nature human behaviour

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41562-019-0541-6

In the format provided by the authors and unedited.

The wisdom of polarized crowds

Feng Shi[1,2,7], Misha Teplitskiy[2,3,7*], Eamon Duede[2,4] and James A. Evans[2,5,6*]

[1]Odum Institute for Research in Social Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [2]Knowledge Lab, University of Chicago, Chicago, IL, USA. [3]Laboratory for Innovation Science, Harvard University, Boston, MA, USA. [4]Committee on the Conceptual and Historical Studies of Science, University of Chicago, Chicago, IL, USA. [5]Department of Sociology, University of Chicago, Chicago, IL, USA. [6]Santa Fe Institute, Santa Fe, NM, USA. [7]These authors contributed equally: Feng Shi, Misha Teplitskiy. *e-mail: mteplitskiy@fas.harvard.edu; jevans@uchicago.edu

Supplementary Tables

Descriptive Statistics of the Dataset

Summary statistics for the three corpora—Politics, Social Issues, and Science articles—are shown in Supplementary Table 1. The distribution of quality of each article, estimated with `wikiclass`, is shown in Supplementary Table 2. Note that a few articles have no text (e.g., removed or redirected) and hence receive no quality ratings.

Supplementary Table 1: Summary statistics of Wikipedia data sets. Article length, # Edits per article, and # Editors per article refer to the averages over all articles in the corresponding corpus. Article length is measured by bytes. Numbers in parentheses are standard deviations.

Corpus	# Articles	Article length	# Edits per article	# Editors per article
Politics				
Conservative	10,909	9,449 (15,013)	177 (808)	80 (294)
Liberal	10,038	8,645 (14,280)	155 (686)	75 (282)
Social Issues	162,085	13,153 (20,847)	265 (819)	122 (337)
Science	49,995	11,193 (17,297)	210 (632)	103 (284)
All	233,027	-	-	-

Supplementary Table 2: Distribution of article quality in each corpus. Number of articles in each quality rating for Politics, Science, and Social Issues pages. The algorithm used to measure article quality is described in Methods: Measurement of page quality in the main text.

Corpus	Quality rating					
	Stub	Start	C	B	Good Article	Featured Article
Politics	2950	5009	3541	485	871	215
Social Issues	30853	55884	48292	14050	10108	3814
Science	12192	16899	12454	5323	1696	966

Polarization and Article Quality

Average polarization score for teams in each article quality level is shown in Supplementary Table 3.

Supplementary Table 3: Average polarization score of teams in each article quality level (Stub=lowest, FA=highest) for Politics, Social Issues, and Science articles. Numbers in parentheses are 95% confidence intervals around the means estimated from 1000 bootstrap iterations.

	Stub	Start	C
Politics	.042 (.040, .043)	.061 (.060, .062)	.084 (.083, .086)
Social Issues	.066 (.065, .067)	.067 (.067, .068)	.078 (.078, .079)
Science	.060 (.059, .061)	.068 (.068, .069)	.077 (.077, .078)
	B	GA	FA
Politics	.102 (.099, .106)	.097 (.095, .100)	.102 (.098, .106)
Social Issues	.087 (.086, .087)	.084 (.083, .085)	.093 (.092, .094)
Science	.083 (.082, .084)	.077 (.075, .079)	.090 (.088, .093)

Descriptive Statistics of the Variables

Summary statistics of the variables over all pages under study are shown in Supplementary Table 4.

Supplementary Table 4: Summary statistics of variables used in the statistical analysis.

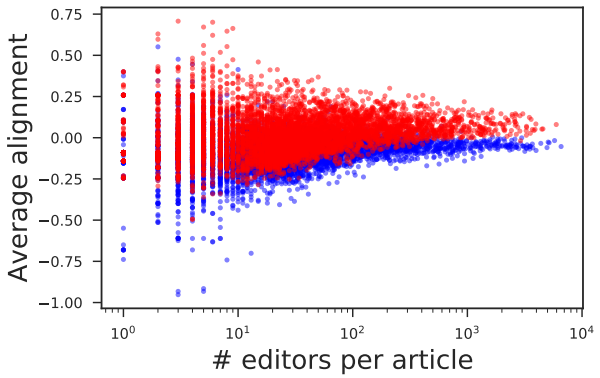
		N	Min	Mean	Median	Max	Std
	Average Align.	205744	-0.962	-0.006	-0.006	0.966	0.103
	Polarization	205744	0	0.074	0.071	0.822	0.048
	Pre Edits	205744	42	496161	399773	5921042	401623
	# Edits	205744	1	42	6	45068	368
	# Editors	205744	1	14	5	6357	52
Article	Length	205744	2	4943	623	468866	13238
	Lexical Div.	205744	50	5050	2436	446390	8171
	Semantic Div.	205744	0.557	3.737	3.74	5.139	0.297
	# Edits	205744	1	267	65	36613	809
	# Editors	205744	1	124	34	13305	336
	Length	205744	21	13088	6464	538466	20315
Talk Page	Lexical Div.	205744	0.799	2162	237	217609	5940
	Semantic Div.	205744	0	3.556	3.595	5.203	0.408
	Temperature	205744	0	0.006	0.004	0.991	0.0196
	Policy	19470	1	3.434	2	86	4.636
	Guideline	20398	1	3.01	2	101	4.108

Supplementary Methods 1

Computational Ideological Alignment Measure

Many Wikipedia editors carry out general copy editing and curatorial tasks that contribute very few bytes to articles, which leads us to estimate an editor's ideological alignment, p, through a conservative, Bayesian framework. Our approach is designed to down-weight data from these numerous curatorial editors to avoid introducing uncertainty in our overall estimation of ideological preference. We use a neutral prior that assumes every editor behaves randomly to reduce small-sample effects. Mathematically, p is assumed to have a prior distribution $P(p) = Beta(a, b)$ before observing any data, such that everyone is assumed to randomly contribute to red or blue articles. Next, the distribution is updated by Bayes' law with observations X (bytes contributed to conservative articles) and K (bytes contributed to political articles): $p|X \sim Beta(a + X, b + K - X)$. Finally, ideological alignment is defined as the posterior mean of p: $E[p|X] = (X + a)/(K + a + b)$.

For presentation purposes, we rescale alignment scores linearly onto [-1,1] with 0 as the neutral point. Thus, ideological alignment is a scalar between -1 and 1. Casual editors with few contributions will be close to neutral (alignment ≈ 0). Editors with many, consistent contributions to liberal articles will be close to the liberal pole of our scale (alignment ≈ -1), and those with intensive, exclusive contributions to conservative articles will be close to conservative pole (alignment ≈ 1). This alignment measure allows us to quantify the political perspective each editor brings to an editing team or community and, in turn, an edited article. For example, an average alignment score close to 0 for a group of editors suggests balanced participation from both conservatives and liberals in the group.



Supplementary Figure 1: Average alignment of editors as a function of the number of editors for each conservative (red) or liberal (blue) article. Some of the red points obscure the overlapping blue points "underneath."

Supplementary Methods 2

Recursive Ideological Measure

Our alignment measure assumes that conservative (liberal) editors will be especially attracted to conservative (liberal) articles. Figure 1A in the main text confirms these editor-level tendencies to cluster their edits within political articles of a given ideology. To evaluate the consistency of this assumption, we investigated how editors invest their edits across pages. Supplementary Figure 1 plots average team alignment as a function of team size for each conservative (red) or liberal (blue) article. As expected, conservative (liberal) articles have a much larger chance of attracting conservative (liberal) editors than the converse.

We also find that articles attracting more attention tend to have more balanced engagement. Editors' choices of what to edit are influenced by many other article attributes, including their broader cultural significance. Most articles on Wikipedia are edited by 10-100 people and so those edited by multitudes stand out for a reason. They involve topics, events and personalities of general interest, who have typically influenced or touched many people. Although we do not measure specifically what makes popular pages popular, popular pages are more likely to attract individuals across the ideological spectrum, making them less informative indicators of ideology. For example, although Ronald Reagan can be classified as a conservative figure ideologically, he is empirically a broad/neutral topic because his significance attracts people across the spectrum. To investigate how sensitive our findings are to the *dichotomous* coding of political articles, we extend the Bayesian framework to measure the alignment of political articles, while at the same time recursively measuring the alignment of editors. Under this continuous measure of article alignment, popular articles with balanced attention from both sides will tend to have alignment scores close to 0 (neutral) and contribute little information to the estimation of editor alignment.

Specifically, an editor's alignment is defined formally the same as above, but the number of bytes contributed to conservative articles (X) and the number of bytes contributed to political articles (K) are calculated differently. Previously, one byte contributed to a conservative article counted as one byte; now we weight this byte by the alignment of the article to which the byte was contributed. This way, if a conservative article has an alignment score close to zero, then bytes contributed to this article count little towards the "conservative bytes" (X) of the contributing editor. The number of bytes contributed to liberal articles are similarly weighted by the alignment scores of the articles. Simultaneously, we estimate a Bayesian model that defines the alignment of political articles. For each political article, we define its alignment as its probability of attracting conservative editors

(u) and model the number of conservative editors Y editing the article as a random variable with the binomial distribution: $Y \sim Bin(E, u)$ where E is the total number of editors on this article. Wikipedia's classification of an article (liberal or conservative) is used as this model's prior such that the alignment of an article has a prior distribution $Beta(a, 0)$ if it is a conservative article and $Beta(0, a)$ if liberal. Having observed the editors of an article, the distribution of u is updated using Bayes rule, then the alignment score of this article is calculated as $(Y + \delta(u)a)/(E + a)$, where $\delta(u) = 1$ if u is initially coded as a conservative article and $\delta(u) = 0$ otherwise, encoding our prior information about the article's alignment.

Putting models for editor alignment and article alignment together, we arrive at a recursive definition of political alignment: the alignment of an editor depends upon the alignment of articles she edits, and the alignment of an article in return hinges on the alignment of editors who contribute to the article. In this way, we calculate the editors' alignment scores iteratively according to the following algorithm. First, each political article is assigned an alignment score of either -1 (liberal) or 1 (conservative) according to Wikipedia's classification. Second, we calculate editors' alignments using the new Bayesian model above (with the scores rescaled to fall in [-1, 1]). Third, we recalculate the alignment scores of political articles following the new Bayesian model (again with the scores rescaled to fall in [-1. 1]), and repeat. After 10 iterations, the distribution of alignment scores appears stable and we stop at 20 iterations. We find that the alignments of popular political articles approach 0, as expected. For example, the alignment for the page "Ronald Reagan" is initially 1 because Ronald Reagan is a conservative figure and his page is categorized under conservative pages by Wikipedia. In the end, its alignment score drops to 0.014025 because it attracts a large number of liberal and conservative editors.

We then used this new set of editor alignment scores to measure team polarization and predict article quality. We estimated an ordinal logistic regression model at the article level with article quality as the outcome and team polarization as the main predictor. Regression results are provided in Supplementary Table 5. See section "Effects on Quality" in the main text for results using the original set of alignment scores. Comparing the results here with Table 1 in the main text, we find that team polarization is still positively and statistically significantly associated with article quality. Moreover, polarization has the largest effect in the Politics corpus, followed by Social Issues and then by Science, consistent with our main findings in Table 1. Hence, our results are robust to the assumption that the political articles are dichotomous (either liberal or conservative). Because of the simplicity of the dichotomous assumption and its associated model, we use the original measure in all analyses unless otherwise stated, and present findings with the original measure in the main text.

Supplementary Table 5: Odds ratios from ordinal logistic regression models predicting article quality. The alignment scores are calculated with the recursive model that estimates alignments for editors and pages simultaneously.

	Dependent variable: article quality					
Independent variable	*Politics*	*p*	*Social Issues*	*p*	*Science*	*p*
polarization	47.65	$< .001$	3.47	$< .001$	2.85	$< .001$
| alignment |	1.20	.33	1.06	.29	0.97	.82
editing experience	1.04	.05	1.06	$< .001$	1.01	.24
number of editors	0.41	$< .001$	0.52	$< .001$	0.56	$< .001$
article length	33.73	$< .001$	47.85	$< .001$	56.35	$< .001$
number of edits	3.24	$< .001$	1.69	$< .001$	1.67	$< .001$
N	12,570		161,070		49,995	

Measuring polarization of social issues and science teams

Our measurement of editors' ideology derives from their activity in the politics corpus. Measuring the ideological composition of teams editing politics is thus straightforward, as the ideology of each editor is known[1]. However, in the social issues and science corpora, many editors never edit any political articles, and their ideologies are thus unobserved. In particular, while we measure ideology for 26.0% of all politics editors[2], we measure ideology of only 1.5% and 4.3% of all editors in social issues and science, respectively. The ideological composition of teams editing social issues and science is thus measured with more noise than teams editing politics.

In regression analyses, higher noise in measurement may result in attenuation bias, such that coefficients estimated on the noisy data are lower than their true values [1, Section 4.7.5]. To test for the possibility of differing amounts of attenuation bias across corpora, we re-estimated the results presented in Table 1 from the main text using only the subset of editors who edited both politics *and* science. This ensures that the amount of measurement error is comparable across corpora. The change in coefficients for polarization is negligible (less than 4%).

Supplementary Methods 3

Significance of Ideological Polarization

To statistically test whether the observed variance of ideological alignments is greater than expected from chance, we simulate editors who have a given "budget" of edits and choose to allocate them to liberal or conservative articles at random. In every simulation, each editor contributes her actual edits to either liberal or conservative articles with a probability proportional to the total size of each set of articles. From 100 such simulations, we construct a distribution of the variance of editor alignments (Supplementary Figure 2). We find that the empirical probability of observing a variance larger than 0.04 in the 100 random simulations is 0; hence, the bootstrap $p < .001$ no matter it is 1- or 2-tailed. The observed variance of ideological alignments (0.04) of real Wikipedia editors is statistically significantly larger than the variance of alignments from simulated editors.

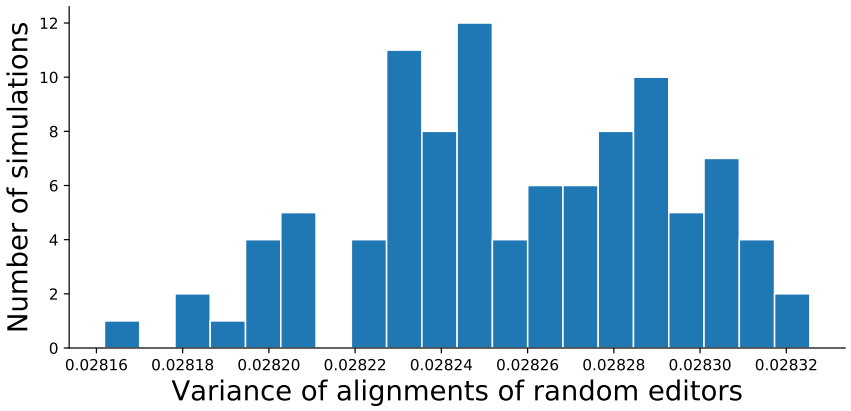
Supplementary Methods 4

Lexical and Semantic Diversity

We measured the lexical diversity of a page by the sum of `tf-idf` scores for distinct words on the page, normalized by page length (i.e., word frequency divided by logged document frequency, normalized by page length in bytes). The tf-idf score down-weights common words that appear on many talk pages and up-weights words that distinguish those pages. To measure semantic diversity, we first projected each word onto a high-dimensional space inscribed by all English language Wikipedia articles, preserving distances between words that reproduce the differences between all their contexts across all articles. This was done by training a popular text processing tool, `fastText`, on all English Wikipedia articles [2] using the skip-gram model, estimated with negative sampling and default parameters (e.g., word window=5, character-level n-grams=3-6, etc.) The latent space approach and associated tools have rapidly become a staple of computational semantics [3]: this particular embedding posts a 72% correlation with surveyed word associations [2, 4]. We assessed

[1]Note that we exclude from analysis unregistered users, whose edits Wikipedia tracks only through their IP addresses.

[2]This represents 100% of the registered users.



Supplementary Figure 2: Distribution of the variance of editor alignments from random simulations. There are 605,359 editors included in the simulations. In every simulation, each editor contributes her actual edits to either liberal or conservative articles with a probability proportional to the total size of each set of articles. We then calculate the alignments of those random editors and the variance of their alignments. This process is repeated 100 times to construct a distribution of variance in alignment.

the semantic diversity of a page by calculating the average distance between each word and the centroid of that page in the semantic embedding space, such that topically narrow talk pages, which discuss fewer Wikipedia issues, post smaller average distances, while topically broad talk pages, which discuss more Wikipedia issues, post larger average distances.

Supplementary Methods 5

Bivariate Correlation

We explored relationships between the variables considered in the study by calculating the Pearson correlation between every pair; the table of pairwise correlations is shown in Supplementary Figure 3.

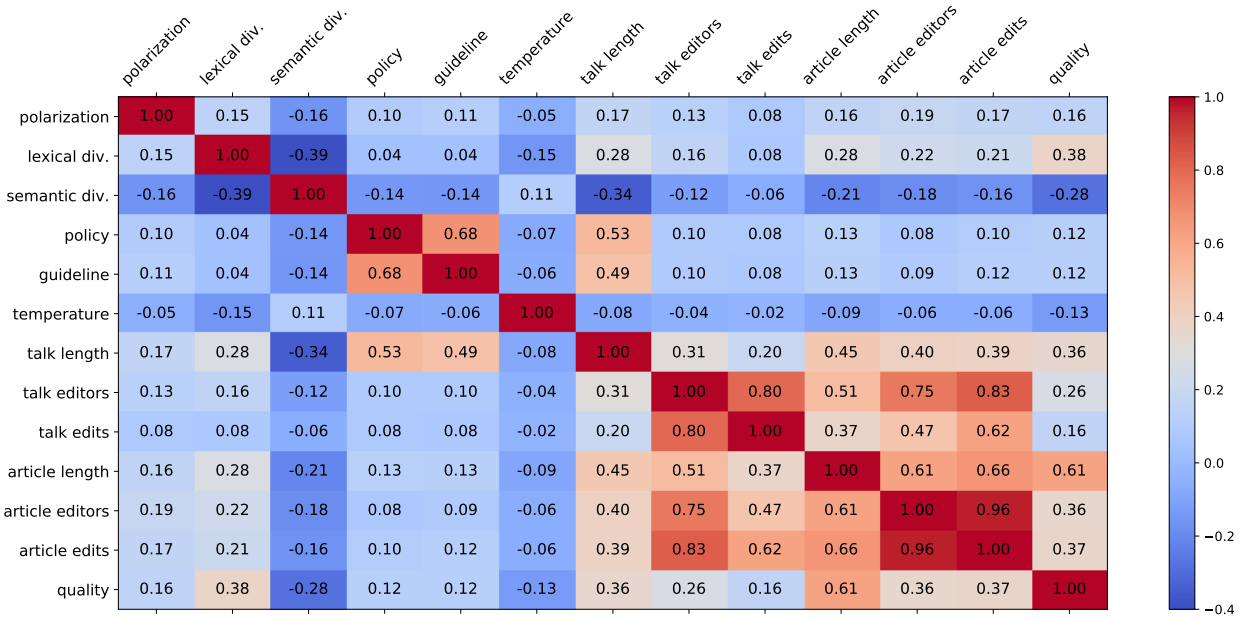
From the table we can see that polarization (i.e., variance of ideological alignments) is positively correlated with quality[3] ($\rho = 0.15, t(205742) = 68.82, p < .001, CI = [0.146, 0.154]$). Besides, polarization is correlated with all other variables in directions consistent with hypothesized mechanisms of collaboration:

- positively with lexical diversity ($\rho = 0.15, t(205742) = 68.82, p < .001, CI = [0.146, 0.154]$), and negatively with semantic diversity ($\rho = -0.16, t(205742) = -73.52, p < .001, CI = [-0.164, -0.156]$), suggesting that polarization will focus debates on fewer contested, ideologically relevant topics, but frame them in competing ways.

- positively with all talk page and article activities such as talk page length ($\rho = 0.16, t(205742) = 73.52, p < .001, CI = [0.156, 0.164]$), number of edits ($\rho = 0.08, t(205742) = 36.40, p < .001, CI = [0.076, 0.084]$), and number of editors ($\rho = 0.13, t(205742) = 59.47, p < .001, CI = [0.126, 0.134]$), suggesting that polarization increases debate volume.

- negatively with talk page temperature ($\rho = -0.05, t(205742) = -22.71, p < .001, CI = [-0.054, -0.046,]$), suggesting that polarization, resulting from balanced engagement, polices and decreases emotional aggression and violent debate.

[3]Here quality is treated as an integer between 1 and 6.



Supplementary Figure 3: Pearson correlation between every pair of variables, with color denoting the value of correlation. The correlations are calculated from 205,744 pages except for correlations involving policy (19470 pages) or guideline (20398 pages) mentions because not every page contains such mentions.

- positively with policy ($\rho = 0.10, t(205742) = 45.59, p < .001, CI = [0.096, 0.104]$) and guideline mentions ($\rho = 0.11, t(205742) = 50.20, p < .001, CI = [0.106, 0.114]$), suggesting increased use of Wikipedia institutions among polarized teams.

The table also reveals that most variables are substantially correlated with each other. Correlation between polarization and other variables could be caused by confounding factors. Therefore, we carried out two further statistical analyses that estimate the conditional effects of polarization, holding other confounding variables constant. First, we estimated several individual regressions, linking polarization to quality through all proposed mechanisms. Then we estimated a structural equation model, which evaluates all effects simultaneously, enabling us to compare their relative strength.

Supplementary Methods 6

Regression Analysis

To assess mechanisms of polarized collaboration, we estimated multiple linear regression models with polarization as the main predictor. Specifically, we tested how polarization (i.e., variance of alignments) affected, respectively, talk page volume (i.e., number of talk page edits, talk page length), semantic and lexical diversity, policy and guideline mentions, and talk page "temperature", controlling for Wikipedia talk and article page features. Models and regression results are shown in Supplementary Table 6. The models are estimated on the three corpora - politics, social issues, and science - pooled together.

All models control for number of talk page editors to make sure that the effects are not simply caused by the number of people involved. Talk and article page lengths are also accounted for whenever possible, so the effects of polarization cannot exclusively be due to the sheer number of edits and words. Lexical and semantic diversity are normalized by page length so it is not necessary

to include page length in those models; instead, we control for lexical and semantic diversities of the articles to account for heterogeneous article content.

Note that we do not include every variable in every model because of substantial collinearity between some variables (see the correlation table above). Simpler models are more interpretable and preferable from a statistical point of view. For example, number of editors, number of edits and page length are highly correlated with each other (Supplementary Figure 3); when all are included in a single model, it is hard to explain the direction of their effects (e.g., in the 2nd model in Supplementary Table 6, number of editors shows a negative effect on page length when number of edits is present). These encourage us to create factors from highly correlated and conceptually similar variables. We did this, and simultaneously evaluated the combined impact of polarization on article quality through estimation of a structural equation model as described below.

Supplementary Table 6: Regression results from 7 models that predict characteristics of Wikipedia talk page deliberation with polarization, controlling for relevant talk and article page features. All variable coefficients shown in the table are statistically significant at the $p < 0.001$ level. Some dependent and independent variables are log-transformed so that their distributions approximate Normality.

	Dependent variable						
Independent var	*# talk edits*	*talk page length*	*lexical diversity*	*semantic diversity*	*temperature*	*policy*	*guideline*
polarization	0.10	1.28	1.43	-0.44	-0.17	0.06	0.09
# talk editors	1.03	-0.09	-0.09	0.15	0.25	-0.28	-0.25
# talk edits		1.13	1.39	-0.21	-0.03	0.29	0.26
talk page length	0.09				-0.21	0.05	0.05
article length	0.05	0.24			-0.09	-0.01	-0.004
article lexical d.			0.10				
article semantic d.				0.34			
R^2	0.95	0.71	0.68	0.21	0.17	0.27	0.26

Supplementary Methods 7

Structure Equation Modeling

The structural equation model we designed is detailed in Figure 4 of the main text, and estimated by the R package "lavaan" [5]. Detailed specifications and estimation results are as follows.

- Latent variables: debate volume, institution, article activity.

$$\begin{aligned} \text{debate volume} &\sim \text{number of talk edits} + 1.15\ \text{talk page length} \\ \text{institution} &\sim \text{policy mentions} + 0.96\ \text{guideline mentions} \\ \text{article activity} &\sim \text{number of article editors} + 2.1\ \text{article length} + 1.48\ \text{number of article edits} \end{aligned}$$

- Regressions

$$\begin{aligned} \text{quality} &\sim -1.19\ |\text{average alignment}| - 0.05\ \text{editing experience} + \\ &\quad 2.14\ \text{article activity} + 0.33\ \text{debate volume} + \\ &\quad 0.50\ \text{lexical diversity} - 0.35\ \text{semantic diversity} + \\ &\quad 0.18\ \text{institution} - 0.23\ \text{debate temperature} \\ \text{debate volume} &\sim 0.38\ \text{polarization} + 1.19\ \text{number of talk editors} \end{aligned}$$

$$
\begin{aligned}
\text{lexical diversity} &\sim 0.20\ \text{polarization} + 0.17\ \text{number of talk editors} \\
\text{semantic diversity} &\sim -0.65\ \text{polarization} - 0.1\ \text{number of talk editors} \\
\text{institution} &\sim 0.20\ \text{polarization} + 0.14\ \text{number of talk editors} \\
\text{debate temperature} &\sim -0.60\ \text{polarization} - 0.12\ \text{number of talk editors}
\end{aligned}
$$

- Effects of polarization on quality through the following paths:

$$
\begin{aligned}
&\text{polarization} \rightarrow \text{debate volume} \rightarrow \text{quality} : 0.125 \\
&\text{polarization} \rightarrow \text{lexical diversity} \rightarrow \text{quality} : 0.098 \\
&\text{polarization} \rightarrow \text{semantic diversity} \rightarrow \text{quality} : 0.230 \\
&\text{polarization} \rightarrow \text{institution} \rightarrow \text{quality} : 0.035 \\
&\text{polarization} \rightarrow \text{talk temperature} \rightarrow \text{quality} : 0.136
\end{aligned}
$$

- Combined effect of polarization on quality through all the paths: 0.624.

The model is well specified and the estimation procedure converged quickly (125 iterations). All estimated parameters in the model are significant at $p < 0.001$, agreeing with the other statistical analyses.

Because the sample size is very large (205,749 observations), a χ^2 test cannot be used to evaluate the fitness of this model. (In fact, the p-value of a χ^2 is approximately 0.) The CFI and NNFI indexes for the model are 0.78 and 0.71, respectively. We do not expect the indexes to be very high because the model is designed to test the effects of polarization through various mechanisms rather than fitting all covariances in the data. For example, the correlation table reveals that article activity is highly correlated with debate volume. If we add a regression of article activity on debate volume to the current model, the CFI index boosts to 0.89, but the interpretation becomes less clear as a new path is introduced through article activity.

Supplementary Discussion 1

Survey Measures of Offline Preferences

Survey 1 (n = 118) was distributed in November 2017 and focused on political party identification. Survey 2 was distributed in April 2018 and focused on ideological self-identity. In Survey 1 we sampled Wikipedia editors having made recent edits (< 1 month) to at least 1 article in the Politics and/or Science corpus of articles, while Survey 2 sampled editors with at least 1 recent edit in the Politics corpus. Wikipedia required that surveys be individualized, solicitations made personally on editors' talk pages, not as a batch, and that we engage all individuals solicited in conversation regarding any questions or concerns. This limited the number of surveys that could be performed. In sum, in Survey 1 we sent out surveys to 500 editors and received 118 responses; in Survey 2 (n = 100) we sent out surveys to 327 editors and received 100 responses. Participants were shown a consent script prior to any questions. All questions on the survey were optional. The surveys presented no more than minimal risk of harm to subjects, and involved no procedures for which written consent is normally required outside of a research context. The surveys' methods were approved by the University of Chicago's Institutional Review Board (IRB17-0679).

Survey 1 – political party

Survey 1 first asked whether the respondent resides in the United States. Those responding in the affirmative were then asked "*Do you generally think of yourself as a Republican, Democrat, Independent or something else?*". The (mutually exclusive) answer choices were 1=*Strong Democrat*, 2=*Not Strong Democrat*, 3=*Independent, Near Democrat*, 4=*Independent*, 5=*Independent, Near Republican*, 6=*Not Strong Republican*, 7=*Strong Republican*, 8=*Other (Please explain)*. Although the computational measure ranges from "Liberal" to "Conservative," the survey question focused instead on specific political party affiliation for concreteness and its long-standing use in survey research. Some respondents chose to instead write-in a political party and, in some cases, mentioned being registered as either Republican or Democrat.

54% of respondents (64/118) reported living outside of the United States. These respondents were then asked "Which political party, if any, do you generally identify with?" Some respondents provided parties that they themselves compared with U.S. Democratic or Republican parties, while Other responses could not be unambiguously aligned with Democratic or Republican parties. Hence, only responses from US-based editors were included in the calculations.

Supplementary Table 7 assesses the relationship between computational ideological alignment and self-reported political identification, where different subsets of survey responses are used. Respondents choosing "Independent" were not used because "Independent" was assumed to reflect ideological alignments that do not clearly align on a Conservative-Liberal spectrum.

We first calculated the Pearson correlation coefficient between computational ideological alignment and self-reported political identification. Because the sample size is small and responses are not quite Gaussian, we conducted a permutation test to assess the p-value of the correlation, which is the probability that one would by chance observe a value larger than the calculated correlation. Specifically, we randomly permuted the computational ideological alignments of the respondents 10000 times, and for each permutation, the Pearson correlation coefficient between permuted ideological alignments and self-reported political identifications was calculated. Then the fraction of correlation coefficients larger than the observed correlation among all permutations is used as the p-value of the observed correlation. The permutation test was 1-sided because we have a clear directional hypothesis that the correlation is positive.

Because each individual might have a subjective scale of alignment, the linear correlation between computational alignments and self-reported alignments pooled together may underestimate their association. Hence, we built a logistic regression model using computational alignment to predict a binary classification of self-reported alignment (Democrat or Republican). Predictive performance of the model is measured by AUC, which is the chance that the model makes a correct classification when given a random Democrat and a random Republican from the dataset. A random classifier would achieve an AUC of 0.5 and a perfect classifier one of 1.0. We further adopted a leave-one-out validation framework to prevent bias from influential data points and to ensure the robustness of our results. In particular, we fit the model to N subsets of responses where the i-th subset includes all responses but the i-th one. For each fitted model, we calculated its AUC score and reported the average and standard deviation of AUC scores for all models in Supplementary Table 7.

Comparison (a) is strictest, using raw responses from US-based editors alone. The correlation, based on n=21 responses, is 0.31 (p=0.083 from a 1-sided permutation test), and the AUC score is 0.72 (with a standard deviation of 0.05). Comparison (b) adds to (a) responses from US-based editors that selected "Other" for political party identification and whose provided comments could be unambiguously recoded to the either "Near Republican" or "Near Democrat". Examples from the 7 recoded responses include "In between Libertarian and Republican"→*Near Republican* and "Social Democrat"→*Near Democrat*. The correlation is 0.35 (n=28, p=0.035 from a 1-sided permutation test), and AUC is 0.71 (with a standard deviation of 0.02).

Supplementary Table 7: Associations between computational ideological alignment and self-reported political identification. Editors' locations are self-reported. Row (a) uses raw survey responses (no recoding). Row (b) adds 7 "*Other (Please explain)*" responses recoded to either "*Independent, Near Democrat*" or "*Independent, Near Republican.*" Correlations are Pearson correlation coefficients with p-values shown in the parentheses beside them. Because the sample size is small and the responses are not quite Normal, we conducted a permutation test to assess the p-value, which is the probability that one would by chance observe a value larger than the calculated correlation. The AUC evaluates the predictive performance of our computational alignment measure in predicting a binary version of the self-reported political identification (Democrat or Republican) using a Logistic regression model. Specifically, AUC is the probability that the model makes a correct classification when it is given a random Democrat and a random Republican from the dataset. The number in the parentheses besides the AUC score is the standard deviation of all AUC's from a leave-one-out validation.

	Editors' location	"Other party" re-coding	# Responses	Correlation	AUC
(a)	US-based	N	21	0.31 (p=0.083)	0.72 (SD=0.05)
(b)	US-based	Y	28	0.35 (p=0.035)	0.71 (SD=0.02)

These comparisons show that our computational measure of ideological alignment correlates moderately well ($\rho = 0.31$ and $\rho = 0.35$) with self-reported political identification, and that it is significantly predictive of political affiliation (AUC=0.7), suggesting that it is not only an online behavior measure, but also a noisy but valid indicator of editors' offline ideological preferences.

Survey 2 – Ideological self-identity

Survey 2 was structured similarly, except that instead of political party identification, respondents were asked, "*Where would you place yourself on this USA based ideological scale?*". The mutually exclusive answer choices were 1=*Extremely liberal*, 2=*Liberal*, 3=*Somewhat liberal*, 4=*Moderate*, 5=*Somewhat conservative*, 6=*Conservative*, 7=*Extremely conservative*, and *Not Applicable*. Because Survey 2 was fielded more than a year and a half after the beginning of the project, we recalculated (computational) ideological alignments using a Wikipedia database dump from 4/1/2018 which is more contemporaneous with the Survey.

We posted 327 solicitations and received 100 responses. 59 (60%) of the respondents reported living outside the USA and were dropped from the analysis. Supplementary Table 8 reports the correlation between computational and self-reported ideological alignments for US-based editors, and the predictive power of the computational alignment measure (AUC), as in Supplementary Table 7.

Results from Survey 2 show slightly weaker associations between online ideological alignment and ideological self-identity. Estimated correlations is 0.25 for US-based editors and the AUC is 0.65. Several considerations may affect these results. First, only 5 individuals who were labeled as solidly conservative by our computational measure ($alignment > 0.4$) responded to the survey, whereas 12 solidly liberal individuals responded. Similarly for self-identification, 7 individuals self-identified as extremely liberal while no one self-identified as extremely conservative. Consequently, the resulting data sample was unbalanced. Second, Survey 2 was fielded after the inauguration of Donald Trump as President, which may have affected popular definitions of liberal and conservative, individuals' self-identities, or willingness to report them in surveys. Lastly, ideological self-identity may be constructed in reference to one's peers rather than being anchored along a country-level spectrum, resulting in increased measurement error.

Supplementary Table 8: Associations between computational and self-reported ideological alignments. Editors' locations are self-reported. The correlation is the Pearson correlation coefficient with p-value shown in the parentheses. AUC evaluates the predictive performance of the computational alignment measure in predicting a binary version of the self-reported ideological identification (liberal or conservative) using a Logistic regression model. Specifically, AUC is the probability that the model makes a correct classification when it is given a random liberal and a random conservative from the dataset. The number in parentheses beside the AUC score is the standard deviation of all AUC's from a leave-one-out validation.

Editors' location	# Responses	Correlation	AUC
US-based	41	0.25 (p=0.06)	0.65 (SD=0.018)

Nevertheless, both surveys provide tentative evidence that online editing preferences are directly correlated with, and significantly predictive of, offline party affiliations and ideological preferences. Further research is necessary to establish these relationships with greater certainty.

Supplementary Discussion 2

Effects of Page Protections

In cases where edits made to a Wikipedia article are regarded as damaging, such as those involved in edit wars, Wikipedia administrators have the ability to place restrictions on who is allowed to edit the page. Such restrictions are called protections. The most commonly used protections are full-protection, where only administrators can edit, and semi-protection, where only administrators and confirmed editors are allowed. Detailed information on protections can be found at the page https://en.wikipedia.org/wiki/Wikipedia:Protection_policy.

By contacting Wikipedia staff and other researchers, we obtained records of which articles were protected by Wikipedia administrators at each time point in Wikipedia history. Then we constructed two relevant dummy variables and added them to our ordered logistic regression model for all articles. The first indicates whether an article has ever been fully protected through the time of our data snapshot and the other indicates semi-protection analogously. The coefficient for full protection is -0.31 ($p < .001$, 2-tailed Wald test), and that for semi-protection is -0.58 ($p < .001$, 2-tailed Wald test). Both effects are statistically significant, but do not weaken the relationships between polarization and quality when protections were not included (the coefficients for polarization are 0.87 ($p < .001$, 2-tailed Wald test) and 0.86 ($p < .001$, 2-tailed Wald test) with and without protections included, respectively). These coefficients suggest that when norms and policies break down, and supervisors enforce editorial bans on violations by excluding the diverse crowd of editors, the quality of those pages is substantially and significantly impacted.

To assess the effect of protections on collaboration mechanisms, we added a dummy variable indicating whether or not an article had ever been protected (full or semi) through the time of our data snapshot, to the structural equation model. We found that polarization maintains a positive effect on quality through protection, while changes to other coefficients of the model are negligible. Details follow:

$$\text{protection} \sim 0.11\ \text{polarization} + 0.075\ \text{number of talk editors}$$

$$\text{polarization} \rightarrow \text{protection} \rightarrow\ \text{quality} : 0.012$$

Supplementary References

[1] W. H. Greene, Econometric Analysis (7th Edition), Pearson, 2011.

[2] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, arXiv preprint arXiv:1607.04606 (2016).

[3] D. Jurafsky, J. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall Series in Artifi, Pearson Prentice Hall, 2009.

[4] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, E. Ruppin, Placing search in context: The concept revisited, in: Proceedings of the 10th International Conference on World Wide Web, WWW '01, ACM, New York, NY, USA, 2001, pp. 406–414.

[5] Y. Rosseel, lavaan: An r package for structural equation modeling, Journal of Statistical Software, Articles 48 (2012) 1–36.