Large Language Models Can Argue in Convincing Ways About Politics, But Humans Dislike AI Authors: Implications for Governance

Alexis Palmer[*] Arthur Spirling[α]

March 25, 2024

Abstract

All politics relies on rhetorical appeals, and the ability to make arguments is considered perhaps uniquely human. But as recent times have seen successful large language model (LLM) applications to similar endeavors, we explore whether these approaches can out-compete humans in making appeals for/against various positions in US politics. We curate responses from crowdsourced workers and an LLM and place them in competition with one another. Human (crowd) judges make decisions about the relative strength of their (human v machine) efforts. We have several empirical "possibility" results. First, LLMs can produce novel arguments that convince independent judges at least on a par with human efforts. Yet when informed about an orator's true identity, judges show a preference for human over LLM arguments. This may suggest voters view such models as potentially dangerous; we think politicians should be aware of related "liar's dividend" concerns.

Introduction

[*] Department of Politics, New York University, New York, NY 10012.
[α] Department of Politics, Princeton University, Princeton, NJ 08544.

What persuades an audience to accept a particular argument may be the oldest and most studied political science question of all. And Aristotle's *Rhetoric* arguably remains the standard for understanding this process. In that account, speakers have three resources to convince their listeners: the speaker's own personal character (*ethos*), the emotional feelings of their audience (*pathos*) and the quality of the logic in the argument itself (*logos*). Perhaps the most obvious example of these concepts is when politicians compete for votes by debating in front of the electorate, but we often see leaders convincing citizens to do other things. These include living healthier lives or signing up to new policy schemes.

A natural assumption historically is that the entity making the argument is *human*; however, recent technical advances means that this need not be the case. That is, we now have access to generative "large language models" (LLMs) that allow computers to produce human-like text in response to user prompts (see e.g. Dai and Radford, 2023; Halterman, 2022 for recent political science applications). For social scientists interested in persuasion, a fundamental question is whether these machines can out-perform Aristotle's "political animal" (i.e. mankind) in their rhetorical interactions with other humans.

This matters for several reasons. First, because it teaches us something inherently interesting about arguments—what works and what doesn't. And because these machines may then be a useful tool in making the public case for a position. Indeed, scholars in the discipline are already applying related 'chat' technologies to potentially alter citizen perceptions of interventions (e.g. Rosenzweig and Offer-Westort, 2022).

Second, it matters because LLMs used in this way may pose a threat to democracy *per se*. Social scientists have written extensively on the possibility that new generative technologies could lead to a "liar's dividend" (Chesney and Citron, 2019), by which voters cannot tell whether the messages they receive are true data about the world or misleading misinformation. This potentially helps those actively seeking to destabilize polities and may be bad for accountability more generally (Schiff, Schiff and Bueno, 2022). The historical

focus in that literature has been on "deep fake" images, audio and video, but obviously it could apply to the text product of an LLM too. In this scenario, LLMs may be a way to flood the discourse with "fake rhetoric" (rather than "fake news" in the sense of Lazer et al, 2018), thereby confusing citizens as to the genuine empirical or ideological arguments for policy. This could conceivably lead to bad actors convincing voters to do things against their own interests, or to do things that harm more vulnerable members of society. This is all the more true if citizens show no inherent preference for human generated reasoning, assuming they ever discover the genesis of the arguments they read.

For these reasons, we ask not merely whether LLMs can construct an appealing argument in terms of content (logos), but also how an audience responds to their *ethos*—that is, the knowledge that the orator is a machine rather than a human. In this way we connect longstanding questions of political philosophy to those of political science, via the methods of computer science. We use an open source LLM—the Meta OPT-30B model (Zhang et al., 2022)—and prompt it to make arguments for and against common positions in contemporary US politics and society.[1] These same argument prompts are then given to *humans*, specifically large numbers of crowd workers. The pairs of responses (one human, one LLM) are then shown to a set of independent human judges. Those judges must decide whether the machine or human argument for a position is the more convincing. To be clear, we curate both (human and LLM) sets of responses to ensure the contests are between the "best" quality outputs. In that sense, our headline findings are "possibility" results. Importantly for assessing causal claims about *ethos*, we randomize whether crowd respondents are informed about the identity—machine or human—of the argument producer. That is, in some cases respondents are aware which position statement was produced by the LLM, and in some cases they are not.

[1] See e.g. Spirling (2023) for discussion as to why open source LLMs may be generally preferable to proprietary efforts.

Our findings are first, that LLMs are capable of producing human-style arguments for different positions on subjects as varied as abortion, guns, immigration, and organ donation. In terms of convincing human judges, they can out-perform human authors, though this varies by topic. Second, we show that the structure and style of LLM arguments differs from those offered by humans. Specifically, LLM arguments tend to have lower levels of reading difficulty and differ somewhat in substance from human positions. Finally, when informed of orator identify, human judges show a small but statistically significant preference for human producers for arguments—though this is partly driven by certain issues. In terms of our concerns about democratic health, the results are mixed. On the one hand, we anticipate that producing compelling arguments for propositions---whatever their welfare effects---will be straightforward for LLMs. That may be good or bad depending on the use-case. That is, it may yield healthier, happier citizens who make better choices after being convinced on the merits, or disengaged cynics who trust little of what they read. On the other hand, informed citizens are somewhat wary of machine output *per se*, which may encourage politicians to steer clear of using LLM generated arguments for their own personal purposes, or to regulate their use in politics more generally.

Results

Our first goal is to assess whether and to what extent LLMs can make arguments—and how well they can do this relative to humans. We begin by demarcating the five issue positions for which the arguments should be made. Three of these issues are known to be some of the most polarizing matters in contemporary US politics (see, e.g., Grumbach, 2018), namely abortion laws, gun rights and immigration. These "polarized" prompts are, respectively:

1. Recently, there has been a lot of discussion in the US about gun rights and gun control. Some people favor more gun control, and others do not want to add restrictions. From your perspective, what is the best argument for [against] more gun control?
2. Abortion is a heavily debated topic in the US. Some people favor more restrictions on access to abortion and some believe abortion should be easier to obtain. From your perspective, what is the best argument for easier access to [more restrictions on] abortion?

3. There are many diverse opinions on immigration to the US. From your perspective, what is the best argument for increasing [restricting] immigration to the US?

A respondent—either a human or the LLM—randomly receives either the prompt as is, or with the relevant position (underlined above) substituted with the contents of the square brackets. Note that, slightly differently in each case, the prompts make reference to current debates or discussion about these matters. They are written in language similar, but not identical, to that used by public opinion researchers.[2] The fourth issue was deliberately chosen to be low salience and of no particular partisan association—namely, the use of an "opt-in" versus an "opt-out" scheme of organ donation after death:

4. In some countries, organ donation after death is the default: people must explicitly 'opt out' of the scheme while alive. In the United States it is not the default, and people must explicitly 'opt in' for their organs to be donated after death. What is the best argument for an 'opt in'[opt out] system?

The fifth prompt is about policy priority (rather than preference) and asks "What social, economic or political problems do you think will be most important in the United States over the next 10 years?" In every instance, respondents have a total of 300 characters (maximum, counting down on-screen as respondents typed) to make their case. So: there are a total of nine possible prompts: the pro/anti for issues (1)–(4), and then the fifth prompt which has only one mode.

Our setup is that respondents to the prompts are of one of two types: they are the machine (LLM) or human crowdworkers. The outcome was a set of 30 LLM provided arguments[3] and 25 crowdworker arguments for each of the nine prompts. We curated these response sets

[2] For example, Gallup has historically asked "In your view, should immigration be kept at its present level, increased or decreased?"

[3] For 4 prompts, a large run of 300 arguments was generated. We did this to get a general sense of performance, in terms of how many 'usable' arguments we should expect: see Supporting Information (SI) C.

such that the resulting contests would be between the highest quality outputs from both types of producers.[4]

In the second stage of the analysis, an independent set of 760 crowd workers were asked to adjudicate between two arguments for a particular position (say, both in favor of expanded gun rights, or both in opposition to more immigration)—one of which was produced by a machine, and one by a human. These arguments were randomized in terms of their order on the page. We designed a lightweight web application for this purpose. These judges were asked to "click the one that you find more convincing". In some (randomized at respondent level) cases, the judges were told (truthfully) which argument was produced by a human, and which was produced by "an Artificial Intelligence algorithm". In all cases, workers were shown a total of 10 sets of arguments (one LLM, one human written for each prompt) for prompts randomly chosen from our set of nine. Therefore, each prompt with a pair of arguments was shown 840 times and each individual argument approximately 280 times.

LLMs can make *convincing* arguments

We say an argument is "convincing" to the extent that independent human judges prefer it to another. The structure of the tasks above means that the relevant comparison is statistically simple, and in Figure 1 we show the probability that the human-generated (as opposed to LLM) argument was chosen by crowdworkers. This was calculated from a linear probability model, both overall (All) and for each prompt. We provide a 95% confidence interval on each value. In two cases—arguing for opt-in organ donation and for more restrictions on abortion—the human written arguments were consistently preferred to the LLM written ones. Put differently, for every other argument, there was no statistically significant difference between the LLM and the human writers, in terms of their ability to convince a judge (everything overlaps with a 0.5 probability). The actual data and *p*-values are included

[4] SI D gives more information on the curation process.

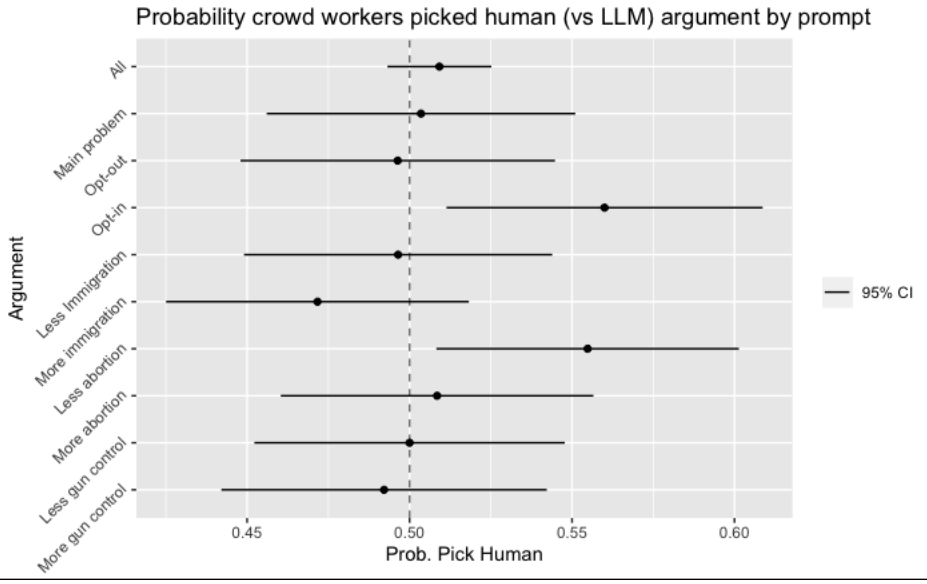

Figure 1: LLMs and humans are generally equally able to convince independent judges as to the merits of arguments for positions.

in SI F. In terms of individual arguments, based on a qualitative reading, the most preferred arguments were those that were more moderate and displayed some amount of nuance. This is true across implied partisan stances (i.e. whether for or against a given position). This suggests that crowdworkers are indeed considering the persuasiveness of the response, rather than (only) judging quality as a function of their personal stance on an issue. Overall, the quantitative "no rhetorical edge" result is interesting, but it does not mean there are no substantive differences between human and LLM arguments. We now turn to these.

LLMs can make *novel* arguments

We say a set of arguments is "novel" to the extent that it differs in some well-defined qualitative or quantitative way from another set. Here, our interest is how arguments produced by the LLM—irrespective of their ultimate popularity—have properties in common with each other, and different to those of the humans.

In SI E we give full details of that analysis, but our summary is as follows. First, LLM arguments are easier to read (in terms of traditional Flesch Reading Ease measures). Second,

LLMs produce arguments that are consistently more positive in tone than human ones—as measured by sentiment dictionaries. In addition, LLM arguments exhibit lower variance on both characteristics (i.e. generally more similar to each other) than human efforts.

The LLM produced more coherent arguments on topics where (we believe) it has access to copious training data—e.g. social media posts for abortion discussion. For instance, tracking terms on Reddit—one of the sources of training data for the Meta OPT model—shows the phrase "gun control" appears at least ten times more often than "organ donation". Consequently, for the more obscure topic of organ donation, the model produces fewer unique arguments and less 'human-like' text as discussed below.

Though admittedly a judgement call, we note that the LLM tends to argue in more simplistic, direct, less nuanced ways than humans do. For example, an argument against (more liberal) abortion (laws) written by the LLM was "I think the best argument for more restrictions on abortion is that it's murder. I think that's pretty clear." Crowdworkers do not view such claims as favorably as more subtle human-produced cases. These differences help explain the aggregate performance contrast between the LLMs and human writers. While the latter had a higher mean performance, they also exhibited lower variance in the appeal of their arguments. More specifically, there were two arguments written by humans that crowdworkers picked at least $\frac{2}{3}$ of the time in the control condition, and the worst human written argument was picked 38% of the time. Conversely, the most preferred LLM argument was picked 59% of the time and the least only 26% of the time.

In terms of the "partisan" nature of judge preferences, our results are mixed. Earlier research (e.g. Motoki et al, 2023) finds that LLMs have a consistent liberal/left political bias. Given this and given that crowdworkers might also be disproportionately liberal/left, we could imagine that this alignment of politics influences the judges' decisions. That is, we might be concerned that, even within topic, workers are responding to the political tone of the arguments rather than their rhetorical appeal. This does not seem to be the case. In

particular, crowdworkers seem to like arguments from both the (traditional) left and right of the spectrum. For example, the second most popular LLM argument in the control condition (in terms of how often it was picked) was an obviously conservative position:

> A lot of immigrants take jobs away from Americans and that we should focus on the Americans doing those jobs, before immigration. Also that they take up valuable resources such as healthcare and government services.

For a final and more general comparison we created document embeddings for all of (i.e. the superset of) the arguments from both groups after some filtering. Specifically, we dropped responses that made no sense given the prompt (e.g. actually made the opposite argument) and/or were unintelligible. About 15-25% of text produced by the LLM was both coherent and unique (not a direct repeat of a previous argument). We focus on the most complete sets of responses; in practice: the responses to "most important problem", more restrictions on abortion, more gun control, opt-in organ donation. In Figure 2, we display the results of reducing these document embeddings to two dimensions and plotting each argument in that space. In addition, we clustered all (embedded) points using k-means, where k=2. In each cluster, the majority class is either LLM or human. Where the particular point, i.e. argument, is actually from an LLM (human) and is in the LLM (human) majority class cluster, we say it is "correctly" classified. The points labelled "misclass. LLM" were human written arguments that were placed by the k-means algorithm with the (majority) LLM written ones based on these embeddings; vice versa for "misclass. Human". Where there are many misclassifications, we have evidence that humans and machines make very similar arguments; specifically, they are sufficiently similar that we cannot easily tell them apart in the embedding space clustering. Where there are few misclassifications, we have evidence that humans and machines make different types of arguments and can be separated via text alone.

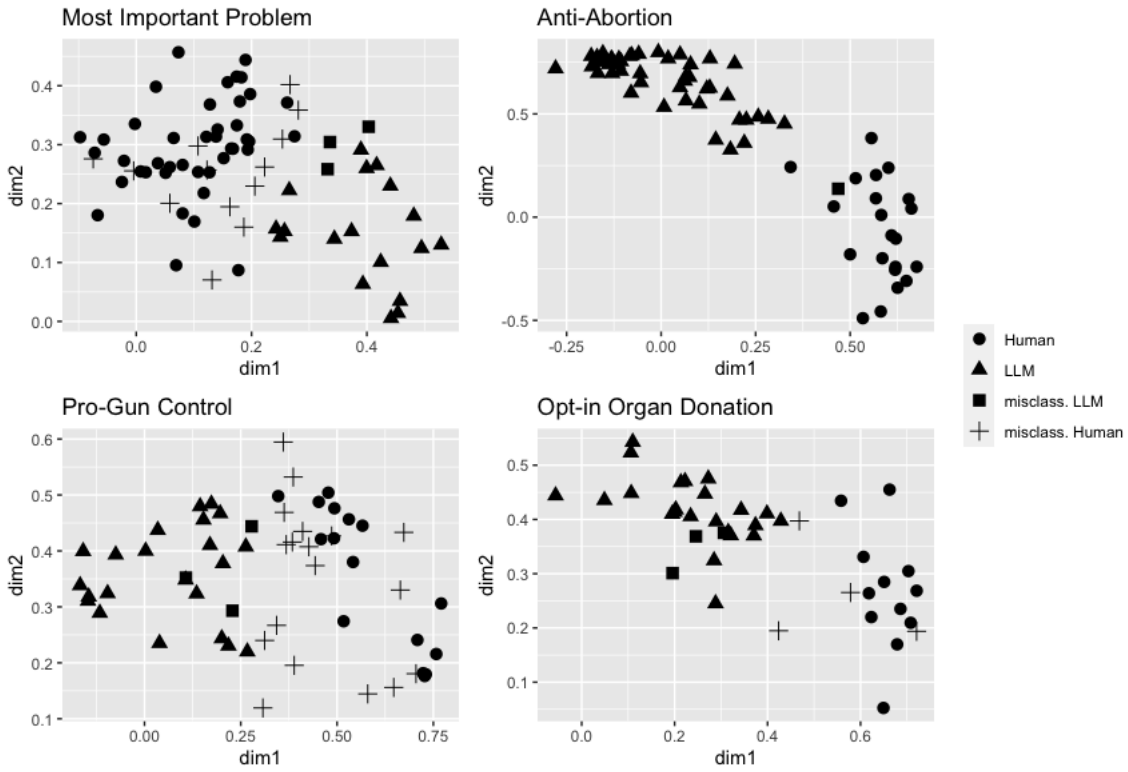

Figure 2: LLMs make more distinct (from humans) arguments on some topics than others. Specifically, the LLM anti-abortion and opt-in organ donation responses read differently in general to human prompt responses.

The clearest differences—i.e. the topics for which the LLM and human arguments are most different—are for anti-abortion prompts and on opt-in organ donation. These are much larger than the differences on "most important problem" and gun control. From qualitative inspection, another observation is that misclassification typically goes one way. That is, while it is relatively often the case that LLM arguments mimic exactly the ones our human crowdworkers make, the LLM is also prone to unusual phrasing (e.g. repetition) that humans are not. For instance, this argument generated by the LLM in favor of increasing immigration was rarely picked by judges:

> I think the best argument is that we need more people to keep our economy going. We need more people to work, pay taxes, and buy things. We need more people to pay for our social security and medicare. We need more people to pay for our schools and roads. We need more people to pay for

> our military. We need more people to pay for our police and fire departments. We need more people to pay for our parks and libraries. We need more people to pay for our courts and jails.

Importantly, the types of LLM arguments that are misclassified as human tend to be more popular with judges in the control condition. Put crudely, judges like the machine to "sound human".

Man v Machine: Humans prefer Human Orators

Finally, we ask whether knowing the identity (LLM or human) of the author of a particular argument had a causal effect on how convincing an audience found it. We did not have strong *a priori* beliefs: on the one hand, an LLM may be viewed as less biased or having access to a greater amount of information and therefore preferred. On the other, given the sensitive and nuanced nature of some prompts, a human perspective could be seen as more valuable and perhaps less "dangerous" or more trustworthy.

To address this, we assigned crowdworkers to either a control condition where they saw only the arguments (377 people) or a treatment condition where workers were told who (LLM or human) wrote each argument, with the order they were presented randomized (388 people).[5] Figure 3 shows the treatment effect of knowing the author on the relative probability workers preferred the human written argument, with fixed effects for each unique argument in all regressions. That is, we are estimating the author effect holding the actual text shown constant. Standard errors were clustered by prompt. The relevant data is included in the Supporting Information. The total treatment effect is positive and significant but small, resulting in an additional 5 percentage point probability that crowdworkers would

[5] The task was fielded through MTurk with a random treatment assignment. Through random chance, the control group was slightly larger than the treatment. We also had several more people in the treatment group vs. the control fail to complete the task adequately to be included in the analysis.

pick the human written argument. That is, overall, the causal effect of being told whether an argument was produced by an LLM or human is to prefer the human effort—but not by much.

The aggregate effect represents a consistent positive effect of knowing the author for all prompts, however it is only significant for three: the argument for reducing immigration, increasing gun control, and restricting abortion. For the first two of these, there was no preference for either author in the control condition. This implies the treatment effect is not, for instance, learning poor text is written by the LLM and this somehow increasing the dislike for the machine-produced content. Nor is it that judges are learning well-written arguments are written by an LLM and thus feeling ''betrayed". Further, though the abortion arguments written by the LLM were textually distinct per Figure 2, the arguments made in favor of gun control were often misclassified as human. Indeed, in the treatment condition, the arguments which were misclassified as human in the previous section are still picked less than half the

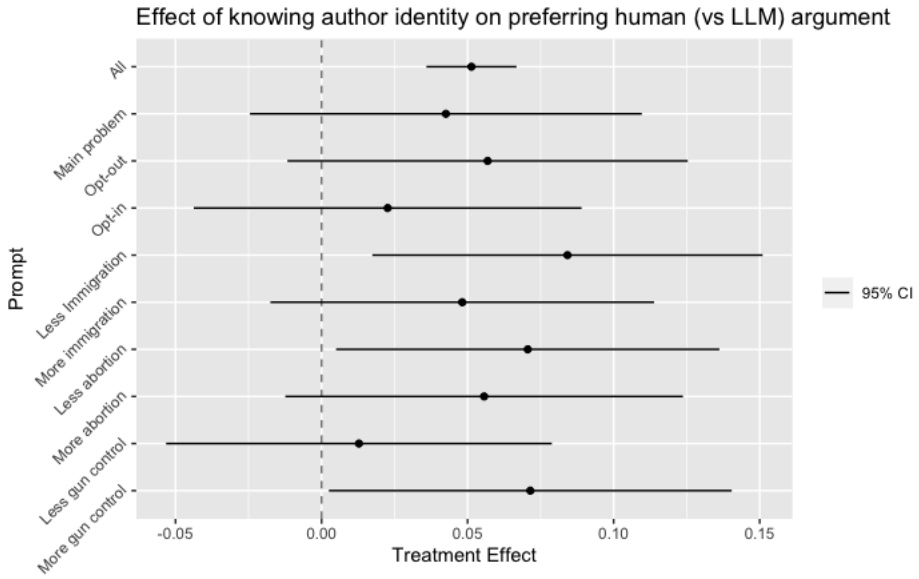

Figure 3: Causal effect of knowing author identity on judge preferences for a human-produced argument (relative to an LLM-produced one) is positive and statistically significant overall. There is considerable heterogeneity by topic, however.

time when the crowdworkers know it was written by an LLM. This suggests that argument quality, in the sense of being distinguishable, is not driving our results.

In SI G we give an alternative (but equivalent), regression-based assessment of this difference. To summarize that analysis: in the control group, judges are about 2 percentage points more likely to pick the human-written arguments (than LLM arguments) on average, though this is not significant. When informed about the author, they are about 10 percentage points less likely to pick the LLM written argument—a substantial difference over the control condition.

Discussion

For Aristotle, the purpose of rhetoric is to assist the orator in persuading their listeners (Rapp, 2009). This need not help with the communication of knowledge or finding of fact: a "good" argument by these standards is one that convinces a public, non-expert audience of the correctness of a position. This idea informed our experiments above, and we found that humans are not unique in terms of rhetorical abilities. On the matter of *logos*—i.e. the content of arguments—we showed that LLMs perform equivalently to humans in suggesting the phrasing for particular issue positions. This was true on both controversial and more banal matters, albeit for a curated "best of" set of arguments. On *ethos*—that is, the appeal arising from the nature of the speaker themselves—our findings suggest that machines have less appeal than humans as orators. This may give elected (and unelected) officials pause as they contemplate open use of such technology to help them in the democratic marketplace. And, given implied citizen preferences, they may seek to regulate the use of such models for argumentation more generally.

We did not explore the use of *pathos*—that is, the manipulation and exploitation of the emotions of the audience. Or rather, it was bundled with the content of the arguments. Future

studies might try to separate this out more than we have done, though we sound two cautionary notes. First, there are ethical concerns with (re)training and instructing LLMs to psychologically manipulate humans, not least because humans may not be able to detect machine-generated language (Jakesch, Hancock and Naaman, 2023). Second, and an issue that affects our work here too, is our "audience" was one of convenience—meaning lessons about *pathos* may be hard to generalize. While we know that our crowdworker judges are based in the United States, we have no reason to believe they are representative of, say, the American voting population (though see, e.g., Coppock, Leeper and Mullinix, 2018, for discussion of why this may present fewer problems than initially supposed). The same is true of our prompt writers. Presumably neither group meets the highest levels of human rhetorical creativity or analysis. So the next steps in such work might be to compare the LLM's abilities to those of true domain experts, like elected politicians.

The broader implications of our work apply to both politics and policy. As we noted, LLMs are not always popular with audiences *per se*. But in any case, while the LLM was able to suggest texts that human coders did not, we did not observe wholly new ideas to justify particular positions. This does not mean models will never be capable of such things: this is a fast-moving area, and there are already products available that outperform the model we used here (e.g. Touvron et al., 2023). Where the problem is to convince the public of the merits of some extant policy, the use of LLMs is more immediate: our experiments on opt-in/opt-out organ donation are in-line with this claim. But as we discussed in the Introduction, such abilities also bring dystopic visions of voters who do not know what or whom to trust. One potential result is that citizens become skeptically inured to all argumentation, and never update their personal beliefs about the merits of a particular position. Much worse, they may be convinced to act in a way that harms themselves and others, though previous work on fake news suggests that the causal effects of such messages are muted in the aggregate (e.g. Allen et al, 2020).

To reiterate, humans seem wary of machines as authors—even when they would otherwise like the content of the output. Future work might helpfully investigate how general this human distrust of machine composition is, and what its genesis might be. One natural extension of our work would be to use deception; that is, to lie to (some) respondents about whether a human or LLM produced the (same) statement, and to see if that author treatment affects the respondents' perceptions of the merits of the argument, holding the content identical. Indeed, one might be interested in whether respondents are more bothered by misrepresentation of human messages as being created by machines, or the false portrayal of machine output as human. Given our study, our belief is that they will find the latter considerably more concerning. In addition, future work might give broader, more open-ended prompts and use topic or contest models. These could be used to assess what factors (what content) characterize LLM versus human argumentation and makes those statements more or less successful when judged (see e.g. Loewen et al, 2012 for a related approach).

Perhaps as LLMs become more familiar, humans will relax regarding their efforts. We anticipate ethical challenges in the work ahead, for example over whom voters can hold responsible for machine-generated rhetorical appeals that lead to normatively undesirable outcomes. Put more simply, this new technology is political, and requires ongoing study of political philosophy.

References

Allen, J., Howland, B., Mobius, M., Rothschild, D. and Watts, D.J., 2020. Evaluating the fake news problem at the scale of the information ecosystem. Science advances, 6(14), p.eaay3539.

Bisbee, James, Joshua Clinton, Cassy Dorff, Brenton Kenkel and Jennifer Larson. 2023. "Artificially Precise Extremism: How Internet-trained LLMs Exaggerate Our Differences.".

Chesney, Bobby, and Danielle Citron. "Deep fakes: A looming challenge for privacy, democracy, and national security." Calif. L. Rev. 107 (2019): 1753.

Coppock, Alexander, Thomas J Leeper and Kevin J Mullinix. 2018. “Generalizability of heterogeneous treatment effect estimates across samples.” *Proceedings of the National Academy of Sciences* 115(49):12441–12446

Dai, Yaoyao and Benjamin J Radford. 2023. “Large Language Models for Measuring Contested and Multi-dimensional Concepts” Paper presented at Summer Political Methodology Meeting, 2023.

Grumbach, Jacob M. 2018. “From backwaters to major policymakers: Policy polarization in the states, 1970–2014.” *Perspectives on Politics* 16(2):416–435.

Halterman, Andrew. 2023 "Synthetically generated text for supervised text analysis." arXiv preprint arXiv:2303.16028

Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 168–177.

Jakesch, Maurice, Jeffrey T Hancock and Mor Naaman. 2023. “Human heuristics for AI-generated language are flawed.” *Proceedings of the National Academy of Sciences* 120(11):e2208839120.

Lazer, David MJ, et al. "The science of fake news." Science 359.6380 (2018): 1094-1096.

Loewen, Peter John, Daniel Rubenson, and Arthur Spirling. "Testing the power of arguments in referendums: A Bradley–Terry approach." Electoral Studies 31.1 (2012): 212-221.

Motoki, Fabio, Valdemar Pinho Neto, and Victor Rodrigues. "More human than human: Measuring chatgpt political bias." *Available at SSRN 4372349* (2023).

Martindale, Colin. 1990. *The clockwork muse: The predictability of artistic change.* Basic Books.

Rapp, Christof. 2009. The Nature and Goals of Rhetoric. In *A Companion to Aristotle*, ed. Georgios Anagnostopoulos. Wiley Online Library pp. 577–596.

Rosenzweig, Leah R and Molly Offer-Westort. 2022. “Testing interventions to address vaccine hesitancy on Facebook in East and West Africa.” *Open Science Framework* .

Schiff, Kaylyn Jackson, Daniel S. Schiff, and Natalia Bueno. "The Liar’s Dividend: Can Politicians Use Deepfakes and Fake News to Evade Accountability?." (2022).

Spirling, Arthur. 2023. “Why open-source generative AI models are an ethical way forward for science.” *Nature* 616(7957):413–413.

Stone, Philip J, Dexter C Dunphy and Marshall S Smith. 1966. *The General Inquirer: A computer approach to content analysis.* MIT press.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timoth´ee Lacroix, Baptiste Rozi`ere, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave and Guillaume Lample. 2023. “LLaMA: Open and Efficient Foundation Language Models.”.

Zhang, Susan, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang and Luke Zettlemoyer. 2022. “OPT: Open Pre-trained Transformer Language Models.”