Article

new media & society
1–20

DOI: 10.1177/14614448231160526
journals.sagepub.com/home/nms
SAGE

The potential impact of emerging technologies on democratic representation: Evidence from a field experiment

Sarah Kreps and Douglas L. Kriner
Cornell University, USA

Abstract
Advances in machine learning have led to the creation natural language models that can mimic human writing style and substance. Here we investigate the challenge that machine-generated content, such as that produced by the model GPT-3, presents to democratic representation by assessing the extent to which machine-generated content can pass as constituent sentiment. We conduct a field experiment in which we send both handwritten and machine-generated letters (a total of 32,398 emails) to 7132 state legislators. We compare legislative response rates for the human versus machine-generated constituency letters to gauge whether language models can approximate inauthentic constituency voices at scale. Legislators were only slightly less likely to respond to artificial intelligence (AI)-generated content than to human-written emails; the 2% difference in response rate was statistically significant but substantively small. Qualitative evidence sheds light on the potential perils that this technology presents for democratic representation, but also suggests potential techniques that legislators might employ to guard against misuses of language models.

Keywords
Astroturfing, communication, language models, politics, representation

In spring 2017, the Federal Communications Commission (FCC) invited public commentary regarding its proposed rollback of rules regulating how broadband providers treat content. The sheer scale of comment on its website was the first flag of misconduct.

Corresponding author:
Sarah Kreps, Cornell University, Ithaca, NY 14850, USA.
Email: sarah.kreps@cornell.edu

A later examination found that millions of comments were generated by natural language models that used past public comments to predict plausible-sounding content on the new regulations. Further indicators of misuse included repetitive and convoluted language structures.[1] Nonetheless, the experience proved an important concept: that technology could threaten a pillar of democracy, representation through the process of deliberation and bottom-up citizen influence in the policy process. A central feature of democratic policymaking is citizen input. Inviting citizens to participate in deliberations gives them policy buy-in, making the policies more legitimate and sustainable. Mobilizing public participation is legitimate, but impersonating citizens at scale to artificially create a swarm of support for or opposition to a policy challenges democratic legitimacy by manufacturing policy positions that do not correspond to those of real constituents. Ultimately, if these effects are sizable or credible enough, they may produce policy change either legislatively or via executive-branch regulation.

In this research, we investigate whether natural language models—which use machine learning to predict the next words based on previous words and context—can generate letters that legislators perceive to be authentic constituency correspondences. We do so with a field experiment on more than 7000 state legislators to compare legislative responses to both human and machine-written correspondences. In the experiment, we randomized whether the letters were written by humans or by the language model GPT-3 (trained on letters written by humans); the substantive topics of those letters (using one of six different policy issues); as well as whether the ideological slant of the letter was liberal or conservative. We then assessed legislative response rates to human versus artificial intelligence (AI)-generated emails, and whether this varied across issue areas; the length of legislator replies to human versus AI-generated emails; and whether legislators were more or less responsive to ideologically consonant or dissonant AI-generated communications. These analyses afford insight into the extent to which malicious actors can influence the process of democratic representation through machine-manufactured advocacy letters that gain the attention and potentially the action of legislators.

In the aggregate, we found that legislators were modestly less likely to respond to machine-generated content than to human-written emails; however, although the difference in response rate was statistically significant, it was substantively small—less than 2%. On some topics, there were no differences in response rates to the human and machine-generated emails; and for one topic machine-generated correspondences elicited a higher response rate than human emails, although this difference did not reach statistical significance. Furthermore, a sizable number of AI-written correspondences elicited lengthy and personal responses suggesting that legislators believed that they were responding to constituents. While not a perfect predictor of legislative action, responsiveness is a valuable proxy for legislative priorities given the demands on legislators' time (Butler and Broockman, 2011; Costa, 2017; Einstein and Glick, 2017), and the decision to answer a constituent letter implies a calculus about the importance of responsiveness to that individual or issue (Bol et al., 2020).

Nonetheless, the potential malicious use of machine-generated constituency content manifested limitations. Although the content was coherent on average, it sometimes created inconsistencies that a careful reader could identify. A common error, for example, consisted of right-wing gun control machine-written letters that both emphasized the

need to protect Second Amendment rights while also recommending more background checks. Furthermore, as in the FCC case where machine-generated content was flagged because of the use of identical one-sentence supportive comments with similar email constructs, the mass use of AI-written advocacy letters also requires aliases that can, through repeat exposure to legislators from small districts who know many of their constituents, sound contrived or inauthentic.

Overall, however, the nearly equivalent responsiveness and attentiveness to both human and AI correspondences point to the threat emerging technologies pose to democratic representation. Machine-generated content can generate large volumes of emails that are neither form letters nor boiler plate, thus avoiding the type of detection that flagged inauthenticity in the FCC comments case and potentially operating in a manner that can create an erroneous sense of mass sentiment.

The rest of the article is organized as follows. First, we discuss how technologies that can mass produce content that passes for constituency preferences have the potential to distort fundamental pillars of democratic representation. Second, we describe our research design employing a field experiment in which state legislators were randomly sent both machine-generated and human-written emails across six issues and from the right and left of the political spectrum. Third, we present the findings which show that legislators were somewhat less responsive to AI-generated emails than to human-written emails in three policy areas, but equally responsive across three others. We then discuss qualitative evidence about when and why legislators were able to detect differences. We close by discussing the implications of this rapidly evolving technology for democratic representation.

Threats to democratic representation and deliberation

Offering his insights about government through participation and contestation, Robert Dahl observed that "a key characteristic of a democracy is the continuing responsiveness of the government to the preferences of its citizens" (Dahl, 1971: 1). Representation is nested in a broader process of governance, in which societal demands lead to political interaction, which produces government, leading to policy choice, and finally implementation of the policy. In this model, individuals, groups, and political parties create the social environment and inputs for political leaders, and political leaders have incentives to heed the preferences they hear expressed by constituents. In a democratic system, government decisions about policy should respond to changes in public sentiment. Government leaders seek to discern public opinion, and the public follows government policies and responds. Beyond expressing preferences explicitly through public opinion surveys, the public also participates in the political process of expressing preferences via letters, rallies, and elections, with political elites dynamically representing changes in those constituent preferences (Stimson et al., 1995). In particular, scholars suggest that members are more responsive to the preferences of voters than non-voters (Griffin and Newman, 2005) and those who feel strongly enough about the issue to write an advocacy letter than those who do not (Congressional Management Foundation [CMF], 2011).

Written correspondence is a vehicle through which constituents can communicate their opinion in ways that inform legislative priorities (Butler and Nickerson, 2011). In a

study of federal spending, Andrea Campbell (2003) found that surges in constituent mail from seniors caused Congress to back away from Social Security cuts in the 1980s. The underlying mechanism is that high voting constituencies generate more mail, which prompts responsiveness and in turn fosters favorable turnout. Indeed, empirical study lends support to the notion that high participation districts are also high turnout districts. One study of mail across districts showed that at the upper end of district turnout, legislative offices received and responded to twice the amount of mail as districts at the lower end of turnout (Martin and Claibourn, 2013).

Dahl's account of democratic representation has not aged seamlessly. Scholars have observed increasing public skepticism toward representative government and rising opposition to established political parties on the basis that they are "unrepresentative and unresponsive" (Disch, 2019: 2). As Castiglione and Pollak (2018: 1) characterize the challenge of democratic representation, "its institutional machinery is often regarded as inadequate to deal with the intensified speed and complexity of decision-making in the politics of the global age." Individuals, according to this view, are not withdrawing from political life but expressing their political preferences differently, for example, through Occupy movements or non-establishment political parties. One reason, according to Tormey (2015: 125), is that members of society are less trustful of politicians to govern on their behalf than they have ever been in the past.

As a result of this erosion of trust, critics argue that disaffected citizens have turned to other modes of political representation. This theoretical turn, based on observations about the decline in voting rates, for example, calls for "stretching" the idea of representation (Saward, 2008) away from the more rigid and potentially problematic conceptualization of it solely as how representatives, once elected, act as agents of the people and respond to their preferences.

Despite the potential intuitive appeal of this conceptualization, recent data from the United States suggest that many Americans continue to seek to influence their representatives and public policy through direct outreach to elected officials. Public correspondence with elected officials has soured in recent decades. As the Congressional Management Foundation (CMF; 2005: 25) suggests, online advocacy campaigns have rendered emailing an elected official easy compared with the more costly, in time and money, endeavor of sending a letter. Members of Congress have had to devote more resources and staff to email management in recent years as volume has increased; however, the CMF nonprofit found that 50% of emails are not even opened. Indeed, perhaps the two accounts above are compatible with each other. As the ease of correspondence has increased, incoming volume has increased, making it difficult for legislative offices to respond to constituent concerns, reducing trust in representativeness of democratic institutions. Indeed, CMF's advice to legislators, suggesting that members send autoresponses indicating they are listening and tallying constituent concerns, could only exacerbate the downward spiral. The advice concludes, "Public polling shows that about only 1 in 10 Americans think Congress cares what their constituents think. You and your office can help bust that myth and emphatically declare: we're listening."[2]

Potentially exacerbating this dynamic is the emergence of new technologies. Advances in AI have created text prediction technologies that have the potential to distort democratic representation at scale. The 2017 example from the FCC offers an illustrative case

of how text prediction tools could inundate comment lines for proposed net neutrality reversal. In that case, repeated phrases and large-scale, concurrent comment submission offered straightforward clues about differences between real and inauthentic posts. Earlier machine learning algorithms often generated outputs that had grammatical, typographical, or factual errors (Jakesch et al., 2019; Kreps et al., 2022). Yet more powerful models have emerged that may minimize the features that would previously have been markers of machine-generated texts. One particular large language model, GPT-3, increased in parameter size—a calculation in a neural network that weights different aspects of the data, tuning to create more efficient learning and optimize model results—10-fold in a 1-year period. Trained on a corpus of almost 1 trillion words scraped from the Internet, GPT-3 models the style and substance of its inputs. Given snippets of a poem, product review, or news article, it can generate new, analogous, convincing content (Brown et al., 2020).

One of the notable improvements of recent models such as GPT-3 over predecessors is that they can engage in in-context learning compared with previous models that are pretrained and cannot adapt to requested tasks. Earlier models, for example, could simply respond to a natural language prompt, showing an acuity at taking snippets from poems, product reviews, or news articles and mimicking the style and substance in the outputs. The newest models engage in "in-context learning" (Brown et al., 2020), taking the text from the pretrained language model, learning tasks based on a few demonstrations, and then recognizing and completing the new task. These newer models have performed well on reading comprehension tasks, writing creative fiction or nonfiction stories, and general reasoning, but face challenges in natural language inference tasks in which the model must implicitly craft hypotheses about the relationships between two parts of a sentence. Contextual reasoning, in other words, is not a strength of these language models (Liu et al., 2021). Thus, being capable of generating creative fiction may not translate into an ability to generate ideologically consistent advocacy letters across a range of policy issues. A machine learning algorithm may therefore face challenges, even with training, in replicating the subtleties of ideological reasoning.

Nonetheless, as Buchanan et al. (2021: 6) point out, these language models provide opportunities for malicious actors to amplify disinformation and generate new content at scale, what the authors refer to as narrative reiteration. This could allow a malicious actor or a set of actors to violate the one person, one vote, or voice principle of democratic representation (Balinski and Young, 2010; Hayden, 2003). Individuals have a right to be heard, but they do not have a right to artificially, illegally, or illegitimately inflate that vote or voice. In the United States, the 1962 *Baker v. Carr* decision codified the principle of one person one vote by mandating that state legislators must be apportioned on the basis of population, a decision that had major consequences for the redistribution of public expenditures (Ansolabehere et al., 2002).

An analogous argument can be made in the context of voicing rather than voting preferences. In forms of political participation such as protest, social media expression, or advertising, the practice of misrepresenting one's identity and disingenuously magnifying the apparent number of supporters has a name: "astroturfing." This phenomenon is far from new. An earlier version of astroturfing consisted of non-constituents or outside groups masquerading as constituents sending fake mail to members of Congress with the

aim of tilting the scales of support in favor of their preferred policy. For example, in the 1950s a US senator received 100 letters urging that he support a higher minimum wage; when he followed up with the senders he discovered that only 33 were registered voters in the district. In the era of letters, an obvious way to diagnose fraud was that out of any batch of letters, about 5% would enclose the request of the group that had coopted the letter writing (Dexter, 1956).

Technology may make it easier to misrepresent one's political identity, whether individually or *en masse*. Individuals or groups can more easily mass produce and then, with a click, distribute emails or posts that are perceived as constituent letters or sentiment of some type while not being from actual constituents at all. In 2009, the American Coalition for Clean Coal Electricity indirectly hired a lobbying firm that sent fake letters opposing climate change legislation. The legislative recipient, Representative Edward Markey, designated the tactic as astroturfing, in this case an effort to make one group's voice come across as a groundswell of supporters.[3] In this case, the firm had to hand-write each letter, limiting the volume and therefore potential impact. In the field experiment described and carried out below, we investigate whether malicious actors could exploit natural language processing models to write ideologically oriented constituency letters that elicit the same degree of engagement from legislators as human-written letters.

Research design

To evaluate the potential effect of AI-based communication tools on democratic representation, we employed a field experiment on state legislators, sending a range of email communications that varied in terms of whether the text was human or machine-generated, the nature of the policy issue engaged, and the ideological valence of the position advanced. We then evaluated whether legislators responded at different rates to human versus machine-generated communications. Our design builds on previous field experiments that use legislative responsiveness to constituent opinion (Butler and Broockman, 2011; Butler and Dynes, 2016; Butler et al., 2012; Putnam, 1992: 73) as an indicator of democratic representation and an indication of priority in a context of legislators with limited time and resources as an indicator of legislative priority. In our case, we use responsiveness to probe the possibility that AI could purport to be human constituents and distort the process of democratic representation.

As with many other field experiments, we conducted our study with state legislators, because they provide a large sample size, more than 7200 rather than the 535 for the US Congress. Furthermore, as Nyhan and Reifler (2015) note, because state legislators have smaller staffs than members of the US Congress and receive less mail, state legislators should be more likely to encounter the mail we send and therefore be more sensitive to the differences between a human-written correspondence and one that is machine-generated. Thus, if we find that state legislators cannot discern the differences, then national legislators will be unlikely to do so either.

We selected six issues that are high salience in politics: reproductive rights; gun control; policing/crime; tax levels; public health; and education.[4] Drawing on the approach of Butler et al. (2012), we then recruited undergraduate students to draft emails to state legislators, working through the university Political Union to tap into individuals' policy

Dear _____,

I hope this letter finds you well. I would first like to thank you for the job you've done in representing us in this time of national turmoil. The pandemic and social unrest sweeping the nation right now have made life difficult for all, and many citizens, myself included, fear that our Second Amendment right to firearm ownership may become more and more of a necessity in daily life.

However, there are many in this country who would like to strip us of the rights guaranteed to us by the Constitution to protect ourselves and our families in a time when such rights are needed the most. Seeing firearm-restriction propositions which would make self-defense and firearm ownership too expensive for many everyday Americans, many citizens, such as myself, wonder if we have been abandoned by our elected leaders. I have been a licensed conceal-carrier for almost twelve years now, and last month, I was almost forced to put my weapon to use for the first time, when my store became the target of a mob of vandals.

Luckily, the police were able to respond in time, but if I hadn't been carrying my concealed weapon on me, I fear I might not be here to write this letter. Sadly, a few people in this country, including many of your colleagues, would like to do away with every American's inalienable right to firearm ownership and leave the most vulnerable citizens, such as myself, defenseless against the predation of criminals. Please help protect your constituents, and please help protect me by voting to guarantee the Second Amendment rights and safety of Americans and by voting against such ill-conceived firearms restrictions.

Thank you.

Sincerely,
Right-Wing Constituent

(a)

Dear ___

Gun control is a highly contentious topic in American society. As such, I feel that it's important to make my opinion known to you as your constituent, so you have a broader sense of the sentiments within your constituency. It has recently come to my attention that some constituents have asked you to introduce legislation that would restrict certain firearms, as well as background checks for individuals attempting to purchase firearms. I'd like to write to you today to express my disgust with that request.

Guns are a right of the people, and not a privilege. Restricting an individual's right to protect themselves and their family goes against everything this country stands for. We are a nation built on the belief that all men and women are created equal and endowed by their creator with certain unalienable rights. In order to preserve that right, we need to keep firearms in the hands of those who would stand up for those rights, and to be kept away from those who would abuse their liberties.

As the late senator Robert Byrd put it, "When you're marching up that hill, and the enemy is shooting at you, and there's shells falling all around you, there's not much time for prudence and discretion, and you don't worry too much about the Constitution. You'd just as soon come home and say, 'Hey, I've just as much right to live as anyone else.'" In short, we cannot ask our citizens to give up their rights to fight for those rights. Guns are an invaluable tool in preserving the freedoms and rights that this country was founded upon.

Sincerely,
Right-Wing Constituent

(b)

Figure 1. (a) Human-written right-wing constituent letter and (b) AI-written right-wing constituent letter.

background. We focused on advocacy letters because of our theoretical interest in whether machines can manipulate the process of representation. Rather than providing templates which we feared might compromise authenticity, we simply requested that the student write an email "advocating for the right/left-wing position" on their assigned issue.

Next, we produced the machine-generated constituency letters using GPT-3. Using the "few shot learning" approach of taking initial training examples—in our case, a tranche of ideologically directional constituency letters—to infer how to write content and ideology for constituent letters (Brown et al., 2020), we took 12 letters (six issues, right and left for each) and gave the instruction to, for example, "write a conservative letter for topic 1 (e.g. gun control): [show the conservative topic 1 letter]." For each of the ideologies and topics, we generated 100 different outputs. We conducted light editing for the purposes of internal validity, that is, to ensure that the human and AI letters were both 3 paragraphs long so that length differences would not be the basis for response differences. To be sure, the finessing created more polish than the cruder outputs that might result without editing, although our use case involves malicious intent where an actor would seek to mislead legislators and would correspondingly do the type of light editing that we conducted. Figure 1(a) and (b) provides letters representative of the human and GPT-3-written constituent letters, in this case, an ideologically conservative email addressing the issue of gun control.

To send constituent letters to state legislators, we generated aliases for the email addresses and senders, using first name, middle initial, last name, and month/year (MM/YY format) as the standardized gmail address. We consulted the Social Security Administration for the most common first names and the Census Bureau for the most common last names and randomized the combinations of popular names. For example, one alias was MargaretTThomas208@gmail.com, another was NicoleWWilson522@gmail.com, and so on, to include popular first and last names and "birth" month and year.

We selected all female names to control for gender but also because we expected that sending a reproductive rights letter from a female constituent would be more expected than a male constituent. We cross-randomized the names so that we did not inadvertently associate one particular name with an ideology or advocacy group and generated standardized email aliases for all 24 name combinations.

We then collected information to determine an appropriate email timing cadence. Our concerns here were twofold. From an ethical standpoint, we did not intend to inundate the state legislator's inbox with responsibilities. Furthermore, we did not want to prompt concerns about whether large volumes of mail were suspicious. We reached out by phone to 34 state legislators in 7 different states, with a response rate of 35%, to inquire about their daily email volume. Although the number varied, it averaged about 100 correspondences per day. Based on this volume, we developed an email cadence in which we sent one email every other business day, randomly selecting one of the 100 letters generated by the application programming interface (API). Using Python, we then developed the system of sending emails automatically, randomly alternating between human and GPT-3-written emails across the six different substantive issues from conservative and liberal perspectives. Altogether, the program sent 32,398 emails, approximately 5 per legislator. Because of the random distribution of content, some legislators received four GPT-3 email and one human-written email, but on average, a legislator would have received a roughly even number of human and AI-written emails. Similarly, not every legislator would have received a right-wing gun control letter but the random dissemination across the large volume of legislators provides a way to evaluate the relationship between issue, partisanship, and perception that the content was written by a real citizen.

To assess whether legislators reacted differently to AI-generated versus human-produced communications, we first compare the response rates of the AI versus handwritten emails. Following the approach of other field experiments, we look at response rates as an indicator of how legislators spend time and effort and as a reflection of priorities (Butler and Broockman, 2011). We also assess responsiveness by examining the length of legislative replies to human versus AI-generated letters.

Ethical considerations

Although our university's Institutional Review Board (IRB) approved this study, the design of the field experiment does warrant thoughtful engagement of ethical considerations. We drew on the approach of previous studies that have used aliases to gauge legislative responsiveness (e.g. Butler and Broockman, 2011; Butler and Dynes, 2016; Butler et al., 2012; Putnam, 1992: 73). As these studies have noted, members of Congress may have socially desirable reasons to assert virtuous positions so, for example, gauging whether they exhibit racially biased behaviors when it comes to representation requires finding ways that go beyond asking those members whether they are racist. Field experiments using Hispanic, Black, or Muslim aliases and assessing legislative responsiveness have provided useful for understanding whether legislators actually respond at lower rates to minorities than to whites. An audit study affords similar research benefits in answering our question.

Nevertheless, we acknowledge the ethical critiques of deception studies, which include concerns about whether the belief that the emails are authentic causes legislators to take action, even if just to reply to an email or actually to take policy action. With those critiques in mind, we considered possible alternatives. For example, Landgrave (2020) proposes asking legislators directly or requesting that they participate in a study, in our case a study that would evaluate the verisimilitude between AI and human-generated constituency content. If asked directly, legislators might claim to discern the difference between machine and human-generated content. However, we would not know in practice whether in fact they can distinguish one from the other. A field experiment that randomizes human and machine-written emails to legislators and compares then the response rates is the best way to know in practice to establish whether and the conditions under which legislators can tell the difference.

Similarly, asking legislators to participate in a subsequent study in which they know they will have to make distinctions between types of content would prime them to be more attentive than they might be in a more normal, non-experimental setting. In the real world, if a malicious actor engaged in this sort of exercise of sending "constituency" letters, legislators would not have prior awareness. Our study design is the closest to a real-world scenario in which legislators receive actual emails, possibly from malicious actors. Nonetheless, in line with IRB guidance, we sent debrief emails to legislators notifying them of the study design and invited them to remove their data from the study.

More generally, while we are sensitive to concerns that this research could aid malicious actors, the 2016 election, in which the Russian Internet Research Agency to manipulate Americans on social media, shows that malicious actors have already used open-source techniques in nefarious ways. This research is not involved in devising new algorithms. Rather, we are using what is tantamount to commodity technology in the spirit of pragmatism both to uncover the potential threats but also to spark policy conversations about how to combat those threats. We believe that our methodological approach, which follows in the model of all prior audit studies, is essential for accurately assessing the potential threats of emerging technology to democratic institutions and hope that its findings will benefit legislators by helping them better understand the potential risk of technologies that can skew the democratic process and encourage efforts to guard against that risk.

Results

To assess whether state legislators and their staffs were able to distinguish between human versus AI-generated content, we first compare the response rates to the AI versus human-written emails.[5] Figure 2 presents average legislator response rates to human versus machine-drafted constituent emails. Across all six issue areas, state legislators responded to 17.3% of the human-written emails. By contrast, the average response rate to machine-generated emails was 15.4%. This difference is statistically significant ($p < .001$, two-tailed test). However, the gap is substantively modest, suggesting that many state legislators and their offices did not dismiss machine-generated content as inauthentic, but rather responded at a relatively comparable rate as to human-written content.

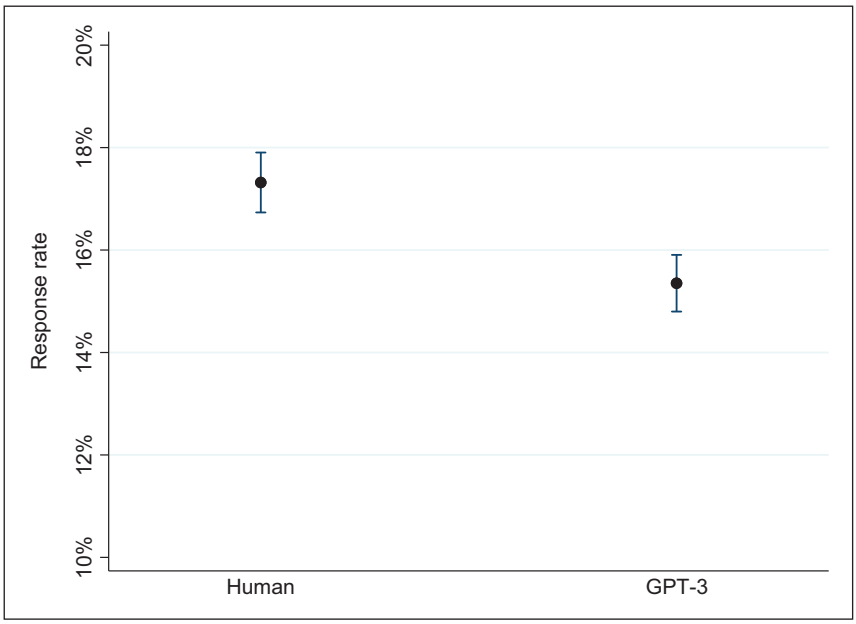

Figure 2. Legislative response rates to human versus AI emails.
Note: I-bars present 95% confidence intervals.

We note that both response rates are lower than those observed in many previous field experiments on state legislators (Costa, 2017); however, as Butler et al. (2012) observe, response rates are significantly lower for policy emails than for constituency-service requests.[6] Furthermore, the pandemic meant that almost all legislators were working from home and may not have had efficient access to their work emails. We also conducted the study during an acrimonious election year (2020), and its aftermath created the potential for greater volumes of constituency emails and competing demands on state legislators' time, which may also have contributed to lower response rates.

The aggregate difference in response rates between human and machine-generated policy advocacy emails presented in Figure 2 masks significant variation across issues. As shown in Figure 3, on two issues—gun control and health policy—we found virtually identical response rates for human versus machine emails. And on a third, education policy, we observed a higher response rate for the machine-generated emails (17.4% vs 15.8%), though the difference is not statistically significant. This suggests that on these issues GPT-3 succeeded in producing content that was almost indistinguishable in the eyes of state legislative offices from human content. By contrast, legislators were less responsive to machine-generated communications on three issues: policing; reproductive rights; and taxes.

Turning from response rates to the characteristics of legislative responses, Figure 4 offers little evidence of substantively meaningful differences in response length to human versus machine emails. The median legislative response to a human-written email was

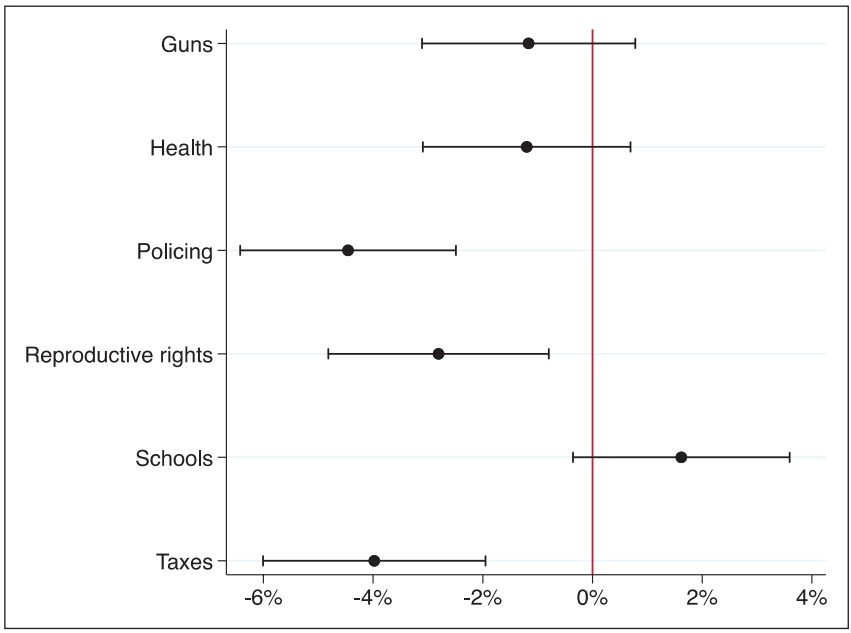

Figure 3. Differential response rates (AI—human emails) by topic.
Note: I-bars present 95% confidence intervals around each difference in means.

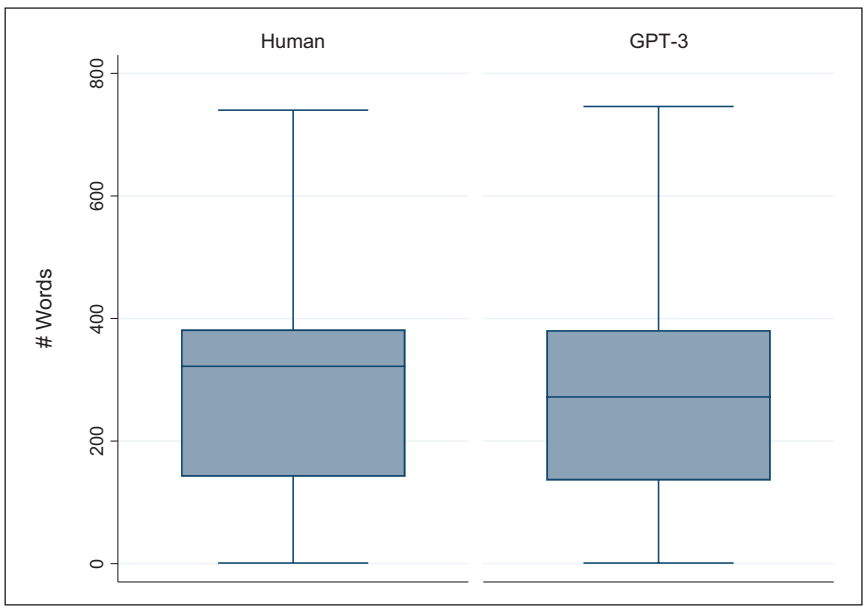

Figure 4. Differences in response length to human versus AI emails.
Note: Outlying values excluded.

Table 1. Partisanship and the response to ideologically congruent messages.

	Democrat	Republican	Democrat	Republican
GPT-3	−0.08*	−0.21***	−0.08	−0.16**
	(0.04)	(0.04)	(0.06)	(0.06)
GPT-3 × Conservative			0.01	−0.08
			(0.09)	(0.09)
Conservative slant	−0.11**	0.27***	−0.12*	0.31***
	(0.05)	(0.04)	(0.06)	(0.06)
Topic: Health	−0.04	−0.22***	−0.04	−0.22***
	(0.08)	(0.08)	(0.08)	(0.08)
Topic: Policing	0.10	−0.13*	0.10	−0.13*
	(0.08)	(0.08)	(0.08)	(0.08)
Topic: Reproductive	0.07	0.01	0.07	0.01
	(0.08)	(0.07)	(0.08)	(0.07)
Topic: Schools	0.12	−0.04	0.12	−0.04
	(0.08)	(0.07)	(0.08)	(0.07)
Topic: Taxes	0.26***	0.03	0.26***	0.03
	(0.07)	(0.07)	(0.07)	(0.07)
Constant	−1.53***	−1.69***	−1.53***	−1.71***
	(0.07)	(0.07)	(0.08)	(0.07)
Observations	14,604	16,723	14,604	16,723

Logistic regressions; standard errors clustered on legislator in parentheses; all significance tests are two-tailed.
***$p < .01$; **$p < .05$; *$p < .10$.

longer, 322 words versus 272 for the median response to a machine-generated email.[7] However, the interquartile ranges of response length are almost identical (human: 142 to 382; GPT-3: 136 to 381). Taken together, there is little evidence from the length of responses to suggest that legislators were less responsive to machine-generated emails than human emails. To the extent that time spent is an indicator of priorities, the comparable lengths indicate comparable degrees of engagement (Butler et al., 2012: 482).

Finally, we investigated whether legislators were better able to detect (and therefore not respond to) AI-generated emails that were ideologically congruent with their partisan priors than ideologically dissonant emails. To avoid triple interactions and ease interpretation, we estimate separate regressions for Democratic and Republican state legislators. In the first pair of models in Table 1, the primary independent variable of interest is an indicator variable identifying AI-generated constituent emails. In the second pair of models, this indicator is interacted with another dummy indicating email with a Conservative ideological slant.

Models 1 and 2 show that both Democratic and Republican legislators were significantly less likely to respond to AI-generated emails than to human-generated emails; however, the gap in response rate was larger for Republicans. Models 1 and 2 also show that legislators are more likely to respond to ideologically congruent emails; Republican legislators were more likely to respond to emails with a conservative slant than those with a liberal slant, and vice versa for Democrats. However, as shown in the interactions

included in Models 3 and 4, neither Democrats nor Republicans were more responsive to ideologically congruent AI-generated content than to ideologically dissonant AI-generated emails.[8]

Qualitative evidence: how to detect AI-generated text

As outlined above, many legislators responded to the machine-generated emails at levels nearly as high as the human-generated emails. Qualitative analysis of replies suggests some characteristics of AI-generated content—for example, emails that invoked personal stories and experiences—that was most likely to be mistakenly accepted as genuine by legislators. One particularly poignant email from a "Margaret Thomas" to one of the legislators read as follows:

> I am writing to you about the reproductive rights issue. I am a 15 year old girl, and I have a good friend who became pregnant and had an abortion. It is illegal to have an abortion in this state. But the reproductive rights of women need to be protected. I agree with freedom of choice. Everyone has their own right to live their own life. My friend was not mentally or physically ready for a child, so she chose an abortion.

The member of Congress wrote back with a personal salutation and thanked Margaret for

> urging support for legislation that broadens access to reproductive health care, and thank you for your patience in receiving a response. I am a proud supporter of the bodily and reproductive health rights of women, men, the LGTBQ community, and gender nonconforming individuals, including the right to affordable reproductive health coverage like birth control and abortion care. That's why I voted in favor of HB 1608 (aka the Protecting Patient Care Act) which passed earlier this year.

The legislator spoke about the sexual education legislation she had endorsed for the state and closed by wishing Margaret's friend well. In this case, the apparent authenticity of the letter and content meshed well with the legislator whose website showcases her commitment to reproductive rights and gender equality, creating a natural ideological affinity that perhaps drew the legislator to sympathize and humanize what was a machine-written letter.

Other exchanges suggest a number of inconsistencies that helped legislators flag machine-generated content. For example, the combination of algorithm and randomized alias was imperfect and sometimes produced some logical inconsistencies that raised red flags for attentive legislative readers. In one right-wing machine-generated gun control email, a "Rebecca Johnson" wrote that

> My name is Rebecca Johnson, and I am a single father raising a daughter. I am very grateful that President Trump has appointed Neil Gorsuch to the Supreme Court and will appoint more judges just like him. Gorsuch is a supporter of the second amendment and will reverse the position of his predecessor on this critical issue. Our 2nd amendment was not made for hunting or even self-defense. It was put into place to ensure that you and I would always have a way to

> defend the rights of our fellow Americans and our posterity from those who would see us destroyed . . . The very existence of government renders the preservation of liberty an absolute necessity . . . A well-armed populace is the final and essential safeguard against tyranny.

The legislator responded to the email, saying,

> Hello Rebecca, I am confused. You say you are a single father? Just want to be correct. Is your name Rebecca? I know several people with names which can be either male or female like Corey and Leslie.
>
> I appreciate your vote of confidence re the 2nd Amendment. We are on the same page.
>
> I would enjoy meeting you prior to the election as long as you are comfortable to talk. What town are you in. I am knocking on doors up and down the District. Please share your contact info so we can touch base. My info is below.
>
> Thanks again.
>
> Be Well

The member correctly noted that the machine-generated content, when mapped onto the alias, had some potential inconsistencies and followed up with the sender by way of eliciting more information.

Reviewing the right-wing machine-generated gun control emails suggests that these communications in particular were problematic, advocating for *more* gun control about 50% of the time when a conservative position would typically advocate for *less* gun control. In retrospect, asking the tool to generate "right-wing gun rights" email often produced emails more consistent with a left-wing agenda. One example among many included the following ideological incompatibilities bordering on nonsense:

> I am a rifle owner and hunter. I support universal background checks because, if allowed, that would make a world where the possibility of obtaining a weapon as a non-political citizen (by the identity of the alleged criminal) through the automatic purchase of an NRA-branded rifle would also be sufficient to satisfy the burden of carrying a gun that does not belong to anyone, including my children or those close to me. I also believe that the purchase of a firearm is a hobby for anyone, not a constitutional right. In addition, a gun owner has a right to not be threatened or victimized by those that are not law abiding citizens, and to not be forced into being a stranger's slave. So I ask that you and your colleagues understand this as the situation of a reasonable gun owner, and not as something that has been created by the NRA with an agenda driven by profit.

Malicious actors could conceivably curate the outputs and jettison the problematic ones, but doing so would come at a cost of efficiency, which would run counter to the astroturfing objective, which requires scale.

Per IRB guidelines, at the conclusion of the experiment we sent follow-up emails to legislators informing them of the study and its research objectives. In response, we

received several informative replies that shed light on both the potential perils that these models present for democratic representation but also the potential techniques that legislators might employ to guard against AI-sourced astroturfing.

Follow-up emails pointed to features of correspondence that are difficult to carry out through a language model, therefore providing clues to legislators about how to guard against machine-initiated astroturfing. First, we heard from a number of legislators that if they did not see a marker that the individual resided in the district or questioned the authenticity of the email, they would not respond. A state legislator from Charlottesville, Virginia, who agreed to be quoted by name for research purposes, responded:

> If we get a message from someone who identifies as a resident of "Charlottesville," "Albemarle County," or a set of nearby zip codes, we respond. . .If the initial message does not contain any residency information, we reply and request their zip code . . .. If the initial message identifies the sender as a resident of another state or region of Virginia we don't respond

except in rare exceptions where the email overture is related to their work.[9] AI-generated content is less likely to include authentic geographic information because the model is trained on content from the broader Internet. Legislators would therefore be wise to continue using address as a marker of authenticity.

Second, and relatedly, some legislators noted their ability to filter out inauthentic content on the basis of legislators' degree of direct knowledge of constituents. One legislator indicated that he covers a small district, so he knows virtually everyone. When he received several similarly crafted emails, he shared his theory with colleagues about a suspected research project, which he thought could have led to the observable implication that both his and his colleagues' responses declined after the initial overtures.[10]

Another legislator from Vermont also sent substantive feedback and observed that the

> House districts in Vermont are so small that it's pretty easy to spot and identify email that is not from a constituent just by looking at the name. We also don't have staff, so in my case at least, the most questionable email won't get an answer.

He revisited the email from "Ms. Johnson" and remarked at the "incoherent justification" for a policy prescription she provided.[11] While such comments were limited to legislators from small districts, it suggests that legislators from larger districts at the state or federal level would be wise to be attentive to the other markers of machine generation, whether ideological discrepancies or the absence of specific addresses.

Third, some members noted that knew their type of constituent and were therefore able to identify prose that did not use that locality's vernacular. One member from a less affluent district responded that his constituents "write like they talk" and since the AI-written letters were less colloquial and more formal, he had flagged them as spam or from outside the district and thereby not worth consideration or a response. This analysis suggests that legislators should maintain and even augment the human connection with constituents in ways that will directly prop up trust in democratic representation and provide context that allows them to guard against inauthentic campaigns.

Conclusion

Emerging technologies are increasingly able to blur the line between fact and fiction. Deepfakes have proliferated online in ways that can impugn the reputation of public officials and celebrities and are increasingly able to evade human detection (Ternovski et al., 2021). Natural language models can develop news text that is as credible as newspapers of record. Whether elected officials are susceptible to these technologies has not yet been examined. In particular, whether legislators can discern differences between constituent concerns and machines that aim to manufacture and amplify preferences has not been studied despite the verisimilitude of generated text and the legislative body's importance for laws, oversight, and representation.

In this research, we implemented a field experiment to expose legislators to both human-written advocacy letters and machine-generated letters trained on those same human-written letters. Overall, we find that legislators were less likely to respond to AI-generated content than to human-drafted emails, but by less than 2%. Qualitative evidence both in the original responses and in follow-up correspondences with legislators who participated in the study corroborates that a number of legislators flagged the confusing or inconsistent formulations of machine-generated text. Nonetheless, while legislators were marginally less responsive to AI-generated messages, a nontrivial number was persuaded enough to respond to machine-written communications and to do so in substantively meaningful ways. To the extent that those particular members were either vocal or powerful in their respective legislators, the technology may therefore have consequences for democratic representation.

Limitations of the design might actually understate the potential of the technology. Closer scrutiny would perhaps have caught the ideological inconsistencies yielded by the machine learning tool that raised skepticism among some legislators. Subsequent research could consider additional editing to ensure that the machine-written emails were all coherent—more of a human editor in the loop—although adding the human would also erode the advantages of the tool other than producing original, creative, and non-plagiarized content at scale.

Another potential limitation is that we studied state rather than national-level legislators. In our follow-up correspondences, many legislators reviewed the litany of ways they discerned whether a letter writer is a constituent; many of these methods are likely effective because the districts were small or homogeneous. These tools may be less effective in helping national legislators root out malicious machine-generated content. In other respects, however, national legislators may be more insulated from at least some attempts at astroturfing. Letters contain postmarks, and those from outside the state would be obvious flags of non-constituent correspondence. Sending emails is also more challenging since national-level legislators do not have emails listed online like the state legislators but rather buttons to click. Transparency groups have initiated open-source code for emailing national legislators, which might paradoxically increase the potential for malicious actors to astroturf. Subsequent research might therefore consider fielding the experiment at the national level. Constituents may have more anonymity in the larger national-level districts; the sheer level which correspondence received may also mean less discernment, but this might perhaps be offset by the greater professionalization of

the full-time national-level staff. Given these differences, it is not clear whether we can generalize from our results on state legislators to federal legislators.

Thus, while our field experiment offered a meaningful proof of the concept for whether and how natural language models can generate credible constituency emails and potentially manipulate the process of representation, subsequent studies might consider ways to build on our design and better understand the scope of the threat. The 2016 presidential election raised the prospect that outside actors might try to use open institutions in a democracy to manipulate public opinion. We go beyond studying the mass public or social media to show that elected leaders themselves are potentially susceptible to large-scale manipulation through standard, open channels of political communication. Our analysis thereby speaks to longstanding questions of democratic representation (Butler and Broockman, 2011; Butler et al., 2012; Putnam, 1992) while updating it for the context of emerging technologies and outside influence.

Our research does raise questions about how to protect against the potential impacts of these technologies on democratic representation. If machines can pass as human constituents, then it is possible that they could successfully astroturf members of Congress and create a sense of mass support for a particular policy unless proper guardrails are established. One solution for machine-generated text that seeks to mislead is technology itself. Because neural networks helped generate these language models such as GPT-3, then the same machine learning algorithms might be adapted to identify machine-generated text. Researchers (Gambini et al., 2022; Zellers et al., 2019) have shown that the accuracy of machine-based detection of generated text can reach accuracy rates of close to 90%. Human detection is more difficult because the markers of generated text are subtle: grammatical errors, repetition, and typographical errors. However, awareness of those hallmarks through greater digital literacy of generated text offers a potentially fruitful path for detecting differences between human and machine content in analogous ways to identifying online deepfakes (Groh, 2020). As these technologies evolve, so too must digital literacy to serve as a counterweight to these technologies' potentially corrosive effects on society and democracy.

Author Note

We trained the tool on student-written letters and then chose not to edit the content heavily because our misuse case involved malicious actors who might not have the language skills or time to finesse the content.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Sarah Kreps https://orcid.org/0000-0002-0924-4234

Supplemental material

Supplemental material for this article is available online.

Notes

1. Lapowsky (2017).
2. "How to Deal with a High Volume of Incoming Communications," https://www.congressfoundation.org/office-toolkit-home/improve-mail-operations-menu-item-new/1320
3. Strom (2009).
4. John Haughey, "17 Issues Facing State Legislatures," https://info.cq.com/resources/17-issues-facing-state-legislatures-in-2019/.
5. Our results are robust to alternate operationalizations of response rates that screened out various types of replies. See the Supporting Information (SI) for more discussion and a series of robustness checks (SI Figures 1–3 and SI Tables 1–3).
6. For example, Butler et al. (2012: 479) found that state legislators responded to 51% of their constituency-service letters but only 28% of policy letters.
7. While substantively small, the difference is statistically significant. A nonparametric *k*-sample test on the equality of medians rejects the null of equality, $p = .001$, two-tailed test. More broadly, a Wilcoxon rank-sum test also rejects the null of two independent samples, $p < .001$, two-tailed test.
8. Additional analyses examining responses only in the three issue categories with significant differences in response rates across human versus AI-generated emails show similar results; see SI Table 2.
9. Sally Hudson (D-VA), 12 December 2020, email correspondence.
10. Josh Boshee (D-ND), 20 December 2020, email correspondence.
11. David Durfee (D-VT), 20 December 2020, email correspondence.

References

Ansolabehere S, Gerber A and Snyder J Jr (2002) Equal votes, equal money: court-ordered redistricting and public expenditures in the American states. *American Political Science Review* 96(4): 767–777.

Balinski M and Young HP (2010) *Fair Representation: Meeting the Ideal of One Person, One Vote*. Washington, DC: Brookings Institution Press.

Bol D, Geschwend T, Zittel T, et al. (2020) The importance of personal vote intentions for the responsiveness of legislators: A field experiment. *European Journal of Political Research* 60(2): 455–473.

Brown TB, Mann B, Ryder N, et al. (2020) Language models are few-shot learners. Available at: https://arxiv.org/abs/2005.14165

Buchanan B, Lohn A, Musser M, et al. (2021) How language models could change disinformation. Center for Security and Emerging Technology. Available at: https://cset.georgetown.edu/publication/truth-lies-and-automation/

Butler DM and Broockman DE (2011) Do politicians racially discriminate against constituents? A field experiment on state legislators. *American Journal of Political Science* 55(3): 463–477.

Butler DM and Dynes AM (2016) How politicians discount the opinions of constituents with whom they disagree. *American Journal of Political Science* 60(4): 975–989.

Butler DM and Nickerson DW (2011) Can learning constituency opinion affect how legislators vote? Results from a field experiment. *Quarterly Journal of Political Science* 6(1): 55–83.

Butler DM, Karpowitz CF and Pope JC (2012) A field experiment on legislators home styles: service versus policy. *The Journal of Politics* 74(2): 474–486.

Campbell A (2003) *How Policies Make Citizens: Senior Political Activism and the American Welfare State*. Princeton, NJ: Princeton University Press.

Castiglione D and Pollak J (2018) *Creating Political Presence: The New Politics of Democratic Representation*. Chicago, IL: The University of Chicago Press.

Congressional Management Foundation (CMF) (2005) *Communicating with Congress: How Capitol Hill Is Coping with Citizen Advocacy*. Washington, DC: CMF.

Congressional Management Foundation (CMF) (2011) *Communicating with Congress: Perceptions of Citizen Advocacy on Capitol Hill. Partnership for a More Perfect Union*. Washington, DC: CMF.

Costa M (2017) How responsive are political elites? A meta-analysis of experiments on public officials. *Journal of Experimental Political Science* 4(3): 241–254.

Dahl R (1971) *Polyarchy: Participation and Opposition*. New Haven, CT: Yale University Press.

Dexter LA (1956) What do congressmen hear: the mail. *Public Opinion Quarterly* 20(1): 16–27.

Disch L (2019) *Constructivist Turn in Political Representation*. Edinburgh: Edinburgh University Press.

Einstein KL and Glick DM (2017) Does race affect access to government services? An experiment exploring street-level bureaucrats and access to public housing. *American Journal of Political Science* 61(1): 100–116.

Gambini M, Fagni T, Falchi F, et al. (2022) On pushing DeepFake Tweet Detection capabilities to the limits. In: *14th ACM web science conference*, pp. 154–163. Available at: https://dl.acm.org/doi/abs/10.1145/3501247.3531560

Griffin J and Newman B (2005) Are Voters Better Represented? *Journal of Politics* 67(4): 1206–1227.

Groh M (2020) Detect deepfakes: how to counteract misinformation created by AI. MIT Media Lab. Available at: https://www.media.mit.edu/projects/detect-fakes/overview/

Hayden G (2003) The false promise of one person, one vote. *Michigan Law Review* 102(2): 213–266.

Jakesch M, French M, Ma X, et al. (2019) AI-mediated communication: how the perception that profile text was written by AI affects trustworthiness. In: *CHI 2019: proceedings of the 2019 CHI conference on human factors in computing systems*, 4–9 May. Available at: https://dl.acm.org/doi/10.1145/3290605.3300469

Kreps SE, McCain M and Brundage M (2022) All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science* 9: 104–117.

Landgrave M (2020) Can we reduce deception in elite field experiments? Evidence from a field experiment with state legislative offices. *State Politics & Policy Quarterly* 20(4): 489–507.

Lapowsky I (2017) How bots broke the FCC's public comment system. *WIRED*, 28 November. Available at: https://www.wired.com/story/bots-broke-fcc-public-comment-system/

Liu H, Cui L, Liu J, et al. (2021) Natural language inference in context—investigating contextual reasoning over long texts. *Proceedings of the AAAI Conference on Artificial Intelligence* 35(15): 13388–13396.

Martin P and Claibourn M (2013) Citizen participation and congressional responsiveness: new evidence that participation matters. *Legislative Studies Quarterly* 38(1): 59–81.

Nyhan B and Reifler J (2015) The effect of fact-checking on elites: a field experiment on U.S. state legislators. *American Journal of Political Science* 59(3): 628–640.

Putnam RD (1992) *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton, NJ: Princeton University Press.

Saward M (2008) Representation and democracy: revisions and possibilities. *Sociology Compass* 2(3): 1000–1013.

Stimson J, Mackuen M and Erikson R (1995) Dynamic representation. *American Political Science Review* 89(3): 543–565.

Strom S (2009) Coal group is linked to fake letters on climate bill. *The New York Times*, 4 August. Available at: https://www.nytimes.com/2009/08/05/us/politics/05charity.html

Ternovski JK, Kalla J and Aronow P (2021) Deepfake warnings for political videos increase disbelief but do not improve discernment: evidence from two experiments. Available at: https://osf.io/dta97/

Tormey S (2015) *The End of Representative Politics*. New York: John Wiley & Sons.

Zellers R, Holtzman A, Rashkin H, et al. (2019) Defending against neural fake news. Available at: https://arxiv.org/pdf/1905.12616.pdf

Author biographies

Sarah Kreps is the John L Wetherill Professor of Government at Cornell University and Director of the Cornell Tech Policy Lab. Her research focuses on the intersection of technology, politics, and international relations. She is the author, most recently, of *Social Media and International Relations* (Cambridge University Press, 2020).

Douglas L. Kriner is the Clinton Rossiter Professor in American Institutions at Cornell University and the faculty director of the Institute of Politics and Global Affairs. He has authored five books, most recently (with Dino Christenson) *The Myth of the Imperial Presidency: How Public Opinion Checks the Unilateral Exeuctive* (University of Chicago Press, 2020).