How Artificial Intelligence Can Enrich Our Understanding of Organizational Culture

Amir Goldberg, Stanford University **Sameer B. Srivastava,** University of California, Berkeley

Amir Goldberg and Sameer B. Srivastava provide three concrete illustrations of how managers can use AI to better understand and more effectively manage organizational culture.

Culture is arguably the most nebulous and complicated construct in the social sciences. The innate tendency of humans to create and perpetuate culture is inherent to our ability to feel, imagine, and interpret the world around us.

For most managers, artificial intelligence (AI) is comparably opaque and complex. To many workers, AI is an unknown set of technologies that threaten their knowledge, skills, and livelihoods.

Algorithmic technologies can be powerful tools, helping you to understand cultural processes and the management of culture.

So what do culture and AI have to do with each other? You may believe that AI is propelling us down an oppressive path or you may feel that negative sentiments and expectations about it are shortsighted. We hope, with the help of our recent research, to persuade you that algorithmic technologies can be powerful tools, helping you to understand cultural processes and the management of culture.[1]

Algorithms and Culture

Recent advances in artificial intelligence have led to the release of several large language models (LLMs)—including OpenAI's GPT 4.0 and Google's Gemini—which exhibit unprecedented conversational abilities. Scholars continue to debate whether these abilities are evidence of nascent machine intelligence. Regardless, it is clear that LLMs are not intelligent in the way humans are. They perform well on certain analogical tasks, for example, but fail at other forms of trivial reasoning. It remains unclear whether these algorithms understand, in the intuitive sense of the word, or merely reproduce statistical patterns drawn from copious data.[2]

Yet, while they may not be actually intelligent, contemporary AI algorithms are extremely good at prediction.[3] In many cases, given the right data, they are far better at making predictions than humans. These algorithms can thus be immensely beneficial in solving difficult problems, so long as their human minders correctly construe the problem as a prediction task. Indeed, the breathtaking recent advances in natural language processing algorithms were driven by the fact that they were trained for simplified prediction tasks. For example, they might be asked to predict a randomly masked word in human-generated text using the surrounding unmasked words as contextual clues. A decade or so ago, just such an approach produced a family of linguistic algorithms known as word embedding models.[4]

With access to a sufficiently large set of training data, a word embedding algorithm gradually "learns" how members of a group communicate with one another.

With access to a sufficiently large set of training data, a word embedding algorithm gradually "learns" how members of a group communicate with one another. Of course, it does not understand the group's language the way a human would; it just predicts a masked word. The algorithm's learning process relies on developing a numeric representation of the group's language, with each word represented by a set of numbers, usually a few hundred of them. This is a terribly impoverished representation of language compared to the amazingly complex linguistic cognition of humans. Nevertheless, it approximates human semantic cognition quite well. Think of these numeric representations of words as coordinates in a multidimensional space. Provided that the algorithm was trained on a large enough body of data to provide a comprehensive sample of English, the distance between words in this space reflects people's subjective perception of the semantic similarity between words. "Cat" and "dog," for example, are likely to be much closer to each other than "cake" and "transportation."

Word embedding models can reveal the complexity of people's perceptions without asking them directly what they think.

Word embedding models can thus reveal the complexity of people's perceptions without asking them directly what they think. In one early demonstration, researchers used word embedding models trained separately on texts from different historical periods to trace the changing meaning of the word "gay."[5] Models examining the beginning of the twentieth century placed "gay" closest to words such as "happy" or "jovial." Those using texts from the end of the century placed it closest to words describing gender and sexual orientation. This progression mirrored the word's etymological evolution from meaning happy to meaning homosexual. In a different study, researchers used word embedding to evaluate gender bias in English.[6] They found that feminized professions such as "librarian" are closer in embedding space to the word "woman," while masculinized professions like "carpenter" are closer to the word "man."

How does all this relate to culture? While there are many definitions of culture, they all encompass the beliefs and perceptions shared by a group of people, whether that group is a nation of millions or a startup firm with only a few dozen employees. The operative term here is "shared."

And if word embeddings can help us measure people's perceptions, they can also help us assess the extent to which those perceptions are shared.

The main goal of traditional approaches to the study of culture is to understand the various beliefs and perceptions unique to a given culture. They find, for example, that Americans espouse individualism more than any other nation in the world, or that some firms emphasize moving fast and breaking things, while others, in keeping with the environment in which they operate, are more cautious. This is commonly referred to as the *content* approach to studying culture. Content analyses of culture are interpretative; they rely on the researcher's ability to understand the substance of the group's shared perceptions. A typical algorithmic agent is unable to perform such a task well.

Instead, we use word embedding models to apply a *distributive* approach to studying culture. This approach allows us to use word embeddings to evaluate the extent to which people share perceptions and along what dimensions. We don't need our algorithmic agent to understand; we only need it to reliably measure semantic similarities.

Three Studies

We recently conducted three studies using word embeddings to shed new light on the cultural dynamics of organizations.

In the first study, we used the online software development platform Gigster. Together with Katharina Lix and Melissa Valentine, we analyzed the communication between members of 117 teams on the instant messaging platform Slack.[7] We aimed to determine whether these teams performed better when each member thought about her team's task differently, or when all members were on the same page.

Alignment of thinking among team members is a double-edged sword. On one hand, if everyone interprets the team's goals similarly, they can easily coordinate their activities. On the other hand, similar thinking can readily turn into groupthink, encouraging team members to quickly converge on a suboptimal idea without anyone challenging it.

Recall that a word embedding model assigns each word a position in a multidimensional space. We computed the location of each speaker in that space by computing the average position of all the words she wrote during a day of interaction. This mapping allowed us to compute the semantic distance between each pair of speakers. We defined a team's discursive diversity—the degree to which the meanings conveyed by group members in a set of interactions diverge from one another—as the average pairwise semantic distance between all team members. In other words, discursive diversity reflects the extent to which the thinking of team members is aligned or divergent.

The teams that performed best were those that could adapt their discursive diversity to the task they were engaged in.

We found that the teams that performed best were those that could adapt their discursive diversity to the task they were engaged in. During periods of coordination, when members allocated and executed responsibilities, high-performing teams were discursively aligned. During periods of ideation, when they were focused on devising solutions, these same teams diverged discursively, with members expressing a range of different ideas. To reap the performance rewards of dynamic alignment, it was important for members of these teams to synchronize these shifts and for their leaders to facilitate this dynamic. They diverged from or converged with one another in unison.

For managers, these findings have three core implications. First, diversity, equity, inclusion, and belonging (DEIB) initiatives typically seek to expand the range of thoughts and ideas within an organization. Yet it has heretofore been difficult to systematically measure cognitive diversity and how it varies in different groups and over time. Given the ubiquity of digital trace data and the accessibility of word embedding models, reliable indicators of cognitive diversity—including, but not limited to, the discursive diversity measure we developed—can soon be at the fingertips of every organization's leaders. Second, in constructing teams, the key is not simply to maximize cognitive diversity; for many tasks, the most effective teams will be able to modulate their expressed diversity in accordance with the requirements of their tasks. Finally, executives should select and develop leaders who are attuned to the cognition of their teams and who know how to adjust cognitive diversity while keeping group members in sync.

In the second study, in collaboration with Paul Gouvard, we used the same methodology to analyze the communication in quarterly earnings calls conducted by executives of publicly traded firms. During these calls members of the C-suite, and often the CEO (chief executive officer), discuss their firms' financial performance and strategy with securities analysts. Executives seek to give analysts

positive impressions of their firm's potential. Some do that by diverging from the ways their competitors tend to talk about their businesses. This tactic may give them the advantage of appearing unique, but it may also cause them to seem frivolous or incompetent.

When executives from a given firm speak differently from their counterparts in competing firms, analysts tend to become overly and unjustifiably optimistic about the focal firm's future performance.

We evaluated the extent to which the executives of each firm used typical or atypical language in a quarterly earnings call by measuring the extent to which their discourse diverged from that of their competitors. We found that analysts are usually swayed by atypical performances. When executives from a given firm speak differently from their counterparts in competing firms, analysts tend to become overly and unjustifiably optimistic about the focal firm's future performance. This response generally results in a negative earnings surprise, with the firm's future earnings failing to meet analysts' expectations. Not all atypical calls lead analysts to become overly bullish, though. Rather, analysts are particularly receptive to atypicality when it emulates the speech of celebrated trailblazers. In other words, analysts interpret uniqueness as a signal of potential performance only when it conforms to popular notions of innovation.

Managers should draw at least two key lessons from this study. First, although quarterly earnings calls and other forms of engagement with external stakeholders are performances that can have evaluative consequences, the scope of these performances goes beyond carefully crafted talking points and frequently asked questions. Audience members will judge all facets of this communication—including how executives frame firm strategy and performance, how they engage with and build upon one another's ideas, and how they respond to or subtly dodge questions—according to normative expectations that are defined by their perception of the firm's peer group. Second, although executives generally assume that firms benefit from differentiating themselves from their peers, they should be alert to the unintended negative consequences of differentiation. Positive evaluations that stem from atypical performances can portend a negative earnings surprise.

In the final study, with Lara Yang, we once again used word embedding models to measure similarities and divergences. Instead of measuring distances between people, however, we measured the distances between the words of individual speakers, focusing on "I" and "we." We reasoned that the closer these two words are in embedding space, the more strongly an individual identifies with the collective. Building on this intuition, we developed a novel measure of how strongly an employee identifies with her organization, as reflected in internal communication with colleagues. We fine-tuned separate word embedding models for each employee and calculated the distance between these two pronouns separately for each three-month period. This allowed us to evaluate variations in the strength of each person's identification with the firm over time.

Managers have traditionally used engagement surveys to measure how much employees identify with their organization. It is impractical to conduct these surveys on a frequent basis, though. Our approach allows us to measure identification unobtrusively and trace its evolution over time. Using this method at three different organizations, we found that, irrespective of where they work, people's sense of identification changes continually. As one would expect, employees' identification gradually increases as their tenure lengthens. But their views are also influenced by the people with whom they interact. Tight-knit, strongly interconnected networks encourage people to identify with their organization. The more people build connections with colleagues in different parts of the organization, the stronger their organizational identification. So people identify with their organization not just because of their personalities and preferences, but also in response to their shifting position within its internal structure. And their organizational identification influences their motivation and commitment.

People identify with their organization not just because of their personalities and preferences, but also in response to their shifting position within its internal structure.

Managers should draw at least three lessons from this study. First, that pairing digital trace

data and AI tools with traditional survey instruments is extremely valuable. The former can map behavioral indicators over time to reveal foundational constructs like organizational identification. The latter can yield validated measures of how people think and feel about the organization. By combining them, we can not only validate the new language-based measures, but also begin to grasp how people are thinking and feeling without surveying them repeatedly. Second, we can learn to fine tune general-purpose algorithms that have been trained on large data sets drawn from a broad cross section of individuals and groups to extract information about a specific organization, period, or person. Finally, the documented changes in network structure arising from the abrupt shift to remote work during the COVID-19 pandemic may also have had secondary consequences for organizational identification.[8] Because remote or hybrid work causes workplace networks to become more siloed, it may also fray the ties that bind people to the broader organization, leaving them to identify only with their immediate work or social group.

Data-Driven Management and Culture

Whether you like it or not, algorithms are already changing organizations, from supply chain management to marketing. Even the aspects of management that draw heavily on social and emotional skills are not immune to the benefits of algorithmic analysis. No one can afford to wait on the sidelines until every debate about the trajectory and social consequences of AI is resolved.

Managers who learn how to ethically harness algorithmic technologies to better understand and manage their culture are most likely to gain a competitive advantage.

Managers who learn how to ethically harness algorithmic technologies to better understand and manage their culture are most likely to gain a competitive advantage. Using word embedding in a distributive approach can illuminate why some teams perform better than others, how external stakeholders evaluate a firm's performance, and which employees are likely to identify with their organization.

The tools to develop and broadly deploy such measures are already readily available, but managers must learn to use them effectively. Few, if any, employees are enthusiastic about having their communications constantly analyzed by a Big Brother algorithm. Indeed, if people believe that how they communicate digitally will affect their prospects, they will change their behavior to fit their (likely incorrect) understanding of what the algorithms in question measure. This situation is a recipe for unintended, deleterious consequences and the erosion of trust.

We believe that these technologies are best implemented as self-empowering tools, which employees can decide whether to engage with, keeping the ability of employers to monitor them in check. Imagine, for example, a conversational bot that occasionally asks, "Do you really want to send this message?" as the user types an email or instant message. "You may not have intended this, but your message might be interpreted as aggressive or hostile," the bot might tell the user. Employed correctly, such bots could help to foster psychologically safe and productive working environments.

Technological innovations are often greeted with passion, whether it be the enthusiasm of those who see them as tools of efficiency and empowerment or the hostility of those who fear they will become instruments of oppression. Word embeddings could be both. Technology is morally neutral. Whether it is liberating or repressive depends on how we choose to use it. The possibility of dire misuse is far from hypothetical and managerial decisions can powerfully affect people's livelihoods and sense of worth.

Whether cultural algorithms become tools of coercion or empowerment is, ultimately, the responsibility of organizational leaders. Culture has long been proven to be a powerful source of competitive advantage or disadvantage. The upside of using algorithms for cultural management is therefore immense, but needs to be continually managed. A poorly thought-out implementation can easily backfire, alienating employees and destroying healthy cultures that took years to build. The conflict that often arises between business and ethics is therefore obviated: doing the right thing by ethically rolling out AI-based approaches to measuring culture is also the right business decision. ■

Author Bios

Amir Goldberg is a professor of organizational behavior and (by courtesy) sociology at the Stanford Graduate School of Business, where he is a founding codirector of the Computational Culture Lab. His research uses computational methods to understand and model cultural processes in organizations and beyond.

Sameer B. Srivastava is the Ewald T. Grether Professor of Business Administration and Public Policy at UC Berkeley's Haas School of Business. He is cofounder and codirector of the Computational Culture Lab and the Berkeley Center for Workplace Culture and Innovation. His research uses computational methods to unpack the complex relationships between group culture, individual cognition, and interpersonal networks and examine how they relate to individual attainment and organizational performance.

Endnotes

1. Matthew Corritore, Amir Goldberg, and Sameer B. Srivastava, "The New Analytics of Culture," *Harvard Business Review* 98, no. 1 (2020): 76-83.
2. Kyle Mahowald et al, "Dissociating Language and Thought in Large Language Models," *Trends in Cognitive Science* 28, no. 6 (June 2024): 517-540
3. Ajay Agrawal, Joshua Gans, and Avi Goldfarb, *Prediction Machines: The Simple Economics of Artificial Intelligence* (Boston, MA: Harvard Business Review Press, 2018).
4. Tomas Mikolov et al., "Distributed Representations of Words and Phrases and their Compositionality," in *Neural Information Processing Systems Advances 26*, eds. C. J. Burges et al., (Red Hook, NY: Curran 2013), 3111–3119.
5. William L. Hamilton, Jure Leskovec, and Dan Jurafsky, "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change," in *The 54th Annual Meeting of the Association for Computational Linguistics* (Berlin, Germany: Association for Computational Linguistics), 1489–1501.
6. Nikhil Garg et al., "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes," *Proceedings of the National Academy of Sciences* 115, no. 16 (2018): E3635–E3644.
7. Katharina Lix et al., "Aligning Differences: Discursive Diversity and Team Performance," *Management Science* 68, no. 11 (November 2022): 8430–8448.
8. Longqi Yang et al., "The Effects of Remote Work on Collaboration among Information Workers," *Nature Human Behavior* 6, (2022): 43-54.