Mosaic image of particle and vapor plumes ejected from the ice-covered surface near the south pole of Saturn's moon Enceladus. The image has been rotated 180° from its original orientation.

state of chemical disequilibria. If correct, this observation has fundamental implications for the possibility of life on Enceladus; chemical disequilibrium that is known to support microbial life in Earth's deep oceans is also available to support life in the Enceladus ocean.

The detection of H_2 in Enceladus plumes represents a window into processes regulating the composition of its ocean. Although numerous processes can produce H_2 on Enceladus, Waite *et al.* present convincing arguments that point to water-rock reactions in the silicate core as the most likely source. Thus, liquid water on Enceladus is not only a requirement for life-sustaining biochemical processes, but may also be essential for geochemical processes responsible for the production of H_2. Indeed, fluid flow and associated water-rock reactions are ubiquitous on Earth in a diverse range of submarine environments (*7*) that continues to expand as we study the vast unexplored regions of our oceans. Many of these environments produce H_2 during hydrothermal alteration of rocks that contain ferrous iron and/or organic matter. The lower temperature limit for H_2 generation during fluid-rock interaction on Earth is poorly constrained, but highly relevant to assessing the viability of fluid-rock reactions as a source of H_2 on Enceladus, because the availability of heat represents a key variable that may limit the temperature of H_2 generation in the silicate core (*8*). In the context of Enceladus' geochemical evolution, the importance of water-rock reactions extends far beyond H_2 generation. As is the case on Earth, where circulation of seawater through the oceanic lithosphere regulates the chemistry of seawater, hydrothermal processing of Enceladus' silicate core has been postulated as a control on the pH, salinity, and abundance of silica in the Enceladus ocean (*9*–*11*).

Waite *et al.*'s results represent an important advance in assessing the habitability of Enceladus. Many questions remain, however, regarding geological processes on Enceladus that lack Earth analogs. For example, unlike Earth, where plate tectonics delivers magmatic heat and continuously supplies unaltered ultramafic rocks to near-seafloor environments readily accessed by hydrothermal fluids, there is no a priori evidence for plate tectonics or magmatic activity on Enceladus. Sustained H_2 generation on Enceladus requires that hydrothermal fluids have access to organic- and ferrous-iron–bearing rocks in the entire silicate core. What are the mechanisms for the formation of permeability and heat that allows the flow of aqueous fluids through the silicate core and back to the ocean? The accumulation of H_2 in the Enceladus ocean is conspicuous in the context of an Earth analog, where H_2 delivered to oxygenated oceans from submarine hot springs is rapidly consumed by pervasive microbial populations in seawater. Is the presence of H_2 in the Enceladus ocean an indicator for the absence of life, or is it a reflection of the very different geochemical environment and associated ecosystems on Enceladus? We still have a long way to go in our understanding of processes regulating the exchange of mass and heat across geological interfaces that define the internal structure of Enceladus and other ice-covered planetary bodies. Future missions to explore oceans beyond Earth will answer many of these questions and further constrain the possibility of life elsewhere in our solar system. ■

REFERENCES

1. J. H. Waite *et al.*, *Science* **356**, 155 (2017).
2. L. Iess *et al.*, *Science* **344**, 78 (2014).
3. W. B. McKinnon, *Geophys. Res. Lett.* **42**, 2137 (2015).
4. O. Cadek *et al.*, *Geophys. Res. Lett.* **43**, 5653 (2016).
5. P. C. Thomas *et al.*, *Icarus* **264**, 37 (2016)
6. K. Raymann, C. Brochier-Armanet, S. Gribaldo, *Proc. Nat. Acad. U.S.A.* **112**, 6670 (2015).
7. C. R. German, W. W. Seyfried Jr., in *Treatise on Geochemistry*, H. D. Holland, K. K. Turekian, Eds. (Elsevier, Oxford, ed. 2, 2014), vol. 8, pp. 191–233.
8. W. Bach, *Front. Microbiol.* **7**, 107 (2016).
9. F. Postberg *et al.*, *Nature* **459**, 1098 (2009).
10. C. R. Glein, J. A. Baross, J. H. Waite, *Geochim. Cosmochim. Acta* **162**, 202 (2015).
11. H.-W Hsu *et al.*, *Nature* **519**, 207 (2015).

10.1126/science.aan0444

ARTIFICIAL INTELLIGENCE

An AI stereotype catcher

An artificial intelligence method identifies implicit human biases such as gender stereotypes

By **Anthony G. Greenwald**

Those who converse regularly with their smartphones know that the language skills of computing devices have emerged from a lengthy childhood. On page 183 of this issue, Caliskan *et al.* unveil a new language achievement of artificial intelligence (AI) (*1*). In large bodies of English-language text, they decipher content corresponding to human attitudes (likes and dislikes) and stereotypes. In addition to revealing a new comprehension skill for machines, the work raises the specter that this machine ability may become an instrument of unintended discrimination based on gender, race, age, or ethnicity.

Caliskan *et al.* used a word-embedding method implemented in Global Vectors for Word Representation (GloVe) (*2*), an algorithm that represents each word in a large vocabulary (2.2 million words, each a distinct case-sensitive letter sequence) as a vector of 300 semantic dimensions. These dimensions are derived from word co-occurrence counts in a corpus of 840 billion tokens (roughly, words) obtained from a large-scale crawl of the web. GloVe's implementation of co-occurrence is based on two specific vocabulary words occurring within 10 words of each other.

Pennington *et al.* (*2*) previously established GloVe's usefulness in solving word analogies, a familiar item type in verbal aptitude tests. Bolukbasi *et al.* (*3*) further showed that word embedding could pick up meanings corresponding to gender stereotypes. Caliskan *et al.* now go further to show that GloVe's vectors capture human implicit biases (such as gender stereotypes) in a fashion parallel to a psychological method, the Implicit Association Test (IAT).

After Greenwald *et al.* (*4*) introduced the IAT in 1998, many findings soon indicated

Department of Psychology, University of Washington, Seattle, WA, USA. Email: agg@u.washington.edu

Data mining in massive bodies of text captures spatial proximities that match pervasive gender stereotypes, such as the proximity of "male" to "leader" and that of "female" to "helper."

that IAT-measured racial and ethnic biases were pervasive in all groups tested (*5*). The IAT's procedure requests computer key presses to classify words from four categories, such as female names, male names, words for leader (such as director or chief), and words for helper (such as assistant or employee). When the categories of male and leader are assigned to one key and female and helper to a second, this writer is strikingly faster than when female and leader share one key, with male and helper on the other (*6*). This speed difference indicates the presence of an implicit gender stereotype that associates male (more than female) with leadership.

Caliskan *et al.*'s Word-Embedding Association Test (WEAT) algorithm uses cosine similarity (a correlation-like indicator) between word vectors in different word categories, much as the IAT uses response latencies; greater cosine similarity corresponds to faster IAT responding. The authors show that the WEAT and the IAT agree in a wide variety of domains. For example, both show strongly greater association of racial white than racial black with pleasant, contrasting with the near absence of race preference found in survey studies that use self-report measures of the associations (*7*).

The attitudes and beliefs associated with race, gender, age, and ethnicity revealed by IAT measures often deviate substantially from attitudes and beliefs that the same persons express publicly. These IAT measures are described as capturing implicit biases, which are distinguished from endorsed (explicit) attitudes. People are often unaware of such implicit biases. Caliskan *et al.* show that, for their text corpus and (mostly) American IAT respondents, attitudes and beliefs tapped by the WEAT correspond more closely to the IAT's implicit biases than to explicit (self-report–measured) attitudes and beliefs.

The agreement between the IAT and the WEAT suggests that language might be a source of the implicit biases that the IAT reveals. Although that theory was previously conceivable, it was untestable due to lack of an appropriate method. Also, the theory that language causes implicit bias is in part questionable because of its similarity to the controversial Sapir-Whorf hypothesis (*8*, *9*), according to which cultural differences in thought patterns are rooted in language differences. Caliskan *et al.*'s WEAT method may create new research opportunities for testing the Sapir-Whorf hypothesis.

There are two ready alternatives to a language-causes-implicit-bias interpretation of Caliskan *et al.*'s findings. One of these reverses the causal direction, treating human mental biases as the source of biases identifiable in human-produced text. A more plausible alternative theory is that both WEAT-measured and IAT-measured biases have other societal sources.

The WEAT will be very useful as a proxy for the IAT when IAT measurement is impossible. For example, IAT measures collected over the past 15 years at the Project Implicit website (*10*) show a slow but steady reduction in implicit bias against gays and lesbians. The WEAT should be able to analyze relevant bodies of text generated not only during these same 15 years but also before the IAT existed. Finding that the WEAT shows the same changes as the IAT would reinforce belief in the connection between the WEAT and the IAT. If changes in the WEAT precede those in the IAT, it is plausible that language is influencing implicit biases. If changes in the WEAT follow those in the IAT, it is plausible that implicit biases are influencing language. If the changes are concurrent, then it is plausible that some other variables are influencing both.

The WEAT can also be used—in ways that the IAT cannot—to diagnose differences in bias content between broadcast and print media; entertainment, news, and sports media; or media targeted at racially or ethnically different audiences. Differences in bias content among these media categories, and changes in them over time, may shed light on correlates of a widely recognized increasing liberal-conservative divide in the United States and elsewhere.

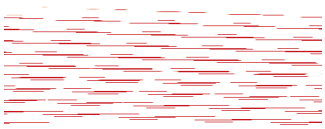
It seems likely that the WEAT method will generalize to languages related to English, such as Spanish, French, German, and Italian. Tests on text corpora in other languages will further clarify the generalizability of the WEAT method.

Considering that AI software might unintentionally perpetrate gender discrimination, Bolukbasi *et al.* (*3*) suggested computational methods to gender-debias AI text analyses. Caliskan *et al.*'s findings further encourage pursuit of this challenging task. Computational debiasing necessarily entails some loss of meaning, and gender is just one dimension on which AI text analyses might be debiased. How much useful meaning may disappear in the process of debiasing simultaneously for the legally protected classes of race, skin color, religion, national origin, age, gender, pregnancy, family status, and disability status? Hopefully, the task of debiasing AI judgments will be more tractable than the as-yet-unsolved task of debiasing human judgments. ■

REFERENCES AND NOTES

1. A. Caliskan, J. J. Bryson, A. Narayanan, *Science* **356**, 183 (2017).
2. J. Pennington, R. Socher, C. D. Manning, *EMNLP* **14**, 1532 (2014).
3. T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, *Adv. Neural Inf. Proc. Syst.* **2016**, 4349 (2016).
4. A. G. Greenwald, D. E. McGhee, J. L. Schwartz, *J. Pers. Soc. Psychol.* **74**, 1464 (1998).
5. B. A. Nosek *et al.*, *Eur. Rev. Soc. Psychol.* **18**, 36 (2007).
6. Interested readers can test their own responses for implicit bias at this website: http://research.millisecond.com/agg/leaderhelperiat.forscience.25mar2017.web
7. M. R. Banaji, A. G. Greenwald, *Blindspot: Hidden Biases of Good People* (Delacorte Press, New York, 2013), Appendix 1.
8. B. L. Whorf, *Language, Thought, and Reality* (MIT Press, Cambridge, MA, 1956).
9. A. G. Greenwald, *Perspect. Psychol. Sci.* **7**, 99 (2012).
10. Project Implicit; https://implicit.harvard.edu.

10.1126/science.aan0649



Editor's Summary

An AI stereotype catcher

Anthony G. Greenwald (April 13, 2017)

Science **356** (6334), 133-134. [doi: 10.1126/science.aan0649]