The language of opinion change on social media under the lens of communicative action

Corrado Monti[1], **Luca Maria Aiello**[2,3,*], **Gianmarco De Francisci Morales**[1], **and Francesco Bonchi**[1,4]

[1]CENTAI, Torino, Italy
[2]IT University of Copenhagen, Copenhagen, Denmark
[3]Pioneer Centre for AI, Copenhagen, Denmark
[4]Eurecat, Barcelona, Spain

ABSTRACT

Which messages are more effective at inducing a change of opinion in the listener? We approach this question within the frame of Habermas' theory of communicative action, which posits that the *illocutionary intent* of the message (its pragmatic meaning) is the key. Thanks to recent advances in natural language processing, we are able to operationalize this theory by extracting the latent social dimensions of a message, namely archetypes of social intent of language, that come from social exchange theory. We identify key ingredients to opinion change by looking at more than 46k posts and more than 3.5M comments on Reddit's `r/ChangeMyView`, a debate forum where people try to change each other's opinion and explicitly mark opinion-changing comments with a special flag called *delta*. Comments that express no intent are about 77% less likely to change the mind of the recipient, compared to comments that convey at least one social dimension. Among the various social dimensions, the ones that are most likely to produce an opinion change are knowledge, similarity, and trust, which resonates with Habermas' theory of communicative action. We also find other new important dimensions, such as appeals to power or empathetic expressions of support. Finally, in line with theories of constructive conflict, yet contrary to the popular characterization of conflict as the bane of modern social media, our findings show that voicing conflict in the context of a structured public debate can promote integration, especially when it is used to counter another conflictive stance. By leveraging recent advances in natural language processing, our work provides an empirical framework for Habermas' theory, finds concrete examples of its effects in the wild, and suggests its possible extension with a more faceted understanding of intent interpreted as social dimensions of language.

1 Introduction

The "public sphere", as theorized by Jürgen Habermas [1], is the public arena wherein the democratic discourse develops. It plays a crucial role in the healthy functioning of a democracy, as it allows citizens to shape public opinion, thus influencing policies and decisions [2]. Such a role, today, is arguably filled in large part by the Internet and social media [3, 4].

Habermas sees shared understanding, achieved through rational arguments, as the necessary pre-condition for social integration, and thus for democracy [5]. The underlying assumption is that communication—and language in particular—is the only way to reach this shared understanding through grounded reasoning [6]. In particular, Habermas is interested in language from a formal pragmatics point of view, which differs from the socio-linguistic one. The focus is not on grammar, semantics, style, or sentiment, but rather on the interpretation of utterances [7]. The meaning of an utterance is not found in the sentence itself, as per the encoding/decoding paradigm of language, but in the intent of the speaker and in its reconstruction by the receiver, as per the intentionalist and dialogic paradigms [8]. By loading language with intent, the speaker exercises an illocutionary force that can effectively change the hearer's mind [9] and eventually achieve cooperation based on a shared understanding of reality. This process, which Habermas calls *communicative action*, can be triggered by potentially many different types of illocutionary forces [8], but especially by virtue of *"shared knowledge, mutual trust, and accord with one another"* [6].

The theory of communicative action is in stark contrast with how opinion dynamics has been traditionally operationalized. Models of opinion change fail to take into account nuances of communication, let alone the intent of the actors involved, as they describe social interactions as one-dimensional events [10]. These models have been necessarily oversimplified due to the complexity of quantifying social interactions. However, thanks to recent advances in natural language processing, we are now able to operationalize these concepts and measure the illocutionary force of an utterance, i.e., its intent. On one side, social theorists have identified universal hallmarks of communication that capture fundamental *social dimensions* of pragmatics, namely archetypes of social intent that language can express (Table 1), on the other side computer scientists have developed methods to identify those dimensions from conversational text automatically and accurately [11]. In this work, we employ these operationalizations to corroborate Habermas' hypothesis, and ask *"which types of illocutionary intents are more effective at inducing a change of opinion in the listener?"*.

We seek to answer this question in the wild. While there have been some attempts in a similar direction [12], our approach is general and widely applicable, employs automated coding (instead of a manual one), and makes use of large-scale data, obtained from a public discussion forum. In particular, we analyze data from Reddit's `r/ChangeMyView`, an on-line forum where people debate positions and try to change each other's opinion and reach some form of consensus: a success of the communicative action. This forum is particularly suited for our purposes as it provides a ground truth of opinion change. The debaters, in fact, explicitly recognize convincing arguments that changes their mind, thus producing a sort of consensus. Moreover, people participating in the forum have an epistemic goal [13], as they are asked to approach the discussion 'in an effort to understand other perspectives on the issue' and 'with a mindset for conversation' (https://www.reddit.com/r/changemyview). Therefore, the forum presents close to an ideal Habermasian communication situation [13], whereby perlocutionary acts and strategic actions [6] have no reason to be [14]. The pragmatic illocutionary intent, i.e., the social dimensions we employ, are therefore representations of the communicative action happening among participants of the forum.

Our results align with the hypothesis of Habermas: the three most important social dimensions for a convincing argument are the conveying of *knowledge*, the appeal to *similarity*, and the expression of *trust*. Conversely, messages that do not clearly convey a social intent are exceedingly unlikely to change someone's view (77% less, as illustrated in Figure 1). In addition, we identify particular intents that are more or less effective at changing opinions when replying to a specific expressed intent. For instance, responding to posts containing tones of conflict by signaling group identity is less effective at changing the mind of the poster. Conversely, we find that expressing conflict actually improves the odds of changing someone's mind, especially when replying to an already conflictive stance.

While Habermas' theory is vast and far-reaching, in this study we focus on the aspects related to the pragmatic intent of language, and its effects on opinion change. Our contribution lies in the operationalization of this aspect through the lens of social dimensions of pragmatics inspired by social exchange theory [15]. We find empirical support to the theory by looking at text in the wild from a popular online discussion forum, and by using fully-automated means of coding.

2 Research design

Gathering public sphere discussions from Reddit

Reddit is an online forum organized in topical communities, called *subreddits*. Inside a subreddit, users can publish *posts*, or *comment* in response to other posts or comments. In the `r/ChangeMyView` subreddit, posters express a point of view that commenters attempt to change. Comments that succeed in doing so receive from the poster a token of merit called *delta* (Δ) to symbolize a successful attempt at opinion change. We limit the scope of our study to `r/ChangeMyView` posts dealing with sociopolitical issues. This way, we bring the object of our study closer to the public sphere discussion as conceptualized by Habermas, while also ensuring a higher topical homogeneity. Following the operative definition given by Moy and Gastil [16], we define a post as *sociopolitical* if it is about at least one of the following categories: (i) political figures, parties or institutions; (ii) broad cultural and social issues (e.g., civil rights, moral values); (iii) national issues (e.g., healthcare, welfare).

To automatically categorize `r/ChangeMyView` posts as sociopolitical or not, we train a supervised classifier on Reddit posts (details in Section 5.1). Out of the 65 727 `r/ChangeMyView` posts with textual content, we identify 46 046 as sociopolitical: 20 239 of these have at least one comment with Δ (P_Δ), whereas 25 807 posts do not have any ($\overline{P_\Delta}$). Those 46 046 sociopolitical posts received 3 690 687 comments, which we split in two sets: one containing the 38 165 comments that received a Δ (C_Δ) and one containing the remaining comments ($\overline{C_\Delta}$). Summary statistics about the Reddit dataset are provided in Figure SI1.

Capturing social intent from language

To infer the social intent that Reddit messages convey, we ground our analysis in a theoretical model of *social dimensions* that reflect fundamental social aspects of the pragmatics of language (Table 1). In ordinary conversations, these dimensions are often verbalized to signal social intent, for example to confer appreciation or to give emotional support. These dimensions have been identified through an extensive survey of social science research [17] and they are comprehensive of some of the most influential categorizations of social interactions [18, 19, 20]. Thus, we use our analysis as a test bed for a novel natural language processing model [11], able to capture with high accuracy expressions of the social dimensions in conversational language (details in Section 5.2). Given an input message m and a social dimension d from Table 1, the tool produces a score $s_d(m)$ that represents the likelihood that message m contains social dimension d. To ease the interpretation of the results, we binarize the scores to split messages between those that carry dimension d with high probability and those that do not (see Section 5.3).

Our study relies on estimating the effect of the intent extracted from Reddit comments on the behavior of the user who reads them. As such, it can be sorted under the 'text-as-treatment' umbrella within the causal inference literature [21]. Feder et al. [21]

Table 1. Social dimensions of relationships historically identified in social sciences and surveyed by Deri at al. [17].

Dimension	Description
Knowledge	Exchange of ideas or information; learning, teaching [22]
Power	Having power over behavior and outcomes of another [15]
Status	Conferring status, appreciation, gratitude, or admiration [15]
Trust	Will of relying on the actions or judgments of another [23]
Support	Giving emotional or practical aid and companionship [22]
Similarity	Shared interests, motivations or outlooks [24]
Identity	Shared sense of belonging to the same group [25]
Fun	Experiencing leisure, laughter, and joy [26]
Conflict	Contrast or diverging views [27]

specifically identify three requirements for proper causal inference under the potential outcomes framework: *ignorability*, *positivity*, and *consistency*. The *ignorability* assumption requires the treatment assignment to be independent of the realized counterfactual outcomes. In our case, while the treatment is not randomly assigned, the author of the post does not have control on who writes an answer. This fact ensures that there is no selection bias due to the choices of the poster. The focus on a narrow topic reduces the possibility of unobserved confounders (e.g., a specific social dimension is more present on a topic for which it is easier to change opinion). The interaction between users happens exclusively via the text, so it is unlikely that there are other unobserved confounders that have effect both on the social dimensions and on the opinion of the original author of the post (e.g., body posture or voice pitch). One possible source of confounding is the profile of the poster which contains their posting history, which, if viewed, might influence the other users, and naturally influences the opinion of the original author. While we cannot completely exclude this confounder, judging a poster by anything other than their argument goes against the spirit of `r/ChangeMyView`, thus we assume that this possible confounder plays a negligible role. *Positivity* is the assumption that the probability of receiving treatment is strictly bounded between 0 and 1. This assumption is easily verified empirically in our case. Finally, *consistency* requires that the observed outcome at a given treatment status for an individual is the same as would be observed if that individual was assigned to the treatment. In practice, for the purpose of assuming consistency and making valid inferences, it is necessary to develop the measure of the treatment with different data than the data used to estimate the causal effect, such that there is no interference. In our case, the measure of social intent we use has been developed separately [11] and is not related to the opinion change outcome.

Quantifying communicative action

To assess the interplay between social intent and opinion change, we define suitable probabilities and odds ratios which are detailed in Section 5.4. We consider the odds ratios (OR) of finding dimension d_i in comments with a Δ (or posts that awarded a Δ), compared to those without Δ. To study the interactions between posts and comments, we need a more elaborate model to account for their complex relationship, which we base on the dialogic interpretation of language (see Section 5.4). We therefore compute the increase of the probability—compared to random chance—that a comment with dimension d_i would receive a Δ given that its corresponding post contains dimension d_j.

To account for possible confounders and to assess the significance of our findings, we complement the analysis of probabilities with logistic regression models. In particular, we consider the original ideological leaning of the two involved individuals because it is a factor that several opinion models take into account [10]. As a proxy for such leaning, we extract different sets of variables from individuals' participation in partisan sociopolitical groups [28]. We then use this information to test three alternative sets of confounders. First, *individual political sides*: whether each of the two involved individuals participates to a left- or right-wing group. Second, *interaction of political sides*: whether the two individuals both participate to polarized groups, and whether they are on opposing sides. Third, we check *shared groups*, i.e., whether the two individuals participate in exactly the same political and polarized group. Finally, to make sure that the signal captured by the social dimensions is not mere sentiment polarity, we add positive and negative sentiment of the message as controls. The considered confounders are summarized in Table 3 and further detailed in Section 5.5.

3 Results

We design our study to answer three research questions:

RQ1. Are messages that convey a social dimension more likely to change the opinion of the reader?

RQ2. Which *types* of social dimension are more often present in opinion-changing messages?

RQ3. Which *combination* of intents expressed by the poster and the commenter are more likely to result in an opinion change?

To find whether expressions of social intent matter in the process of opinion change (RQ1), we compute the odds ratios $OR_{C_\Delta}(d_i)$ of a social dimension d_i being conveyed by comments with Δ, compared to comments with no Δ (Figure 1a).

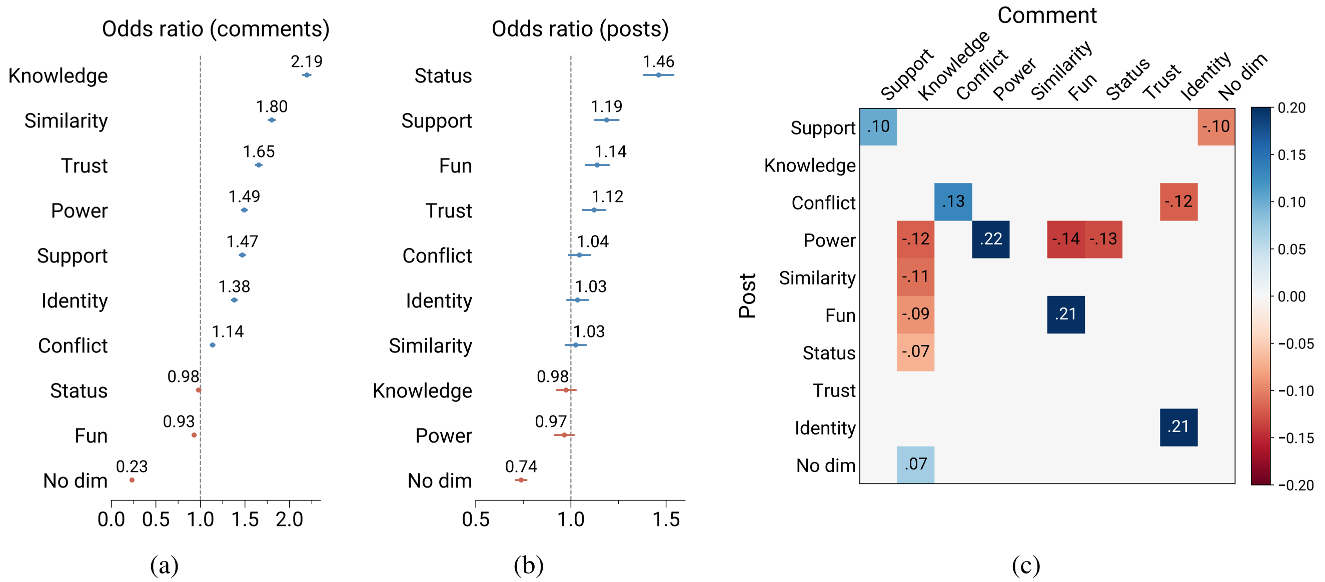

Figure 1. Odd ratios of containing a dimension (a) in comments that were successful in changing the poster's opinion versus those that were not, and (b) in posts expressing opinions that were changed by other community members versus posts that did not experience any opinion change. Error bars represent 95% confidence intervals. On the right (c), we report only the statistically significant odds ratios ($p < 0.01$) for interactions between dimensions in comments and posts. Cells represent the variation of the probability of achieving a Δ given a combination of dimensions.

Comments that express no intent exhibit an odds ratio of 0.23, meaning that they are about 77% less likely to change the mind of the recipient, compared to comments that convey at least one social dimension. To confirm the significance of this result, we apply a logistic regression model by using all the dimensions as independent variables, and whether the comment received a Δ as dependent variable.

We find (Table SI4) that all dimensions have a significant association with the message being considered as view-changing, with the notable exception of *fun*, that we therefore remove from further analysis. All the other dimensions emerge as significant. In general, social intents are positively associated with opinion change, except for *status*, that exhibits a significant yet negative relationship instead. Status captures admiration, appreciation, and praise. Because these expressions are typically more associated with agreement than with disagreement, it is likely that comments conveying status are meant to support the point of view of the original poster rather than attempting to change it. Thus, these comments are less likely to receive a Δ.

To check the robustness of the association between social dimensions and Δ to the inclusion of other factors, we fit regression models that use the sets of confounders we presented (Table 2). The confounders do not change the main result, and social dimensions keep consistently emerging as significant: for instance, in model E we observe that the inclusion of sentiment scores do not alter the significance of social dimensions. This is true even if we consider also the length of the message (Table SI5). Moreover, for each set of confounders, we test what happens with and without considering the social dimensions. Beside statistical significance, we assess quality of fit as measured by the adjusted Pseudo-R^2 metric. All the models that use social dimensions (models E-H) have a higher quality of fit than those that only use the ideological positions of the authors (models A-D). In particular, the best model that does not consider social dimensions but only sentiment and political group (model D) has a quality of fit that is roughly half of the model when with social dimensions (model H). The results presented in Table 2 are robust to data imbalance (Table SI6); also, the significance of the coefficients is not caused by random fluctuations in the data: a randomized regression model with reshuffled variables across examples loses all statistical significance (Table SI7).

In summary, messages that convey a social intent are more likely to be associated with opinion change in the intended reader, even after controlling for confounding factors. This finding backs our main hypothesis that considering social dimensions is essential to correctly model opinion change.

To assess what types of social intent are associated with comments that are successful in changing people's opinion (RQ2), we compare the magnitude of the odds ratios across the different social dimensions (Figure 1a). Comments that received a Δ are exceedingly more likely to convey *knowledge* than those with no Δ (+119%). To find illustrative examples of argumentative nuances associated with comments conveying different dimensions, we inspected manually the messages with Δ that were

Table 2. Odds ratios obtained by logistic regression. Each column corresponds to a model with a specific set of variables. A description of each confounder is given in Table 3. We indicate with asterisks the statistically significant correlations (with one, two or three asterisks corresponding to $p < 0.05$, $p < 0.01$ and $p < 0.001$ respectively). P-values are corrected according to the Benjamini-Hochberg procedure [29], to reduce the chance of spurious correlation emerging because of the high number of factors we consider.

	A	B	C	D	E	F	G	H
Adj. Pseudo-R^2	0.00305	0.00446	0.00720	0.00935	0.01643	0.01780	0.02050	0.02265
Intercept	0.010***	0.011***	0.011***	0.011***	0.010***	0.010***	0.010***	0.010***
Support	....	....	....	....	1.096***	1.095***	1.097***	1.098***
Knowledge	....	....	....	....	1.217***	1.216***	1.216***	1.216***
Conflict	....	....	....	....	1.024***	1.026***	1.025***	1.024***
Power	....	....	....	....	1.097***	1.099***	1.097***	1.096***
Similarity	....	....	....	....	1.110***	1.108***	1.110***	1.112***
Status	....	....	....	....	0.930***	0.929***	0.934***	0.937***
Trust	....	....	....	....	1.143***	1.143***	1.144***	1.144***
Identity	....	....	....	....	1.085***	1.086***	1.086***	1.086***
Sentiment Pos.	1.174***	1.173***	1.175***	1.177***	1.110***	1.110***	1.110***	1.112***
Sentiment Neg.	1.080***	1.082***	1.081***	1.081***	1.056***	1.058***	1.058***	1.057***
Comment Left-Wing	....	1.014	....	....	....	1.005	....	....
Comment Right-Wing	....	0.790***	....	....	....	0.795***	....	....
Post Left-Wing	....	0.867***	....	....	....	0.873***	....	....
Post Right-Wing	....	0.852***	....	....	....	0.849***	....	....
Both Polarized	....	....	0.321***	....	....	....	0.324***	....
Both Polarized & Diff. Side	....	....	2.555***	....	....	....	2.510***	....
Diff. Side	....	....	1.022	....	....	....	1.023	....
Shared Group	....	....	....	0.286***	....	....	....	0.287***

classified with high confidence. We report these examples next.

The classifier assigns high knowledge scores to messages that provide logical reasoning (e.g., *"The NHS isn't free, it's free at the point of use, we pay for it through tax contributions"*), refer to factual evidence (*"Door levers are more likely to fail over time since they require springs."*), cite sources (*"History shows us that the benefits of widespread vaccinations greatly outweigh the risks of possible resistant mutations"*), and present points of view that might be debatable yet stem from factual observations of the world (*"The automobile market is oversaturated with cars that run solely on gasoline"*).

Successful comments are 80% more likely to allude at *similarity* between the stance of the poster and the commenter (*"I'm glad to know we agree on this"*) or between their experiences (*"My friends used to live in a large city in Asia too"*), and 65% more likely to contain language that discloses *trust* towards entities relevant to their argument (*"I believe what they're saying"*, *"The people causing problems are a tiny minority compared to the reasonable majority"*). Appeals to power (*"The only rights which exist in objective reality are legal rights"*), expressions of support (*"I can understand being disappointed, but ..."*, *"I'd feel sympathy for their situation"*), and language markers of group identity (*"They are members of their tribe, they don't necessarily want to be part of the larger nation."*) are also more frequent in comments that received a Δ. These results are corroborated by the coefficients of the regression models. In all of our models (Table 2) the coefficients obtained by each social dimension are extremely stable: each odds ratio varies by less than 0.01.

In summary, our results provide empirical evidence for the presence of the *"shared knowledge, mutual trust, and accord with one another"* that Habermas identified as the founding pillars of communicative action. Indeed, the three social intents of *knowledge*, *trust*, and *similarity* characterize messages with a Δ more than all the other intents.

Messages written in an attempt to change someone's opinion are not isolated entities; rather, they are part of an on-going conversation between two parties. We study the simplest form of such interactions in Reddit, namely the relationship between the intent expressed by the post which starts the conversation and that expressed by the aspiring view-changing comment (RQ3). Interestingly, posts by people who change their view are characterized by different social dimensions from those found in view-changing comments (Figure 1b). Most prominently, conversation-starting messages written by individuals who end up awarding a Δ are 46% more likely to convey *status*—words of appreciation or gratitude. The presence of status is an indication of approaching the dialogue with respect, either towards the subject of the discussion (*"I have nothing but the utmost respect for service men and women, but ..."*) or the discussion itself (*"I'd really appreciate if someone could help me quantify and qualify my views"*). Conversely—and in agreement with the observation that partisans abiding to power have a less objective view of reality [30]—people who introduce their opinion by appealing to power or mentioning power dynamics (*"If people were required to vote, they would take more of an interest in the political situation"*) are the least likely to grant a Δ.

Discussions that lead to a change in opinion are often those where the intent of the poster receives a response motivated by a similar intent. Figure 1c shows a matrix of interaction between the intent of the poster and that of the commenter. Cells represent the variation of the probability of achieving a Δ given a specific combination of intents. Comments that receive a Δ are more likely to express an intent that matches that of the original poster, at least for five out of all the social dimensions that our tool captures. For example, when a post intends to convey a *power* dynamic, the most effective response is to make a similar appeal to power ($+22\%$ more likely of getting a Δ). This observation is in line with the general principle of *reciprocity* [31], and with the interpretation of conversations as social exchanges that occur under the assumption that a contribution of a certain type should be matched by a response of a similar type [15].

Some combinations of dimensions that break this symmetry are less likely to reach an agreement. When the poster expresses *power*, comments replying with *status* are 14% less likely to receive a Δ. Power-status dynamics occur frequently in social relationships. Individuals entertaining relationships with people who have power over them tend to maintain those relationships stable by means of providing status back [15] (e.g., employees expressing admiration for their manager). Contrary to typical social norms, the aim of `r/ChangeMyView` is not to maintain stability, but rather to disrupt ideas, a process that is not well-supported by power-status dynamics. Knowledge and fun are also less effective responses to power (-12% and -13% respectively).

Responding with comments containing markers of group identity is less effective in changing the mind of the poster, if the discussion was initiated with tones of conflict (-12%). Identity and conflict are tightly coupled in human societies. When brought up in debates, expressing identity often sparks conflict with those who do not feel (or are not entitled to) the same sense of belonging. Symmetrically, identity is also a typical way to oppose conflict, as it is signaled as a way to manifest in-group defense in response to an out-group aggression [27]. Such dynamics are more likely to originate disagreement than to facilitate convergence of ideas.

Last, despite knowledge-based reasoning being the most effective approach to changing someone's view (Figure 1a), its effectiveness is highest when responding to posts that are not strongly characterized by a social intent, such as those that plainly state an opinion (e.g., *"I believe that the individual is not more vital then the global scale."*). When responding to posts that are strongly characterized by an intent (especially those expressing power), knowledge-conveying comments are less effective. Even though our tools of analysis cannot ascertain the root cause of this phenomenon, we speculate that posts conveying a clear intent might indicate stronger motivated reasoning, or stronger ideological biases of the poster, both of which are shields to persuasion.

4 Discussion

We have shown that the pragmatic intent of the dialogue between two parties has an important effect on the success of communicative action, as theorized by Habermas, and in reaching agreement. In particular, the social dimensions that are most indicative of a successful communicative action are the ones originally indicated by Habermas: knowledge, similarity, and trust. Results show robustness across different settings (see Figure SI5 and Table SI5). These findings echo the interpretation of Habermas' theory provided by Terry [32]: in an ideal speech situation *"all participants must [...] refer to facts and knowledge with which all are familiar, contribute to the discussion in an open, honest way, and be prepared to place themselves in the position of others in order to understand the latter's point of view"*. However, other dimensions that have not previously been identified as relevant also have a positive effect, such as appeals to power or empathetic expressions of support. In line with theories of constructive conflict [33], yet contrary to the popular characterization of conflict as the bane of modern social media, we found that in the context of a structured public debate, voicing conflict can promote integration, especially when it is used to counter a conflictive stance. We find indeed that belonging to different ideological sides increases the chances of one of the parts to change their opinion (Table 2, models G and H). This finding is further detailed in Figure SI6. Finally, we find that sentiment, as commonly measured in natural language processing, by itself is an unreliable indicator for opinion change. Indeed, the definition of sentiment commonly used ('the underlying feeling, attitude, evaluation, or emotion associated with an opinion' [34]) is too broad for our purposes.

These results, combined, point towards an extension of the original Habermasian theory which includes a more faceted understanding of intent, interpreted as social dimensions of language. In particular, the original theory of communicative action is quite broad in scope, but only provides a few concrete examples. This study contributes to materialize it, by providing an empirical descriptive framework for it, and by finding concrete examples of its effects in the wild. In fact, while Habermas' theory aims at a high level of generality and abstraction, able to unify different levels of the theory of communication, it still places value on its ability to receive empirical confirmations. As noted by Bohman [35], practical verification of the consequences of such theoretical constructs should be the main route to solve the problem, posed by pluralism, of choosing between alternative theories. Thanks to the explosive growth of available data and of our ability to process it, we are able to connect such large and general theories to empirical and falsifiable claims [36]. We focus on one specific aspect of the more general theory of communicative action; specifically, on understanding which types of illocutionary intents are more effective at reaching consensus.

Other works have previously explored how to operationalize habermasian ideas by measuring the nature of discourse, typically in the study of deliberative democracies [37]. However, these works present two profound differences with our approach. First, we focus on how the pragmatic intent of language affect communication between social media users, rather than communication performed in controlled settings or by professional politicians [38]. Second, in order to achieve this goal on the large scale of social media data, we apply machine learning techniques to automatically classify intents, rather than relying on human evaluators as previously done in the literature [39]. We leverage such data availability of the Web in order to operationalize the study of communication as it happens in the wild, studying how social media contribute to the public sphere. In this sense, we follow Habermas' idea of communication operating with similar mechanisms at different levels.

The link between social exchange theory [15], on which the social dimensions are based, and the theory of communicative action, has been missing from the pragmatics literature, but is clearly worth exploring further in light of the results presented here. Social exchange theory is focused explaining an equilibrium in social relationships, and provides an economics-based framework for understanding social behavior. Conversely, the theory of communicative action explains (among other things) a change—i.e., reaching a consensus—and insists that strategic actions based on cost-benefit analysis are counter to this ultimate goal. Our study, based on social media data, provides empirical support for both theories, and yet hints at a larger picture: a theory that is a synthesis of both sources while resolving the tension existing due to their different goals. Creating such a theory is a challenging endeavor, but also an exciting one given its potential broad impact. Our study is but a first step in this direction that shows a working, proof-of-concept operationalization of social dimensions extracted from language, and its importance in understanding the consensus-reaching process. Our work can help guiding the design of communication campaigns (e.g., against vaccine hesitancy) by better understanding the potential effectiveness of alternative strategies through the analysis of their social intent.

Future work could improve on our analysis in five main aspects. First, our social dimensions model alone is not predictive of whether a comment will get a Δ, as this is not the focus of our work. Such prediction is hard to make not only because of sparsity (successful comments are overwhelmingly outnumbered by unsuccessful ones), but mainly because of the intrinsic limit to predictability of complex social phenomena [40]. Moreover, several orthogonal factors influence the outcome of the debate including the author's reputation, the quality of argumentation, the discussion topic, and the societal and historical context around the discussion. As studies accumulate evidence supporting the role of different factors in opinion change [41], future work can incorporate the social dimensions into more comprehensive and predictive models.

Second, the classifiers we use to detect political posts and to extract the social dimensions are accurate, yet not exhaustive nor error-free. The social dimension classifier is based on a conceptualization of social relationships that is broader than existing theoretical models [18, 19, 20], and its proponents have shown empirically [11] that it accounts for key dimensions of traditional psycholinguistic models [42, 43]. Yet, future research should strive for models that are more accurate (i.e., lower error rate), more comprehensive (i.e., more dimensions), and especially more detailed (i.e., different nuances of a given dimension). In fact, many of the social dimensions that we use in our work have been charcterized by previous research as complex superpositions of different psychological constructs; for example, trust comprises both cognitive and affective components [44] that the tool we use in this work is not able to disentangle. Given the rapid progress of natural language processing technologies, we expect that improved models for language understanding can soon replace the ones used in this study. New methods could also attempt to qualify social dimensions in ways that we have not considered in this study, for example accounting for the *directionality* of the social intent: our classifier detects the presence of a social intent in an utterance but cannot determine who is the subject that expresses the intent (e.g., *"you trust me"* and *"I trust you"* are both classified as trust, and are considered equivalent in our analysis).

Third, `r/ChangeMyView` is a platform with unique qualities: it is designed to attract members who are keen on participating in public discussions and who are open to change their opinion; it also provides clear way to track interactions and opinion changes. It is therefore likely that participants self-select to be particularly open to mutual understanding. On the one hand, these properties—akin to an idealized setting for communication—allow us to work on clean data, while preserving a good degree of ecological validity compared to studies conducted in the lab. On the other hand, replicating our experimental setup on multiple platforms is needed to support the generality of our findings.

Fourth, we analyze conversations with one exchange only—one post and one response—mainly because these make for the vast majority of interactions on `r/ChangeMyView`. With more extensive data at hand, future studies could look into how the role of different social dimensions changes as the dialogue progresses. More broadly, our results can provide a basis for operationalizing complex psychological theories of communication (e.g., transactional analysis [45]).

Fifth, and last, one of the main goals of this work is to propose a framework that can overcome the oversimplified design of opinion dynamics models by adding nuance to the types of social interactions considered. In most opinion dynamics models, interactions are either binarized, or associated with a polarity, to represent interactions with positive/negative sentiment, or between actors with same/different ideological stances [10]. Our regression results show that neither sentiment nor political side explain opinion change as well as speaker intent, as operationalized by the social dimensions exchanged.

Table 3. Explaination of the confounders used in the logistic regression models (Table 2). More details are provided in Section 5.5.

Variable		Description
Sentiment	Sentiment Pos. .... Sentiment Neg.	Positive and negative sentiment in $[0,1]$ of the `r/ChangeMyView` comment, as extracted by Vader [46].
Political side (individual)	Comment Left-Wing ...... Comment Right-Wing ..... Post Left-Wing .... Post Right-Wing	Boolean variables describing whether the author of the comment (resp. post) ever participated in a subreddit that we identified as left-wing (resp. right-wing) at the time of their submission to `r/ChangeMyView`.
Political side (interaction)	Both Polarized	Equal to 1 if both the author of the comment and the author of the post ever participated in one of the subreddits that we identified as left-wing or right-wing.
	Both Polarized & Diff. Side	Equal to 1 if the author of the comment participated in a left-wing subreddit and the author of the post in a right-wing subreddit, or vice-versa.
	Diff. Side	Equal to 1 if the author of the comment participated in a left-wing (or right-wing) subreddit and the author of the post did not, or vice-versa.
Political group	Shared Group	Equal to 1 if the author of the comment and that of the post ever participated in the same polarized subreddit.

5 Materials and methods

5.1 Classification of sociopolitical posts

To focus on a homogeneous set of posts, we develop a supervised classifier that recognizes posts with a sociopolitical topic, according to the definition given by Moy and Gastil [16]. In addition, given the focus on opinion change, we wish to exclude posts that discuss strictly factual statements. To train such a classifier, we manually categorize the 2000 most popular (non-access-restricted) subreddits in 2019 as sociopolitical or not by looking at their description and a sample of their posts, and find 51 sociopolitical subreddits. Then, we use such classification to build a training set for our supervised classifier. Specifically, we take a random sample of 50 posts per month for each sociopolitical subreddit, from 2011 to the end of 2019; if a subreddit does not have at least 50 posts in a month, we take all available posts for that month. We also take a random sample of equal total size for each month from all the non-sociopolitical subreddits. This way, we obtain a training set composed by 104 292 sociopolitical posts (stratified over subreddits and time) and the same number of non-sociopolitical ones (stratified over time). Using the same sampling procedure, we collect a test set of equal size. We use the training set to train a logistic regression model with L1 regularization, to distinguish between sociopolitical and non-sociopolitical posts. As features, we employ $\{1,2,3\}$-grams, and keep only the 10 000 most frequent ones in the whole dataset.

We evaluate the results of the classifier in two ways. First, on the test set, on which the classifier gets an average F1 score of 89.5%. (detailed results in Table SI1 and Table SI2).

Then, we manually build a validation set of `r/ChangeMyView` posts by randomly selecting 500 posts from the subreddit and labelling each one as sociopolitical or not by manual inspection, according to the definition given before (Table SI3 shows an excerpt). Out of the 500 posts, 269 are labelled as sociopolitical (53.8%). On this validation set, our classifier obtains an F1-score of 75% if we consider all posts, including the ones for which the main text of the post was subsequently removed by the author and only the title is available. If we consider only the 206 posts where text is present (120 of which sociopolitical), the classifier obtains an F1-score of 82%. We report other accuracy metrics in Table SI2. For this reason, in the rest of this work we consider only posts with text present. We then proceed to classify all `r/ChangeMyView` posts by using our sociopolitical classifier. Finally, we extract the comments of the posts that are recognized as sociopolitical.

To check how the precision of this classifier impacts our results, we repeat all our experiments using a different threshold to classify sociopolitical posts. Specifically, in this alternative setting we choose to classify posts as sociopolitical if the classifier assigns them a score of 0.75 (instead of 0.5). This experiment leads to 40 296 posts classified as sociopolitical instead of 46 046 (i.e., 87.5%). Also, by using this higher threshold on the validation data set, we obtain a higher precision of 83% (+7%) and a lower recall of 71% (−17%). Then, we repeat all of the experiments presented in this work. We observe that none of our results change substantially: for example, in Table 2 the significance of results stays unchanged, and the odds ratios change only in the third decimal digit in most cases (see Table SI8).

5.2 Extracting social dimensions from text

To extract the social dimensions from the set of sociopolitical posts and their respective comments, we leverage a previously developed model [11] with a publicly-available Python implementation.[1]

Given a textual message m and a social dimension d, the model estimates the likelihood that m conveys d by giving a score from 0 (least likely) to 1 (most likely). The model is not a multi-class classifier, rather it includes a set of independently-trained binary classifiers C_d, one per each dimension, i.e., it is a multi-label classifier. This choice is driven by the theoretical interpretation of the social dimensions [17], as any sentence may potentially convey several dimensions at once (e.g., a message expressing both trust and emotional support). Each classifier is implemented by using a Long Short-Term Memory neural network (LSTM) [47], a type of Recurrent Neural Network (RNN) that is particularly effective in modeling both long and short-range semantic dependencies between words in a text, and it is therefore widely used in a variety of natural language processing tasks [48]. Similarly to most RNNs, LSTM accepts fixed-size inputs. This particular model takes in input a 300-dimensional embedding vector of a word, one word at a time for all the words in the input text. Embedding vectors are dense numerical representations of the position of a word in a multidimensional semantic space learned from large text corpora. This model uses GloVe embeddings [49] learned from Common Crawl, a text corpus which contains 840B tokens.

The dimensions classifiers C_d are trained by using about 9k sentences manually labeled by trained crowdsourcing workers. Most of these sentences are taken from Reddit, which makes it the ideal platform to apply the model on. The models achieve very good classification performance which averages to an Area Under the Curve (AUC) of 0.84 across dimensions. AUC is a standard performance metric that assesses the ability of a classifier to rank positive and negative instances by their likelihood score, independent of any fixed decision threshold (the AUC of a random classifier is expected to be 0.5, whereas the maximum value is 1).

In practice, the classifier estimates a score for each sentence S in m and returns the maximum score, namely: $s_d(m) = \max_{S \in m} s_d(S)$. By using the maximum score, we consider a message as likely to express dimension d as its most likely sentence, thus avoiding the dilution effect of the average. This reflects the theoretical interpretation of the use of the social dimensions in language [17]: a dimension is conveyed effectively through language even when expressed only briefly.

5.3 Binarization and normalization of social dimension scores

To conduct our analysis, we binarize the classifier scores $s_d(m)$ via an indicator function that assigns dimension d to m if $s_d(m)$ is above a certain threshold θ_d:

$$d(m) = \begin{cases} 1, & \text{if } s_d(m) \geq \theta_d \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

We use dimension-specific thresholds because the empirical distribution of the classifier scores s_d varies across dimensions, which makes the use of a fixed common threshold unpractical. We conservatively pick the value of θ_d as the 85^{th} percentile of the empirical distribution of the scores s_d, thus favoring high precision over recall. This effectively reduces the number of messages marked with each dimension to 15% of the total number of messages.

When assigning a dimension d to a message based on the single sentence that most prominently expresses that dimension, the probability of being labeled with d naturally increases with the length of the message. Figure SI3 shows that such increase is roughly linear with the number of words. To mitigate the length bias, we design a length-discounting factor. The typical length of messages may vary considerably across dimensions (Figure SI2), therefore, to avoid excessively penalizing some dimensions over others, we use a dimension-specific discounting factor. Given a message m with length $len(m)$ (measured in number of words), and labeled with dimension d (i.e., such that $d(m) = 1$), we proceed as follows. First, we standardize $len(m)$ with respect to the length distribution of all messages labeled with d: given $\mu_{len}(d)$ and $\sigma_{len}(d)$ as the average and standard deviation of the length distribution of messages with d, the standardized value of length is $zlen_d(m) = \frac{len(m)-\mu_{len}(d)}{\sigma_{len}(d)}$. We then redefine $d(m)$ as follows:

$$d(m) = \begin{cases} \frac{1}{1+zlen_d(m)} & \text{if } s_d(m) \geq \theta_d \wedge zlen_d(m) >= 0 \\ 2 - \frac{1}{1-zlen_d(m)} & \text{if } s_d(m) \geq \theta_d \wedge zlen_d(m) < 0 \\ 0, & \text{if } s_d(m) < \theta_d \end{cases} \tag{2}$$

This length-discounted value is bounded between 0 and 2. It is equal to 1 when the message length is equal to the average length of messages with dimension d. It approaches 0 as the length of the message increases, and it gets closer to 2 as the length

[1] http://www.github.com/lajello/tendimensions

approaches 0, thus effectively weighting more those messages whose length is shorter than what is expected for the dimension considered (and vice versa). In Figure SI4, we show the effect that the weight discounting has in reducing the cross-correlation between pairs of dimensions (i.e., in increasing their orthogonality).

5.4 Odds ratios

The length-discounted prior probability of a message (either a post or a comment) being labeled with dimensions d is

$$p(d) = \frac{\sum_{m \in M} d(m)}{2 \cdot |M|}, \tag{3}$$

where M is the set of messages and the factor 2 is used to limit $p(d)$ between 0 and 1, as the values of $d(m)$ range from 0 to 2.

We define the conditional probability that a comment contains d, given that it received a Δ as

$$p(d|\Delta) = \frac{\sum_{m \in C_\Delta} d(m)}{2 \cdot |C_\Delta|}, \tag{4}$$

where C_Δ is the set of comments with Δ. We use an analogous formulation for the set of messages $\bar{\Delta}$. The odds ratio between d and Δ, which is a measure of the strength of their association, is defined as

$$OR(p(d|\Delta), p(d|\bar{\Delta})) = \frac{p(d|\Delta)/(1 - p(d|\Delta))}{p(d|\bar{\Delta})/(1 - p(d|\bar{\Delta}))}. \tag{5}$$

We compute the conditional probability of a comment containing dimension d_i, given that its corresponding post contains dimension d_j, as

$$p(d_i(comment)|d_j(post)) = \frac{\sum_{c \in C(P_{d_j})} d_i(c)}{2 \cdot |C(P_{d_j})|}, \tag{6}$$

where P_d is the set of posts with d, $C(P_d)$ is the corresponding set of comments. When considering only the set C_Δ of comments with Δ (and equivalently for $\bar{\Delta}$), the formula becomes

$$p_\Delta(d_i|d_j) = \frac{\sum_{c \in C_\Delta(P_{d_j})} d_i(c)}{2 \cdot |C_\Delta(P_{d_j})|}. \tag{7}$$

As the joint distribution of the dimensions in comments and posts is different between the case of Δ and $\bar{\Delta}$, to make the $p_\Delta(d_i|d_j)$ and $p_{\bar{\Delta}}(d_i|d_j)$ values comparable, we offset them by their expected value under a randomized null model. Specifically, we create a shuffled dataset via a random permutation r, so that the association between posts and comments is randomized, but the dimension-message association is unchanged. This null model destroys the association between posts and comments, thus reflecting the baseline probability that a post with d_j would receive a comment with d_i just by chance. We calculate $p^r_\Delta(d_i|d_j)$ and $p^r_{\bar{\Delta}}(d_i|d_j)$ in this randomized model, and we then calculate the odd ratios between the real data and the random baseline:

$$OR(p_\Delta(d_i|d_j), p^r_\Delta(d_i|d_j)) = \frac{p_\Delta(d_i|d_j)/(1 - p_\Delta(d_i|d_j))}{p^r_\Delta(d_i|d_j)/(1 - p^r_\Delta(d_i|d_j))}. \tag{8}$$

We indicate this odds ratio with $OR_\Delta(d_i, d_j)$, and we analogously define $OR_{\bar{\Delta}}(d_i, d_j)$.

The 95% confidence intervals of the odds ratios are calculated as:

$$ci = 1.96 \cdot \sqrt{\frac{1}{|C_{d,\Delta}|} + \frac{1}{|C_{d,\bar{\Delta}}|} + \frac{1}{|C_{\bar{d},\Delta}|} + \frac{1}{|C_{\bar{d},\bar{\Delta}}|}}, \tag{9}$$

where 1.96 is the critical value of the Normal distribution at $\alpha/2$ (with $\alpha = 0.05$) and $|C_{\bullet,\bullet}|$ represents the cardinality of the set of comments with or without a given dimension (d or $\bar{d}$) and with or without a delta (Δ or $\bar{\Delta}$).

5.5 Confounders

Beside the social dimensions, we test possible confounders in our regression models. To model message-level confounders, we include in our models the positive and negative sentiment of the message as measured by Vader [46]. We also consider the length of the message $len(m)$ (Table); specifically, we use the standardized value of $\log(len(m))$, to account for the fat-tailed distribution of message lengths.

To estimate the ideological leaning of the participants to the conversation, we consider subreddits in which a user has participated with a submission at any point in time *before* their message in `r/ChangeMyView` under consideration. From the posting history of each pair of poster and commenter, we extract three sets of variables that capture, respectively, the political leaning of each of the two users, the interaction between their leanings, and their similarity.

The first set describes whether each of the two users has participated in a right-wing or a left-wing subreddit. We manually select a set of 10 right-leaning subreddits (e.g., `r/The_Donald`, `r/Conservative`) and a set of 15 left-leaning subreddits (e.g., `r/SandersForPresident`, `r/Socialism`). In the selected groups, rules typically suggest that participants adhere to the ideological view of the subreddits; as such, participation can be taken as a meaningful proxy of a political ideology.

The second set expresses whether both involved users have participated in any politically-identified subreddit, and whether they participated on the same side. We use two boolean variables and their interaction to encode this information.

The last set considers whether the two users have participated in precisely the same subreddit among the 25 subreddits selected to assess political leaning, plus a list of 14 additional subreddits whose members express homogeneous political views, but that are not necessarily well-positioned on the traditional left-right spectrum (e.g., `r/atheism`, `r/conspiracy`, or `r/NeutralPolitics`).

Table 3 summarizes all the considered confounders.

Acknowledgements

LMA acknowledges the support from the Carlsberg Foundation through the COCOONS project (CF21-0432). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability

The raw Reddit data is freely available through the `pushift.io` API. The data that we used to for the experiments, and the sociopolitical classifier are available at: `https://github.com/corradomonti/10-dim-of-op-change`. The social dimensions classifier is available at: `https://github.com/lajello/tendimensions`.

Author contributions

CM collected the data, CM and LMA conducted the experiments. CM, LMA, GDFM, and FB conceived the experiments, wrote and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Habermas, J. *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society* (MIT Press, Cambridge, Mass, 1962).

2. Habermas, J. Further reflections on the public sphere. *Habermas public sphere* **428** (1992).

3. Gimmler, A. Deliberative democracy, the public sphere and the internet. *Philos. & Soc. Crit.* **27**, 21–39 (2001).

4. Fuchs, C. Social media and the public sphere. *tripleC: Commun. Capitalism & Critique. Open Access J. for a Glob. Sustain. Inf. Soc.* **12**, 57–101 (2014).

5. Habermas, J. *Lifeworld and System: A Critique of Functionalist Reason*. No. Jürgen Habermas. Transl. by Thomas MacCarthy ; Vol. 2 in The Theory of Communicative Action (Beacon, Boston, 1981).

6. Habermas, J. *Communication and the Evolution of Society*, vol. 572 (Beacon Press, 1979).

7. Habermas, J. *Reason and the Rationalization of Society*. No. Jürgen Habermas. Transl. by Thomas MacCarthy ; Vol. 1 in The Theory of Communicative Action (Beacon, Boston, 1981).

8. Krauss, R. M. & Chiu, C.-Y. Language and Social Behavior. In *Handbook of Social Psychology*, vol. 2 (McGraw-Hill, 1998), fourth edn.

9. Austin, J. L. *How to Do Things with Words: The William James Lectures Delivered at Harvard University in 1955* (Harvard Univ. Press, Cambridge, Mass, 1962).

10. Grabisch, M. & Rusinowska, A. A survey on nonstrategic models of opinion dynamics. *Games* **11**, 65 (2020).

11. Choi, M., Aiello, L. M., Varga, K. Z. & Quercia, D. Ten social dimensions of conversations and relationships. In *Proceedings of the Web Conference 2020*, 1514–1525 (2020).

12. Rosenberg, S. The empirical study of deliberative democracy: Setting a research agenda. *Acta Polit.* **40**, 212–224, DOI: 10.1057/palgrave.ap.5500105 (2005).

13. Habermas, J. Interview with Jürgen Habermas. In Bächtiger, A., Dryzek, J. S., Mansbridge, J. & Warren, M. (eds.) *The Oxford Handbook of Deliberative Democracy*, 870–882, DOI: 10.1093/oxfordhb/9780198747369.013.60 (Oxford University Press, 2018).

14. Heng, M. S. & De Moor, A. From habermas's communicative theory to practice on the internet. *Inf. Syst. J.* **13**, 331–352 (2003).

15. Blau, P. M. *Exchange and Power in Social Life* (Transaction Publishers, 1964).

16. Moy, P. & Gastil, J. Predicting deliberative conversation: The impact of discussion networks, media use, and political cognitions. *Polit. Commun.* **23**, 443–460 (2006).

17. Deri, S., Rappaz, J., Aiello, L. M. & Quercia, D. Coloring in the Links: Capturing Social Ties as They are Perceived. *Proc. ACM on Human-Computer Interact.* **2**, 1–18, DOI: 10.1145/3274312 (2018). 1902.04528.

18. Wellman, B. & Wortley, S. Different strokes from different folks: Community ties and social support. *Am. journal Sociol.* **96**, 558–588 (1990).

19. Fiske, A. P. The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychol. review* **99**, 689 (1992).

20. Spencer, L. & Pahl, R. *Rethinking Friendship* (Princeton University Press, 2018).

21. Feder, A. *et al.* Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *arXiv preprint arXiv:2109.00725* (2021).

22. Fiske, S. T., Cuddy, A. J. & Glick, P. Universal dimensions of social cognition: Warmth and competence. *Trends cognitive sciences* **11**, 77–83 (2007).

23. Luhmann, N. *Trust and Power* (John Wiley & Sons, 1982).

24. McPherson, M., Smith-Lovin, L. & Cook, J. M. Birds of a feather: Homophily in social networks. *Annu. review sociology* **27**, 415–444 (2001).

25. Tajfel, H. *Social Identity and Intergroup Relations* (Cambridge University Press, 2010).

26. Argyle, M. *The Psychology of Happiness* (Routledge, 2013).

27. Tajfel, H., Turner, J. C., Austin, W. G. & Worchel, S. An integrative theory of intergroup conflict. *Organ. Identity* (1979).

28. Massachs, J., Monti, C., De Francisci Morales, G. & Bonchi, F. Roots of Trumpism: Homophily and Social Feedback in Donald Trump Support on Reddit. In *WebSci '20: 12th International ACM Web Science Conference* (2020).

29. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).

30. Keltner, D. & Robinson, R. J. Defending the status quo: Power and bias in social conflict. *Pers. Soc. Psychol. Bull.* **23**, 1066–1077 (1997).

31. Gouldner, A. W. The norm of reciprocity: A preliminary statement. *Am. sociological review* 161–178 (1960).

32. Terry, P. R. Habermas and education: Knowledge, communication, discourse. *Curriculum Stud.* **5**, 269–279, DOI: 10.1080/14681369700200019 (1997).

33. Follett, M. P. Constructive conflict. *Sociol. Organ. Struct. Relationships* **417** (2011).

34. Cambria, E., Das, D., Bandyopadhyay, S. & Feraco, A. *A practical guide to sentiment analysis* (Springer, 2017).

35. Bohman, J. Theories, practices, and pluralism: A pragmatic interpretation of critical social science. *Philos. social sciences* **29**, 459–480 (1999).

36. Buyalskaya, A., Gallo, M. & Camerer, C. F. The golden age of social science. *Proc. Natl. Acad. Sci.* **118** (2021).

37. Steiner, J., Bächtiger, A., Spörndli, M. & Steenbergen, M. R. *Deliberative politics in action. Analysing parliamentary discourse* (Cambridge University Press, 2005).

38. Steenbergen, M. R., Bächtiger, A., Spörndli, M. & Steiner, J. Measuring political deliberation: A discourse quality index. *Comp. Eur. Polit.* **1**, 21–48 (2003).

39. Gerber, M., Bächtiger, A., Fiket, I., Steenbergen, M. & Steiner, J. Deliberative and non-deliberative persuasion: Mechanisms of opinion formation in europolis. *Eur. Union Polit.* **15**, 410–429 (2014).

40. Salganik, M. J. *et al.* Measuring the predictability of life outcomes with a scientific mass collaboration. *Proc. Natl. Acad. Sci.* **117**, 8398–8403 (2020).

41. Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C. & Lee, L. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-Faith Online Discussions. *Proc. 25th Int. Conf. on World Wide Web - WWW '16* 613–624, DOI: 10.1145/2872427.2883081 (2016). 1602.01103.

42. Plutchik, R. A general psychoevolutionary theory of emotion. In *Theories of emotion*, 3–33 (Elsevier, 1980).

43. Tausczik, Y. R. & Pennebaker, J. W. The psychological meaning of words: Liwc and computerized text analysis methods. *J. language social psychology* **29**, 24–54 (2010).

44. Dowell, D., Morrison, M. & Heffernan, T. The changing importance of affective trust and cognitive trust across the relationship lifecycle: A study of business-to-business relationships. *Ind. Mark. Manag.* **44**, 119–130 (2015).

45. Stewart, I. & Joines, V. *TA Today: A New Introduction to Transactional Analysis* (Lifespace Pub., 1987).

46. Hutto, C. J. & Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, ICWSM, 216–225 (AAAI, 2014).

47. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9**, 1735–1780 (1997).

48. Sundermeyer, M., Schlüter, R. & Ney, H. LSTM neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association*, Interspeech (2012).

49. Pennington, J., Socher, R. & Manning, C. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP, 1532–1543 (Association for Computational Linguistics, 2014).