How Can AI Enrich Our Understanding of Organizational Culture?

Amir Goldberg, Stanford University

Sameer B. Srivastava, University of California, Berkeley

September 2022

Forthcoming in *Management and Business Review*

(Special issue on "Managing Organizational Culture")

Abstract:

Culture is one of the most nebulous and complicated constructs in the social sciences, and artificial intelligence is comparably opaque. Yet algorithmic technologies can be useful for understanding cultural process and the management of culture. We focus on one family of deep-learning algorithms known as word embeddings. A word embedding algorithm can, with access to a sufficiently large corpus, be trained to "learn" how people in a social group communicate with one another. Such models can be invaluable tools in taking a distributive approach to studying culture. Rather than trying to interpret a given organizational culture, analysts can instead use word embeddings to assess the extent to which organizational members' perceptions are *shared* and along what dimensions. Three recent studies, all of which use word embeddings, shed new light on cultural dynamics in organizations: These models can help illuminate why some teams perform better than others, how external stakeholders evaluate firm performance, and which employees are more likely to identify with their organization. More generally, firms that learn how to harness algorithmic technologies in ethical ways to better understand and more effectively manage their culture stand the greatest chance of gaining competitive advantage in a rapidly evolving work environment.

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

For most managers, artificial intelligence (AI) is comparably opaque and complex. Indeed, AI represents for many workers an unknown set of technologies that they find threatening to their knowledge, skills, and livelihoods.

What do culture and AI have to do with each other? Whether you believe that AI is propelling us down an oppressive path or that harboring negative sentiments and expectations about it is short-sighted, we will try to persuade you—using examples from our recent joint research—that algorithmic technologies can be powerful tools for understanding cultural process and the management of culture (see also Corritore, Goldberg, and Srivastava [2020]).

Algorithms and Culture

Existing machine learning algorithms are not intelligent in any intuitive sense of the word. Despite their name, they do not "understand" or "learn" in the way humans do. Alan Turing, widely considered the progenitor of modern computer science, famously suggested the following test for determining whether an algorithm qualifies as artificially intelligent. Imagine a human interrogator conversing in written text with two agents, one human and the other algorithmic, and not knowing which one is which. The algorithm passes the Turing Test if the average evaluator cannot distinguish between the two interrogators. If you have ever found yourself conversing with a virtual assistant on your smartphone, you have probably noticed that it can fail the test quite spectacularly. Even the most advanced language algorithms such as GPT-3—in fact, so advanced and potentially pernicious that OpenAI, its developer, decided not to release it publicly for fear that it would be used to generate fake news at scale—will fail the test quite quickly if you ask the right questions.[1]

If not for actual intelligence, then what is “AI” good for? Contemporary algorithms are especially good at prediction (Agrawal et al. 2018). In fact, on many tasks, they are far better than humans in making predictions—so long as they are given the right data to make that prediction with. These algorithms can be immensely beneficial in solving difficult problems—assuming the problem is correctly construed as a prediction task. Indeed, the breathtaking recent advances in the performance of natural language processing algorithms have been driven by the fact that they were trained to solve simplified prediction tasks. For example, they might be asked to predict a randomly masked word in human-generated text, using the other unmasked words surrounding it for contextual clues. Employing such an approach, a family of linguistic algorithms known as word embedding models were introduced a decade or so ago (Mikolov et al. 2013).

With access to a sufficiently large set of training data, a word embedding algorithm gradually “learns” how group members communicate with one another. Of course, it does not understand the group’s language the way a typical human does. All it does is predict a masked word. Yet, during this learning process the algorithm develops a numeric representation of the group’s language. Each word is represented by a set of numbers, usually a few hundred of them. This is a terribly impoverished representation of language compared to humans’ amazingly complex linguistic cognition. Nevertheless, it turns out that it approximates humans' semantic cognition quite well. Think of these numeric representations of words as coordinates in a multi-dimensional space. The distance between words in embedding space appears to reflect people's subjective perception of the semantic similarity between words (provided that the algorithm was trained on a large enough data corpus that represents a comprehensive sample of English). "Cat" and "dog," for example, are likely to be much closer in such a space than are the words "cake" and "transportation."

Word embedding models can reveal different facets of people's perceptions without asking them directly what they think. In one early demonstration, for example, researchers used word embedding models, trained separately on texts produced in different historical periods, to show the changing meaning of the word "gay" (Hamilton et al. 2016). Whereas in the beginning of the twentieth century the word was closest to words such as "happy" or "jovial," by the end of the century the words closest to it were those that denote gender and sexual orientation. This mirrored the word's etymological evolution from denoting happiness to indicating homosexuality. In a different study, researchers used word embeddings to evaluate gender bias in language (Garg et al. 2018). They found that feminized professions such as "librarian" are closer in embedding space to the word "woman" than professions that are conventionally perceived as masculine, such as "carpenter." The latter are much closer in space to the word "man."

How does all this relate to culture? The many definitions of culture that analysts commonly use are unified in assuming that culture relates to beliefs and perceptions that are shared among a group of people, whether that group is a nation comprising hundreds of millions of people or a startup firm with only a few dozen employees. The operative term here is “shared.” If word embeddings can help us measure people’s perceptions, then they can be useful in assessing the extent to which these perceptions are shared.

Traditional approaches to the study of culture see their main goal as understanding the various beliefs and perceptions that uniquely characterize 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 inculcate a more cautious culture, as prescribed by the strategic environments in which they operate. 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. This is not a task that a typical algorithmic agent can perform well.

We instead use word embedding models in the context of a *distributive* approach to studying culture. Rather than trying to interpret a culture, we rely on word embeddings to evaluate the extent to which people’s perceptions are shared and along what dimensions. Our algorithmic agent does not need to understand. All it is required to do is to reliably measure semantic similarities.

Findings of Three Studies

In what follows, we share findings from three recent studies, all of which use word embeddings to shed new light on cultural dynamics in organizations. The first explores software development teams' collaborative work, the second examines executives' interactions with outside stakeholders, and the third assesses employees' identification with their organization.

Our first study used the online software development platform Gigster. Together with Katharina Lix and Melissa Valentine (Lix et al. 2022), we analyzed the communication between members of 117 distributed teams using the instant messaging platform Slack. Did these teams perform better when each member thought about her respective team's challenge differently, or did their performance peak when all members were on the same page?

Alignment between team members' thinking is a double-edged sword. On the one hand, if everyone interprets the team's goals very similarly, it is easy for members to coordinate their activities. On the other hand, similar thinking runs the risk of groupthink, whereby team members quickly converge on a sub-optimal idea without anyone challenging its usefulness.

We used word embeddings to evaluate the extent to which team members' thinking diverged or converged. Recall that a word embedding model assigns each word a position in a multidimensional space. We computed the location of an individual speaker in that space by computing the average position of all the words this team member wrote during a day of interaction. This 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 team members are aligned or divergent in their thinking.

We found that the teams exhibiting the highest performance were the ones that were able to modulate their discursive diversity as a function of the task they were engaged in. During periods of coordination, when members allocated responsibilities or engaged in execution, high-performing teams were discursively aligned. During periods of ideation, in contrast, when they were focused on coming up with solutions, these teams diverged discursively such that different people expressed different ideas. Moreover, it was important for team members to synchronize this pattern of modulation. To reap the performance rewards of dynamic alignment, they needed to diverge from or converge with one another at the same time. Finally, we discovered that it was especially important for team leaders to facilitate this dynamic.

For managers, the study has three core implications. First, diversity, equity, inclusion, and belonging (DEIB) initiatives typically seek to expand the range of thoughts and ideas that circulate through an organization. Yet it has heretofore been difficult to systematically measure cognitive diversity and how it varies across 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 every organization's fingertip. Second, in constructing teams, the key is not simply to maximize levels of cognitive diversity; rather, for many tasks, the most effective teams will be ones that exhibit a capacity to modulate their levels of expressed diversity to match their shifting task requirements. Finally, organizations should select and develop leaders who are attuned to group-level cognition and who know how to orchestrate the necessary adjustments in cognitive diversity while keeping group members in sync with each other as they make these shifts.

In the second study, in collaboration with Paul Gouvard, we used the same methodology to analyze the communication in quarterly earnings calls conducted by publicly traded firms. These calls feature executives, typically the CEO or other C-suite members, who discuss their firms' financial performance and strategy with securities analysts. Executives seek to positively shape analysts' impressions of their firm's future potential. Some do that by diverging from the ways their competitors conventionally talk about their businesses. This carries the potential advantage of appearing unique but also the risk of seeming frivolous or incompetent.

We evaluated the extent to which executives of a given firm use typical or atypical language in a quarterly earnings call by measuring the extent to which their discourse diverges from their competitors'. We found that analysts are systematically swayed by atypical performances. When executives from a given firm speak differently from their counterparts in competitor firms, analysts tend to become overly and unjustifiably optimistic about the focal firm's future performance. This results in a negative earnings surprise, namely, the firm's future earnings underperform analyst expectations. We found, moreover, that not all atypical calls lead analysts to become overly bullish. Rather, analysts are especially positively receptive to atypicality when it emulates celebrated trailblazers' speech. In other words, only when uniqueness conforms to popular notions of innovation is it interpreted by analysts as a signal of potential performance.

This study has at least two key implications for managers. First, although it is well understood that quarterly earnings calls and other forms of engagement with external stakeholders represent a type of performance that can have evaluative consequences, the scope of these performances go beyond carefully crafted talking points and frequently asked question guides. All facets of such communication—including how firm strategy and performance are framed, how executives engage with and build upon one another's ideas, and how they respond to or subtly dodge questions from audience members—represent a type of performance that audience members will judge relative to normative expectations that are defined in relation to the firm's perceived peer group. Second, whereas it is generally assumed that firms benefit from differentiating themselves from their peers, differentiation can, in some cases, also have unintended negative consequences. As our results indicate, the positive evaluations that stem from performative atypicality can portend a negative earnings surprise.

In the final study, together with Lara Yang, we once again used word embedding models to measure similarities and divergences. In this project, however, instead of measuring distances between people, we measured distances between words within the same speaker. We focused on two words in particular: "I" and "we." The closer these two words are in embedding space, we reasoned, the more strongly an individual identifies with the collective. Building on this intuition, we developed a novel measure of employee's strength of organizational identification, 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 3-month period. This allowed us to evaluate within-person variation in identification strength over time.

Organizational identification is traditionally measured using engagement surveys. It is impractical to conduct these surveys on a frequent basis. Our approach allows us to measure identification unobtrusively and trace its changes over time. Using this method across three different organizations, we found that, irrespective of which organization one works for, people's sense of identification changes quite a lot as time goes by. As one would expect, identification gradually, but modestly, grows as employees' tenure increases. But it is also influenced by the people with whom they interact. The more employees are part of tight-knit networks in which people are all highly interconnected, the more they identify with their organization. In addition, the more people are connected to colleagues who hail from different subcommunities within the organization, the stronger their organizational identification. Thus, the extent to which people identify with their organization, which in turn influences their motivation and commitment to the organization, is not just a function of their personalities and preferences. It is also a result of their shifting positions within internal network structure.

This study provides at least three takeaways for managers. First, it demonstrates the value of pairing of digital trace data and the tools of AI with traditional survey instruments. The former can yield time-varying, behavioral indicators of such foundational constructs as organizational identification, while the latter 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 draw inferences about how people are thinking and feeling without having to survey them repeatedly. Second, the method we develop illustrates how general-purpose algorithms that have been trained on large data sets compiled across a broad cross-section of individuals and groups can be fine-tuned to extract organization-, time-, and even person-specific insights. Finally, findings from this study suggest that the documented shifts in network structure stemming from the abrupt shift to remote work during the COVID-19 pandemic (Yang et al.

2022) may also have had knock-on consequences for organizational identification. Insofar as remote or hybrid work causes workplace networks to become more siloed, it may also lead to a fraying of ties that help bind people to the broader organization beyond their immediate work or social group.

Data-Driven Management and Culture

What does this all mean for those who run real-life organizations? Whether you like it or not, algorithms are already changing organizations, from supply-chain management to marketing. But even the "softest" aspects of management—those that draw heavily on social and emotional capabilities—are not immune to the analysis of algorithmic agents. No one can afford to wait on the sidelines until the various debates about the likely trajectory and social consequences of AI get resolved.

Firms that learn how to harness algorithmic technologies in ethical ways to better understand and more effectively manage their culture stand the greatest chance of gaining competitive advantage in a rapidly evolving work environment. As we have illustrated, using word embedding in a distributive approach can help illuminate why some teams perform better than others, how external stakeholders evaluate firm performance, and which employees are more likely to identify with their organization.

The tools to develop and deploy such measures at scale are already readily and freely available. The challenge lies in knowing how to implement them effectively. Few if any employees embrace the idea of having their communications consistently analyzed by a "big brother" algorithm. Moreover, if people believe that how they speak in digital communication media will affect their prospects in an organization, they will change their behavior to fit their (most likely incorrect) impressions of what the algorithms in question measure. This is a recipe for unintended deleterious consequences and the erosion of trust.

We believe that the best implementations of these technologies will be those that function as self-empowering tools, which employees can opt in to or out of, and that constrain the monitoring abilities of employers. Imagine, for example, a conversational bot that occasionally asks, "Do you really want to send this message?" before one hits "send" on one's email or instant message. "You may not have intended this, but your message might be interpreted as overly aggressive or hostile," the bot might tell the user. Employed correctly, such bots may prove to be immensely useful in helping to foster psychologically safe and productive working environments.

Technological innovations such as AI are often greeted with passion, either enthusiastically by those who see them as tools of efficiency and empowerment or with suspicion by those concerned that they will be employed as instruments of oppression. Word embeddings are potentially both things. Technology is morally neutral. Whether it becomes a liberating or repressive force depends on how it will be used. Managerial decisions can have serious consequences for people's livelihoods and sense of worth. The possibility of dire misuse is far from hypothetical.

Whether cultural algorithms become tools of coercion or empowerment is, ultimately, the responsibility of organizational leaders. Culture has long been known to be a potentially powerful source of competitive advantage. The upside of using algorithms for cultural management is therefore immense; however, it needs to be managed as an ongoing process, not as a turnkey event. A poorly thought-out implementation can easily backfire, alienating employees and destroying healthy cultures that took years to build. Thus, the conflict that often arises between business and ethics is obviated: Doing the right thing ethically in the roll-out of AI-based approaches to measuring culture is, fortunately, also very likely to be the right business decision.

References

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2018. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Boston, MA: Harvard Business Review Press.

Corritore, Matthew, Amir Goldberg, and Sameer B. Srivastava. 2020. "The New Analytics of Culture." Harvard Business Review (January-February issue).

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018 "Word embeddings quantify 100 years of gender and ethnic stereotypes." *Proceedings of the National Academy of Sciences* 115(16):E3635–E3644.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1489–1501. Berlin, Germany: Association for Computational Linguistics.

Lix, Katharina, Amir Goldberg, Sameer B. Srivastava, and Melissa A. Valentine. "Aligning Differences: Discursive Diversity and Team Performance." *Management Science*, February 2022. ISSN 0025-1909. doi: 10.1287/mnsc.2021.4274.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. "Distributed Representations of Words and Phrases and their Compositionality." In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors. *Advances in Neural Information Processing Systems* 26:3111–3119. Curran Associates, Inc.

Yang, Longqi, David Holtz, Sonia Jaffe, Siddharth Suri, Shilpi Sinha, Jeffrey Weston, Connor Joyce, Neha Shah, Kevin Sherman, Brent Hecht, and Jaime Teevan. 2022. "The effects of remote work on collaboration among information workers." *Nature Human Behavior* 6:43-54.

Endnotes

[1] https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html