nature human behaviour

Review article

https://doi.org/10.1038/s41562-024-01938-0

How developments in natural language processing help us in understanding human behaviour

Received: 22 December 2023

Accepted: 1 July 2024

Published online: 22 October 2024

Check for updates

Rada Mihalcea[1], Laura Biester[2], Ryan L. Boyd[3], Zhijing Jin[4], Veronica Perez-Rosas[1], Steven Wilson[5] & James W. Pennebaker[6]

The ways people use language can reveal clues to their emotions, social behaviours, thinking styles, cultures and the worlds around them. In the past two decades, research at the intersection of social psychology and computer science has been developing tools to analyse natural language from written or spoken text to better understand social processes and behaviour. The goal of this Review is to provide a brief overview of the methods and data currently being used and to discuss the underlying meaning of what language analyses can reveal in comparison with more traditional methodologies such as surveys or hand-scored language samples.

Language is the common currency of most human communication. We have all overheard snippets of conversation between two strangers while out walking or sitting in a restaurant or on a plane. In very little time, we can detect the speakers' general ages, possible relationship, education and social class, and much more. We also can determine the conversational topic and even the ways the people are thinking. In fact, these same clues are apparent in people's text messages, personal emails and transcripts from random Zoom meetings. Many of the same linguistic signals we hear can also be detected through computational analyses.

Over the past 20 years, thousands of studies have documented that everyday language can tell us about the content of people's thoughts; their relationships with their audience; their motives, goals and values; and essential aspects of their personality. These insights now go far beyond a single person or conversation. Through the analysis of social media and other large language datasets, we can begin to understand small groups, organizations and entire societies over days, years or even centuries.

The links between language, psychological state and behaviour

There is a long philosophical and scientific history that has attempted to disentangle the distinctions between language and thought[1]. People often talk about issues they think about. At the same time, the mere act of verbalizing issues can influence the ways people think about those issues. One common approach, the attention/language model, posits that people naturally talk about objects, people, emotions or events to which they are attending. If people hear or read about a disease, they often report feeling sensations of the disease and are more likely to visit a physician to see whether they have the disease (for example, when President Gerald Ford's popular wife, Betty, was diagnosed with breast cancer, there was a surge of women who subsequently were tested for breast cancer[2]). If people in your social network mention losing weight, buying a particular brand of clothing or liking a particular political candidate, the odds of you thinking about these issues and changing your behaviours go up accordingly[3].

It is a small leap to understand how the language of individual people or even groups of people can provide clues to their thoughts. Hungry people tend to talk about food; friends who are madly in love have difficulty not talking about it. Just as language can reflect our thoughts, our behaviours are often driven by both our thoughts and language. Language, psychological state, environment and behaviours are intertwined[4]. As researchers, if we are able to track people's psychology through natural language, we are better able to understand who they are, what they are thinking and how they might behave.

The psychological and computational approaches to language

For the purposes of this paper, we focus only on digitized verbal language—either transcribed oral or written. Language, of course, is

[1]University of Michigan, Ann Arbor, MI, USA. [2]Middlebury College, Middlebury, VT, USA. [3]University of Texas at Dallas, Richardson, TX, USA. [4]Max Planck Institute for Intelligence Systems, Tübingen, BW, Germany. [5]Oakland University, Rochester, MI, USA. [6]University of Texas at Austin, Austin, TX, USA. e-mail: mihalcea@umich.edu

infinitely complex and can be broken down by letter, morpheme, word, phrase, sentence, thought unit, paragraph or entire text. Rather than focusing on language itself, our primary interest is in understanding people's everyday thinking patterns, emotions and social connections through their use of natural language. Computationally, this generally means collecting and analysing words from as many people as possible. In this paper, our goal is to review the space of methods for inferring human behaviour from language in a way that is relevant to both psychologists and computer scientists, particularly to the ones working in natural language processing (NLP; see Box 1 for terminology)—a field that combines computational linguistics with statistical and machine learning models to understand and generate language. The paper therefore covers lines of work from both disciplines, eventually converging towards a unified approach to synthesize insights on human behaviour through the analysis of language. Figure 1 is an overview of our paper, moving from people to the language they use to the inference of behaviours at the individual, interpersonal and group levels.

Developments in NLP

The recent growth of NLP has been primarily driven by advancements in large language models (LLMs). These LLMs, including GPT4 (ref. 5), Llama[6] and Mistral[7], are trained on massive datasets of text, images and code, leading to major advances in our ability to perform automatic language understanding and generation, often across multiple languages. Breakthroughs in dialogue systems and the ability to generate human-quality conversation have been facilitated by techniques such as word embeddings, which map words to high-dimensional vectors capturing semantic relationships[8], and the Bidirectional Encoder Representations from Transformers (BERT) architecture for contextual embeddings[9,10]. To test these models, often in comparison with human abilities, much of the recent work has focused on probing tasks using survey-like instruments or benchmark evaluations. We are now seeing human-like performance on many tasks[11,12], such as question answering and textbook knowledge, while at the same time revealing major limitations on skills such as mathematical reasoning[13] and cultural common sense[14]. To address some of these limitations, a research area that has seen extensive growth is the integration of knowledge in LLMs (also known as retrieval augmented generation)[15], as well as the alignment of LLMs to given sets of values or opinions[16].

Behaviour of the individual

Scholars have long understood language as a gateway to the mind. Over the past century, the empirical study of language has begun to reveal people's motives, states, traits, social connections, identities and more—often tapping into aspects of the self about which the person is largely unaware[17]. Today, a growing body of research is pointing to ways in which language can be used to reliably quantify the individual person's psychology, ranging from moment-to-moment changes in psychological states (such as emotions and cognitive processes) to broad, holistic and systematic variations between people that are key to defining who we are as people (for example, identity, personality, values and lifestyles).

Personality

'Personality' refers to the relatively stable ways that individuals think, feel and behave across time and contexts. Independently and together, people's traits, values and stories provide a general understanding of who they are[18]. Many, if not most, of these dimensions have important links to natural language[4].

In the field of personality, traits are distinct characteristics. The dominant trait approach in academic psychology is the Big Five model, which assumes that most self-reported traits can be placed within a five-factor space: openness, conscientiousness, extraversion, agreeableness and neuroticism[19,20].

BOX 1

Terminology

NLP: a field that combines computational linguistics with statistical and machine learning models to understand language.

Dictionary-based (or lexicon-based) methods: text analysis approaches that rely on mappings between words and a set of corresponding predefined categories.

Machine learning: a group of computational and statistical modelling approaches that enable computer systems to recognize patterns and generalize from existing data to new data without explicit instructions.

Neural networks (also referred to as deep learning): a class of machine learning models, loosely based on interconnected neurons, that rely on connected layers of linear and nonlinear transformations of data and whose parameters are updated via the backpropagation algorithm.

Text encoder: a model that transforms raw text into a numeric representation that can then be used for analysis, clustering, input to a text classifier and so on.

Representation learning: methods for transforming words or entities into high-dimensional vectors in a way that captures their contextual and semantic relationships.

Word embeddings: vector representations of words.

Contextual embeddings: word embeddings that differ on the basis of the context of the word.

LLMs: models based on contextual embeddings that can be used to generate text.

Latent Dirichlet allocation: a statistical model to automatically extract topics from text.

The majority of NLP-driven personality research is built on this framework. For example, computational studies of language that correlate self-reported Big Five scores often report that people scoring higher in extraversion tend to talk more, focus more on social topics and use a greater number of words indicative of positive emotions[21–23]. Other studies of personality disorders adopt a similar approach, bringing NLP methods to bear on the psychological mechanisms underlying maladaptive interpersonal traits[24–26]. Other NLP work can differentiate psychological constructs such as empathy from compassion[27] and depression from loneliness[28].

Earlier work often used established psychological measures of language paired with machine learning algorithms to estimate personality traits from real-world data[29], but over the years such methods have become heavily vocabulary driven, ranging from the use of *n*-grams to topics derived by topic modelling methods such as latent Dirichlet allocation[30] to the most recent innovations in contextual embeddings[31]. Future work will probably see similar advances as it extends such methods to more objective forms of personality assessment, such as behavioural and life outcome data[32,33].

Values and behaviours

A related approach to personality explores how people navigate their daily lives in terms of goals, motives, values and personal strategies. Personal values represent the core beliefs and guiding principles that shape an individual's attitudes, behaviours and decision-making processes.

https://doi.org/10.1038/s41562-024-01938-0

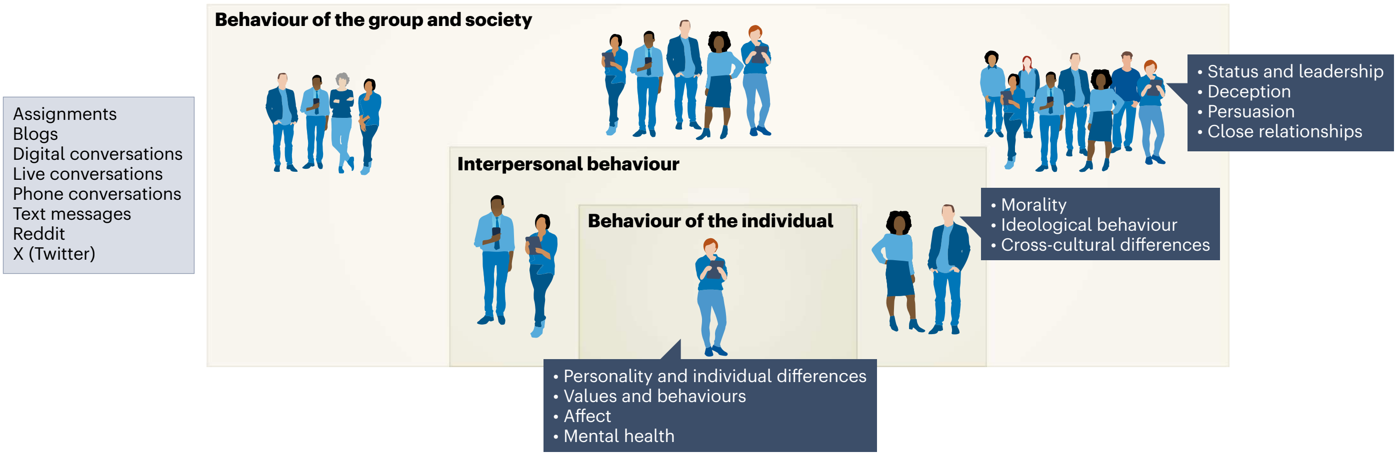

Fig. 1 | Layers of human behaviours. An overview of the paper, covering the inference of behaviours at the individual, interpersonal, group and society levels.

To uncover values from language, earlier NLP methods relied on analyses of open-ended responses to prompts eliciting personal concerns[34] or those crafted to promote reflections on personal values[35]. Alternative approaches consist of topic modelling methods, such as the meaning extraction method[36], which relies on identifying patterns in word co-occurrence to reveal recurring themes within texts. Compared with traditional forced-choice surveys, this open vocabulary approach often results in a set of value dimensions that exhibit meaningful correlations with a wider range of everyday behaviours as measured in both behaviours extracted from self-reflective journaling and naturalistic instances of social media behaviours[35]. Subsequent work aimed to develop general-purpose dictionary-based resources[37,38], or more complex representations using deep-learning-based text encoders[39,40].

Similar strategies are currently underway for abstracting individual patterns of behaviour (for example, voting behaviours and health behaviours) at the community and society levels. For instance, analyses of language through social media can reveal high-level institutional processes such as political leadership[41], or general patterns of human behaviour that reflect adaptations to external demands and pressures, such as mental health support-seeking[42], addiction[43] and violence[44].

Thinking styles

Just as people differ in their traits and broader values, they also vary in the ways they think. Thinking styles can be reflected in the degree to which people rely on logical and analytic thinking and the degree to which they naturally construct stories to understand their world. Psychologically, the ability to build a coherent narrative about who they are and where they came from is generally associated with better adjustment[18]. There is a vast literature about the value of writing about emotional upheavals and how writing narratives about these experiences is associated with improved psychological and physical health[45–48].

Early research on language-based markers of narrative by Graesser and his colleagues relied on latent semantic analysis and other methods to identify features of coherent narratives[49]. His later development of Coh-Metrix[50] popularized his technique to identify narrativity, which was helpful in the assessment of texts and in early attempts to evaluate narrativity in student essays. Other researchers have attempted to identify narrative progression by tracking the emotional arcs within stories[51,52]. Another approach has been to focus more on the cognitive shifts that unfold over the course of a story with particular focus on the language of a story's climax[53].

The development of LLMs is ushering in a new perspective on human-like intuitive behaviours[54], including the construction of narratives[55–57] that are often indistinguishable from stories written by people. From a psychology perspective, the challenge for future research will be to determine what features of a story best reflect the personality of the writer—and, more broadly, the degree to which future LLMs write our stories for us.

Affect

Affect has an essential role in shaping our decisions, relationships and overall well-being, influencing the way we perceive and interact with the world. Numerous studies in psychology have shown how emotion language[58,59], self-reported feelings, the neural substrate of emotions[60,61] and perceptions of emotion by others are related[62].

Research in NLP has primarily focused on emotion recognition and sentiment analysis[63], where 'sentiment' is defined as a coarse division of positive and negative affective states. The two seminal papers in sentiment analysis were simultaneously published in 2002 and involved categorizing reviews on the basis of their sentiment[64,65], building off earlier work that explored the sentiment of phrases using syntactic rules[66]. Soon afterwards, the first large digital lexicons of emotions (WordNet Affect[67]) and sentiment (SentiWordNet[68]) were developed, and annotated datasets have been introduced[69].

Earlier research heavily relied on lexicons used in conjunction with rule-based systems checking for the presence of affective words[70–72]. They were soon replaced by statistical methods trained on data drawn from product reviews, blogs or other online sources. Much of this work has leveraged neural approaches, including distributional embeddings and techniques for representation learning, which involve methods for transforming words or entities into high-dimensional vectors in a way that captures their contextual and semantic relationships, including, among others, algorithms such as recurrent neural networks[73], attention[74,75] and adversarial learning[76]. Work currently underway is considering affective reasoning[77,78], which aims for a deeper understanding of the implications of affective state, as well as the integration of sentiment and emotion analysis methods into end applications or even emotion-aware robots[79,80].

Mental health

A separate set of methods has emerged to infer psychological states related to mental health. In the 1980s, researchers first attempted to infer mental health diagnoses through computational methods[81]. Laboratory-based studies of language associated with depression continued into the 2000s, with studies in psychology using computational tools[82] to build on and confirm earlier findings on the heightened use of first-person singular pronouns by individuals with depression[83].

In the past decade, the quantity of data and the power of NLP tools have increased. Three studies stand out as highly influential to this shift: a data-driven analysis of college-related essays written by students who experienced depression, where 'I' words were found to be the best predictors of depression[82]; a study demonstrating the feasibility of predicting depression prior to its onset using Twitter data and the potential value of tools to identify people who need to be connected to help[84]; and finally, a new data collection method where pattern-matching was used to identify social media users with clinical depression[85].

Building on these studies, work in NLP has expanded the scope of language sources and methods to identify depression and anxiety[86], including the use of NLP and temporal models to explore behavioural patterns on mental health forums during major events[87,88]. Recent work has adopted modern NLP methods based on contextualized embeddings[89] and adopted additional signals beyond text, such as temporal information when detecting depression[90,91] and audio for the detection of stress[92]. Studies are also underway to use modern methods for both more accurate and more interpretable mental health predictions[93–96].

Interpersonal behaviour

Language is contextual. In conversations with others, we adjust our speaking style depending on our shared backgrounds, our relative status, our conversational goals and the degree to which we like or trust them. It is often not what we say that conveys these styles but how we speak. Indeed, some of the most promising advances in computational language research have explored markers of status and leadership, deception and the dynamics of close relationships.

Status and leadership

Humans form social hierarchies that help them to negotiate their interactions with others. We naturally defer to those with higher status and, at the same time, adjust our interactions with people of lower status. Even when two strangers talk for the first time, they quickly can discern their relative status.

Some languages, such as Korean and Japanese, have linguistic markers of status embedded so deeply that it is almost impossible for two strangers to have a conversation without knowing their relative status[97,98]. In most other common languages (Hindi, Russian, Mandarin and Spanish), relative status can be conveyed through the use of a formal versus informal 'you', a distinction that has largely disappeared in English[99].

Studies in psychology have discovered subtle word markers of social status in analyses of everyday spoken interactions as well as written correspondence. Across multiple studies of two-person interactions, the individual who uses fewer first-person singular pronouns (for example, 'I', 'me' and 'my') is the one with higher social status[100]. As the disparity in relative power increases, so does the difference in 'I'-word usage[101]. Lower rates of 'I' words are also apparent in texts written by people who are older and have higher education[102].

Closely allied with status and power is the nature of leadership. Consistent with the status hierarchy findings, people who assume leadership positions drop in their use of 'I' words. Interestingly, this pattern holds for people who are randomly assigned to leadership positions in laboratory studies[100]. In recent years, many NLP studies have been conducted using big data methods to study CEOs' language during quarterly earnings calls[103] or the language of US presidents or other world leaders over time[104,105]; these studies have confirmed the previous findings on the use of first-person pronouns.

Deception

Interpersonal relationships, including status and leadership, are often based on trust, but deceptive interactions are also common. One of the earliest comprehensive analyses of computational methods for deception detection identified as many as 79 linguistic deception cues and obtained a robust analysis of verbal deceptive indicators[106]. Following those early studies, later work in NLP and social psychology has covered both in-person and online deception. Interpersonal deception has been studied in personal essays[107,108], legal statements[109] and police interrogations[110]. Online deception studies have covered a variety of communication outlets, including email[111], social media platforms[112], dating profiles[113], identity deception[114], consumer review sites[115], online games[116] and news[117,118]. Recent developments in LLMs have motivated the need for computational methods that can also detect machine-generated deception[119].

Early NLP approaches on automatic deception detection focused on statistical text analysis to identify verbal cues associated with deceptive behaviour, which included basic linguistic representations such as counts of words and sentences, word diversity, positive and negative words and self-references[107,120], and more complex linguistic features derived from syntactic trees and part-of-speech tags[121,122]. Other research explored the inclusion of psycholinguistic aspects related to the deception process using lexicon-based resources such as the LIWC lexicon to build automatic deception models[123,124]. Although most approaches initially leveraged these linguistic cues using statistical machine learning approaches to build deception detection models, later work used deep learning approaches and word-vector representations (for example, word embeddings) to add semantic information.

Persuasion

A behaviour dimension orthogonal to trust and deception is persuasion, which involves an individual or entity trying to induce another party (the persuadee) into believing or disbelieving an idea or performing an action. The NLP community has been increasingly interested in the automatic identification of persuasive discourse to understand how people express and form opinions, how to represent them and ultimately how to predict whether a given text is persuasive or not, in areas such as congressional debates, online debate forums and news.

Early work focused on linguistic characteristics of persuasive text to study what makes an argument convincing[125,126]. Other research explored aspects related to the view holder and the audience, from several perspectives, such as prior beliefs and personality[127,128]. More recent work has addressed the task of generating persuasive dialogue focusing on enhancing the persuasiveness of a message. These systems often seek to change users' thoughts, opinions or behaviours through natural conversations by influencing their cognitive and emotional responses. Work on this area has incorporated aspects such as politeness, empathy and emotion[129–131] to improve users' acceptance of specific arguments—for instance, persuading individuals to improve their diet or health behaviours[132,133], participate in charity events[134] or counter vaccine misinformation[135].

Close relationships

A central topic in the social sciences involves close relationships. Some of the earliest work on analysing language to identify how two people are connecting with one another relied on the similarity of the content of their speech. Later work focused on the linguistic style of speech, assessed through the use of function words, including pronouns, prepositions, articles and related short and frequently used words that supplement and shape language content. By calculating the relative use of function words between two or more texts, it is possible to determine the degree to which the texts match. Across multiple studies, language style matching was applied to the language of two people to determine the degree that they were socially connected[136–138], or even as a predictor of young dating relationships[139].

With the rise of social media platforms, NLP methods allow us to track close relationships over months and years at a scale that was unimaginable a decade ago. A recent study on thousands of social media posts identified and analysed the language of people who underwent a major relationship breakup[140]. Over 6,800 people who had posted about their own breakup on a breakup subreddit were studied. In Fig. 2, the date of each person's first post on the breakup is at week 0. All participants' posts across all of Reddit from the year before to the year after their first post were analysed and aggregated by week.

 https://doi.org/10.1038/s41562-024-01938-0

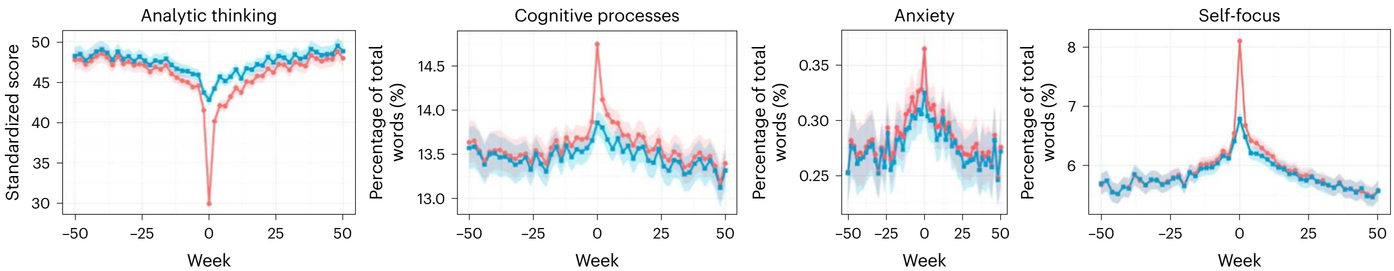

Fig. 2 | Change in language usage from one year before to one year after a first breakup Reddit post. Analytic thinking is a factor-analytically derived metric of logical, formal and hierarchical thinking (scaled from 0 to 100). 'Cognitive' refers to word dictionaries reflecting the act of 'working through' a problem. Anxiety words reflect words related to anxiety, nervousness or fear. 'I' words (self-focus) are based on first-person singular pronouns. The cognitive, anxiety and 'I'-word variables reflect the percentage of total words within each post, using language dimensions based on the LIWC2015 dictionary.

The red line depicts all Reddit posts, and the blue line reflects only those posts that were not in a relationship subreddit. The separate graphs reflect analytic or logical thinking, cognitive processing or working through thinking styles, anxiety and self-focus. Across the various text dimensions, signs of the impending breakup begin to emerge between six weeks and four months before the breakup, and the effects last, on average, for about six months after the first breakup post. The findings highlight the power of computational methods to uncover patterns over thousands of occurrences of a social behaviour, which was not possible before.

Behaviour of the group and society

People naturally seek out groups. Across the lifespan, we routinely join and often leave multiple in-person and virtual groups. With the advent of large-scale text analysis, we can now track the language of people in virtual groups in a way that reveals some of the internal dynamics of group processes—including the moral values and norms of the group, their ideological behaviours or even how different groups compare among themselves.

Morality

Moral judgements, values and norms are not just individual cognitive functions but are deeply entrenched in the fabric of group interactions and collective identities. Social identity theory finds that group membership shapes individual moral perspectives[141]. Moral foundations theory posits that moral reasoning is influenced by innate, modular foundations that are heavily shaped by cultural and social contexts[142]. The group-level dynamics of morality have also been explored through the lens of intergroup conflict, cooperation and contact. The classic Robbers Cave experiment highlighted how completely randomly assigned social allegiances can sway moral decisions and intergroup attitudes[143]; recent work on intergroup contact has additionally shown that moral judgements are often influenced by our exposure to more diverse social groups[144,145].

Building on this understanding of morality within group psychology, linguistic analysis offers a nuanced lens to detect and interpret the moral leanings of individuals and groups. The way people express themselves—their choice of words, metaphors and narratives—can be an indicator of their moral frameworks and allegiances[146]. With advancements in NLP technology, there has emerged a variety of models for automatically categorizing the types of moral foundations underlying a piece of text[147,148]. These computational linguistic analyses have used diverse text sources, including partisan news articles[149,150], microblog political discourse[151] and Twitter[152–154]. Further research extends from general coarse-grained moral value classification to more nuanced analyses of stances towards political entities[155] and cross-domain classification of moral values[156]. These advances in NLP research on morality are paving the way for high-level analyses of how morality changes[157] and impacts society[158].

Ideological behaviour

Morality can shape ideological behaviour, guiding individuals in their adherence to principles and belief systems. Researchers have tracked how ideological leanings influence the ways people process information[159]. In the early stages, political scientists recognized the untapped potential of NLP methods in harnessing text as an invaluable data source. This led to a substantial integration of computational methods into political analysis[160–163], a confluence of disciplines that fostered the emergence of a robust 'text-as-data' community within political science[160,164].

Early work on the inference of policy positions through textual analysis involved a meticulous examination of political texts, focusing on topic detection and the analysis of stylistic elements that shape the political narrative. Resources such as legislative speech[165], Senate press releases[166] and electoral manifestos[167] have been instrumental in these endeavours. With the availability of various political texts, NLP models have enabled automatic classification of ideology in news articles[168,169], political speech[170,171] and even social media posts[162,172] and academic writings[173].

Currently, various laboratories are pretraining specific language models on political text corpora to enable better ideology identification[174,175], together with techniques to make models perform better on unfamiliar text sources[176]. The inherently metaphor-rich and emotionally charged nature of political rhetoric poses unique challenges for NLP technologies[177], necessitating the development of specialized models and approaches[178,179]. Additionally, there is growing concern regarding the ethics of LLM usage. One example is political microtargeting with LLMs to generate personalized persuasive text[163,180]. Another aspect is that these models might be limited due to ideological biases present in NLP models' training data. Studies have revealed that these biases, often rooted in natural text written by humans with different ideologies, have leaked into NLP model behaviours, affecting their accuracy[181] and fairness[182].

Cross-cultural differences

People's cultural backgrounds can shape their values, attitudes and ideology in ways that impact behaviour[183,184]. Computational cross-cultural analyses of language have led to many insights into cultural differences, addressing slang terms[185] or emotional expressions and markers of depression[186], the linguistic features of politeness and taboos[187], expectations and stereotypes[188], and the attributes of social roles[189]. It is worth noting that many studies applying NLP methods to study human behaviours are limited to language that comes solely from WEIRD (Western, educated, industrialized, rich and democratic) samples that are not representative of the world's population[190]. This can have a detrimental effect on the generalizability of research findings even in the field of artificial intelligence[191].

Topic modelling analyses of open-ended surveys and blogs have found that cultural background has a critical role in moderating the relationships between values and everyday activities[192]. For instance,

 https://doi.org/10.1038/s41562-024-01938-0

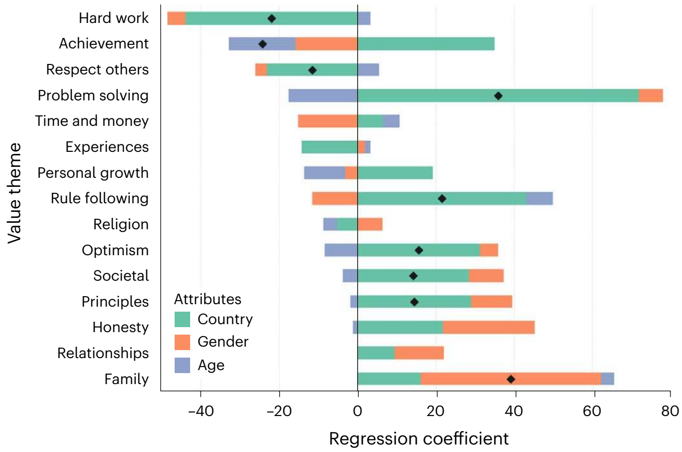

Fig. 3 | Data-driven cross-cultural values and their group associations inferred from blog text samples from the USA and India. Negative values indicate that theme usage was more associated with the USA, men and younger respondents for the country, gender and age variables, respectively. Positive values indicate that theme usage was associated with India, women and older respondents. The black diamonds indicate significant differences between categories ($P < 0.001$).

Fig. 3 illustrates the values that were identified in these analyses, along with their association with culture, gender and age as captured through a linear regression model where value theme usage was predicted from only author-level demographic variables.

Lexicon-based approaches have also been used to study short-text responses from respondents in a variety of countries[193], finding that measurements of many of the values expressed through language were positively correlated with quantitative data obtained from the World Values Survey[194]. Later work analysed the information encoded in pretrained multilingual language models such as multilingual BERT[9] and found that these models also weakly capture cross-cultural differences as measured by the World Values Survey[195].

Challenges and future directions

Our ability to analyse large corpora of text across millions of people is providing a tool to help us to understand and predict human behaviour with a precision and at a scale never previously imagined. We are witnessing a paradigm shift in the computer and social sciences that is reverberating across societies all over the world[196]. Many of our recent discoveries raise both exciting and sometimes disturbing questions about the very nature of data, theory, privacy and the roles language analyses may have in our lives[197].

Data use and unintended consequences

The inventions of the printing press and the automobile transformed the entire world. Despite their great advancements at the time, these breakthroughs ultimately contributed to wars, nationalism and cultural polarization, and climate change. The dual-use problem—wherein an invention leads to unintended uses that can cause great harm—is inherent in almost every finding we have discussed in this Review. We are now gradually learning how to handle technology with dual use. The active and growing field of ethical artificial intelligence[198] is uncovering the weak points of our technologies with strategies to address them. Among others, this includes clear ethical statements associated with research publications and, at a broader level, governmental regulations regarding how personal data can and cannot be used[199] or what are unacceptable applications of technology[200].

Data collection traps

Data-hungry big data and deep learning approaches require extremely large-scale datasets, which allow for the training of complex and powerful models. Collecting data at scale from social media APIs often results in unreliable content that may not represent the platform or the population at large[201]. Because of the difficulty of checking the quality of such large datasets, a host of problems are likely, including accepting data that are identifiable[202], unreliable[203] and explicitly prohibited[204]. Moving forward, privacy concerns in text analysis may become even more difficult to trace given the potential for LLMs to leak personally identifying information that was present in their training data, necessitating tools that are able to help individuals to understand how their data are being used to train these models[205].

Bias

Word representations[206] and language models[207] have been shown to replicate and even amplify[208] the social biases and stereotypes that exist in their training data. These biases can be particularly strong for certain demographic groups, such as intersectional identities[209]. When models are trained to infer human behaviour, underlying bias in the training data (either for the downstream models or for lower-level models such as LLMs) can lead to unreliable models that perform poorly or are even harmful when applied to certain groups of people. For example, language-based classifiers to predict whether users have depression or post-traumatic stress disorder were found to perform on par with classifiers that used inferred age as the only input variable[210]. Awareness of these potential confounds and the resulting model and data biases can help to address them, so that the NLP models of human behaviour do not discriminate and are beneficial to all users.

Interpretability and transparency

NLP systems are becoming increasingly good at predicting human behaviours. However, as their predictive powers increase, our ability to understand the inner workings of our tools is decreasing. This is particularly problematic when it comes to inferring human behaviour from language. A simple behaviour prediction (for example, that a particular piece of text is deceptive) does not provide any insights into what linguistic features may be driving this prediction. Digging deeply into the algorithm for deception detection may yield the finding that the text was particularly low in 'I' words. This would be an unsatisfying answer unless we knew that 'I' words signal self-reflection, something that deceptive speakers avoid. As we move forward, it is important to acknowledge that interpretable models contribute to fairness, accountability and ethical deployment and facilitate compliance with regulations and ethical standards, especially in fields where clear explanations are crucial (such as health care).

Connecting social and computer science

On the surface, social and computer scientists are interested in language for similar reasons. Both believe that understanding the ways people use language can help us to understand and predict human behaviour. Despite these seemingly similar views, collaborations between these disciplines can often be complicated. Most social psychologists, for example, mainly focus on people's behaviours and use language as a way of understanding how people think and feel. By contrast, most computer scientists aim to predict behaviours—that is, whereas the psychologist wants to understand the behaviour, the computer scientist wants to predict it.

A good example of the different approaches concerns the different ways women and men use common function words. Counter to many people's expectations, women use more 'I' words, social words (for example, 'he', 'she' and 'they') and cognitive words (for example, 'think', 'wonder' and 'understand') than men do. Men, by contrast, tend to use more articles ('a', 'an' and 'the') and prepositions (for example, 'to', 'of' and 'for') than women. Interestingly, there are very few sex differences for 'we' words and emotion words[211]. For a psychologist, this information tells us how women and men differ in looking at their worlds. Women tend to be more self-reflective, socially oriented and cognitively engaged about social topics. Men are more focused on objects and things (articles and prepositions are typically used when referring to concrete nouns). Computer scientists, however, look at these results and ask whether these groups of words will help their predictions in identifying the gender of the person who generated the written or spoken text. Working together, social and computer scientists can maximize their understanding of language and prediction of behaviour at the same time.

On the horizon, we see the new wave of research brought up by the rise of LLMs, which are increasingly entering many areas of life. They are bringing a new set of questions that consider not only human language but also the language automatically generated by the LLMs. In addition to using these models to infer human behaviour[212–214], we increasingly see methods referred to as 'prompting' and 'probing' that borrow strategies produced by decades of research on inferring human behaviour from language to gain insights into the NLP systems themselves[215,216]. We are thus continuing the virtuous cycle of discovery: the early work in psychology has fuelled research in NLP, which in turn has led to new discoveries in psychology, which are now being used to gain new insights into the NLP systems themselves. With the two fields informing and propelling each other forward, the future of this research space is bright.

References

1. Gentner, D. & Goldin-Meadow, S. *Language in Mind: Advances in the Study of Language and Thought* (MIT Press, 2003).
2. Dubriwny, T. N. Constructing breast cancer in the news: Betty Ford and the evolution of the breast cancer patient. *J. Commun. Inq.* **33**, 104–125 (2009).
3. Pentland, A. *Social Physics: How Good Ideas Spread—the Lessons from a New Science* (Penguin, 2014).
4. Boyd, R. L. & Schwartz, H. A. Natural language analysis and the psychology of verbal behavior: the past, present, and future states of the field. *J. Lang. Soc. Psychol.* **40**, 21–41 (2021).
5. OpenAI et al. GPT-4 technical report. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2303.08774 (2023).
6. Touvron, H. et al. LLaMA: open and efficient foundation language models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2302.13971 (2023).
7. Jiang, A. Q. et al. Mistral 7B. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2310.06825 (2023).
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **26**, 3111–3119 (2013).
9. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers)* (eds Burstein, J. et al.) 4171–4186 (Association for Computational Linguistics, 2019).
10. Reimers, N. & Gurevych, I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In *Proc. 2019 Conf. Empir. Methods Nat. Lang. Process.,* 3980–3990 (Association for Computational Linguistics, 2019).
11. Gilardi, F., Alizadeh, M. & Kubli, M. ChatGPT outperforms crowd workers for text-annotation tasks. *Proc. Natl Acad. Sci. USA* **120**, e2305016120 (2023).
12. Argyle, L. P. et al. Out of one, many: using language models to simulate human samples. *Polit. Anal.* **31**, 337–351 (2023).
13. Hong, P. et al. Caught in the quicksand of reasoning, far from AGI summit: evaluating LLMs' mathematical and coding competency through ontology-guided interventions. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2401.09395 (2024).
14. Shen, S. et al. Understanding the capabilities and limitations of large language models for cultural commonsense. *Proc. 2024 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2024* (eds Duh, K. et al.) 5668–5680 (Association for Computational Linguistics, 2024).
15. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inf. Process. Syst.* **33**, 9459–9474 (2020).
16. Sun, Z. et al. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Adv. Neural Inf. Process. Syst.* **36**, 2511–2565 (2023).
17. Schönbrodt, F. D. et al. Measuring implicit motives with the picture story exercise (PSE): databases of expert-coded German stories, pictures, and updated picture norms. *J. Pers. Assess.* **103**, 392–405 (2021).
18. McAdams, D. P. *The Stories We Live By: Personal Myths and the Making of the Self* (Guilford, 1993).
19. Digman, J. M. Personality structure: emergence of the five-factor model. *Annu. Rev. Psychol.* https://doi.org/10.1146/annurev.ps.41.020190.002221 (2003).
20. John, O. P. in *Personality Psychology: Recent Trends and Emerging Directions* (eds Buss, D. M. & Cantor, N.) 261–271 (Springer, 1989).
21. Mehl, M. R., Gosling, S. D. & Pennebaker, J. W. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *J. Pers. Soc. Psychol.* **90**, 862–877 (2006).
22. Hirsh, J. B. & Peterson, J. B. Personality and language use in self-narratives. *J. Res. Pers.* **43**, 524–527 (2009).
23. Yarkoni, T. Personality in 100,000 Words: a large-scale analysis of personality and word use among bloggers. *J. Res. Pers.* **44**, 363–373 (2010).
24. Entwistle, C. & Boyd, R. L. Uncovering the social-cognitive contributors to social dysfunction in borderline personality disorder through language analysis. *J. Pers. Disord.* **37**, 444–455 (2023).
25. Entwistle, C. et al. Natural emotion vocabularies and borderline personality disorder. *J. Affect. Disord. Rep.* **14**, 100647 (2023).
26. Berry-Blunt, A. K., Holtzman, N. S., Donnellan, M. B. & Mehl, M. R. The story of 'I' tracking: psychological implications of self-referential language use. *Soc. Pers. Psychol. Compass* **15**, e12647 (2021).
27. Yaden, D. B. et al. Characterizing empathy and compassion using computational linguistic analysis. *Emotion* https://doi.org/10.1037/emo0001205 (2023).
28. Liu, T. et al. Head versus heart: social media reveals differential language of loneliness from depression. *NPJ Ment. Health Res.* **1**, 16 (2022).
29. Iacobelli, F., Gill, A. J., Nowson, S. & Oberlander, J. in *Affective Computing and Intelligent Interaction* 568–577 (Springer Berlin Heidelberg, 2011).
30. Schwartz, H. A. et al. Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS ONE* **8**, e73791 (2013).
31. Jain, D., Kumar, A. & Beniwal, R. Personality BERT: a transformer-based model for personality detection from textual data. In *Proc. International Conference on Computing and Communication Networks* 515–522 (Springer Nature Singapore, 2022).
32. Boyd, R. L., Pasca, P. & Lanning, K. The personality panorama: conceptualizing personality through big behavioural data. *Eur. J. Pers.* **34**, 599–612 (2020).
33. Jose, R. et al. Using Facebook language to predict and describe excessive alcohol use. *Alcohol. Clin. Exp. Res.* **46**, 836–847 (2022).

34. Chung, C. K., Rentfrow, P. J., & Pennebaker, J. W. in *Geographical Psychology: Exploring the Interaction of Environment and Behavior* (ed. Rentfrow, P. J.) 195–216 (American Psychological Association, 2014).
35. Boyd, R. et al. Values in words: using language to evaluate and understand personal values. *ICWSM* **9**, 31–40 (2015).
36. Chung, C. K. & Pennebaker, J. W. Revealing dimensions of thinking in open-ended self-descriptions: an automated meaning extraction method for natural language. *J. Res. Pers.* **42**, 96–132 (2008).
37. Wilson, S. R., Shen, Y. & Mihalcea, R. in *Social Informatics* 455–470 (Springer International, 2018).
38. Ponizovskiy, V. et al. Development and validation of the Personal Values Dictionary: a theory-driven tool for investigating references to basic human values in text. *Eur. J. Pers.* **34**, 885–902 (2020).
39. Wilson, S. & Mihalcea, R. Predicting human activities from user-generated content. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics* (eds Korhonen, A. et al.) 2572–2582 (Association for Computational Linguistics, 2019).
40. Sorensen, T. et al. Value kaleidoscope: engaging AI with pluralistic human values, rights, and duties. *Proc. AAAI Conf. on Artificial Intelligence* **38**, 19937–19947 (2024).
41. Jordan, K. N., Sterling, J., Pennebaker, J. W. & Boyd, R. L. Examining long-term trends in politics and culture through language of political leaders and cultural institutions. *Proc. Natl Acad. Sci. USA* **116**, 3476–3481 (2019).
42. Saha, K., Yousuf, A., Boyd, R. L., Pennebaker, J. W. & De Choudhury, M. Social media discussions predict mental health consultations on college campuses. *Sci. Rep.* **12**, 123 (2022).
43. Sarker, A., Gonzalez-Hernandez, G., Ruan, Y. & Perrone, J. Machine learning and natural language processing for geolocation-centric monitoring and characterization of opioid-related social media chatter. *JAMA Netw. Open* **2**, e1914672 (2019).
44. Ni, Y. et al. Finding warning markers: leveraging natural language processing and machine learning technologies to detect risk of school violence. *Int. J. Med. Inform.* **139**, 104137 (2020).
45. Sloan, D. M. & Marx, B. P. *Written Exposure Therapy for PTSD: A Brief Treatment Approach for Mental Health Professionals* (American Psychological Association, 2019).
46. Guo, L. The delayed, durable effect of expressive writing on depression, anxiety and stress: a meta-analytic review of studies with long-term follow-ups. *Br. J. Clin. Psychol.* **62**, 272–297 (2023).
47. Gerger, H., Werner, C. P., Gaab, J. & Cuijpers, P. Comparative efficacy and acceptability of expressive writing treatments compared with psychotherapy, other writing treatments, and waiting list control for adult trauma survivors: a systematic review and network meta-analysis. *Psychol. Med.* **52**, 3484–3496 (2021).
48. Pennebaker, J. W. Expressive writing in psychological science. *Perspect. Psychol. Sci.* **13**, 226–229 (2018).
49. Graesser, A. C., Singer, M. & Trabasso, T. Constructing inferences during narrative text comprehension. *Psychol. Rev.* **101**, 371–395 (1994).
50. Graesser, A. C. et al. Coh-Metrix measures text characteristics at multiple levels of language and discourse. *Elem. Sch. J.* **115**, 210–229 (2014).
51. Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M. & Dodds, P. S. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Sci.* **5**, 1–12 (2016).
52. Berger, J., Kim, Y. D. & Meyer, R. What makes content engaging? How emotional dynamics shape success. *J. Consum. Res.* **48**, 235–250 (2021).
53. Boyd, R. L., Blackburn, K. G. & Pennebaker, J. W. The narrative arc: revealing core narrative structures through text analysis. *Sci. Adv.* **6**, eaba2196 (2020).
54. Hagendorff, T., Fabi, S. & Kosinski, M. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nat. Comput. Sci.* **3**, 833–838 (2023).
55. Chu, H. & Liu, S. Can AI tell good stories? Narrative transportation and persuasion with ChatGPT (2023).
56. Sap, M. et al. Quantifying the narrative flow of imagined versus autobiographical stories. *Proc. Natl Acad. Sci. USA* **119**, e2211715119 (2022).
57. Begus, N. Experimental narratives: a comparison of human crowdsourced storytelling and AI storytelling. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2310.12902 (2023).
58. Vine, V., Boyd, R. L. & Pennebaker, J. W. Natural emotion vocabularies as windows on distress and well-being. *Nat. Commun.* **11**, 4525 (2020).
59. Ong, D. C., Zaki, J. & Goodman, N. D. Computational models of emotion inference in theory of mind: a review and roadmap. *Top. Cogn. Sci.* **11**, 338–357 (2019).
60. Mattavelli, G., Celeghin, A. & Mazzoni, N. *Explicit and Implicit Emotion Processing: Neural Basis, Perceptual and Cognitive Mechanisms* (Frontiers Media SA, 2021).
61. Barrett, L. F., Mesquita, B., Ochsner, K. N. & Gross, J. J. The experience of emotion. *Annu. Rev. Psychol.* **58**, 373–403 (2007).
62. Lange, J., Heerdink, M. W. & van Kleef, G. A. Reading emotions, reading people: emotion perception and inferences drawn from perceived emotions. *Curr. Opin. Psychol.* **43**, 85–90 (2022).
63. Poria, S., Hazarika, D., Majumder, N. & Mihalcea, R. Beneath the tip of the iceberg: current challenges and new directions in sentiment analysis research. *IEEE Trans. Affect. Comput.* **14**, 108–132 (2023).
64. Turney, P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics* 417–424 (Association for Computational Linguistics, 2002).
65. Pang, B., Lee, L. & Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. In *Proc. 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)* 79–86 (Association for Computational Linguistics, 2002).
66. Hatzivassiloglou, V. & McKeown, K. R. Predicting the semantic orientation of adjectives. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics* 174–181 (Association for Computational Linguistics, 1997).
67. Strapparava, C. & Valitutti, A. WordNet Affect: an affective extension of WordNet. In *Proc. Fourth International Conference on Language Resources and Evaluation (LREC'04)* (eds Lino, M. T. et al.) (European Language Resources Association, 2004).
68. Esuli, A. & Sebastiani, F. SentiWordNet: a publicly available lexical resource for opinion mining. In *Proc. Fifth International Conference on Language Resources and Evaluation (LREC'06)* (eds Calzolari, N. et al.) (European Language Resources Association, 2006).
69. Strapparava, C. & Mihalcea, R. SemEval-2007 Task 14: affective text. In *Proc. Fourth International Workshop on Semantic Evaluations (SemEval-2007)* 70–74 (Association for Computational Linguistics, 2007).
70. Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M. Lexicon-based methods for sentiment analysis. *Comput. Linguist.* **37**, 267–307 (2011).
71. Pérez-Rosas, V., Banea, C. & Mihalcea, R. Learning sentiment lexicons in Spanish. In *Proc. Eighth International Conference on Language Resources and Evaluation (LREC'12)* 3077–3081 (European Language Resources Association, 2012).

72. Kiritchenko, S. & Mohammad, S. Happy Accident: a sentiment composition lexicon for opposing polarity phrases. In *Proc. Tenth International Conference on Language Resources and Evaluation (LREC'16)* 1157–1164 (European Language Resources Association, 2016).
73. Socher, R. et al. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. 2013 Conference on Empirical Methods in Natural Language Processing* 1631–1642 (Association for Computational Linguistics, 2013).
74. Wang, Y., Huang, M., Zhu, X. & Zhao, L. Attention-based LSTM for aspect-level sentiment classification. In *Proc. 2016 Conference on Empirical Methods in Natural Language Processing* 606–615 (Association for Computational Linguistics, 2016).
75. Phan, H. T., Nguyen, N. T. & Hwang, D. Convolutional attention neural network over graph structures for improving the performance of aspect-level sentiment analysis. *Inf. Sci.* **589**, 416–439 (2022).
76. Karimi, A., Rossi, L. & Prati, A. Adversarial training for aspect-based sentiment analysis with BERT. In *2020 25th International Conference on Pattern Recognition (ICPR)* 8797–8803 (IEEE, 2021).
77. Ghosal, D., Shen, S., Majumder, N., Mihalcea, R. & Poria, S. CICERO: a dataset for contextualized commonsense inference in dialogues. *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2022).
78. Zhang, M., Liang, Y. & Ma, H. Context-aware affective graph reasoning for emotion recognition. In *2019 IEEE International Conference on Multimedia and Expo (ICME)* 151–156 (IEEE, 2019).
79. Korn, O., Akalin, N. & Gouveia, R. Understanding cultural preferences for social robots: a study in German and Arab communities. *J. Hum. Robot Interact.* **10**, 1–19 (2021).
80. Yang, J. et al. AI-enabled emotion-aware robot: the fusion of smart clothing, edge clouds and robotics. *Future Gener. Comput. Syst.* **102**, 701–709 (2020).
81. Oxman, T. E., Rosenberg, S. D., Schnurr, P. P. & Tucker, G. J. Diagnostic classification through content analysis of patients' speech. *Am. J. Psychiatry* **145**, 464–468 (1988).
82. Rude, S., Gortner, E.-M. & Pennebaker, J. W. Language use of depressed and depression-vulnerable college students. *Cogn. Emot.* **18**, 1121–1133 (2004).
83. Bucci, W. & Freedman, N. The language of depression. *Bull. Menninger Clin.* **45**, 334–358 (1981).
84. De Choudhury, M., Gamon, M., Counts, S. & Horvitz, E. Predicting depression via social media. *ICWSM* **7**, 128–137 (2013).
85. Coppersmith, G., Dredze, M. & Harman, C. Quantifying mental health signals in Twitter. In *Proc. Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (eds Resnik, P. et al.) 51–60 (Association for Computational Linguistics, 2014).
86. Zirikly, A., Resnik, P., Uzuner, Ö. & Hollingshead, K. CLPsych 2019 Shared Task: predicting the degree of suicide risk in Reddit posts. In *Proc. Sixth Workshop on Computational Linguistics and Clinical Psychology* (eds Niederhoffer, K. et al.) 24–33 (Association for Computational Linguistics, 2019).
87. Biester, L., Matton, K., Rajendran, J., Provost, E. M. & Mihalcea, R. Understanding the impact of COVID-19 on online mental health forums. *ACM Trans. Manage. Inf. Syst.* **12**, 1–28 (2021).
88. Park, A. & Conway, M. Longitudinal changes in psychological states in online health community members: understanding the long-term effects of participating in an online depression community. *J. Med. Internet Res.* **19**, e71 (2017).
89. Ji, S. et al. MentalBERT: publicly available pretrained language models for mental healthcare. In *Proc Thirteenth Language Resources and Evaluation Conf.* (eds. Calzolari, E. et al.) 7184–7190 (European Language Resources Association, 2022).
90. Tsakalidis, A. et al. Overview of the CLPsych 2022 Shared Task: capturing moments of change in longitudinal user posts. In *Proc. Eighth Workshop on Computational Linguistics and Clinical Psychology* (eds. Zirikly, A. et al.) 184–198 (Association for Computational Linguistics, 2022).
91. Liu, Y., Biester, L. & Mihalcea, R. Improving mental health classifier generalization with pre-diagnosis data. *ICWSM* **17**, 566–577 (2023).
92. Yao, Y., Papakostas, M., Burzo, M., Abouelenien, M. & Mihalcea, R. MUSER: Multimodal Stress Detection Using Emotion Recognition as an Auxiliary Task. In *Proc. 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds Toutanova, K. et al.) 2714–2725 (Association for Computational Linguistics, 2021).
93. Nguyen, T., Yates, A., Zirikly, A., Desmet, B. & Cohan, A. Improving the generalizability of depression detection by leveraging clinical questionnaires. In *Proc. 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Muresan, S. et al.) 8446–8459 (Association for Computational Linguistics, 2022).
94. Varadarajan, V. et al. Archetypes and entropy: theory-driven extraction of evidence for suicide risk. In *Proc. 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)* (eds Yates, A. et al.) 278–291 (Association for Computational Linguistics, 2024).
95. Lee, A., Kummerfeld, J. K., An, L. & Mihalcea, R. Micromodels for efficient, explainable, and reusable systems: a case study on mental health. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (eds Moens, M.-F. et al.) 4257–4272 (Association for Computational Linguistics, 2021).
96. Wang, Y., Inkpen, D. & Kirinde Gamaarachchige, P. Explainable depression detection using large language models on social media data. In *Proc. 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)* (eds Yates, A. et al.) 108–126 (Association for Computational Linguistics, 2024).
97. Battiste, M. in *Reclaiming Indigenous Voice and Vision* (ed. Battiste, M.) 192–208 (Univ. British Columbia Press, 2000).
98. Sohn, H.-M. *Korean Language in Culture and Society* (Univ. Hawaii Press, 2005).
99. Schiffman, H. *Linguistic Culture and Language Policy* (Routledge, 2012).
100. Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M. & Graesser, A. C. Pronoun use reflects standings in social hierarchies. *J. Lang. Soc. Psychol.* **33**, 125–143 (2014).
101. Hagiwara, N., Slatcher, R. B., Eggly, S. & Penner, L. A. Physician racial bias and word use during racially discordant medical interactions. *Health Commun.* **32**, 401–408 (2017).
102. Pennebaker, J. W. & King, L. A. Linguistic styles: language use as an individual difference. *J. Pers. Soc. Psychol.* **77**, 1296–1312 (1999).
103. Pollock, T. G., Ragozzino, R. & Blevins, D. P. Not like the rest of us? How CEO celebrity affects quarterly earnings call language. *J. Manage.* 01492063221150629 (2023).
104. Ahmadian, S., Azarshahi, S. & Paulhus, D. L. Explaining Donald Trump via communication style: grandiosity, informality, and dynamism. *Pers. Individ. Dif.* **107**, 49–53 (2017).
105. Figueiredo, S., Devezas, M., Vieira, N. & Soares, A. A psycholinguistic analysis of world leaders' discourses concerning the COVID-19 context: authenticity and emotional tone. *Int. J. Soc. Sci.* **9**, 2 (2020).
106. Hauch, V., Blandón-Gitlin, I., Masip, J. & Sporer, S. L. Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. *Pers. Soc. Psychol. Rev.* **19**, 307–342 (2015).
107. Mihalcea, R. & Strapparava, C. The lie detector: explorations in the automatic recognition of deceptive language. In *Proc. ACL-IJCNLP 2009 Conference Short Papers* 309–312 (Association for Computational Linguistics, 2009).

108. Pérez-Rosas, V. & Mihalcea, R. Cross-cultural deception detection. In *Proc. 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* 440–445 (Association for Computational Linguistics, 2014).
109. Fornaciari, T. & Poesio, M. Automatic deception detection in Italian court cases. *Artif. Intell. Law* **21**, 303–340 (2013).
110. Bachenko, J., Fitzpatrick, E. & Schonwetter, M. Verification and implementation of language-based deception indicators in civil and criminal narratives. In *Proc. 22nd International Conference on Computational Linguistics (Coling 2008)* 41–48 (Coling 2008 Organizing Committee, 2008).
111. Zhou, L., Burgoon, J. K. & Twitchell, D. P. in *Intelligence and Security Informatics* 102–110 (Springer Berlin Heidelberg, 2003).
112. Chiluwa, I. E. & Samoilenko, S. A. *Handbook of Research on Deception, Fake News, and Misinformation Online* (IGI Global, 2019).
113. Toma, C. & Hancock, J. Reading between the lines: linguistic cues to deception in online dating profiles. In *Proc. 2010 ACM Conference on Computer Supported Cooperative Work* 5–8 (ACM, 2010).
114. Pérez-Rosas, V., Davenport, Q., Dai, A. M., Abouelenien, M. & Mihalcea, R. Identity deception detection. In *Proc. Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* 885–894 (Asian Federation of Natural Language Processing, 2017).
115. Ott, M., Choi, Y., Cardie, C. & Hancock, J. T. Finding deceptive opinion spam by any stretch of the imagination. In *Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* 309–319 (Association for Computational Linguistics, 2011).
116. Girlea, C., Girju, R. & Amir, E. Psycholinguistic features for deceptive role detection in Werewolf. In *Proc. 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 417–422 (Association for Computational Linguistics, 2016).
117. Rashkin, H., Choi, E., Jang, J. Y., Volkova, S. & Choi, Y. Truth of varying shades: analyzing language in fake news and political fact-checking. In *Proc. 2017 Conference on Empirical Methods in Natural Language Processing* 2931–2937 (Association for Computational Linguistics, 2017).
118. Pérez-Rosas, V., Kleinberg, B., Lefevre, A. & Mihalcea, R. Automatic detection of fake news. In *Proc. 27th International Conference on Computational Linguistics* 3391–3401 (Association for Computational Linguistics, 2018).
119. Schuster, T., Schuster, R., Shah, D. J. & Barzilay, R. The limitations of stylometry for detecting machine-generated fake news. *Comput. Linguist. Assoc. Comput. Linguist.* **46**, 499–510 (2020).
120. Fitzpatrick, E., Bachenko, J. & Fornaciari, T. *Automatic Detection of Verbal Deception* (Springer Nature, 2022).
121. Feng, S., Banerjee, R. & Choi, Y. Syntactic stylometry for deception detection. In *Proc. 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* 171–175 (Association for Computational Linguistics, 2012).
122. Xu, Q. & Zhao, H. Using deep linguistic features for finding deceptive opinion spam. In *Proc. COLING 2012: Posters* 1341–1350 (COLING 2012 Organizing Committee, 2012).
123. Newman, M. L., Pennebaker, J. W., Berry, D. S. & Richards, J. M. Lying words: predicting deception from linguistic styles. *Pers. Soc. Psychol. Bull.* **29**, 665–675 (2003).
124. Almela, Á., Valencia-García, R. & Cantos, P. Seeing through deception: a computational approach to deceit detection in written communication. In *Proc. Workshop on Computational Approaches to Deception Detection* 15–22 (Association for Computational Linguistics, 2012).
125. Habernal, I. & Gurevych, I. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proc. 2015 Conference on Empirical Methods in Natural Language Processing* 2127–2137 (Association for Computational Linguistics, 2015).
126. Rehbein, I. On the role of discourse relations in persuasive texts. In *Proc. 13th Linguistic Annotation Workshop* 144–154 (Association for Computational Linguistics, 2019).
127. Wei, Z., Liu, Y. & Li, Y. Is this post persuasive? Ranking argumentative comments in online forum. In *Proc. 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* 195–200 (Association for Computational Linguistics, 2016).
128. Longpre, L., Durmus, E. & Cardie, C. Persuasion of the undecided: language vs. the listener. In *Proc. 6th Workshop on Argument Mining* 167–176 (Association for Computational Linguistics, 2019).
129. Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J. & Potts, C. A computational approach to politeness with application to social factors. In *Annual Meeting of the Association for Computational Linguistics* (eds Schuetze, H. et al.) 250-259 (Association for Computational Linguistics, 2013).
130. Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C. & Althoff, T. Towards facilitating empathic conversations in online mental health support: a reinforcement learning approach. In *Proc. Web Conference 2021* 194–205 (Association for Computing Machinery, 2021).
131. Samad, A. M., Mishra, K., Firdaus, M. & Ekbal, A. Empathetic persuasion: reinforcing empathy and persuasiveness in dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2022* (eds Carpuat, M. et al.) 844–856 (Association for Computational Linguistics, 2022).
132. Hunter, A., Chalaguine, L., Czernuszenko, T., Hadoux, E. & Polberg, S. Towards computational persuasion via natural language argumentation dialogues. In *KI 2019: Advances in Artificial Intelligence* 18–33 (Springer International, 2019).
133. Zhou, Y. et al. Towards enhancing health coaching dialogue in low-resource settings. In *Proc. 29th International Conference on Computational Linguistics* (eds Calzolari, N. et al.) 694–706 (International Committee on Computational Linguistics, 2022).
134. Wang, X. et al. Persuasion for good: towards a personalized persuasive dialogue system for social good. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics* (eds Korhonen, A. et al.) 5635–5649 (Association for Computational Linguistics, 2019).
135. He, B., Ahamad, M. & Kumar, S. Reinforcement learning-based counter-misinformation response generation: a case study of COVID-19 vaccine misinformation. In *Proc. ACM Web Conference 2023* 2698–2709 (Association for Computing Machinery, 2023).
136. Niederhoffer, K. G. & Pennebaker, J. W. Linguistic style matching in social interaction. *J. Lang. Soc. Psychol.* **21**, 337–360 (2002).
137. Ireland, M. E. & Pennebaker, J. W. Language style matching in writing: synchrony in essays, correspondence, and poetry. *J. Pers. Soc. Psychol.* **99**, 549–571 (2010).
138. Gonzales, A. L., Hancock, J. T. & Pennebaker, J. W. Language style matching as a predictor of social dynamics in small groups. *Commun. Res.* **37**, 3–19 (2010).
139. Ireland, M. E. et al. Language style matching predicts relationship initiation and stability. *Psychol. Sci.* **22**, 39–44 (2011).
140. Seraj, S., Blackburn, K. G. & Pennebaker, J. W. Language left behind on social media exposes the emotional and cognitive costs of a romantic breakup. *Proc. Natl Acad. Sci. USA* **118**, e2017154118 (2021).
141. Tajfel, H., Turner, J. C., Austin, W. G. & Worchel, S. in *Organizational Identity: A Reader* (eds Hatch, M. J. & Schultz, M.) 56–65 (Oxford Academic, 1979).

142. Haidt, J. *The Righteous Mind: Why Good People Are Divided by Politics and Religion* (Vintage, 2013).
143. Sherif, M., Harvey, O. J., White, B. J., Sherif, C. W., & Campbell, D. T. *The Robbers Cave Experiment: Intergroup Conflict and Cooperation* (Wesleyan Univ. Press, 1988).
144. Crystal, D. S., Killen, M. & Ruck, M. It is who you know that counts: intergroup contact and judgments about race-based exclusion. *Br. J. Dev. Psychol.* **26**, 51–70 (2008).
145. Dovidio, J. F., Gaertner, S. L., Niemann, Y. F. & Snider, K. Racial, ethnic, and cultural differences in responding to distinctiveness and discrimination on campus: stigma and common group identity. *J. Soc. Issues* **57**, 167–188 (2001).
146. Lakoff, G. *Moral Politics: How Liberals and Conservatives Think* (Univ. Chicago Press, 2016).
147. Sagi, E. & Dehghani, M. Measuring moral rhetoric in text. *Soc. Sci. Comput. Rev.* **32**, 132–144 (2014).
148. Dehghani, M., Ekhtiari, H., Forbus, K., Gentner, D., & Sachdeva, S. The role of cultural narratives in moral decision making. In *Proc. of the Annual Meeting of the Cognitive Science Society* **31**, 31 (2009).
149. Fulgoni, D., Carpenter, J., Ungar, L. & Preoţiuc-Pietro, D. An empirical exploration of moral foundations theory in partisan news sources. In *Proc. Tenth International Conference on Language Resources and Evaluation (LREC'16)* (eds Calzolari, N. et al.) 3730–3736 (European Language Resources Association, 2016).
150. Shahid, U., Di Eugenio, B., Rojecki, A. & Zheleva, E. Detecting and understanding moral biases in news. In *Proc. First Joint Workshop on Narrative Understanding, Storylines, and Events* (eds Bonial, C. et al.) 120–125 (Association for Computational Linguistics, 2020).
151. Johnson, K. & Goldwasser, D. Classification of moral foundations in microblog political discourse. In *Proc. 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Gurevych, I. & Miyao, Y.) 720–730 (Association for Computational Linguistics, 2018).
152. Hoover, J. et al. Moral Foundations Twitter Corpus: a collection of 35k tweets annotated for moral sentiment. *Soc. Psychol. Pers. Sci.* **11**, 1057–1071 (2020).
153. Roy, S., Pacheco, M. L. & Goldwasser, D. Identifying morality frames in political tweets using relational learning. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing* (eds Moens, M.-F. et al.) 9939–9958 (Association for Computational Linguistics, 2021).
154. Mooijman, M., Hoover, J., Lin, Y., Ji, H. & Dehghani, M. Moralization in social networks and the emergence of violence during protests. *Nat. Hum. Behav.* **2**, 389–396 (2018).
155. Roy, S. & Goldwasser, D. Analysis of nuanced stances and sentiment towards entities of US politicians through the lens of moral foundation theory. In *Proc. Ninth International Workshop on Natural Language Processing for Social Media* (eds Ku, L.-W. & Li, C.-T.) 1–13 (Association for Computational Linguistics, 2021).
156. Liscio, E., Dondera, A., Geadau, A., Jonker, C. & Murukannaiah, P. Cross-domain classification of moral values. In *Findings of the Association for Computational Linguistics: NAACL 2022* (eds Carpuat, M. et al.) 2727–2745 (Association for Computational Linguistics, 2022).
157. Ramezani, A., Zhu, Z., Rudzicz, F. & Xu, Y. An unsupervised framework for tracing textual sources of moral change. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (eds Moens, M.-F. et al.) 1215–1228 (Association for Computational Linguistics, 2021).
158. Rezapour, R., Shah, S. H. & Diesner, J. Enhancing the measurement of social effects by capturing morality. In *Proc. Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (eds Balahur, A. et al.) 35–45 (Association for Computational Linguistics, 2019).
159. Chaiken, S. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *J. Pers. Soc. Psychol.* **39**, 752–766 (1980).
160. Grimmer, J. & Stewart, B. M. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Polit. Anal.* **21**, 267–297 (2013).
161. Jin, Z. & Mihalcea, R. in *Handbook of Computational Social Science for Policy* (eds Bertoni, E. et al.) 141–162 (Springer International, 2023).
162. Törnberg, P. ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2304.06588 (2023).
163. Simchon, A., Edwards, M. & Lewandowsky, S. The persuasive effects of political microtargeting in the age of generative artificial intelligence. *PNAS Nexus* **3**, gae035 (2024).
164. Laver, M., Benoit, K. & Garry, J. Extracting policy positions from political texts using words as data. *Am. Polit. Sci. Rev.* **97**, 311–331 (2003).
165. Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H. & Radev, D. R. How to analyze political attention with minimal assumptions and costs. *Am. J. Polit. Sci.* **54**, 209–228 (2010).
166. Grimmer, J. A Bayesian hierarchical topic model for political texts: measuring expressed agendas in Senate press releases. *Polit. Anal.* **18**, 1–35 (2010).
167. Menini, S., Nanni, F., Ponzetto, S. P. & Tonelli, S. Topic-based agreement and disagreement in US electoral manifestos. In *Proc. 2017 Conference on Empirical Methods in Natural Language Processing* (eds Palmer, M. et al.) 2938–2944 (Association for Computational Linguistics, 2017).
168. Young, L. & Soroka, S. Affective news: the automated coding of sentiment in political texts. *Polit. Commun.* **29**, 205–231 (2012).
169. Baly, R., Da San Martino, G., Glass, J. & Nakov, P. We can detect your bias: predicting the political ideology of news articles. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (eds Webber, B. et al.) 4982–4991 (Association for Computational Linguistics, 2020).
170. Hirst, G., Riabinin, Y. & Graham, J. Party status as a confound in the automatic classification of political speech by ideology. In *JADT 2010: 10th International Conference on Statistical Analysis of Textual Data* (eds Bolasco, S. et al.) 731-742 (2010).
171. Diermeier, D., Godbout, J.-F., Yu, B. & Kaufmann, S. Language and ideology in Congress. *Br. J. Polit. Sci.* **42**, 31–55 (2012).
172. Preoţiuc-Pietro, D., Liu, Y., Hopkins, D. & Ungar, L. Beyond binary labels: political ideology prediction of Twitter users. In *Proc. 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Barzilay, R. & Kan, M.-Y.) 729–740 (Association for Computational Linguistics, 2017).
173. Jelveh, Z., Kogut, B. & Naidu, S. Detecting latent ideology in expert text: evidence from academic papers in economics. In *Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (eds Moschitti, A. et al.) 1804–1809 (Association for Computational Linguistics, 2014).
174. Kawintiranon, K. & Singh, L. PoliBERTweet: a pre-trained language model for analyzing political content on Twitter. In *Proc. 13th Language Resources and Evaluation Conference* (eds Calzolari, N. et al.) 7360–7367 (European Language Resources Association, 2022).
175. Liu, Y., Zhang, X. F., Wegsman, D., Beauchamp, N. & Wang, L. POLITICS: pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Association for Computational Linguistics: NAACL 2022* (eds Carpuat, M. et al.) 1354–1374 (Association for Computational Linguistics, 2022).

176. Chen, C., Walker, D. & Saligrama, V. Ideology prediction from scarce and biased supervision: learn to disregard the 'what' and focus on the 'how'! In *Proc. 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Rogers, A. et al.) 9529–9549 (Association for Computational Linguistics, 2023).
177. Huguet Cabot, P.-L., Dankers, V., Abadi, D., Fischer, A. & Shutova, E. The pragmatics behind politics: modelling metaphor, framing and emotion in political discourse. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (eds Cohn, T. et al.) 4479–4488 (Association for Computational Linguistics, 2020).
178. Bhatia, S. & P, D. Topic-specific sentiment analysis can help identify political ideology. In *Proc. 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (eds Balahur, A. et al.) 79–84 (Association for Computational Linguistics, 2018).
179. Shen, Q. & Rose, C. What sounds 'right' to me? Experiential factors in the perception of political ideology. In *Proc. 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (eds Merlo, P. et al.) 1762–1771 (Association for Computational Linguistics, 2021).
180. Hackenburg, K. & Margetts, H. Evaluating the persuasive influence of political microtargeting with large language models. *Proc. Natl Acad. Sci. USA* **121**, e2403116121 (2024).
181. Guo, M., Hwa, R., Lin, Y.-R. & Chung, W.-T. Inflating topic relevance with ideology: a case study of political ideology bias in social topic detection models. In *Proc. 28th International Conference on Computational Linguistics* (eds Scott, D. et al.) 4873–4885 (International Committee on Computational Linguistics, 2020).
182. Feng, S., Park, C. Y., Liu, Y. & Tsvetkov, Y. From pretraining data to language models to downstream tasks: tracking the trails of political biases leading to unfair NLP models. In *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds. Rogers A. et al.) 11737–11762 (Association for Computational Linguistics, 2023).
183. Singelis, T. M. & Brown, W. J. Culture, self, and collectivist communication: linking culture to individual behavior. *Hum. Commun. Res.* **21**, 354–389 (1995).
184. Matsumoto, D. Culture, context, and behavior. *J. Pers.* **75**, 1285–1319 (2007).
185. Lin, B. Y., Xu, F. F., Zhu, K. & Hwang, S.-W. Mining cross-cultural differences and similarities in social media. In *Proc. 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Gurevych, I. & Miyao, Y.) 709–719 (Association for Computational Linguistics, 2018).
186. Loveys, K., Torrez, J., Fine, A., Moriarty, G. & Coppersmith, G. Cross-cultural differences in language markers of depression online. In *Proc. Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic* (eds Loveys, K. et al.) 78–87 (Association for Computational Linguistics, 2018).
187. Li, M., Hickman, L., Tay, L., Ungar, L. & Guntuku, S. C. Studying politeness across cultures using English Twitter and Mandarin Weibo. *Proc. ACM Hum. Comput. Interact.* **4**, 1–15 (2020).
188. Veale, T., Hao, Y. & Li, G. Multilingual harvesting of cross-cultural stereotypes. In *Proc. ACL-08: HLT* (eds Moore, J. D. et al.) 523–531 (Association for Computational Linguistics, 2008).
189. Dong, M., Jurgens, D., Banea, C. & Mihalcea, R. in *Social Informatics* 157–172 (Springer International, 2019).
190. Henrich, J., Heine, S. J. & Norenzayan, A. Most people are not WEIRD. *Nature* https://doi.org/10.1038/466029a (2010).
191. Prabhakaran, V., Qadri, R. & Hutchinson, B. Cultural incongruencies in artificial intelligence. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2211.13069 (2022).
192. Wilson, S., Mihalcea, R., Boyd, R. & Pennebaker, J. Disentangling topic models: a cross-cultural analysis of personal values through words. In *Proc. First Workshop on NLP and Computational Social Science* (eds Bamman, D. et al.) 143–152 (Association for Computational Linguistics, 2016).
193. Shen, Y., Wilson, S. R. & Mihalcea, R. in *Social Informatics* 143–156 (Springer International, 2019).
194. Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C. & Kizilova, K. *World Values Survey: Round Six-Country-Pooled Datafile Version* (JD Systems Institute, 2014).
195. Arora, A., Kaffee, L.-A. & Augenstein, I. Probing pre-trained language models for cross-cultural differences in values. *Proc. of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)* (eds. Dev, S. et al.) (Association for Computational Linguistics, 2023).
196. Grossmann, I. et al. AI and the transformation of social science research. *Science* **380**, 1108–1109 (2023).
197. Weidinger, L. et al. Taxonomy of risks posed by language models. In *Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency* 214–229 (Association for Computing Machinery, 2022).
198. Kuipers, B. in *The Oxford Handbook of Ethics of AI* 421 (Oxford Univ. Press, 2020).
199. Voigt, P. & von dem Bussche, A. *The EU General Data Protection Regulation (GDPR)* (Springer International, 2017).
200. Biden, J. R. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. *White House* https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ (2023).
201. Morstatter, F., Pfeffer, J., Liu, H. & Carley, K. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's Firehose. *ICWSM* **7**, 400–408 (2013).
202. Stasi, M. L. Social media platforms and content exposure: how to restore users' control. *Compet. Regul. Netw. Ind.* **20**, 86–110 (2019).
203. Goga, O., Loiseau, P., Sommer, R., Teixeira, R. & Gummadi, K. P. On the reliability of profile matching across large online social networks. In *Proc. 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1799–1808 (Association for Computing Machinery, 2015).
204. More about restricted uses of the Twitter APIs. *X Developer Platform*, https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases (accessed March 2024).
205. Kim, S., Yun, S., Lee, H., Gubri, M., Yoon, S. and Oh, S.J., ProPILE: probing privacy leakage in large language models. In *Proc. of the 37th International Conference on Neural Information Processing Systems*, 20750-20762 (2023).
206. Caliskan, A., Bryson, J. J. & Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
207. Liang, P. P., Wu, C., Morency, L.-P. & Salakhutdinov, R. Towards understanding and mitigating social biases in language models. In *Proc. 38th International Conference on Machine Learning* (eds Meila, M. & Zhang, T.) 6565–6576 (PMLR, 2021).
208. Zhao, J., Wang, T., Yatskar, M., Ordonez, V. & Chang, K.-W. Reducing gender bias amplification using corpus-level constraints. *Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing* (eds Palmer, M. et al.) (Association for Computational Linguistics, 2017).
209. Tan, Y. C. & Celis, L. E. Assessing social and intersectional biases in contextualized word representations. *Proc. of the 33rd International Conference on Neural Information Processing Systems* 13230-13241 (2019).

210. Preoțiuc-Pietro, D. et al. The role of personality, age, and gender in tweeting about mental illness. In *Proc. 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* 21–30 (Association for Computational Linguistics, 2015).
211. Pennebaker, J. W. *The Secret Life of Pronouns: What Our Words Say About Us* (Bloomsbury, 2013).
212. Cao, X. & Kosinski, M. ChatGPT can accurately predict public figures' perceived personalities without any training. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/zbhyk (2023).
213. Rao, H., Leung, C. & Miao, C. Can ChatGPT assess human personalities? A general evaluation framework. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1184–1194 (2023).
214. Jin, Z. et al. When to make exceptions: exploring language models as accounts of human moral judgment. *Adv. Neural Inf. Process. Syst.* **35**, 28458–28473 (2022).
215. Shiffrin, R. & Mitchell, M. Probing the psychology of AI models. *Proc. Natl Acad. Sci. USA* **120**, e2300963120 (2023).
216. He, Y. et al. Hi-ToM: a benchmark for evaluating higher-order theory of mind reasoning in large language models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2310.16755 (2023).

Competing interests

J.W.P. and R.L.B. have commercial interests in the LIWC text analysis program. The other authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Rada Mihalcea.

Peer review information *Nature Human Behaviour* thanks Anne-Marie Nussberger, Marshall Taylor and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.