Large Language Models Can Argue in Convincing and Novel Ways About Politics: Evidence from Experiments and Human Judgement*

Alexis Palmer† Arthur Spirling‡

*First draft: February 27, 2023. This draft: March 6, 2023.
†Department of Politics, New York University. ap6100@nyu.edu
‡Department of Politics and Center for Data Science, New York University. arthur.spirling@nyu.edu

Significance

What makes a political argument convincing? Scholars have studied this question since ancient times, and concluded that the answers lies partly in the properties of the message, but also in the nature of the orator. Traditionally, we assumed that the speaker must be human, but the recent advent of large language models (LLM) that can generate human-like text changes this. We conduct experiments to examine whether these models can compete with humans in making arguments, how the arguments they produce differ from humans, and whether judges find them compelling. We show that LLMs can outperform human arguers, but that human judges mildly prefer human-produced arguments, especially when they are told a human was the author.

Abstract

All politics relies on rhetorical appeals. Part creative art, part scientific analysis of what works, the ability to construct persuasive appeals is considered perhaps uniquely human. In recent times however, we have seen successful LLM applications to many such areas of human endeavor. Here, we explore whether these autoregressive transformer approaches can out-compete humans in making political and policy appeals. Our areas of interest include controversial partisan issues in the US, such as abortion and gun rights, but also more banal and open-ended matters. We use a relatively large number of crowdsourced US workers to produce "best" arguments, and then an open-source LLM to compete with them. Human (crowd) judges make decisions about the relative strength of their (human v machine) efforts. Our results are threefold. First, LLMs can produce arguments on a par with humans, at least in terms of convinces independent judges. That is, LLMs can be persuasive. Second, we show that LLMs produce novel arguments insofar as their output has different quantitative and qualitative characteristics to that produced by humans. LLM arguments are typically easier to read, and written with slightly more positive affectation. But LLM arguments can lack nuance—at least if the goal is to convince others of their merits. Finally, we show that judges mildly prefer human arguments on a given topic. This is true when uninformed about the orator's identity—i.e. human or machine—and becomes more pronounced when they are informed in a randomized controlled experiment.

Introduction

What persuades an audience to accept a particular argument may be the oldest and most studied political science question of all [e.g. 1, 5, 14, 11, 17]. And despite literally thousands of years of intervening research, 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 the recent pandemic has seen use of these ideas in convincing citizens to do other things—wear masks, get vaccinations, and return (or not) to their workplaces. Without hyperbole, the outcomes of such appeals are potentially matters of life and death.

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. In particular, we now have access to generative "large language models" (LLMs) that allow computers to produce human-like text in response to user prompts. These machine learning autoregressive approaches have been shown to exhibit competence of varying degrees in many tasks. Though such capabilities are exciting in their own right, their arrival raises several classic questions about what is or is not uniquely human. Most famously this is the central question of Turing's work on "Computing Machinery and Intelligence" but such themes are much older. For Turing, and his eponymous "test", the specific interest is in whether machines can sufficiently imitate humans so as to fool them that those same computers are human. But for social scientists interested in persuasion, a more fundamental question is whether those machines can out-perform Aristotle's "political animal" (i.e. mankind) in their rhetorical interactions with other humans. This matters 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 policy.

Here we investigate whether LLMs can do the core business of democratic politics: convincing humans of the merits of a particular issue position. We ask not merely whether they 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 long-standing questions of political philosophy to those of political science, via the methods of computer science. Specifically, we use an open source LLM—the Meta `OPT-175B` model [21]—and prompt it to make arguments for and against common positions in contemporary US politics and society. For example, to be 'pro' and 'anti' expansion of gun rights. We also prompt the LLM in more open-ended ways. These same argument position prompts are also given to *humans*, specifically large numbers of crowd workers. After some cleaning and quality matching, these pairs of responses are then shown to a set of independent human judges who must decide whether the machine or human argument for a position is the most plausible. 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.

Using this simple research design, we believe we are the first to assess rhetoric this way. 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 routinely 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, even for very similar lengths of texts. Specifically, LLM arguments tend to be written at a lower level 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, specifically abortion.

Results

Our first goal is to assess whether and to what extent LLMs can make arguments that persuade—and how well they can do this relative to humans. We begin by demarcating the five issue position 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., 8, 12], 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 <u>for</u> [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 <u>for easier access to</u> [more restrictions on] abortion?
3. There are many diverse opinions on immigration to the US. From your perspective, what is the best argument for <u>increasing</u> [restricting] immigration to the US?

A respondent—either a human or the LLM—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 researching polling such questions.[1] 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 <u>'opt in'</u>[opt out] system?

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

The fifth prompt is open-ended 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 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 comes in only one mode.

As suggested above, 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[2] and 25 crowdworker arguments for each of the nine prompts. We winnowed these responses by retaining those that met the following criteria: first, the arguments must be 'coherent' or meaningful English; second, they are "on topic"; third, for the crowdworker responses, those that were not copy-and-pasted from existing text online. We then independently rated the remaining responses by their quality and selected the three 'best' arguments from each group based on the joint score per prompt. Ultimately, this results in similar length arguments of comparable quality.

In the second stage of the analysis, an independent set of 400 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) 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 for prompts randomly chosen from our set of nine. Therefore arguments for each prompt were judged approximately 435 times, though by chance 3 (more gun control, more immigration, and less immigration) were judged around 395 times.

[2]For 4 prompts a large run of 300 arguments was generated.

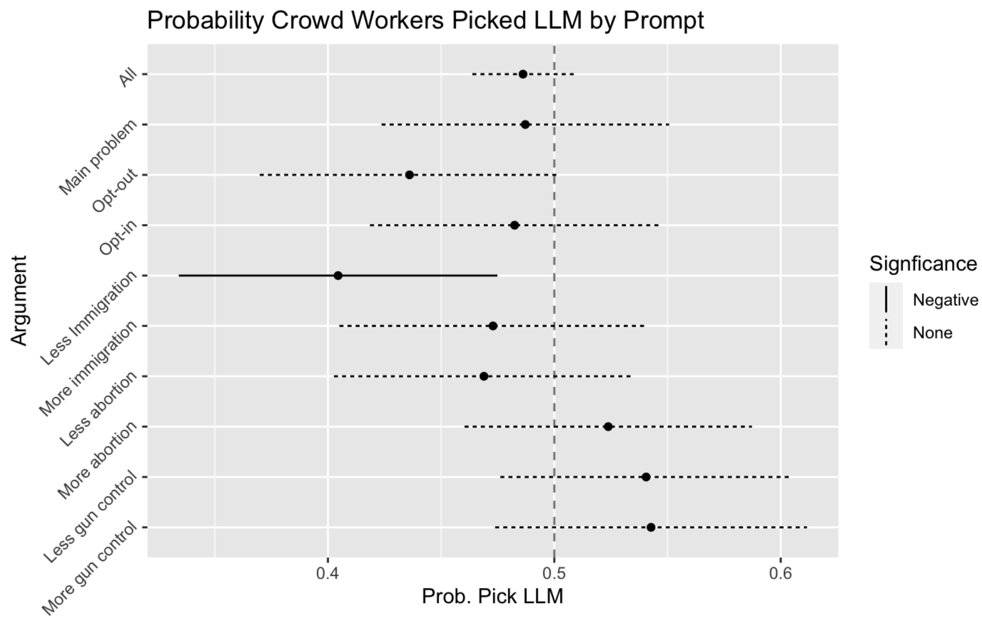

Figure 1: LLMs and humans are generally equally able to convince independent judges as to the merits of arguments for positions.

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 LLM-generated 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 only one case—an argument for less immigration—did the LLM under-perform humans, on average. 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 quantitative "no rhetorical edge" result is interesting, but it does not mean there are no qualitative differences between human and LLM arguments. First, we note that the LLM does better on topics which are frequently discussed online and in the general discourse—e.g.

abortion and gun control. It does worse when asked to defend organ donation policies, for example, often producing only sentence fragments or arguing for unrelated positions. This is unsurprising given where the LLM is trained (i.e. on public documents), but suggests finding automated arguments for more obscure topics is harder.

Second, and related: humans and the LLMs differ on the relative nuance with which topics are discussed. In particular, the LLM at times provides very simplistic arguments on divisive topics such as restricting abortion or immigration, which crowdworkers do not view as favorably as more subtle human-produced cases. 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." Again, this is in keeping with how such models are trained, but suggests LLMs may struggle to anticipate the reaction of human audiences.

These differences help explain the aggregate performance differences 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. By contrast, the most preferred LLM argument was picked 59% of the time and the least only 26% of the time. In general, we note that judges liked arguments that were logically ordered, and appealed to human welfare. For example, the most preferred arguments from each source in the control condition were (from a human and the LLM, respectively):

- **Human:** "There are a ridiculous number of people waiting for organs on the transplant lists that have to wait sometimes years to get said organ, even though people die every day. This is because the dying do not donate their organs enough, so making it default is better for those waiting to continue living."
- **LLM:** "I think the best argument for more gun control is that it is a proven fact that more guns in the hands of more people leads to more gun violence. The United States has more guns per capita than any other country in the world, and we have the highest rate of gun violence in the world."

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.

We start with simple descriptive statistics. First, on “reading ease” in the sense of Flesch [see, e.g., 4, for discussion], LLM arguments are typically higher mean ($p < 0.05$) and lower variance. That is, LLM arguments are easier to read, and tend to be more similar to each other on this metric. Second, while the parts of speech used were very similar across groups, the overall sentiment varied. The LLM was consistently more positive in speech (multiple dictionaries, $p < 0.05$ in one case). However, the human written text covered a wider variety of sentiment. That is, the LLM produced little difference in tone as compared to humans. There was (at most) weak correlation between positive sentiment and judge preference for a given argument.

For a deeper and more general comparison we created document embeddings for all of the coherent and on-topic arguments from both groups (in practice: “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.

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 similarity 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

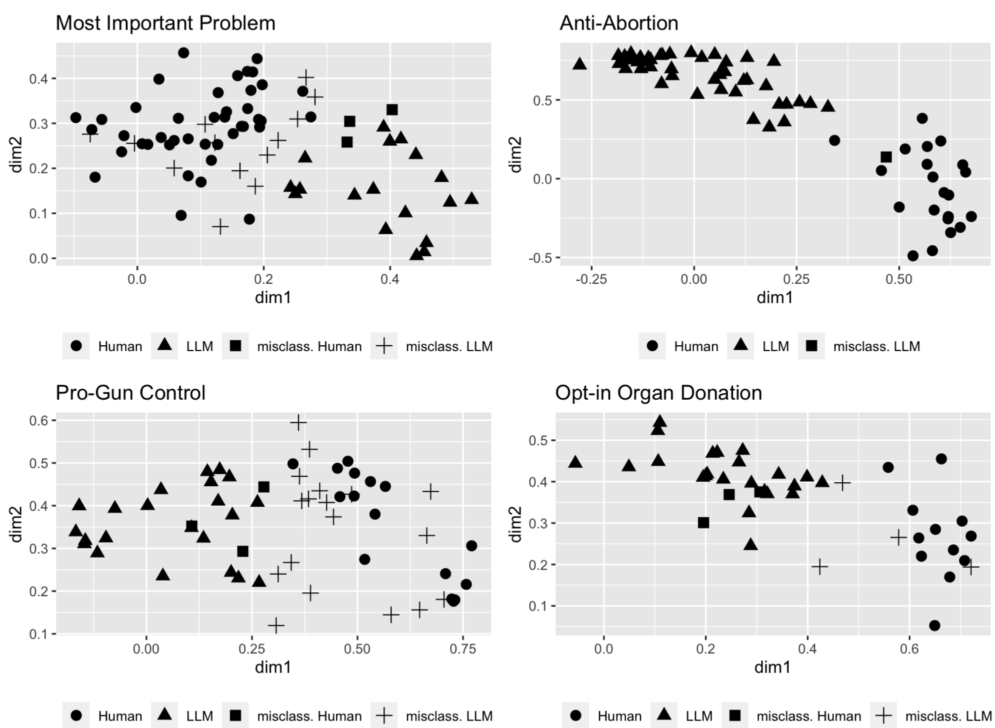

Figure 2: LLMs make more distinct (from humans) arguments on some topics than others. Specifically, the LLM anti-abortion and opt-in organ donation responses read differently in general to human prompt responses.

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.

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”.

To address this, we assigned crowdworkers to either a control condition where they saw only the arguments (effectively 198 people) or a treatment condition where workers were told who (LLM or human) wrote each argument, with the order they were presented randomized (180 people). Figure 3 shows the treatment effect of knowing the author on the likelihood workers preferred the human written argument, with controls for each argument in all regressions and prompt fixed effects for the overall effect. The total treatment effect is positive and significant but small, resulting in an additional 6 percentage point probability that crowdworkers would 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.

Additionally, though the effect is positive for the majority of prompts, it is only significant for three: the “main problem”, the best argument for restricting abortion, and the best argument for less gun control. This last prompt was one of the prompts for which the

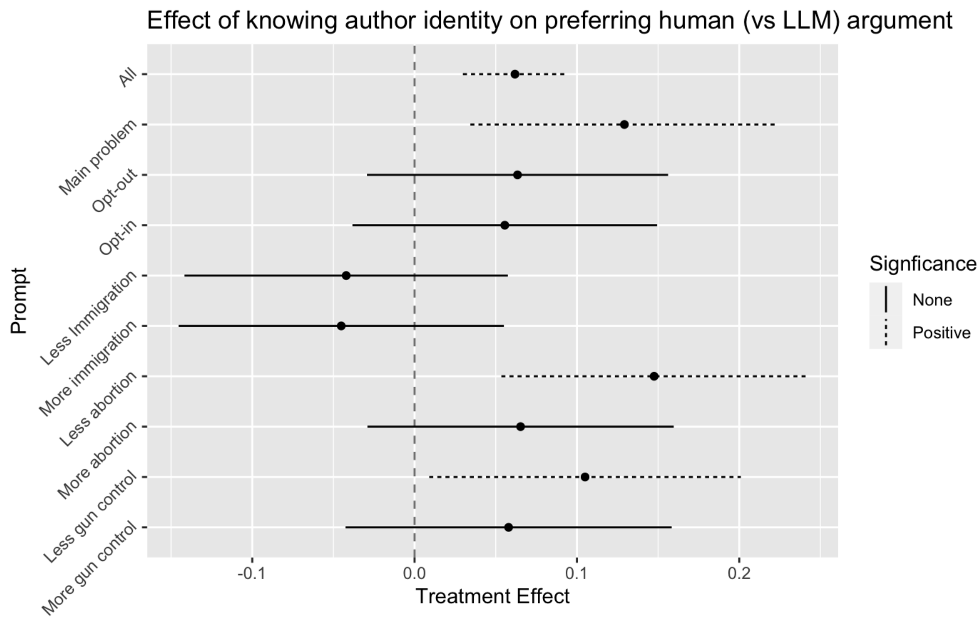

Figure 3: Causal effect of knowing author identity on judge preferences for a human-produced argument (relative to an LLM-produced one) is positive and statistically significant over all. There is considerable heterogeneity by topic, however.

LLM was *most* preferred in the control condition. The two prompts relating to immigration were the only ones to have negative (though not significant) effects as a function of author identity knowledge; they were also two of the prompts for which the LLM arguments were *least* preferred in the control condition, so likely harder to move.

As an alternative way to view the aggregated preferences, consider Table 1. There we give the relevant coefficient estimates for a regression of preferring a human argument ($Y = 1$) on the treatment, which is knowing the identity of the author and the interaction of that with that author being the LLM. The point here is that the interaction is statistically significant, and negative: that is, overall, when judges are told that a given argument is produced by an LLM they are more likely to prefer the human-produced argument.

Table 1: Regression results showing that an LLM wrote the argument causes judges to prefer human offerings over machine ones.

	Dependent variable:
	LLM argument preferred human argument
Knows Author	0.063***
	(0.016)
Argument author is LLM	−0.028*
	(0.016)
Knows Author × Author is LLM	−0.127***
	(0.023)
Prompt FE	Yes
Observations	7,416
R^2	0.012
Adjusted R^2	0.010
Note:	*p<0.1; **p<0.05; ***p<0.01

Based on Table 1, in the control group, judges are about 3% more likely to pick the human-written arguments (than LLM arguments) on average. When informed about the author, they are about 9% more likely to pick the human written argument—a relatively small absolute number but a three fold increase in the preference verses the anonymous condition.

Discussion

For Aristotle, the purpose of rhetoric is to assist the orator in persuading their listeners [16]. 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. On *ethos*—that is, the appeal arising from the nature of the speaker themselves—our findings suggest that machines have generally less appeal than humans as orators. But the differences are not large overall, and on many issues the results are equivocal: that is, judges show no particular preference either way.

We did not explore the use of *pathos*—that is, the manipulation and exploitation of the emotions of the audience. Or rather, to the extent that was a treatment, it was bundled with the content of the arguments. Future studies might try to separate this out more than we have done, though we would sound two cautionary notes. First, there are broad ethical concerns with (re)training and instructing LLMs to psychologically manipulate humans. 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., 6, for discussion of why this may present fewer problems than initially supposed]. The same is true of our original prompt writers. Perhaps neither group is particularly reflective of innate human skill in creating or responding to rhetorical appeals, though our qualitative inspection of their efforts suggested they were astute and sufficiently capable.

The broader implications of our work apply to both politics and policy. On the former, one could imagine politicians using LLMs to help them cultivate and curate argument strategies. That said, while the LLM in this case was able to suggest texts that human coders did not, we did not observe wholly new ideas to justify particular positions. But this does not mean models will never be capable of mimicking "political entrepreneurs" [e.g. 7]. Indeed, we note that this is a fast-moving area, and there are already products available that outperform the model we used here [e.g. 20]. Where the problem is to convince the public of the merits of some extant policy, the use of LLMs is more obviously immediate: our experiments on the opt-in/opt-out possibilities of organ donation are in-line with this claim. In either case though, we might be especially anxious about relying on proprietary products for these citizen-facing tasks and the potential lack of transparency that incurs [19, 10]. This was part of our motivation for using an open-source model in our work above.

Conclusion

We showed that in a narrow but precise sense, LLMs can 'do' political rhetoric, often as well as humans can. In one way, this is not surprising: LLMs perform well at many related written tasks, such as composing letters, scripts or essays [see also 2, 3, 13]. But what makes politics different, in democracies at least, is the need to have popular support for the *person* making the argument. Here at least, humans remain ahead for now—albeit by a slim absolute margin in our study. Future work might helpfully investigate how general this human wariness of machine composition is, and what its genesis might be. We could imagine that as LLMs become more familiar, humans relax regarding their efforts. By contrast, descriptive representation [in the sense of 15] presumably precludes machines ever becoming political agents of citizen principals. In any case, we anticipate ethical challenges in the work ahead, for example over whom voters can hold responsible for rhetorical appeals that lead to normatively undesirable outcomes. Put more simply, this new technology is political, and requires ongoing study of political philosophy.

Materials and Methods

To provide the LLM written arguments, we used Open Pre-trained Transformer Language model from [21]. We generated 15-30 arguments in response to each of our prompts. For four of the nine ("most important problem", more restrictions on abortion, more gun control, opt-in organ donation) we also ran a large batch of 300 responses to asses how often the

LLM produced usable/unique arguments. In the SI we report more details on that analysis.

We did minimal editing to the responses, e.g. deleting repeating clauses and fixing punctuation. We filtered for those we judged to be coherent and 'on topic', and selected the the qualitatively 'best' three of the arguments for each prompt.

To solicit responses from crowdworkers, we developed an app that would provide five prompts for workers to answer with a limit of 300 characters. Each respondent answered only the 'pro' or the 'con' side of the first four prompts described above, and all saw the final, open ended question. They were asked to provide the best argument for the prompts regardless of personal opinion. We used Amazon MTurk to find respondents and we required that they must be in the U.S.; no other information was collected about them. We had a total of 50 participants for a total of 25 responses to each side of the first four topics and 50 responses to the last, open-ended question.

From this set of responses, we again filtered by coherency and 'on topic'-ness. We also filtered out responses that exist in that form online. We (the authors) independently rated the remaining responses by quality and used the arguments with the best joint scores.

To compare which responses are preferred, we created a light-weight app and recruited participants through MTurk as judges. We collected 400 responses, which when filtered for adequate task completion etc. Each judge was randomly assigned 10 pairs of arguments, with the prompt (from nine possible), LLM and human arguments (from three each possible per prompt), and order of the same randomized. They were asked, given the question, regardless of personal opinion, which is the "best" argument. Judges were randomly assigned to either a control condition, where the authors of all arguments were anonymous (198 respondents) or a treatment condition where they were informed whether a human or an LLM wrote the argument (180 respondents).

We used dictionaries from [9] and [18] to estimate sentiment and the R package `udpipe` for parts of speech. We used the R package `doc2vec` to create embeddings and topics from these embeddings on the prompts for which were had generated a large number of responses. These were plotted along two dimensions of topics and clustered using k-means to determine the similarity of the language used.

Treatment and control were compared using a linear regression (1) with "choosing the human argument of a human/LLM pair" as the binary dependent variable ($Y \in \{0, 1\}$) and (2) with each argument instance now an observation, with the dependent variable being whether it was chosen and with the author and treatment condition as the independent variables (plus their interaction)

References

[1] Kevin Arceneaux. Cognitive biases and the strength of political arguments. *American Journal of Political Science*, 56(2):271–285, 2012.

[2] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, pages 1–15, 2022.

[3] Michiel A Bakker, Martin J Chadwick, Hannah R Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew M Botvinick, et al. Fine-tuning language models to find agreement among humans with diverse preferences. *arXiv preprint arXiv:2211.15006*, 2022.

[4] Kenneth Benoit, Kevin Munger, and Arthur Spirling. Measuring and explaining political sophistication through textual complexity. *American Journal of Political Science*, 63(2):491–508, 2019.

[5] Jack Blumenau and Benjamin E Lauderdale. The variable persuasiveness of political rhetoric. *American Journal of Political Science*, 2022.

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

[7] Norman Frohlich and Joe A Oppenheimer. *Political leadership and collective goods.* Princeton University Press, 1971.

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

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

[10] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019.

[11] Joshua L Kalla and David E Broockman. Reducing exclusionary attitudes through interpersonal conversation: Evidence from three field experiments. *American Political Science Review*, 114(2):410–425, 2020.

[12] Emily Kubin, Curtis Puryear, Chelsea Schein, and Kurt Gray. Personal experiences bridge moral and political divides better than facts. *Proceedings of the National Academy of Sciences*, 118(6):e2008389118, 2021.

[13] Jonathan Mellon, Jack Bailey, Ralph Scott, James Breckwoldt, and Marta Miori. Does gpt-3 know what the most important issue is? using large language models to code open-text social survey responses at scale. *Using Large Language Models to Code Open-Text Social Survey Responses At Scale (December 22, 2022)*, 2022.

[14] Diana Carole Mutz, Paul M Sniderman, and Richard A Brody. *Political persuasion and attitude change*. University of Michigan Press, 1996.

[15] Hanna F Pitkin. *The Concept of Representation*. Univ of California Press, 1967.

[16] Christof Rapp. The nature and goals of rhetoric. *A Companion to Aristotle*, pages 577–596, 2009.

[17] William H Riker, William H Riker, and William H Riker. *The art of political manipulation*, volume 587. Yale University Press, 1986.

[18] Philip J Stone, Dexter C Dunphy, and Marshall S Smith. The general inquirer: A computer approach to content analysis. 1966.

[19] Mariarosaria Taddeo and Luciano Floridi. How ai can be a force for good. *Science*, 361(6404):751–752, 2018.

[20] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[21] Susan Zhang, 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. Opt: Open pre-trained transformer language models, 2022.

Supporting Information

Proportion of Useable LLM Arguments

As can be seen in Table when we generated large batches of arguments, there were a maximum of $\frac{1}{6}$ of the sample size that consisted of unique, usable arguments.

	Argument	Total Runs	Usable Answers	Unique Usable Ans.
1	Opt-in Donation	300	29	27
2	More Gun Control	300	42	40
3	More Abortion Restrictions	300	76	49
4	Most Important Problem	300	231	31