The impact of engagement and partisan influence campaigns in an isolated social media environment

Manikya Alister[1] (alisterm@student.unimelb.edu.au)
Keith Ransom[2] (keith.ransom@adelaide.edu.au)
Anthony Lua[1] (anthony.lua@student.unimelb.edu.au)
Andrew Perfors[1] (andrew.perfors@unimelb.edu.au)
[1]School of Psychological Sciences, University of Melbourne
[2]School of Computer and Mathematical Sciences, University of Adelaide

Abstract

Despite growing concerns about the effect of social media engagement on people's beliefs and behavior, estimating the actual impact is difficult. Here we present preliminary results from our own isolated social media platform named Magpie Social. In it, participants could interact with each other like typical social media, but we had control over the platform and measured people's beliefs and behavior before and after using it. This allowed us to more closely approximate the ecological validity of naturally occurring social-media data, while retaining the ability to measure variables and infer causation. Our week-long task had three between-subject conditions (total $N = 311$): a CONTROL in which people engaged on Magpie with no external influence, and two (LEFT and RIGHT) in which a small number of posts were secretly made by us, sharing typical talking points from one political side. We found small but statistically reliable effects suggesting that, relative to the CONTROL, the presence of right-wing trolls resulted in a higher level of right-wing belief and a greater perception of political division in the US. Conversely, the left-wing troll campaign did not appear to have any statistically reliable effect on these measures. We also found considerably more overall engagement in both troll conditions, probably because content with a clear political stance tended to receive more activity. However, participants (especially those on the left) disliked the RIGHT condition more than the others. **Keywords**: social media, belief, consensus, influence, information

Introduction

Malign influence campaigns are deliberate attempts to create controversial or provocative social media content, often using multiple fake accounts controlled by a single user or organization. These "troll" campaigns have many goals, from shifting people's opinions in a particular direction, creating the perception that a specific viewpoint is more broadly supported than it actually is, increasing polarization and discord, or generating engagement. Because of their potential real-world impacts, these campaigns are of increasing concern to people from policymakers to scientists to public agencies (Meta, 2024; Microsoft Threat Analysis Center, 2024; US Department of Defense, 2023). Despite this, understanding the nature and extent of the impact on people's beliefs and attitudes is a challenging and still largely unsolved problem.

The fundamental difficulty is that it is extremely challenging to access the data necessary to measure real-world impact. On the one hand, we can analyze social media data directly; this enables us to do things like trace information flow or to estimate how troll accounts grow, receive engagement, and interact with other users (e.g., Bailo et al., 2024; Cork et al., 2020; Vosoughi et al., 2018). Such approaches are valuable because they leverage large quantities of real-world data, enabling a rich and nuanced characterization of troll campaigns in their natural environment. However, because the data is observational, the conclusions it licenses are limited. It is very difficult to measure the *impact* of troll campaigns (or social media engagement in general) on beliefs and attitudes – much less draw inferences about how they might influence people's subsequent actions or whether different people are influenced in different ways.

On the other hand, we can study how social media affects behavior by using controlled experiments to understand the factors that underlie online influence (e.g., Butler et al., 2024; Jagayat & Choma, 2024; Pennycook & Rand, 2019; Pennycook et al., 2020). These experiments are valuable for isolating and understanding causal relationships and contributing to our understanding of how (and why, and under what circumstances) information presented on social media might affect people's beliefs and behavior. For example, lab-based studies have shown that inflating the extent to which a social media post appears to be supported by a consensus often leads to changes in belief in line with that consensus (Alister et al., 2022; Lewandowsky et al., 2013; Simmonds et al., 2024).

However – valuable though it is to strengthen psychological theory – controlled experiments have the drawback that they simplify or remove many of the features of real social media environments. It is one thing to observe an effect in a study where the relevant factors are systematically varied, do not occur alongside dozens of competing and uncontrolled factors, and the effect measures occur immediately after the manipulation. It is another to do so in a situation more like the real world – a world where people differ dramatically in how they engage with social media, where targeted information is spread thinly and intermixed with everything else, and where effects must persevere much longer than the duration of a typical study in order to have an impact on behavior.

In this paper we present preliminary results from a paradigm designed to retain some of the real-world applicability of observational studies while still directly measuring beliefs and impact. We did this by building our own social media platform (an isolated Mastodon instance, see also Doshi et al., 2024) that we call Magpie Social (or "Magpie" for short). We invited participants to anonymously engage with each other on Magpie for multiple days, as similar to "real" social media as possible. Moreover, we measured many aspects of their beliefs, identity, and attitudes both before and after their interactions on Magpie. This yields a treasure trove of data allowing us to not only measure belief and attitude change, but also explore how it is related to people's experiences and behavior on the platform (and vice-versa).

Because we controlled the platform, we were also able to systematically manipulate key factors between different runs of the experiment: in this case, the presence (and nature) of

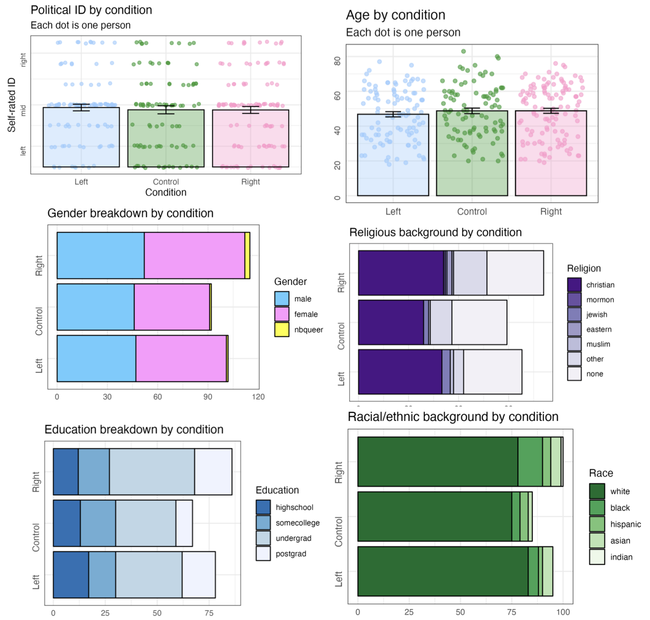

Figure 1: **Demographics**. Participants across conditions were similar in political alignment (based on self-ID on a 7-point scale, and balanced at the midpoint), age, gender, race, religion, and education.

external influence campaigns. In a CONTROL condition, people discussed topics with each other with no interference from us. In two troll conditions, one LEFT and one RIGHT, a handful of the accounts were secretly controlled by us and used to inject content from one side of the political spectrum. By comparing conditions we can explore how troll campaigns affect people's beliefs about the topics under discussion, attitudes toward each other, and engagement with the platform.

Given the large-scale nature of this study, we obtained more data (and have more measures) than is possible to fully analyze in the limited space available. We therefore report on a subset of the data. Our focus here is on whether interacting on Magpie (and/or the troll campaigns) had an effect on overall belief change or perceptions of consensus, as well as how (or if) these were related to people's engagement.[1]

Method

Participants Following a short pre-screening designed to ensure demographically similar and politically balanced pools of participants, 480 people (160 per condition), all from the US, were recruited from Prolific Academic to participate in the week-long study. Of those, 353 responded to the initial invitation. All analyses include the 311 people who completed all parts of the study (93 in the CONTROL, 116 in the RIGHT, and 102 in the LEFT conditions). People were paid $12USD/hr plus a bonus if they completed every component (up to $38.25 each for the entire study). As Figure 1 shows, the demographic characteristics of the final sample did not differ between conditions.

Day	Component	Description
Mon	Before	30-min survey of beliefs, attitudes, and demographics
Tues-Thu	Magpie	Interaction on the social media platform, 30 min/day minimum
Fri	After	45-min survey of beliefs, attitudes, perceptions of Magpie

Table 1: **Experimental procedure**. Each condition followed the same weekly pattern. People were paid for each component they completed and were sent daily reminders about what to do.

Procedure Each condition was run in a separate week between August 5 and September 7, 2024.[2] The procedure and instructions were identical across conditions (Table 1), and consisted of three days of interaction on our social media platform bookended by a surveys Before and After.[3].

Magpie Social Our social media platform, Magpie Social, is a bespoke Mastodon instance that operates much like X or Bluesky. Each person was assigned to an account of their own and given a random positive adjective+animal username (e.g., HappyKangaroo) and corresponding avatar, which they could not change. Participants could see all other people's posts in reverse chronological order and had access to the same functionality as in most social media: making posts of their own (including embedding external links, photos, or videos); replying to, reblogging, or favoriting other posts; visit other profiles; and unfollowing, blocking, or muting others. Since our server was isolated, they could not interact with any Mastodon users other than fellow participants that week.

On their first day with Magpie, participants were given detailed instructions about how to use the platform, the Code of Conduct, the timeline, and the procedure. People were asked to talk primarily about four topics: transgender rights, climate change, AI, and Israel/Palestine. They were told they would be paid if they actively engaged for at least 30 minutes per day (not necessarily all at once) and made at least one post or reply; they were not paid extra for going over 30 minutes but were allowed to do so, and many did.

The instructions, which were identical in all conditions, emphasized that we were "really interested in how people use the platform and what people do naturally" and that "it's fine if discussion meanders somewhat (that's what happens in conversations after all) as long as you mainly try to discuss the list [of four topics] above." In addition, we noted in all conditions that "like real social media, most of the people you will talk to will be genuine, but there's some chance that a person might have a hidden agenda, is lying, or isn't entirely acting in good faith - you should presume that it's pretty similar to real life in that respect."

[1] There is a pre-registration, but it contains many analyses and questions not reported here for space reasons and was exploratory in the sense of not having strong hypotheses. Thus, our results should be interpreted with that in mind. `tinyurl.com/4nvnuj8t`.

[2] Fortuitously, no major political events occurred during this time.

[3] Ethics approval was granted by Anonymous University (Approval #2393v1). The Supplementary Materials contains details about the LLM classification, instructions, stimuli, and measures: `https://tinyurl.com/4nvnuj8t`

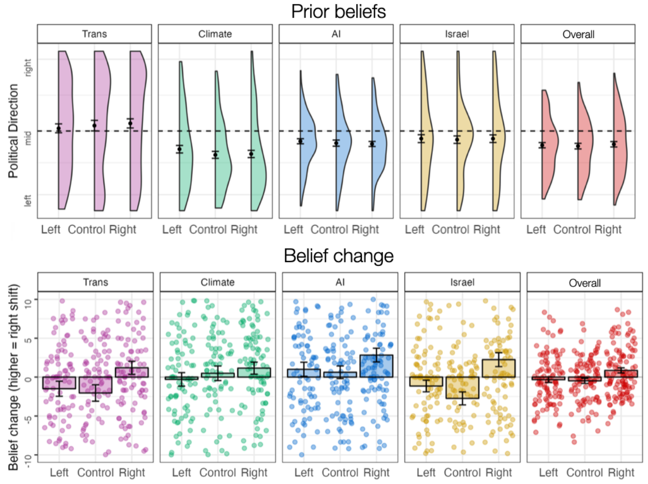

Figure 2: **Beliefs and belief change**. *Top*: Initial beliefs by topic and overall (red violins). There was substantial individual variation but all conditions were slightly left-leaning in general (means and standard errors = block dots and bars), with no initial differences between conditions. *Bottom*: change in beliefs from Before to After (positive = right shift). Beliefs changed in some topics more than others, but overall beliefs only changed significantly (slightly to the right) in the RIGHT condition.

Conditions The CONTROL condition proceeded without any direct involvement from us, but in the RIGHT and LEFT conditions we secretly controlled six accounts that looked no different than the other users (same naming scheme, avatars, etc). Over the course of each experiment, these accounts contributed 100 posts total (25 per topic). They were inauthentic in the sense that they were pre-written to mimic standard talking points from the LEFT or RIGHT side of the political aisle and the content was not generated as a part of organic conversations. For ethical reasons, unlike real trolls, they did not insult people; like many real trolls, they did not favorite other posts or respond when people engaged with them.

Belief measures Both the Before and After surveys contained the same 48 items (presented in random order) designed to measure specific beliefs. There were 6 items for each of the four main topics, and 4 each for a variety of others (health care, economy, race, guns, etc). Each item was formulated as a statement to which participants had to indicate their degree of agreement on a scale of 0 (no agreement at all) to 100 (full agreement). Within each topic, half the items were phrased so that agreement was consistent with left-leaning views, and the other half right-learning.[4] For the analysis, the left items were transformed so that scores closer to zero always indicated more left views. By comparing each person's scores from Before to After, we obtain a measure of belief change over the course of the study.

Perception measures We additionally asked several questions (both Before and After) intended to gauge people's perceptions of the state of discourse. Perceived Consensus was measured by asking "To what extent do you think that most people in the US generally agree with each other about their overall political beliefs?" (0 = no agreement; 100 = full agreement). We also estimated people's Relative beliefs by asking "Where would you classify your overall political beliefs, relative to most people in the US?" (0 = extremely far left; 100 = extremely far right). And finally, on the After survey we asked "Generally speaking, how would you rate the overall experience on Magpie Social?" (0 = I hated it; 10 = it was fantastic).

Predictor	Estimate	SE	t	p
Intercept	-0.18	0.59	-0.31	0.756
$\text{Belief}_{\text{Before}}$	0.99	0.01	86.12	< 0.001
$\text{Condition}_{\text{Left}}$	0.14	0.50	0.28	0.782
$\text{Condition}_{\text{Right}}$	1.32	0.49	2.70	0.007

Table 2: **Linear model for beliefs after being on Magpie**. Full model: `BeliefAfter ~ BeliefBefore + Condition`. Condition estimates are reported in reference to CONTROL. After controlling for the initial beliefs there was no significant difference in the After beliefs between the CONTROL and LEFT condition. However, After beliefs in the RIGHT condition were 1.32 higher (more right wing) than the CONTROL, a difference which was small but significant. These results were replicated using the equivalent Bayesian linear model; in it, both the $\text{Belief}_{\text{Before}}$ and $\text{Condition}_{\text{Right}}$ coefficients had 95% credible intervals that did not overlap with 0.

Results

Participants in all conditions engaged extensively on Magpie. For instance, despite being paid for only 30 minutes per day, nearly 25% of daily contributions totaled 45 minutes or more, and over 2000 posts were generated in each condition. People also liked Magpie a great deal, giving mean ratings of their overall experience of over 7 out of 10. Indeed, multiple people asked to be permitted to continue on Magpie after the ending date because they were enjoying it so much.

Beliefs and belief change

This study allows us to directly examine the extent to which people's beliefs changed after engaging in conversations on social media. As the top panel of Figure 2 shows, in all conditions our participants' initial beliefs were on average slightly left-leaning, with some variation by topic.[5] Our population was also highly heterogeneous, with average beliefs ranging over most of the 0-100 point scale.

As the bottom panel of Figure 2 indicates, people did change their beliefs; the shifts, which are in the expected directions, are small but *notable* given the fact that the discussions were naturalistic and the belief measures separated by many days. A linear model controlling for prior beliefs indicated that the only significant shift in overall beliefs was in the RIGHT condition, which became more right-wing (Table 2).[6]

[4] Items had previously been validated in a pilot experiment to ensure that the left/right encoding was accurate.

[5] A typical feature of American politics is that people's actual beliefs tend to be more left-leaning than their identification (which in our sample is in the center; see top left panel in Figure 1).

[6] We did not include `Topic` as a predictor because given the limited space available, we opted for a simpler analysis over the most important variable (overall beliefs, based on all 48 items).

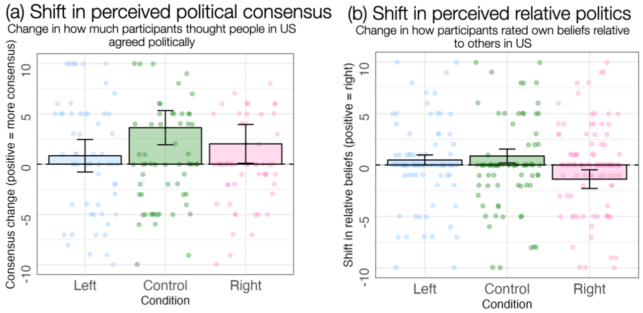

Figure 3: **Changes in perception**. (a) Shifts in participant estimates of how much people in the US generally agree with each other politically. Positive values indicate that people shifted toward perceiving *more* consensus (i.e., less belief polarization), with the largest change occurring in the Control condition. (b) Shifts in people's estimates of where they stand politically relative to others in the US (positive indicates that they felt more right-wing at the end than they did at first). Changes were very small, but those in the Right condition felt that they were more left-wing than they did at the beginning; this is equivalent to thinking that the population was further right (relative to themselves) than they did before. People in the other two conditions did not shift in the same way.

Perception of political environment

Another potential impact of social media is to affect perceptions about what the rest of the population believes. The frequency of particular views or talking points might shape people's sense of how much support those views have, or how out-of-the-mainstream their own views are.

Perceived consensus Our Consensus measure asked people to estimate what proportion of the US they thought agreed with each other; it can thus also be interpreted as a measure of belief polarization. On the Before survey, ratings on this measure were relatively low ($M = 36.9$), consistent with perceptions of a divided population. As Figure 3 indicates, people generally believed there was *more* consensus (i.e., less polarization) after taking part in Magpie, with the shift being most pronounced in the Control condition. This suggests that, contrary to expectation, interacting on Magpie may have decreased people's estimates of belief polarization. We consider reasons for this counterintuitive result in the Discussion.

Table 3 shows the results of a linear model that controlled for people's initial perceptions of consensus. It finds that the Right condition was significantly different from the Control (Table 3), showing the least positive shift. This suggests that those who saw right-wing trolls stayed closest to their initial perceptions of the US as a polarized nation. (Those who saw left-wing trolls also did not shift as much as in the Control, but the difference was not significant).

Relative belief estimation Another measure relevant to people's perceptions about others was our Relative belief score, in which people rated where they thought their own overall political beliefs stood relative to the rest of the population (with 0 indicating far to the left, 50 at the center, and 100 far to the right). The mean rating on the Before survey was 44.0, consistent with our direct belief measures indicating that our participants were slightly left-leaning overall.

Predictor	Estimate	SE	t	p
Intercept	23.03	2.48	9.29	< 0.001
$\text{Consensus}_{\text{Before}}$	0.51	0.05	10.92	< 0.001
$\text{Condition}_{\text{Left}}$	-4.30	2.25	-1.91	0.057
$\text{Condition}_{\text{Right}}$	-4.61	2.21	-2.09	0.038

Table 3: **Linear model: After consensus perception**. Full model: `ConsensusAfter ∼ ConsensusBefore + Condition`. Condition estimates are reported in reference to Control. There was a significant positive shift (Intercept) indicating that people perceived less polarization after engaging on Magpie. After controlling for the initial perceptions of consensus, there was a significant difference between the Right and Control conditions; those who saw right-wing tweets thought there was less consensus (by 4.61 points on the 0-100 scale) than those who saw no troll tweets at all. Those who saw left-wing tweets also thought there was less consensus than in the Control condition, but the difference was not significant. These results were replicated using the equivalent Bayesian linear model; in it, both the $\text{Consensus}_{\text{Before}}$ and $\text{Condition}_{\text{Right}}$ coefficients had 95% credible intervals that did not overlap with 0.

As Figure 3 shows, there was only a small amount of change in this measure of relative belief from Before to After. Those in the Left and Control had positive shifts, indicating that they thought their beliefs were more right-wing (and the population more left-wing) than they did originally. However, neither the overall shift (intercept) or the difference between Left and Control were significant in a linear model (Table 4). Conversely, participants in the Right condition shifted in the opposite direction, rating themselves as more left-wing (and the population relatively more right-wing) after being on Magpie – significantly different from the Control condition. This result suggests that in the presence of right wing trolls, people perceived their own beliefs to be more left-wing than they had previously thought they were. It is consistent with the idea that viewing more right-wing content shifted people's sense of population norms.

Relationship to belief change How are changes in belief related to changes in perception of the political environment? We investigated this by calculating the correlations between our three measures (change in belief, change in perception of consensus, and change in relative belief rating). However, none of the relationships were significant.

Predictor	Estimate	SE	t	p
Intercept	1.78	1.01	1.76	0.079
$\text{RelativeBelief}_{\text{Before}}$	0.98	0.01	65.94	< 0.001
$\text{Condition}_{\text{Left}}$	-0.22	1.06	-0.21	0.837
$\text{Condition}_{\text{Right}}$	-2.17	1.04	-2.09	0.038

Table 4: **Linear model: After perception of relative beliefs**. Full model: `RelativeBeliefAfter ∼ RelativeBeliefBefore + Condition`. Condition estimates are reported in reference to Control. After controlling for people's initial perception of their beliefs relative to others in the US, there was a significant difference between the Right and Control conditions. People who saw right-wing tweets adjusted their perception of their relative beliefs leftward, indicating that they thought the population was more right-wing than they had previously. These results were replicated using the equivalent Bayesian linear model; in it, both the $\text{RelativeBelief}_{\text{Before}}$ and $\text{Condition}_{\text{Right}}$ coefficients had 95% credible intervals that did not overlap with 0.

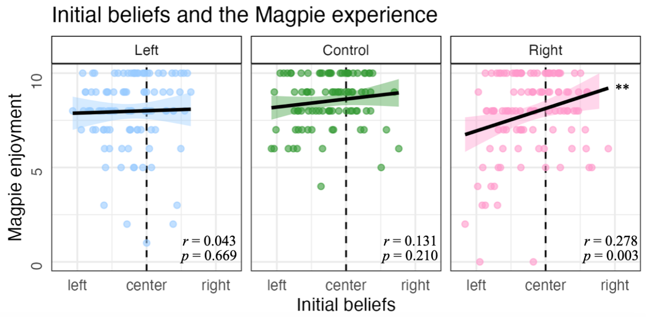

Figure 4: **Relationship between beliefs and enjoyment** of Magpie Social. Correlation between people's initial beliefs (average of 48 items on Before survey, *x* axis) and rating of how much they enjoyed Magpie on the After survey (higher is better, *y* axis). Most people in all conditions liked it, but in the RIGHT condition, more people on the left had a negative experience.

The Magpie Social experience

Our results so far suggest that the interactions and engagement on Magpie had small but significant effects on people's beliefs and perceptions of others, especially in the RIGHT condition. To further understand why and how this happened, we investigate the nature of people's engagement with the platform, the troll accounts, and each other.

Enjoyment One obvious question is whether people enjoyed participating on Magpie, and if some enjoyed it more than others. Enjoyment matters in part because in the real world, people only stay on social media that they enjoy – but also because understanding people's emotions can be a revealing clue about how they participated and what engagement looked like. As Figure 4 shows, most people liked Magpie, and most of the time this was true regardless of their political beliefs. However, in the RIGHT condition, there was a significant positive correlation between initial beliefs and enjoyment: a noticeable number of people, all on the left, did *not* like their experience. This suggests that right-wing propaganda (but not left-wing propaganda) made our social media experience less pleasant for people with the opposite beliefs.

Engagement We can also ask what people talked about and how they engaged with each other and the topics. In order to analyze this, it was necessary to identify both the *topic* that a post was discussing as well as the *political stance* (polarity) that each post took.[7] Identifying post topics was straightforward; it was coded manually by the senior author and checked by the first author. Classifying the political stance of the 6000+ posts was more subjective and time consuming, so we used a LLM (GPT4).[8] The LLM classifier was validated against a small subset of posts (∼15%) that were coded by three of the authors. The LLM obtained a reliability of 71%, which was higher than the inter-rater reliability of the human coders.[9]

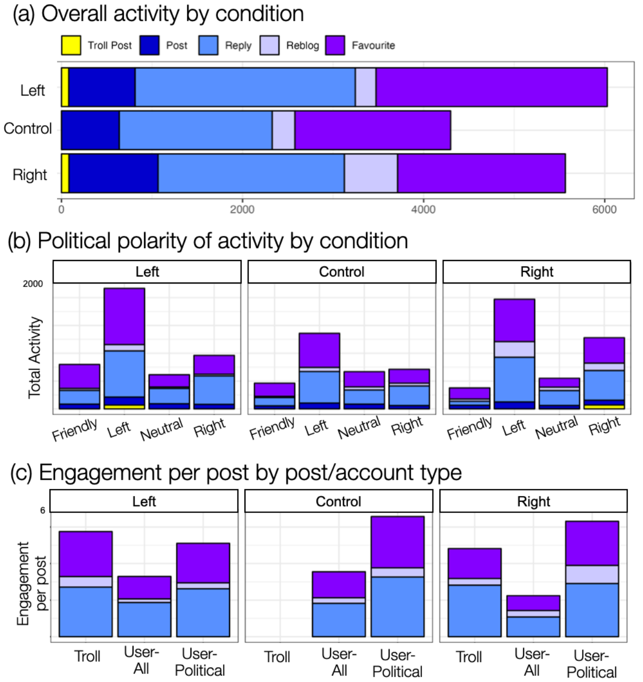

Figure 5: **Breakdown of activity by condition.** (a) There was substantial engagement on Magpie in all conditions, least of all in the CONTROL condition. (b) Posts and replies were classified as Left if the sentiment expressed in them was consistent with left-wing ideas, Right if right-wing, Neutral if the topic was political but there was no direction (e.g, "What do you think about climate change?"), and Friendly if the topic was not political. Favorites and reblogs were given the classification of the post that was being favorited or reblogged. In all conditions, there was somewhat more left-wing content, and far less Friendly content in the Right condition. (c) Troll accounts received more engagement per post than posts generated by people (User-All); however, engagement was similar when compared to just the political posts made by people (User-Political).

Our study allowed us to test whether our troll accounts, which were posting political and often divisive content, led to more overall activity. Figure 5(a) reveals that both the LEFT and RIGHT conditions received considerably more engagement compared to the CONTROL condition, with the LEFT condition showing the most activity overall. Although there was a similar number of original posts in the LEFT condition as the CONTROL condition, there were more replies and favorites in the LEFT. The RIGHT condition contained more posts, replies, and favorites than the CONTROL condition.

Figure 5(b) breaks activity in each condition down by the political direction of each post (reblogs and favorites were classified according to the post they attached to, so a favorite of a left-wing post would count as left-wing). As is evident from the figure, left-wing activity was the most common in all conditions (which makes sense given that our sample was left-leaning), and the amount of engagement was the most balanced in the RIGHT condition. We also identified a category of post that we called Friendly; these were not about any of the conversation topics, but were social interactions between participants ("How's everyone's day going?", "Have a good night", etc.) As Figure 5(b) reveals, the RIGHT condition had strikingly fewer friendly interactions – in absolute

[7] We refer to both original posts and replies as "posts."

[8] A full description of the classifier and the validation procedure can be found in the Supplementary Materials.

[9] The LLM also correctly labeled all of the troll posts, suggesting high accuracy for posts that had an obvious political stance.

numbers, fewer than even the CONTROL condition despite the total engagement in RIGHT being substantially higher.

Why was there more engagement in the two troll conditions? Figure 5(c) shows that on average, troll posts received 43% more engagement than posts by participants (User-All, second bar of each panel). Considering only replies – a more active form of engagement than reblogs and favorites – LEFT trolls received 21% more replies and RIGHT trolls 62% more than participant posts did in their respective conditions. This suggests RIGHT trolls generated content that received more active engagement, perhaps because people in our left-leaning sample were more likely to argue against them (see Rathje et al., 2021).

It is also useful to compare the engagement received by troll posts with the engagement received by *any* political post. Figure 5(c) isolates the political posts contributed by users (User-Political) that have a clear polarity (as opposed to neutral queries like "what does everyone think about X?"). Engagement with these is similar to the engagement with troll posts, suggesting that it is the stance-taking political content that drives activity. Consistent with this, we found a strong positive correlation between strength of political beliefs and biased engagement: people who had stronger left-wing beliefs had a higher proportion of left-wing activity (posting and favoriting more left-wing content), and vice-versa (LEFT: $r = .62$, CONTROL: $r = .46$, RIGHT: $r = .51$, all ps$< .001$).

General Discussion

We observed small but statistically reliable effects suggesting that, relative to the CONTROL condition, exposure to RIGHT-wing trolls in our isolated social media environment resulted in 1) beliefs that were more right-wing on average, 2) a greater perception that the US population was generally politically divided, and 3) a greater perception that people's own political beliefs were more left-wing than the norm. We did not find statistically reliable differences on any of these measures between the CONTROL condition and LEFT.

Although our effect sizes were small compared to those observed in many conventional behavioral experiments, they are noteworthy nevertheless. Our manipulation was extremely small; only 100 posts in each condition were inauthentic, which is less than 5% of the over 2000 posts generated by users. Because the algorithm was reverse chronological, there was no guarantee that everyone even saw them. Moreover, for ethical reasons our trolls did not engage in many kinds of behavior found in real-life influence campaigns – amplifying each other, coordinating messages or attacks, insulting or provoking people in order to spread discord, and so forth. Also, the troll posts were not only spread out over multiple days, but people often responded to them with counter-arguments.

Given these limitations, the fact that there was any effect at all has sobering implications when we consider that in the real world, many platforms (e.g., X) probably have a higher proportion of inauthentic activity and misinformation (Cinus et al., 2024). They also have algorithms that push this activity more strongly (Cinelli et al., 2021), more harmful content, no ethical limitations against coordination or attacking, and repeated, ongoing exposure lasting months or years instead of a few days.

We also found that that the inclusion of trolls (both left-wing and right-wing) substantially increased the amount of activity on the platform. In part, this reflected the fact that the troll accounts only posted content with a clear political stance, which tended to receive more engagement than neutral or friendly posts. However, it is worth noting that participants *themselves* generated more political posts in the RIGHT condition than in the CONTROL, suggesting that there was some spillover from trolls to users that extended beyond engagement on the troll posts themselves. In follow-up analyses we plan to explore this more thoroughly.

Another finding is that engaging on Magpie made people *more* likely to think that people in the US shared consensus views; given the prevailing view that social media experience exacerbates polarization, this came somewhat as a surprise. However, it is perhaps explicable when taken in combination with the fact that most people enjoyed Magpie a great deal; indeed, many explicitly noted that the lack of algorithm, ads, and "mean people" meant that they felt better about polarization and politics than they did before. It is also notable that this effect was smallest in the RIGHT condition, which also was the least enjoyable and contained the smallest amount of friendly exchanges. Future investigation is necessary to determine what features of the platform and messaging have the most effect on people, and how much of the effect occurs indirectly by changing the nature of the discourse.

There are many aspects to our study that require additional scrutiny and follow-up in order to determine both how replicable and robust these findings are, as well as what factors are most determinative. Our participants were anonymous to each other, with no way in their profiles to indicate partisan allegiance; we did this to remove one uncontrolled variable but since people behave differently when anonymous (e.g., Nitschinsk et al., 2022), it is worth noting. Also, in order to provide some conversational guidance and focus our belief measures, we asked people to discuss four specific topics; this was somewhat artificial, especially because they were topics that the participants didn't select themselves.

Although we took care not to incentivize people to post or behave in a certain way, the fact that they were paid for engagement may matter; in the real world, *which* people self-select to engage on social media is no doubt a strong driver of its dynamics. In some ways our paid sample is an advantage: our participant pool was heterogeneous and diverse along many dimensions, and they were given enough freedom in their behavior for us to explore how their individual differences relate to their behavior and the impact of the platform (this, too, is the subject of ongoing investigation). Regardless, our work should be viewed not as "finished science" but rather a preliminary step using a promising approach to studying how people are affected by interaction on social media.

Acknowledgments

Many thanks to Ben Stone, Kelly Kirkland, Morgan Saletta, Ed Hovy, Piers Howe, Yoshi Kashima, Ben Simmonds, and all who helped get MAGPIE off the ground. Manikya Alister was supported by an Australian Government Research Training Program scholarship.

References

Alister, M., Ransom, K. J., & Perfors, A. (2022). Source independence affects argument persuasiveness when the relevance is clear. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44). Retrieved February 1, 2023, from https://escholarship.org/uc/item/5hg4p8cm

Bailo, F., Johns, A., & Rizoiu, M.-A. (2024). Riding information crises: The performance of far-right Twitter users in Australia during the 2019–2020 bushfires and the COVID-19 pandemic. *Information, Communication & Society*, *27*(2), 278–296. https://doi.org/10.1080/1369118X.2023.2205479

Butler, L. H., Lamont, P., Wan, D. L. Y., Prike, T., Nasim, M., Walker, B., Fay, N., & Ecker, U. K. H. (2024). The (Mis)Information Game: A social media simulator. *Behavior Research Methods*, *56*(3), 2376–2397. https://doi.org/10.3758/s13428-023-02153-x

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9), e2023301118. https://doi.org/10.1073/pnas.2023301118

Cinus, F., Minici, M., Luceri, L., & Ferrara, E. (2024, October). Exposing Cross-Platform Coordinated Inauthentic Activity in the Run-Up to the 2024 U.S. Election. https://doi.org/10.48550/arXiv.2410.22716

Cork, A., Everson, R., Levine, M., & Koschate, M. (2020). Using computational techniques to study social influence online. *Group Processes & Intergroup Relations*, *23*(6), 808–826. https://doi.org/10.1177/1368430220937354

Doshi, J., Novacic, I., Fletcher, C., Borges, M., Zhong, E., Marino, M. C., Gan, J., Mager, S., Sprague, D., & Xia, M. (2024, August). Sleeper Social Bots: A new generation of AI disinformation bots are already a political threat. https://doi.org/10.48550/arXiv.2408.12603

Jagayat, A., & Choma, B. L. (2024). A primer on open-source, experimental social media simulation software: Opportunities for misinformation research and beyond. *Current Opinion in Psychology*, *55*, 101726. https://doi.org/10.1016/j.copsyc.2023.101726

Lewandowsky, S., Gignac, G. E., & Vaughan, S. (2013). The pivotal role of perceived scientific consensus in acceptance of science. *Nature Climate Change*, *3*(4), 399–404. https://doi.org/10.1038/nclimate1720

Meta. (2024, August). Taking Action Against Malicious Accounts in Iran. Retrieved January 22, 2025, from https://about.fb.com/news/2024/08/taking-action-against-malicious-accounts-in-iran/

Microsoft Threat Analysis Center. (2024). *Russia, Iran, and China continue influence campaigns in final weeks before Election Day 2024* (tech. rep. No. 5). https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/MTAC-Election-Report-5-on-Russian-Influence.pdf

Nitschinsk, L., Tobin, S. J., & Vanman, E. J. (2022). The Disinhibiting Effects of Anonymity Increase Online Trolling. *Cyberpsychology, Behavior and Social Networking*, *25*(6), 377–383. https://doi.org/10.1089/cyber.2022.0005

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, *31*(7), 770–780. https://doi.org/10.1177/0956797620939054

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. Retrieved January 22, 2025, from https://www.sciencedirect.com/science/article/pii/S001002771830163X

Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26), e2024292118. https://doi.org/10.1073/pnas.2024292118

Simmonds, B. P., Ransom, K. J., & Stephens, R. (2024). Navigating Health Claims on Social Media: Reasoning from Consensus Quantity and Expertise. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*(0). Retrieved October 28, 2024, from https://escholarship.org/uc/item/76j8m25k

US Department of Defense. (2023). *Strategy for Operations in the Information Environment* (tech. rep.).

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559