trust content that if verifiably not bot-created. This approach has several problems. Requiring such detailed information from human users breaks users' expectation of privacy, which users trust regulators to protect. Even if users were able to opt out of the verification process, their content would begin to be mistrusted if they chose to do so. Additionally, many online bots assume the accounts of compromised or inactive users, so a once-verified account could become a bot later, meaning that the forum would have to constantly reverify the user (Murthy et. al., 2016). This approach would require changing the societal norms about what identifying data forum creators like Facebook should possess.

Another approach forum creators could use to tackle the bot problem requires special attention to the motivations of the bot creators. Bot creators are often trying to manipulate real people into interacting with certain content. Among the most prominent usages of bots is to get people to interact with fake news. This is called astroturfing (Kovic et. al., 2018). The forum creators, to address the problem introduced by bots, could take steps to limit fake news in general. However, this can break the user's expectation of freedom of speech, which regulators are also trusted to protect. Additionally, determining what content is true and false is arguably harder than determining what users are bots or humans; it is still an open problem with potentially epistemological barriers (Anderson, Rainie, 2017). This approach is not technically or regulatorily possible in our current society.

Although forum creators are limited by the expectations of their users, it is possible that regulators could play a more direct role. Even for crimes that are hard to detect, governments are often able to decrease the amounts of that crime by enforcing a deterrent. For example, insider trading is difficult to detect when measures are taken to conceal it (Beams, 2002). However, the consequences of being caught are criminal, which is preventative enough for much of the population. A similar approach might be taken to reduce the bot problem. Currently, less advanced, but still effective bots are available for purchase by ordinary users (Murthy et. al., 2016). The existence of a market for bots makes it easier for semi-determined people to use them. These partially determined actors might also be adequately deterred by a substantive prison sentence, if only regulators played a more direct role in preventing bots. However, one limitation of this approach is that the bots available for purchase are less advanced than the emerging bots that pose the biggest threat of infiltrating and garnering interaction from the most humans. Additionally, the most nefarious users of bots are probably not concerned with criminal consequences, since their goals are likely illegal as well. Finally, this solution ignores the fact that bots are becoming increasingly indistinguishable; deterrence is ineffective if there is no chance of being caught.

Paradigm Shift and Reframing

All possible approaches so far have involved human organizations changing their behavior towards other actants. However, they all conflict with some other expectation or societal norm. This conflict, in conjunction with technical impossibility of detecting an agent that uses detectors to learn, like Generative Adversarial Networks, implies that a paradigm shift is occurring. If we assume that we cannot distinguish advanced bots from humans, no solutions arise from the existing actant network, so either the actant network must change or the problem must be reframed.

While reducing freedom of speech to allow for forums to reduce fake news was a potential for mitigating the effects of bots, it was infeasible for epistemological reasons. However, relaxing users' expectation of privacy has potential. If everyone had verified accounts that were tied to their identity, similar to the verified “official” accounts that many celebrities have on popular forums, bots would be uniquely identifiable as the users with no such verification, if they were allowed on the forum at all. However, this would end anonymization online. Anonymous content would be relegated to bot-quality. If users had no expectation of privacy, and were ok with associating their name with everything online, this solution would have no insurmountable problems in the given actor network. Adapting to the paradigm shift in this way would return society to that reminiscent of smaller-tightly knit communities, where anonymization was not assumed; "small communities where membership itself is a form of vouching" (Marx, Caplan, Torpey, 2001). The internet's facilitation of a global community has deviated from this old model, but the reward of returning would be immediate.

The primary problem with this adaptation is that there is not yet any incentive for any single user to verify their account in this way. The benefits are reaped when the majority of the user base is verified, when verification can begin to be used as a signal against bots. Being the first to verify one's account only sacrifices the privacy of the user. The bot problem would need to be more visible to ordinary users for the long-term benefits to be motivating enough for them to start to share that information.

The other profound adaptation to the paradigm shift towards indistinguishable bots would be to accept the presence of the bots and neglect the related problems (Ferrara et. al., 2016). Fundamentally, there is no reason why a human user cannot have a meaningful and productive interaction with an advanced online bot. If bots were an accepted part of online forums, well-meaning bot creators could spread well-meaning bots. Making bots more available and supported on forums could rebalance their amplifying power away from nefarious bot creators.

An obvious consequence of this adaptation would be that forums would be flooded with new content, as every legitimate human viewpoint gets multiplied by an army of bots. This causes technical problems for the forum creators, but more importantly, would hurt human user's ability to find value in the forum content. There would be more reliance on the news feed, search, and recommendations algorithms that are already criticized for being too influential on user's online activity (Geschke, 2019). If the algorithms were successful, forums might function similarly to how they do now. If the algorithms were unsuccessful, users would likely return to smaller circles of personally familiar online connections. There would be a decline in the usage of online forums as a way to bring together disparate and underserved groups.

Ultimately, the problem with online bots is a reflection of the problems with massive online anonymous text-based forums in general. Users are trusting of the content they see, regardless of its source, and news feed, search, and recommendation algorithms trap users in echo chambers (Geschke, 2019). Fake news proliferates without the usage of bots (Allcott, Gentzkow, 2017). Privacy is violated, and freedom of speech is challenged. Bots are only a particularly scapegoated amplification of these underlying wicked problems. It's likely that the best and longest-term solution to bots is to approach the problems with online forums in general with a holistic sociotechnical approach. However, those problems are out of the scope of this paper. How those problems relate to bots in particular is excellent future work for Science, Technology and Society researchers. Quantifiable experiments on how human users react to the adaptations mentioned above would inform forum creators, regulators, and users on how to progress as language models become more advanced and proliferated.

This research is limited by many factors. Many of the potential solutions and adaptations are fairly under-researched, and experiments should be run on human behavior in each of the aforementioned scenarios. Even with that data, however, we can only speculate on how those adaptations would work at the scale of a platform like Facebook, with over 2 billion users. Additionally, future developments in online technology are hard to predict in general. It's possible that developments in areas like differential privacy or other undiscovered technologies will resolve the current bot problem. Finally, this research was strictly limited by the time given to conduct it. As a senior thesis, only nine months were available to collect the vast amount of information available on online forums, bot technology, and human behavior.

Conclusion

Engineers and research scientists have a difficult job ahead of them. The status quo network of actors restricts the response to online bots to a series of optional actions. Ultimately users are the ones who need to adapt to bots. Thus, the engineers and research scientists have a responsibility to ease the transition for the users. This includes fostering trust in the forum and its creators and providing transparency and choice. The technical paradigm shift of indistinguishable online bots demands a change in societal norms and the expectations of users, so regulators and forum creators need to work together to make that change painless. Working to mitigate the larger problems with large online text-based forums will help mitigate the amplified instances of those problems presented by bots. Hopefully, these solutions will preserve the connective power that the internet has delivered since its inception. However, the future of online interaction is hard to predict. All society can do is be mindful in engineering, regulating, and using the forums of the future.

Works Cited

Adelani, D. I., Mai, H., Fang, F., Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Generating Sentiment-Preserving Fake Online Reviews Using Neural Language Models and Their Human- and Machine-based Detection. *ArXiv:1907.09177 [Cs]*. Retrieved from http://arxiv.org/abs/1907.09177

Aiello, L. M., Deplano, M., Schifanella, R., & Ruffo, G. (2014). People are Strange when you're a Stranger: Impact and Influence of Bots on Social Networks. *ArXiv:1407.8134 [Physics]. http://arxiv.org/abs/1407.8134*

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Anderson, J., & Rainie, L. (2017). The Future of Truth and Misinformation Online. *Pew Research Center,* 224.

Beams, J. D. (2002). *Insider Trading: A Study of Motivations and Deterrents*. https://vtechworks.lib.vt.edu/handle/10919/30182

Bellegarda, J. R. (2004). Statistical language model adaptation: Review and perspectives. *Speech Communication*, *42*(1), 93–108. https://doi.org/10.1016/j.specom.2003.08.002

Bessi, A., & Ferrara, E. (2016). *Social Bots Distort the 2016 US Presidential Election Online Discussion* (SSRN Scholarly Paper No. ID 2982233). Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=2982233

Cohen, M. (2015). *Paradigm Shift: How Expert Opinions Keep Changing on Life, the Universe, and Everything*. Retrieved from http://ebookcentral.proquest.com/lib/uva/detail.action?docID=4393934

Cressman, D. (2009). *A Brief Overview of Actor-Network Theory: Punctualization, Heterogeneous Engineering & Translation*. Retrieved from https://summit.sfu.ca/item/13593

Dankert, R. (2011, November 30). *Using Actor-Network Theory (ANT) doing research*. Tips over beleid maken, schrijven en uitvoeren. https://ritskedankert.nl/using-actor-network-theory-ant-doing-research/

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2017, July). *The Rise of Social Bots*. https://cacm.acm.org/magazines/2016/7/204021-the-rise-of-social-bots/fulltext

Geschke, D., Lorenz, J., & Holtz, P. (2019). The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology, 58*(1), 129–149. https://doi.org/10.1111/bjso.12286

Kovic, M., Rauchfleisch, A., Sele, M., & Caspar, C. (2018). Digital astroturfing in politics: Definition, typology, and countermeasures. *Studies in Communication Sciences, 18*(1), 69–85–69–85. https://doi.org/10.24434/j.scoms.2018.01.005

Kuhn, T. S. (1964). *The Structure of Scientific Revolutions*. Retrieved from https://search.lib.virginia.edu/catalog/u3307668

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., … Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Marx, G. T., Caplan, J., & Torpey, J. (2001). *Documenting Individual Identity: The Development of State Practices in the Modern World*. Princeton University Press.

Murthy, D., Powell, A. B., Tinati, R., Anstead, N., Carr, L., Halford, S. J., & Weal, M. (2016). Automation, Algorithms, and Politics| Bots and Political Influence: A Sociotechnical Investigation of Social Network Capital. *International Journal of Communication, 10*(0), 20.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (n.d.). *Language Models are Unsupervised Multitask Learners*. 24.